

The role of treatment crossover adjustment methods in the context of economic evaluation

Nicholas Robert Latimer

BSc, MSc.

Thesis submitted for the degree of Doctor of Philosophy



Health Economics and Decision Science

School of Health and Related Research

October 2012

Acknowledgements

I would like to thank a number of individuals for their help, support and patience over the past three years. Firstly, considerable thanks must go to my supervisors, Ron Akehurst and Mike Campbell. Without their advice and direction I would have ended up with a neverending document that was of little use to anyone. I'm sorry though that my final product is still a little long. Thanks must also go to Keith Abrams, Paul Lambert, Michael Crowther, James Morden and Mike Bradburn, who provided me with statistical support that was critical to my own personal training and development, as well as to the simulation study that makes up a substantial part of my thesis. I'm particularly grateful to Paul Tappenden, who read the whole thing and made excellent comments. Thanks also go to Ian White, Chris Carroll, Lesley Uttley, Suzy Paisley, Andrea Shippam and Allan Wailoo for taking the time to advise, read and review several of my chapters. I also thank the Pharmaceutical Oncology Initiative for their support and input on appropriate simulation study scenarios, and for their provision of the dataset used in Chapter 7.

Finally, thanks to Lizzy and Emma for understanding when I got home late and when I kept on talking about statistics.

Funding source

This PhD thesis has been prepared as part of a Doctoral Research Fellowship funded by the National Institute for Health Research (NIHR) (reference number DRF/2009/02/). The views and opinions expressed therein are those of the author and do not necessarily reflect those of the NIHR.

Publications

I have made efforts to disseminate the research presented in this thesis. The papers listed below have been published and accepted for publication, or are currently under review. I plan to submit papers based upon Chapters 6 and 7 in the coming months, and a future National Institute for Health and Clinical Excellence (NICE) Decision Support Unit (DSU) Technical Support Document on treatment crossover will be based upon my research.

1. Morden JP, Lambert PC, **Latimer NR**, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Medical Research Methodology* 11. 2011.

I helped conceive this project and sought funding for it. I also helped design the simulation study, interpret the results, and revise the paper. James Morden carried out the statistical analysis under the supervision of Paul Lambert and Keith Abrams, and wrote the initial draft of the paper. My thesis extends this study in a number of ways.

2. **Latimer NR**. Survival analysis for economic evaluations alongside clinical trials – extrapolation with patient-level data: Inconsistencies, limitations and a practical guide. (In Press)

This paper has been accepted for publication by the *Medical Decision Making* journal. It is based upon the DSU Technical Support Document listed below. Some of the contents of this paper are used in Chapter 5 of this thesis.

3. **Latimer NR**. NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. Report by the Decision Support Unit. 2011.

This document is available from the NICE DSU website. Some of the contents are used in Chapter 5 of this thesis:

http://www.nicedsu.org.uk/NICE%20DSU%20TSD%20Survival%20analysis_finalv2.pdf

Author's declaration

I declare that this thesis is my original work and that none of the material contained in this thesis has previously been submitted for a degree in this, or any other, awarding institution. The contents and views expressed are my own. All the work has been conducted by myself, with support from those mentioned in my "Acknowledgements". I declare that the STATA program used to generate initial survival times for my simulation study reported in Chapter 6 was written by Mr Michael Crowther, though his original program was adapted in collaboration with myself, Paul Lambert, Keith Abrams, James Morden and Allan Wailoo. Within my simulation study I integrate the use of freely available programs for implementing the Rank Preserving Structural Failure Time Model (RPSFTM), the Iterative Parameter Estimation (IPE) algorithm, and Structural Nested Models (SNM). The programs used were "strbee" (written by Ian White for the RPSFTM and IPE algorithm); and "stgest" (written by Jonathan Sterne and Kate Tilling) and these are referenced in the thesis. I wrote all other statistical code.

Abstract

This thesis investigates the problem of treatment crossover – where patients randomised to the control group of a clinical trial are permitted to cross over onto the experimental treatment at some point during follow-up. Methods commonly used to adjust for treatment crossover within health technology assessments are known to be prone to bias, and these may lead to inconsistent resource allocation decisions. The objective of the thesis is to identify which methods are most appropriate for adjusting for treatment crossover in an economic evaluation context.

If control group patients cross over and benefit from the experimental treatment, an intention to treat analysis will underestimate the “true” survival benefit associated with the new treatment – that is, the benefit that would have been observed had treatment crossover not been allowed. Simple methods for adjusting for crossover, such as excluding or censoring crossover patients, will lead to substantial bias when crossover is associated with prognosis. More complex crossover adjustment methods have been described in the literature and previous research has shown that some of these, such as the Rank Preserving Structural Failure Time Model, perform very well when their key methodological assumptions are satisfied. However, a full comparison of all relevant methods across a range of realistic scenarios – including scenarios where key assumptions are *not* satisfied – has not previously been undertaken. Approaches for incorporating these methods within an economic evaluation – specifically their use in combination with extrapolation modelling – have also not previously been investigated.

In this thesis I demonstrate the importance of the treatment crossover problem, review and assess relevant crossover adjustment methods, and provide an analysis framework to enable the most appropriate method to be identified on a case-by-case basis. Importantly, it is shown that no single method will be satisfactory in all circumstances. In order to identify the method that is likely to provide least bias, consideration must be given to the crossover mechanism, the available trial data, disease and patient characteristics, and the nature of the treatment effect.

Contents

	Page
Part 1: What is the treatment crossover problem?	
<i>Chapter 1: Why treatment crossover is an important factor in health technology assessments, and the relevance of this thesis</i>	1
1.1 Introduction	1
1.2 Motivations for this thesis	2
1.3 Research questions	13
1.4 Thesis structure	14
<i>Chapter 2: Survival analysis and the theoretical framework of economic evaluation</i>	18
2.1 Chapter overview	18
2.2 Allocating scarce resources.....	18
2.3 Theoretical underpinnings	19
2.4 Implications for this thesis	24
2.5 Conclusions	24
Part 2: How important is the treatment crossover problem?	
<i>Chapter 3: Treatment crossover in a practical, health technology assessment context</i>	26
3.1 Chapter overview	26
3.2 Introduction	26
3.3 Search strategy.....	27
3.4 Summary of review findings.....	30
3.5 The prevalence of crossover	32
3.6 Methods used to adjust for crossover	33
3.7 Potential impact of treatment crossover on NICE's decisions	48
3.8 Review update	51
3.9 Discussion.....	55
3.10 Conclusions	57
Part 3: What are the potential solutions?	
<i>Chapter 4: Systematic review of statistical methods for adjusting survival estimates in the presence of treatment crossover</i>	58
4.1 Chapter overview	58
4.2 Introduction	59
4.3 Methods for the review	59
4.4 Methods for the search.....	61
4.5 Search strategy.....	62
4.6 Quality assessment	65
4.7 Data extraction.....	65
4.8 Data synthesis	66
4.9 Search results.....	66
4.10 Results: Narrative synthesis of identified methods	70
4.11 Summary	98
4.12 A novel two-stage method.....	99
4.13 Discussion and conclusions.....	100
<i>Chapter 5: Survival analysis and economic evaluation – implications for treatment crossover methods</i>	103
5.1 Chapter overview	103
5.2 Introduction	104
5.3 Search strategy.....	105
5.4 Modelling methods.....	105
5.5 Model selection process algorithm.....	109
5.6 Extrapolation and treatment crossover adjustment methods	111
5.7 Discussion.....	117

5.8 Conclusions	119
Part 4: How do the treatment crossover adjustment methods perform?	
<i>Chapter 6: Simulation study of methods for adjusting survival estimates to take into account treatment crossover</i>	<i>121</i>
6.1 Chapter overview	121
6.2 Introduction	122
6.3 Relevant previous simulation study	122
6.4 Novel simulation study	124
6.5 Proviso	150
6.6 Results.....	151
6.7 Limitations	195
6.8 Conclusions	198
<i>Chapter 7: A real-world application of treatment crossover adjustment methods.....</i>	<i>200</i>
7.1 Chapter overview	200
7.2 Introduction.....	200
7.3 Background.....	201
7.4 Methods.....	203
7.5 Results.....	208
7.6 Impacts on cost-effectiveness	227
7.7 Limitations	230
7.8 Conclusions	232
Part 5: What should be done in future?	
<i>Chapter 8: Overview, recommendations, discussion, conclusions and future research priorities.....</i>	<i>234</i>
8.1 Chapter overview	234
8.2 Thesis overview	234
8.3 Clinical opinion	239
8.4 Discussion	241
8.5 Recommendations.....	247
8.6 Further research	253
8.7 Conclusions	255
References.....	258
Appendix 1: NICE metastatic and/or advanced cancer appraisals – evidence tables.....	272
Appendix 2: Summary of survival analysis methods used in NICE appraisals of advanced and/or metastatic cancer treatments	373
Appendix 3: Summary of treatment crossover in NICE appraisals of advanced and/or metastatic cancer treatments	385
Appendix 4: Evidence tables.....	387
Appendix 5: Exclusion lists.....	454
Appendix 6: Parametric models summarised.....	465
Appendix 7: STATA ado file code for simulating initial survival times.....	467
Appendix 8: STATA do file code for running simulation study	469
Appendix 9: Scenario parameter input values	502
Appendix 10: Case study STATA do file code	505

List of tables

Table 3.1	NICE technology appraisals included in the review	28
Table 3.2	NICE technology appraisals that included treatment crossover	32
Table 3.3	Methods used to account for crossover in NICE technology appraisals	33
Table 4.1	Framework for the review of identified methods	60
Table 4.2	Initial search terms	63
Table 4.3	Initial (29/09/10) and secondary (01/12/10) search results	66
Table 4.4	Papers included in the systematic review	70
Table 4.5	Methodological categories, sub-categories and papers identified	71
Table 5.1	Methods used to estimate mean survival in NICE technology appraisals	105
Table 6.1	Probability of treatment crossover by prognostic groups and consultation	136
Table 6.2	Simulated scenarios – Parameter values and alternatives tested	137
Table 6.3	Crossover methods for inclusion in simulation study	142
Table 6.4	Overview of simulated scenarios	154
Table 6.5	Bias of RPSFTM and IPE approaches in zero TDC scenarios	169
Table 7.1	Existing analyses of EGF100151 OS data	202
Table 7.2	Alternative parametric model fits to the experimental group ITT data	212
Table 7.3	Methods key	217
Table 7.4	Trial EGF100151 analysis – results when covariates are excluded	219
Table 7.5	Trial EGF100151 analysis – results when covariates are included	220
Table 7.6	Indicative cost-effectiveness results – methods excluding covariates	228
Table 7.7	Indicative cost-effectiveness results – methods including covariates	229

List of figures

Figure 1.1	The potential impact of treatment crossover illustrated	9
Figure 1.2	Diagrammatic representation of the thesis structure	16
Figure 3.1	NICE technology appraisal search results	28
Figure 4.1	Summary of literature search	69
Figure 5.1	Survival model selection process algorithm	110
Figure 6.1	Overall Survival Kaplan-Meier from simulated dataset scenario 1: Without crossover	131
Figure 6.2	Overall Survival Kaplan-Meier – lapatinib+capecitabine for metastatic breast cancer..	131
Figure 6.3	Overall Survival Kaplan-Meier – cetuximab for metastatic squamous cell carcinoma of the head and neck	132
Figure 6.4	Simulated acceleration factor over time	133
Figure 6.5	Overall Survival Kaplan-Meier from simulated dataset scenario 1: With crossover .	137
Figure 6.6	Mean bias (%) across scenarios – ITT analysis	160
Figure 6.7	Mean bias (%) across scenarios – Exclusion and censoring approaches	160
Figure 6.8	Mean bias (%) across scenarios – TDCS and TDCS-Weibull	160
Figure 6.9	Mean bias (%) across scenarios – TDCM and TDCM-Weibull	162
Figure 6.10	Mean bias (%) across scenarios – XOTDCS and XOTDCS-Weibull	162
Figure 6.11	Mean bias (%) across scenarios – XOTDCM and XOTDCM-Weibull	162
Figure 6.12	Mean bias (%) across scenarios – IPCW	165
Figure 6.13	Mean bias (%) across scenarios – RPSFTM and IPE Weibull “survivor function” approaches (no covariates)	165
Figure 6.14	Mean bias (%) across scenarios – RPSFTM and IPE Weibull “extrapolation” approaches (no covariates)	165
Figure 6.15	Mean percentage bias of the IPCW method compared to crossover proportion	166

Figure 6.16 Mean bias (%) across scenarios – RPSFTM and IPE “shrinkage” approaches (with covariates)	168
Figure 6.17 Mean bias (%) across scenarios – SNM with g-estimation.....	168
Figure 6.18 Mean bias (%) across scenarios – Two-stage Weibull	168
Figure 6.19 Bias across zero TDC scenarios – Selected methods	178
Figure 6.20 Bias across zero TDC scenarios – Selected methods with truncated axis	178
Figure 6.21 Mean bias compared to crossover proportion – zero TDC scenarios	179
Figure 6.22 Bias across TDC scenarios – Selected methods	181
Figure 6.23 Bias across additional TDC scenarios – Selected methods.....	181
Figure 6.24 Mean bias compared to crossover proportion – TDC scenarios	182
Figure 6.25 Mean bias compared to crossover proportion – scenarios with an additional time-dependent treatment effect.....	185
Figure 6.26 Coverage across zero TDC scenarios – Selected methods.....	189
Figure 6.27 Coverage across TDC scenarios – Selected methods.....	189
Figure 6.28 Coverage across additional TDC scenarios – Selected methods	189
Figure 6.29 MSE across zero TDC scenarios – Selected methods.....	192
Figure 6.30 MSE across TDC scenarios – Selected methods	192
Figure 6.31 MSE across additional TDC scenarios – Selected methods	192
Figure 6.32 Treatment effect bias across zero TDC scenarios – Selected methods.....	194
Figure 6.33 Treatment effect bias across TDC scenarios – Selected methods.....	194
Figure 6.34 Treatment effect bias across additional TDC scenarios – Selected methods.....	194
Figure 7.1 Overall survival Kaplan-Meier curves from the EGF100151 study	203
Figure 7.2 Log plots to assess observed hazards.....	210
Figure 7.3 Quantile-quantile plot for study EGF100151.....	210
Figure 7.4 Exponential model fitted to experimental group, compared to the Kaplan-Meier curve	212
Figure 7.5 Weibull model fitted to experimental group, compared to the Kaplan-Meier curve	213
Figure 7.6 Gompertz model fitted to experimental group, compared to the Kaplan-Meier curve	213
Figure 7.7 Log-logistic model fitted to experimental group, compared to the Kaplan-Meier curve	213
Figure 7.8 Log normal model fitted to experimental group, compared to the Kaplan-Meier curve.....	214
Figure 7.9 Generalised gamma model fitted to experimental group, compared to the Kaplan-Meier curve	214
Figure 7.10 Generalised gamma model to the ITT control group data, with covariates.....	222
Figure 7.11 RPSFTM counterfactual Kaplan-Meier	225
Figure 7.12 RPSFTM “survivor function” approach compared to “extrapolation” approach (with covariates).....	225
Figure 8.1 Treatment crossover analysis framework	252

List of acronyms

ABPI	Association of the British Pharmaceutical Industry
AF	Acceleration factor
AFT	Accelerated failure time
AG	Assessment group
AIC	Akaike's information criterion
AUC	Area under the curve
AV	Auxiliary variables
BIC	Bayesian information criterion
CBA	Cost-benefit analysis
CEA	Cost-effectiveness analysis
CI	Confidence interval
CUA	Cost-utility analysis
CTCAE	Common terminology criteria for adverse events
DSU	Decision Support Unit
ECOG	Eastern Cooperative Oncology Group
EGFR	Epidermal growth factor receptor
EMA	European Medicines Agency
ERG	Evidence Review Group
FAD	Final Appraisal Determination
FDA	United States Food and Drug Administration
GG	Generalised gamma
GIST	Gastrointestinal stromal tumours
GSK	GlaxoSmithKline
HER2	Human epidermal growth factor receptor 2
HR	Hazard ratio
HRQL	Health-related quality of life
HTA	Health technology assessment
ICER	Incremental cost-effectiveness ratio
IPCW	Inverse probability of censoring weights
IPE	Iterative parameter estimation
IPTW	Inverse probability of treatment weights
ITT	Intention to treat
LRIG	Liverpool Reviews and Implementation Group
MRC	Medical Research Council
MSM	Marginal structural models
MSE	Mean squared error
NCI	National Cancer Institute
NHS	National Health Service
NICE	National Institute for Health and Clinical Excellence
NIHR	National Institute for Health Research
OS	Overall survival
PFS	Progression free survival
PH	Proportional hazards
POI	Pharmaceutical Oncology Initiative
Ppcens	Per protocol, censor crossover patients
PPExc	Per protocol, exclude crossover patients
PPS	Post progression survival
QALY	Quality adjusted life year
RBEE	randomisation-based effect estimator
RCC	Renal cell carcinoma
RCT	Randomised controlled trial

RPSFTM	Rank preserving structural failure time model
SE	Standard error
SNM	Structural nested model
TA	Technology appraisal
TSD	Technical Support Document
TDC	Time dependent confounder
TDCM	Treatment as a time dependent covariate including other covariates using a Cox model
TDCM_We	Treatment as a time dependent covariate including other covariates using a Weibull model
TDCS	Treatment as a time dependent covariate excluding other covariates using a Cox model
TDCS_We	Treatment as a time dependent covariate excluding other covariates using a Weibull model
TSD	Technical Support Document
TTP	Time to progression
UK	United Kingdom
WKM	Weighted Kaplan-Meier
XOTDCM	Treatment crossover as a time-dependent indicator, including other covariates using a Cox model
XOTDCM_weib	Treatment crossover as a time-dependent indicator, including other covariates using a Weibull model
XOTDCS	Treatment crossover as a time-dependent indicator, excluding other covariates using a Cox model
XOTDCS_weib	Treatment crossover as a time-dependent indicator, excluding other covariates using a Weibull model

Chapter 1

Why treatment crossover is an important factor in health technology assessments, and the relevance of this thesis

1.1 Introduction

It is commonplace, in many parts of the world, for new drugs to be assessed formally by Health Technology Assessment (HTA) agencies for their effectiveness and value for money before approval is given for their reimbursement. Typically, the evidence to support the effectiveness of the drug comes from randomised controlled trials (RCT) from which the treatment effect for the intervention, perhaps differentiated among patient groups, is estimated. Clearly, for a fair assessment of the drug, estimating the treatment effect (on for example, overall survival (OS)) is of central importance. RCTs allow a comparison of effects between the novel drug and a comparator, used in separate arms of the trial. However, sometimes patients may cross over from one arm of the trial to the other. This is particularly likely to be the case when a drug is being used after all conventional treatments have been exhausted. This often occurs in trials of treatments for metastatic cancer, although this is not the only circumstance in which it may occur. In these circumstances a simple Intention to Treat (ITT) analysis – that is, a comparison of treatment groups as randomised – of the results may not reflect the true effectiveness of the new intervention as some of its benefit appears in both arms of the trial. Adjustment to allow for crossover is needed. This thesis explores the issues in such analyses, examines the methods available for adjusting for crossover and assesses the circumstances in which some methods are superior to others.

The focus of this thesis is on methods for estimating the effect of a novel treatment on a time-to-event outcome in the presence of treatment crossover. This reflects what is often seen in oncology RCTs, where crossover often occurs and survival outcomes are of critical importance. Treatment crossover as referred to in this thesis is defined as the switching of patients randomised to the control group of a clinical trial on to the experimental treatment some time after randomisation. The focus is on the introduction of novel drugs, and therefore examples and discussion are placed in this context. However, the research presented would also be applicable in any other context in which treatment crossover from a control group into an experimental group might occur.

In this chapter I set the scene for the thesis. Section 1.2 discusses the theoretical, practical and personal motivations for the thesis, demonstrating why accurate estimation of survival benefits is critical to the economic evaluation process, and showing the potential impact of treatment crossover on the estimation of treatment effects. Section 1.3 defines my research questions, and Section 1.4 outlines the thesis structure. Several key terms that are used throughout the thesis are defined. The methodological foundations of economic evaluation to inform resource allocation decisions are briefly discussed in this chapter, and are considered in more detail in Chapter 2.

1.2 Motivations for this thesis

1.2.1 The requirement for economic evaluation

From a direct health perspective any valuable health intervention should either extend life or improve the quality of it, or both. Survival advantages represent a key part of the treatment effect of cancer therapies and it is for these potential benefits that novel cancer drugs are usually developed. The Department of Health's NHS Cancer Plan, published in 2000, outlined the government's strategy for investment and reform of cancer services with the specific aims of extending survival and improving quality of life in cancer patients.¹ However, while it is clearly important that the effect of the novel treatment on both length and quality of life is known reasonably accurately, in a world where resources are constrained demonstrating clinical effectiveness is not enough – cost-effectiveness must also be proven such that decision-makers can appropriately allocate resources.

When a new medical technology is introduced into a health system its cost-effectiveness is often assessed formally, more or less explicitly and transparently, by the relevant HTA authorities. The objective of these authorities is to accurately assess the balance of additional benefits and opportunity costs associated with a novel intervention, compared to the next best alternative. In systems where most of the estimation is done quantitatively and explicitly an incremental cost-effectiveness ratio (ICER) – the incremental costs associated with the new intervention divided by the incremental benefits – is commonly calculated.² This, in combination with an incremental cost-effectiveness threshold, provides decision-makers with useful evidence upon which to base decisions on how best to allocate scarce health care resources. As an example, in the UK the National Institute for Health and Clinical Excellence (NICE) appraises new interventions and typically recommends these for use in the National Health Service (NHS) if they exhibit an ICER of less than £20,000 - £30,000 per quality adjusted life year (QALY) gained.³ The QALY is a measure of health-related benefits and incorporates

both length and quality of life. Its importance in economic evaluation is returned to in Chapter 2. The requirement for economic evaluation dictates that it is essential to accurately estimate the quality and length of life impacts associated with a new cancer drug relative to relevant comparators. If the size of the treatment effect is not estimated accurately – given the decision problem faced by decision-makers – inappropriate resource allocation decisions may be made.

From the outset is important to consider why treatment crossover occurs and why it is considered as particularly important for health economists. Treatment crossover is likely to occur in clinical trials of cancer treatments for several reasons, which are discussed further in Section 1.2.7. However, primary amongst these is the fact that trials are often powered to investigate differences in progression-free survival (PFS) as a primary endpoint, rather than OS. While drug regulatory agencies such as the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) may have preferred OS as a primary endpoint, they have recognised the difficulties associated with obtaining long-term OS data and accepted that PFS or time to progression (TTP) represent acceptable primary endpoints for drug approval.^{4;5} Hence, since it has been possible successfully to license a new product with evidence based upon PFS, there has been less motivation for pharmaceutical companies to ensure that randomised groups are maintained beyond disease progression. On the other hand – as recommended by several HTA bodies around the world – the treatment effect on OS must be included in any economic evaluation that affects survival in order for useful cost-effectiveness estimates to be derived.^{3;6-8} Hence, treatment crossover often only becomes a serious problem when the new treatment becomes the subject of a HTA. Treatment crossover should not represent only a “health economics” problem – all parties should be interested in the “true” survival benefit associated with a new therapy – but due to the different requirements for licensing and HTA bodies it is often health economists who seek to address the problem.

1.2.2 The economic evaluation decision problem

Typically, the decision problem addressed in HTA concerns whether the novel intervention represents a cost-effective use of health care resources compared to the current standard of care in its licensed indication. Hence, economic evaluation involves the construction of an implicit or explicit economic model that compares a state of the world in which the novel intervention exists and is given to a cohort of indicated patients, to a state of the world where the novel intervention does not exist and standard treatments are received. Throughout this thesis I discuss the circumstance in which modelling is explicit. However, even in a system where economic evaluation is not conducted explicitly, whereby a group of decision-makers

assess cost-effectiveness implicitly, through committee, a key part of the evidence they will wish to incorporate is almost certain to be the treatment effect of the new drug on survival and quality of life.

When economic evaluations are designed to inform resource allocation decisions for a health system, it is desirable that the lifetime impacts of a new product on the entire disease population should be analysed in order to account for all costs and benefits associated with introducing the new technology. Therefore, economic evaluations typically take into account the total budget that will be required to treat patients with the disease in question,^{9;10} and similarly the treatment effect measure must reflect the average effect across the whole indicated population¹¹ – that is, mean (rather than median) costs and effects should be used in economic evaluations. A median cost or benefit represents the cost accrued or the benefit achieved by the “middle” patient – it is the value lying at the mid-point of a frequency distribution of values. It therefore does not take into account the shape of the distribution of values. This is particularly important for survival data because survival times often have markedly skewed distributions, and therefore the mean – which takes into account all values for all patients in its estimation of an “average” – is often very different to the median.¹¹ Incorporating lifetime impacts on the entire disease population is particularly important for interventions that affect OS.

Within economic evaluations, estimates of the effect of a novel drug compared to standard treatment are usually taken from Phase III RCTs designed to demonstrate efficacy and safety. An ITT analysis is commonly used, whereby intervention and control groups are compared as randomised – that is, a patient randomised to the experimental treatment is always included in this group in the effectiveness analysis. This is the case even if for some reason the patient does not go on to receive the experimental treatment, or if they quickly discontinue treatment. This is appropriate for the economic evaluation decision problem because such treatment discontinuation is likely to reflect what would happen if the intervention was introduced in the real-world – it represents the imperfect nature of the novel drug in real-world conditions and information on this is relevant for the economic evaluation decision problem.

1.2.3 Analysing survival data

Estimating mean treatment effects can be difficult. To illustrate with the example of OS in trials of treatments for metastatic cancers, it is rare for all patients to experience the event of interest (that is, death) before the end of the trial follow-up period. Hence, time-to-event data

for some patients are incomplete and an ITT analysis limited to the trial period would not adequately reflect the expected treatment benefit. Data are said to be censored for individuals when the end-point of interest has not been observed. In RCTs this is usually due to administrative censoring (that is, censoring that occurs when the event of interest has not been experienced at the end of trial follow-up) or because an individual has been lost to follow-up.¹² Censored data are partially but not completely known; it is known that the event has *not* been experienced up until a certain time-point. Thus censored data are different from data that are missing. It is primarily the existence of censoring that means that survival analysis techniques are required for the analysis of survival data, rather than more standard statistical techniques. Survival analysis techniques allow information from censored individuals to be incorporated in the analysis. If censoring is non-informative on the time scale – that is, it is random and not more likely to happen for one patient than another – the censoring will not bias the survival analysis. However, if censoring is informative – that is, a patient with specific characteristics is more likely to be censored than another (for example, a patient who has poor prognosis discontinues treatment and is censored because of this) – survival analysis will be biased. Informative censoring breaks the randomisation of the trial (because patients in different randomised groups are censored for different reasons, or censoring is related in some way to treatment received), and a comparison of randomised groups becomes prone to bias. Therefore, it is important to consider events that trigger censoring. Administrative censoring and censoring due to loss to follow-up are likely to be non-informative (although loss to follow-up is more questionable), and thus in general informative censoring is usually not a problem in well-designed RCTs. However, any type of censoring causes problems for an economic analysis because mean time to events of interest will not be known based upon the trial period alone.

The implication is that while *trial-based* estimates of outcomes that are completely observed may be appropriate for the economic evaluation, *extrapolation of the trial data* is likely to be required for an appropriate estimate of the treatment effect on outcomes that are subject to censoring. In a metastatic cancer trial an outcome such as PFS (that is, time to disease progression or death, where disease progression is typically defined as at least a 20% increase in the sum of diameters of target lesions, where that sum has increased by at least 5 millimetres, or the appearance of one or more new lesions¹³) may be close to completely observed, but OS may not be. The reasons for trial designs that allow this to happen are expanded upon in Section 1.2.7. For OS a restricted mean could be estimated based on the data observed up until the end of follow-up,¹⁴ but this would represent an underestimate of true mean survival and may not be reflective of the true survival benefit associated with the

novel treatment; hence the need to supplement the trial period ITT analysis with extrapolation for use in an economic evaluation. Extrapolation itself is difficult because it is not straightforward to identify the most appropriate extrapolation approach.

A variety of different parametric survival models are available for extrapolation purposes. These assume that survival times follow a specific underlying probability distribution. These models can be fitted to observed survival data but do not have to end at the final data point; based upon the fit to the observed data they continue to predict survival probabilities until there is zero probability of remaining alive (although occasionally some parametric models asymptote at a survival probability of greater than zero). Hence, these models can be used to extrapolate survival data observed in a clinical trial such that the entire survival distribution can be estimated, allowing an estimate of mean survival to be calculated. However, different parametric models are likely to be appropriate in different circumstances and there are a number of techniques that may be used to assess the fit of these models to identify the most appropriate model. If the approach to assessment of fit is not systematic there is a risk that inappropriate survival models will be chosen, which could lead to important inconsistencies between HTAs. It must be shown not only that the chosen model fits the observed data adequately (often referred to as “internal validity”), but also that the extrapolated portion of the survival curve is plausible (often referred to as “external validity”).¹⁵ The issues associated with extrapolation in the context of cost-effectiveness analyses have been expounded by a number of authors, but existing methods for assessing external validity in particular are under-developed and further research would be extremely valuable.^{11;15-19}

The need for extrapolation is inevitable, hence this is an area of central importance in the economic evaluation of cancer treatments. This is considered in more detail in Chapter 5 of this thesis. It is an area in which a substantial amount of research has been undertaken and is currently ongoing and I do not attempt to add to this – hence I do not review alternative extrapolation methods in exhaustive detail. However, the importance of extrapolation in combination with treatment crossover adjustment methods cannot be ignored, particularly given that this thesis primarily investigates crossover adjustment methods in the context of economic evaluation. Hence Chapter 5 provides an overview of commonly used extrapolation methods, and specifically considers how amenable crossover adjustment methods are to extrapolation.

1.2.4 Treatment crossover

In this thesis, treatment crossover is defined as the switch from control treatment to experimental treatment by patients randomised to the control group of an RCT. As defined here, treatment crossover does not involve experimental group patients switching onto the control treatment, or patients randomised to either group receiving other post-study treatments. The reason that these treatment changes are not included within the treatment crossover definition here is that they can both form part of a realistic treatment pathway that still allows an appraisal of the relevant decision problem. If an experimental group patient discontinues the novel therapy and receives a standard treatment (either that received in the control group or a separate standard treatment) this is likely to have occurred due to treatment failure, toxicity, tolerability, or adverse events. Such events and subsequent treatment switches are likely to occur in reality (and may have been driven by the study protocol which, post-study treatment, may attempt to mimic reality) and therefore they form a relevant part of the analysis of outcomes in the state of the world in which the new treatment exists. Hence, in general, we would not wish to adjust for these treatment changes in our economic analysis. Similarly, if control group patients receive post-study therapies that do not include the experimental treatment, this reflects a realistic treatment pathway and we would not wish to adjust for this in our economic analysis. Even if differential proportions of patients receive different post-study therapies this may reflect appropriate treatment pathways given the initial treatment. Unless this can be shown not to be the case, it would be inappropriate to adjust for these differences in the economic analysis.

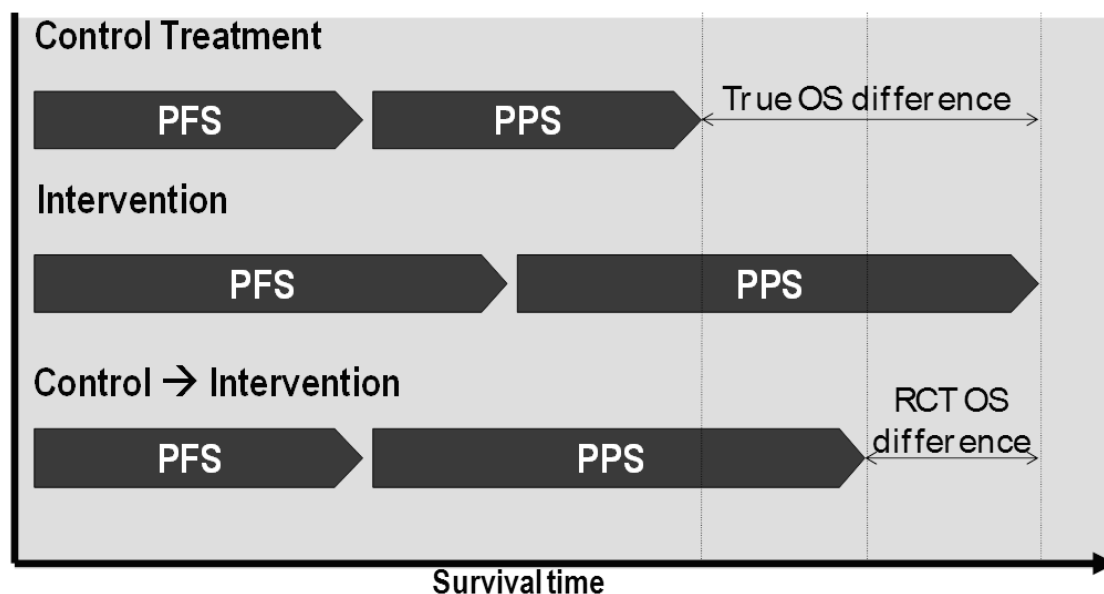
There may be circumstances whereby post-study treatments differ between treatment groups and it is possible to justify why such differences do not represent appropriate or realistic treatment pathways. In these circumstances methods may be used to adjust the analysis for incorporation into the economic model. For example, patients receiving particular post-study therapies might be censored and an inverse probability of censoring weights (IPCW) approach might be taken to adjust for selection bias. This method is discussed fully in Chapter 4 with respect to its use for adjusting for treatment crossover. However, methods to adjust for post-study therapies that do not fulfil the definition of treatment crossover given above are not considered any further.

Treatment crossover is an important problem for economists and decision-makers because, unlike many post-study treatment changes, it typically does not reflect a treatment pathway that is relevant for the decision problem defined in an HTA. Occasionally an economic evaluation might address early treatment with a novel therapy compared to later treatment, in which case treatment crossover may not cause a problem for the analyst. Usually this is not

the case – generally the decision problem defined within an economic evaluation addresses the costs and effects associated with a group of patients who are treated (or, who it is intended will be treated) with the new treatment, compared to the costs and effects associated with similar patients if they are not given the new treatment. When treatment crossover occurs a mismatch arises between what has been studied in the clinical trial and the economic analysts’ decision problem. We are trying to compare two states of the world, one in which the new treatment exists and one in which it does not. The comparator arm, which normally represents the “does not exist” state is contaminated in the presence of treatment crossover.

The bias that may be created by treatment crossover and the theoretical problems that it creates for the economic analysis are illustrated in Figure 1.1. The first two rows (“Control Treatment” and “Intervention”) illustrate the “perfect” trial, where no treatment crossover occurs. Survival time is on the x-axis, and in this example the new intervention extends PFS and also extends post progression survival (PPS). This results in the “True OS difference” identified in the diagram. In this case, an ITT analysis will give us the information that we need for our economic model (ignoring any need for extrapolation). However, the third row (“Control → Intervention”) demonstrates what may happen to survival in the control group if treatment crossover is permitted after disease progression (the reasons for this assumption are explained in Section 1.2.7). PFS is unchanged because crossover is only allowed after disease progression. However PPS is extended compared to the “Control Treatment” comparator, under the assumption that some control group patients cross over and benefit from the new intervention after disease progression. The result of this is that the OS difference observed in the RCT ITT analysis (labelled “RCT OS difference” in Figure 1.1) is smaller than the true OS difference that would have been observed if no treatment crossover had occurred. This demonstrates that a simple ITT analysis will result in bias equal to the difference between the true OS difference and the observed OS difference when treatment crossover occurs. The extent of this bias will be unknown, as the true OS difference will be unobserved. However it is clear that provided crossover patients benefit to any extent from the new intervention, some bias will exist. An economic evaluation that relied upon this ITT analysis would produce inaccurate cost-effectiveness results (in this case the ICER would be over-estimated) and inappropriate resource allocation decisions may be made.

Figure 1.1: The potential impact of treatment crossover illustrated



Notes: PFS = Progression Free Survival; PPS = Post Progression Survival; OS = Overall Survival; RCT = Randomised Controlled Trial

It is important to note what I mean by “bias” when I use the term in this thesis. Typically, a statistician may define a biased estimator as one for which the estimated and observed (“true”) values are systematically different. In the presence of treatment crossover, an ITT analysis will give an unbiased estimate of the treatment effect in groups as randomised but that treatment effect will include the effect of the experimental treatment in control group patients who crossed over. In this thesis I use the term “bias” more loosely, whereby bias exists when there is a systematic difference (error) between the estimated treatment effect and the “true” treatment effect where the “true” treatment effect is that which would have been observed if there had been no treatment crossover. Linked to this, throughout this thesis I refer to “counterfactual” survival times, which are survival times for crossover patients that would have been observed had treatment crossover not occurred.

1.2.5 Simple solutions?

Chapter 3 reviews approaches that have been taken to address treatment crossover in the context of economic evaluations of oncology treatments, and Chapter 4 conducts a further review of these and other methods identified from the literature. However, it is useful to introduce here why simple analytical solutions to the treatment crossover problem do not exist. Chapter 3 demonstrates that commonly used methods to address treatment crossover include censoring patients at the point at which they cross over, or excluding them from the analysis altogether.²⁰ These methods will be subject to selection bias if treatment crossover is

not random. If patients who cross over are chosen purely at random, excluding them from the trial analysis would not result in bias, although reducing the number of patients included in the control group reduces the power of the trial and increases the uncertainty in the results. On the other hand, if the treatment crossover decision is made based upon patient characteristics, excluding or censoring crossover patients represents informative censoring (that is, censoring that is not random) or exclusion and the randomised nature of the trial is broken. Randomised comparisons between treatment groups can no longer be made, and the resulting analysis is subject to selection bias. Economic evaluations that rely upon these methods will provide inaccurate results which may lead to inappropriate resource allocation decisions.

1.2.6 More complex solutions?

More complex solutions to the treatment crossover problem are reviewed in detail in Chapter 4. However, in order to demonstrate the theoretical motivations for this thesis it is useful to introduce key methods here, to demonstrate that there is no definitive answer as to which method is optimal. In a recent paper by Morden *et al* (2011),²¹ on which I am a co-author, we investigated the bias associated with methods for addressing treatment crossover. Simple censoring and exclusion of crossover patients were included in the study as well as more complex methods such as Robins and Tsiatis's Rank Preserving Structural Failure Time Model (RPSFTM) and Branson and Whitehead's Iterative Parameter Estimation (IPE) procedure.^{22;23} The study found that the RPSFTM and IPE methods produced very low bias in predicting true treatment effects in a simulation study in which treatment crossover was applied on a non-random basis. Conversely, the simple censoring and exclusion approaches often produced high bias. However, the study was limited in that it simulated data in such a way that satisfied the assumptions made by the RPSFTM and IPE methods. In particular, it was assumed that patients who crossed over received the same treatment effect (relative to the time for which they took the experimental treatment) as patients initially randomised to the experimental group. This is a key assumption of both the RPSFTM and IPE methods – referred to as the “common treatment effect” assumption – and may not be reasonable given that usually treatment crossover occurs once a patient's disease has progressed; the effect of the new drug might reasonably be expected to be different. Morden *et al* state that the most important limitation of their study was that they did not assess the performance of methods in scenarios that violate their assumptions.²¹ This limitation is addressed by the novel simulation study reported in Chapter 6 of this thesis.

A further limitation of the Morden *et al* study was that not all crossover methods were explored – a systematic search and review to identify methods and assess their relevance was

not undertaken. In particular, the inverse probability of censoring weights (IPCW) method, introduced by Robins and Finkelstein, was not included.²⁴ The IPCW method is explored in detail in Chapter 4 of this thesis, and is included in the simulation study presented in Chapter 6. While IPCW avoids the “common treatment effect” assumption, it has other obvious limitations. In particular, it is reliant on the ability to model the probability of crossover and therefore to avoid bias it requires data on all covariates (baseline and time-dependent) that may influence the crossover decision. This is referred to as the “no unmeasured confounders” assumption, and means that the method is heavily data dependent. It may also be prone to bias if the proportion of control group patients that cross over is extreme – if either very few patients do not cross over, or if very few patients do cross over, it may be difficult accurately to model the probability of crossover. Hence while the IPCW method may have some intuitive advantages over the RPSFTM and IPE methods, it is clear that it may result in important bias in some scenarios.

In addition to these problems, Chapter 4 shows that different crossover adjustment methods provide different statistical outputs – for instance, some produce adjusted hazard ratios (HR), while others produce acceleration factors (AF). A HR is the ratio of the hazard of the event of interest (such as death) at any time for an individual on the new treatment relative to an individual on the standard treatment – a HR of less than 1 indicates that the hazard is lower for the new treatment than the standard treatment.¹² Unlike the HR, an AF works on the time scale rather than the hazard scale. Rather than impacting upon the hazard of an event, it is assumed that the new intervention affects the rate at which an individual proceeds along the time axis. The event-time of an individual on the new treatment is \emptyset times the event-time that the individual would have experienced on the standard treatment, where \emptyset is the AF. If \emptyset is greater than 1 the new treatment extends the time to event.¹² The different interpretations and meanings of HR and AF treatment effect estimates have implications for how extrapolation can be undertaken. Further, some methods produce weighted Kaplan-Meier curves and others produce counterfactual datasets which can be used in different ways for extrapolation purposes. However, presently no research has investigated the alternative methods through which crossover adjustment methods can be combined with extrapolation approaches. Given the importance of extrapolation for economic evaluation, described in Section 1.2.3, this represents a key gap in the literature – Chapter 5 addresses this.

It can therefore be seen that current research on the bias associated with methods to address treatment crossover is of limited use. There is a need to assess the performance of methods in “realistic” scenarios – that is, where data reflect what is commonly seen in practice, rather

than what is required by the methods to produce low bias. There is also a need to bring together all relevant methods within one study to allow meaningful and informed comparisons, and to consider the combination of crossover adjustment methods and extrapolation approaches. This thesis seeks to further current knowledge by addressing these issues.

1.2.7 Applied context motivations

A huge amount of research undertaken by pharmaceutical companies is based upon the development of new cancer therapies. This is reflected by the high proportion of appraisals undertaken by bodies such as NICE that assess new cancer technologies. Chapter 3 reports that between the year 1999 (the inception of NICE) and the end of 2009, 184 technology appraisals were completed; of these, 45 (24%) were in the area of advanced or metastatic cancer. Treatment crossover is not limited to trials in this context, but it is in this context that crossover is most commonly observed. Crossover may occur for a number of reasons, both ethical and practical. Perhaps the most important reason explaining why treatment crossover is allowed to occur in oncology RCTs relates to the fact that new treatments can be licensed based upon PFS data – reducing the incentives to maintain randomisation beyond disease progression, as discussed in Section 1.2.1. This is associated with ethical reasons that provide incentives to allow crossover: when there are no other non-palliative treatments available for patients with late-stage cancer it may be deemed inappropriate to deny control group patients the new treatment if interim analyses indicate a positive treatment effect. Additionally and practically, including the possibility of treatment crossover within a trial protocol is likely to significantly help enrolment as patients (and their clinicians) know that they are likely to receive the novel treatment at some point whichever trial group they are randomised to.

Treatment crossover is likely to represent a continuing problem for health technology appraisal analysts and decision-makers. As described in Section 1.1 it is essential that OS benefits of new cancer products are estimated as accurately as possible in the context of economic evaluation. Given that treatment crossover confounds OS estimates and is certain to be a feature of at least some future clinical trials, there is a clear practical motivation for this thesis. This is supported by the interest of decision-makers and pharmaceutical companies alike in this work. Substantial parts of chapters of this thesis have been used in a NICE Decision Support Unit (DSU) Technical Support Document (TSD) that offers advice on methods for extrapolating survival data,¹⁵ and I have been asked to produce a further TSD specifically addressing the treatment crossover problem. In addition, GlaxoSmithKline (GSK) provided trial data that are used in Chapter 7 of this thesis, and the Pharmaceutical Oncology Initiative (POI)

– a group of pharmaceutical companies with a particular interest in oncology – provided valuable input on the design of the simulation study presented in Chapter 6.

1.2.8 Personal motivations

I have worked as a health economist for 8 years, both in academia and in industry. In that time I have completed several economic evaluations of interventions that impact upon survival. My time spent working for a pharmaceutical company exposed me to situations in which clinical trials of new cancer treatments appeared to show positive effects on PFS, leading to protocol amendments allowing treatment crossover. Often PFS represented the primary endpoint and so OS appeared of secondary importance with regard to the clinical data analysis. However, for the economic analysis, OS was critical yet there were no clear guidelines from HTA authorities on how best to extrapolate censored data, or how to adjust for treatment crossover. Since leaving industry I have continued to see these issues arise in the majority of economic evaluations that tackle an intervention that impacts upon survival. While I myself have written guidelines for the extrapolation of patient-level survival data,¹⁵ and a number of other groups continue to generate valuable research on this topic, no such guidance has appeared on treatment crossover adjustment methods. As will be shown in Chapter 3, naive methods are usually used to adjust for treatment crossover in economic evaluations, and although more complex methods have been used there is no guidance as to which of the more complex methods may be “best”. It is due to this obvious and important gap in the literature that I decided to embark upon the research project reported in this thesis.

1.3 Research questions

The primary aim of my research was to determine how an analyst (or a decision-maker) may decide which method is most appropriate for adjusting for treatment crossover, given a set of trial-specific characteristics and the decision problem. Several methodological issues (including crossover) with respect to the assessment of cancer treatments have been discussed in the literature and it has been suggested that the development of guidelines for cost-effectiveness modelling of cancer therapies would be an invaluable resource,²⁵ hence producing guidelines on the use of crossover adjustment methods in an economic evaluation context constituted an important goal, which has been achieved.

My research question therefore asked which methods for dealing with treatment crossover are most appropriate in a range of different (and relevant) scenarios, given that a state of the world in which the novel intervention exists must be compared to one in which it does not

exist. Linked to this is the hypothesis that the methods commonly used in HTA are likely to lead to biased estimates of the treatment effect and therefore biased cost-effectiveness results. To address this I attempted to identify all relevant statistical methods, and to appraise their assumptions and limitations and assess their performance in a simulation study. A simulation study was required because in order to measure the bias associated with different methods the “truth” must be known, and also because a simulation study allows many different scenarios to be considered. Applying the methods to individual trial datasets only informs us about the set of circumstances present in the trials for which data are obtained, and in these the “truth” is not known. Nevertheless, in Chapter 7 I test the application of identified methods to a real-world dataset, and this presents challenges unforeseen in the simulated analysis.

1.4 Thesis structure

The thesis is structured in five parts. Figure 1.2 illustrates these five parts and the chapters that they incorporate, the influences that the different parts and chapters have upon one another, and the overall structure of the thesis.

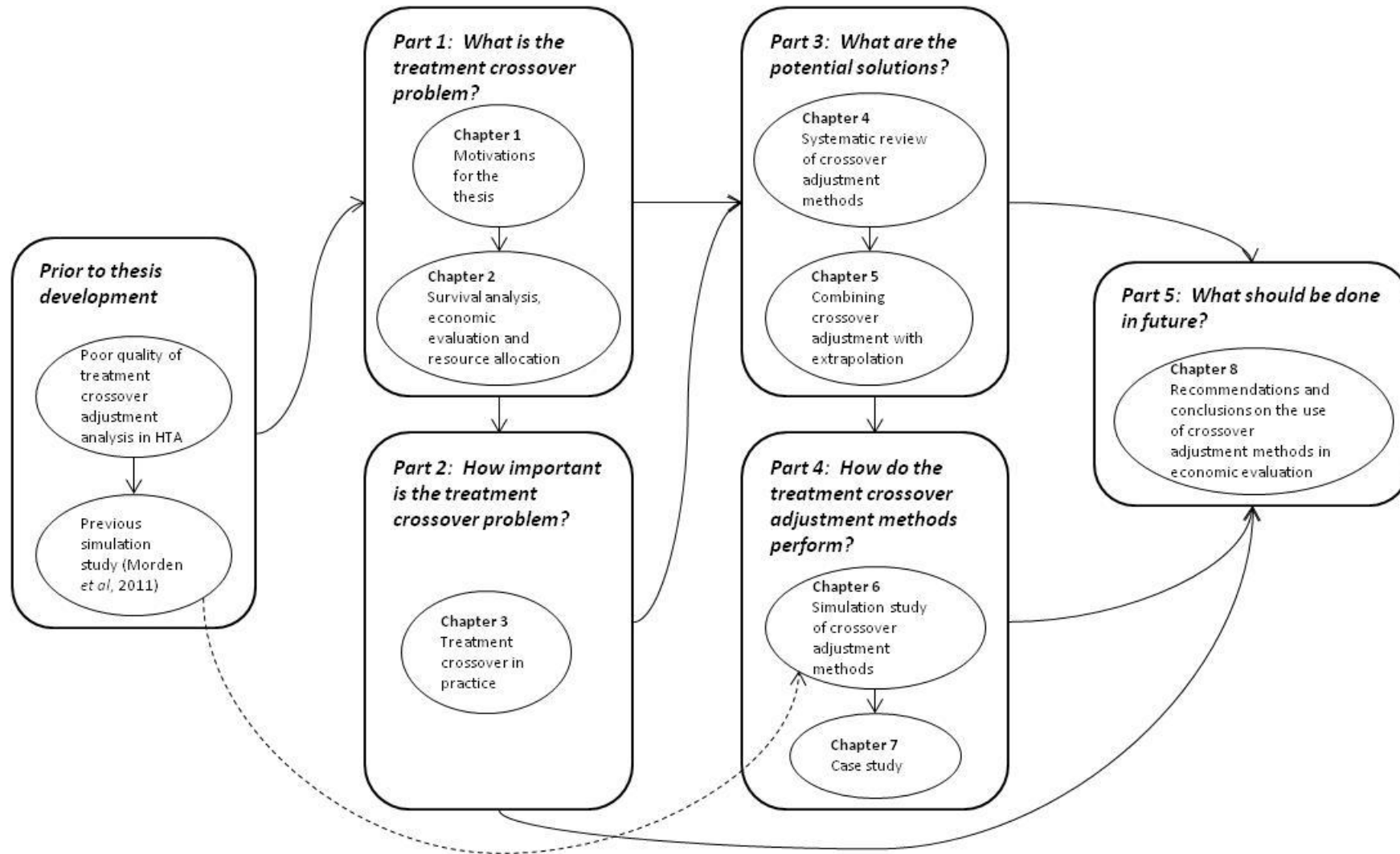
Chapters 1 and 2 form the first part, in which the scene is set for the thesis and the research question. The current chapter has summarised the theoretical and practical motivations behind the research, including a description of the treatment crossover problem and an explanation of the fundamental problems associated with commonly used methods for addressing the issue. In Chapter 2, I discuss the theoretical framework of health economic evaluation in order to set the scene for the importance of the research and to demonstrate that the research question remains important whatever theoretical perspective the economist takes.

The second part consists of Chapter 3 and involves an exploration of the extent of the treatment crossover problem by reviewing its frequency as an issue and the way in which it has been approached in HTAs completed by NICE. A selection of NICE technology appraisals are reviewed in order to identify methods that have been used to adjust survival estimates in the presence of treatment crossover in an applied economic evaluation context. The findings of this part are drawn upon in Part 5 when addressing the hypothesis that commonly used methods are likely to be associated with substantial bias. In circumstances where no action was taken to adjust for identified crossover in the reviewed technology appraisals, or where obviously biased methods were used, Part 2 alone provides evidence demonstrating the biases

associated with methods used for adjusting for treatment crossover in health technology assessments.

The third part consists of Chapters 4 and 5 and involves a review of crossover adjustment methods and their potential applications in economic analysis. The literature is systematically searched and reviewed in order to identify methods available for adjusting survival estimates in the presence of treatment crossover whether they have been used in practice or not. Occasionally, potentially relevant methods may have been developed for a slightly different context (for example, looking at treatment non-compliance rather than treatment crossover), hence the search is broad in scope but the review is specific in that it includes only methods that are directly applicable in a treatment crossover context. The difficulties associated with searching for methodological literature are addressed. Identified methods are considered with respect to their theoretical foundations, and also their relevance for an economic evaluation context. In Chapter 5 broad issues associated with incorporating survival analysis into economic evaluations are discussed, with reference to the treatment crossover methods identified in Chapter 4. Chapter 5 is important because the output derived from crossover methods has important implications for how these can be incorporated into an economic analysis which (as described in Section 1.2.3) is likely to require extrapolation.

Figure 1.2: Diagrammatic representation of the thesis structure



The fourth part of the thesis consists of Chapters 6 and 7 and involves testing the performance of identified crossover adjustment methods. Chapter 6 takes the methods identified in Part 3 and assesses them in a simulation study. The simulation study is carefully designed such that relevant scenarios are tested with the aim of identifying circumstances under which particular methods are preferable. Owing to the detailed description of the data generating methods used and the scenarios tested, and the large amount of results obtained, Chapter 6 is substantial. This level of detail is important for understanding the complexity of the simulation study, the potential impact of this on the crossover adjustment methods, and the performance of the methods across all scenarios. The simulation study builds on previously completed simulation studies in this area²¹ due mainly to the simulation of time-dependent treatment effects and the inclusion of “observational” methods for addressing the crossover problem. A pragmatic approach is taken – given the limitations associated with the identified crossover methods all are likely to result in bias in the majority of scenarios simulated, and data are simulated to reflect “real” data as closely as possible, rather than in a way that satisfies the assumptions of specific crossover methods. Hence it is accepted that a “perfect” method will not be identified. Despite this, from a practical perspective, decisions must be made about which methods to use to adjust for treatment crossover and therefore even if it is known that all methods are likely to result in bias, it is important to know which are likely to consistently lead to comparatively low bias across a range of scenarios. Chapter 7 involves an application of crossover methods to a real-world dataset. An extrapolation exercise is completed so that the potential impact on survival and cost-effectiveness estimates of using different crossover adjustment methods can be illustrated.

The fifth and final part of the thesis consists of Chapter 8 and uses the findings from the previous parts to make recommendations on the use of crossover adjustment methods. The theoretical and practical strengths and weaknesses of alternative crossover adjustment methods are summarised, and the hypothesis that commonly used methods to adjust for crossover are likely to lead to biased cost-effectiveness results is addressed. Where possible, methodological guidance is offered to advise which methods are likely to be most appropriate in different scenarios, thereby answering my research question. An analytic framework, which has been designed to promote the identification of the most appropriate analytical techniques, is presented for the situation where treatment crossover has occurred and an economic evaluation is required. Final conclusions are made, and priorities for future research are suggested.

Chapter 2

Survival analysis and the theoretical framework of economic evaluation

2.1 Chapter overview

In this chapter the theoretical foundations of health economic evaluation are briefly described. This thesis does not investigate the welfare theoretic underpinnings of economic evaluation and therefore this chapter does not represent an all-encompassing description and discussion of the theoretical framework – rather, its aim is to demonstrate that the importance of the research contained within this thesis is not dependent upon the theoretical framework within which economic evaluation is undertaken. Section 2.2 explains briefly why economic evaluation is needed in a society with scarce resources. Section 2.3 discusses the theoretical underpinnings of economic evaluation and uses NICE to illustrate how theory can be put into practice. Section 2.4 considers the implications of the theoretical framework for this thesis and Section 2.5 summarises and specifies how the thesis proceeds from this chapter onwards.

2.2 Allocating scarce resources

Economic evaluations of health technologies are regularly conducted throughout the world as governments and agencies seek to make informed decisions on whether new medicines should be reimbursed. For example, as of May 2012 NICE had completed 256 technology appraisals, mostly although not exclusively of drugs. Economic evaluation forms a central part of technology appraisal because health care resources are scarce compared to human wants, thus resources must be allocated between competing options, and people have differing preferences for different options.²⁶ Resources are scarce because health systems and people in general have limited budgets and cannot afford everything that they may wish to consume. In this situation decisions must be made about what should and what should not be recommended for funding by the health service. To inform these decisions economic evaluations are used because they allow the clinical effects of a new intervention to be weighed against the cost and other impacts. The theoretical basis of economic evaluation is explored in the Section 2.3, with particular reference to the implications for the valuation of survival benefits.

2.3 Theoretical underpinnings

The theoretical basis of economic evaluation has been explored by authors in Welfare Economics. It is impossible to do justice to the vast body of writing in this major area of study in a short chapter in this thesis but a brief overview is helpful. Welfare Economics addresses the issue of how society should allocate its resources to achieve a social optimum. Under Welfarism, the utility of individuals in a society is a function only of goods and services consumed. Social welfare is a function of aggregate individual utility, and individuals are envisioned as utility maximisers who are the best judges of their own utility.²⁷ Pareto Optimality, a position whereby none may be made better off (in their own judgement) without others being made worse off, is arrived at through trading within and between the consumption and production sectors.

Because the trades required to reach a Pareto Optimum require that at no point anyone is made worse off, different initial allocations of goods will lead to different Pareto optimal equilibria.²⁸ Hence whilst Pareto optimality involves an efficient allocation of resources it is silent on distributional issues. A Pareto optimal equilibrium might involve one individual experiencing very low welfare while another individual is much better off, but no re-allocation occurs because the better off individual would be made worse-off.

As virtually all real-world decisions are likely to involve gainers and losers this characteristic of Pareto optimality is practically restrictive. To address this Hicks (1939) proposed the Kaldor-Hicks efficiency criterion using Compensating Variation as a method for ranking resource allocations that are Pareto non-comparable – that is, where a reallocation of resources makes one individual better off but another worse off.^{28;29} Under this criterion a re-allocation of resources is deemed to represent an improvement for society if individuals that gain could, in principle, compensate those who lose and still remain better off themselves. An overall Pareto non-comparable reallocation of resources can thus be evaluated by assessing whether the sum of the Compensating Variations amongst all individuals is positive or negative. Compensating Variation requires that individuals place values (usually monetary) on resource allocation changes, and therefore on their changes in utility caused by the resource re-allocation. This idea – that individuals can reveal their preferences through revealing their monetary valuation of a resource allocation change – leads to a cost-benefit analysis (CBA) approach to allocating resources. Under a CBA approach all matters relevant to the comparison – that is, overall “utility” – between the baseline and alternative resource allocation scenarios must be evaluated in monetary terms by the affected individuals.³⁰

Under a CBA approach the monetary valuations of the impacts of different resource allocations must somehow be derived. In practice both direct questionnaires and indirect methods such as contingent valuation and conjoint analysis have been used to value benefits obtained due to health interventions.²⁸ When deriving these values the expected impact on health (and other dimensions) of the intervention must be known so that the value of this can be estimated. It is due to the requirement for these monetary valuations that CBA has been largely rejected in the health sphere.³¹ Deriving the monetary valuations is fraught with difficulty because individuals may not truthfully reveal their compensating value if there was reason to believe that the higher the value they report, the greater the compensation they would receive.³⁰ In addition, assessing net benefit based upon Compensating Variation implies that the marginal utility of income must be the same for all individuals, no matter their initial level of income.²⁸ This is problematic because additional or lost income may affect individuals in different ways – an extra £1 may make much more difference to a person with a low income compared to a person with a much higher income. Partially in response to this, the “Extra-Welfarist” approach to resource allocation was developed.³²

Extra-Welfarism involves the belief that health itself is the principal output of healthcare, and that it is perfectly acceptable for the analysis of the cost-effectiveness of healthcare interventions to focus upon health rather than upon the maximisation of overall welfare (utility).^{28;32;33} It is argued that health is an important independent argument within the social welfare function, and that a commensurate and quantifiable measure of health benefit could be used for interpersonal comparison to evaluate the effectiveness of healthcare interventions.³² Hence, the focus within an Extra-Welfarist framework is to concentrate upon health as a maximand in the social welfare function. This creates problems from the perspective of traditional Welfarist resource allocation because when additional arguments other than individual utility are included in the social welfare function the consistency between welfare maximisation and Pareto optimality generally will not hold.²⁸ If welfare depends upon utility and upon health, the process of achieving Pareto optimality in consumption and production while also maximising the welfare function becomes much more complex. Pareto optimality may maximise production and consumption but may not maximise a social welfare function that includes utility *and* health. This reflects the view that economic evaluation according to an Extra-Welfarist perspective does not necessarily entail the adoption of a societal welfare perspective because the focus is upon health, rather than utility.³¹ Extra-Welfarism is also seen to permit the use of sources of valuation other than that of the affected individuals, allows the weighting of outcomes according to principles that may not be

preference based, and allows interpersonal comparisons of wellbeing.³¹ For these reasons, Extra-Welfarism may be described as a pragmatic “decision-aiding” approach that helps decision-makers make resource allocation decisions within a framework that allows for distributional concerns.³¹

Extra-Welfarism lends itself to a cost-effectiveness analysis (CEA) type of economic evaluation. CEA allows the use of any measure of effectiveness, and the results of these analyses are reported in terms of the incremental cost per unit of effectiveness gained or event avoided. For example, if the intervention under consideration reduced the probability of a hip replacement being required, the CEA result may be reported in terms of the cost per hip replacement avoided. While CEA falls within the umbrella of Extra-Welfarism it is likely to be inadequate in helping decision-makers because it does not, in general, encompass a broad enough scope of health benefits. Measures of health effect that are not comparable across disease areas do not helpfully inform decisions around the allocation of resources across the health system as a whole.

For decision-making, a more helpful type of CEA is cost-utility analysis (CUA). CUA involves estimating the utility associated with the implementation of a new treatment, where utility is cardinally measured as an indicator of health-related quality of life (HRQL). In some circumstances, whereby the measure of utility includes all relevant terms in the utility function, including non-health and distributional considerations sufficient to create a cardinal scale over all human experience, it may be argued that CUA could be undertaken within a traditional Welfarist framework.³⁴ However, measures of utility within the Pareto optimising Welfare Economics were generally thought of as ordinal and interpersonal comparisons were not possible, thus the reliance on the Pareto Principle for determining preferable re-allocations of resources.³¹ Within a CUA the measure of utility must be cardinal such that numerical values for the outcome can be used within the analysis. Typically it is accepted that utility as measured within a CUA reflects only HRQL rather than all relevant terms within the utility function, and therefore an Extra-Welfarist approach is taken. CUA is potentially much more helpful than CEA for informing resource allocation decisions because the health impact of a new intervention is measured using a generic rather than a disease-specific outcome, and therefore the cost-effectiveness of interventions for different diseases can be compared.

In practice HRQL is in some jurisdictions estimated using the quality adjusted life year (QALY) gain associated with the new treatment.^{35;36} The QALY takes into account both the quality and length of life associated with different treatments. There is debate surrounding whether the

QALY reflects individuals' (and society's) utilities and preferences associated with different health states, or whether it is a measure of health alone, rather than wider utility.³⁴ Typically it is accepted as a measure of health which remains useful for aiding decision-makers within an Extra-Welfarist decision-making framework.

In the UK, NICE represents a prime example of the application of developed rules applying the Extra-Welfarist approach in order to inform decision making on the allocation of health care resources. NICE was established in 1999 as a Special Health Authority (SHA) in order to assess the clinical and cost-effectiveness of new interventions, with an important objective of reducing variation in care across the UK.^{37;38} NICE has several strands and produces public health guidance, clinical guidelines on diseases, technology appraisals, guidance on interventional procedures and guidance on medical technologies (devices and diagnostics). Its greatest influence may lie in its Technology Appraisals Programme since only these involve mandatory funding requirements that must be implemented within the NHS.

Both single and multiple technology appraisals are undertaken (STA and MTA respectively). STAs involve the appraisal of a single novel intervention compared to relevant alternatives, whereas MTAs assess multiple novel therapies within a single appraisal. Within STAs the manufacturer of the intervention provides a submission to NICE on the clinical and cost-effectiveness of the novel therapy and this is reviewed by an independent academic group. The evidence is then assessed by an Appraisal Committee and a decision is made as to whether the intervention should be recommended for use within the NHS, recommended with restrictions or only in research, or not recommended. A similar process is followed within an MTA, except that rather than simply reviewing the analysis undertaken by the manufacturer, the independent academic group constructs its own economic model and provides its own analyses. NICE aims to produce guidance approximately 9 months after initiating an STA, and approximately 12 months after initiating an MTA.^{39;40}

NICE has produced a "Guide to the Methods of Technology Appraisal", which is currently under review.³ The guide includes a "Reference Case" describing which methods should be used when submitting evidence to NICE, in order to promote consistency across appraisals. The Reference Case includes directives on how the economic evaluation should be conducted.³ Importantly, NICE state that cost-effectiveness analysis should be undertaken using QALYs as the measure of health effects (thus the type of economic evaluation might be more accurately described as cost-utility analysis), hence the approach is clearly Extra-Welfarist.

NICE states that below a most plausible ICER of £20,000 per QALY gained treatment recommendations are likely to be based upon the cost-effectiveness estimate.³ At this level of ICER, the intervention is likely to be recommended. As the ICER increases in the £20,000 to £30,000 range the Appraisal Committee will increasingly rely on three key factors: the degree of certainty around the ICER (a more certain ICER is preferable); whether there are strong reasons to suggest that the assessment of change in HRQL has been inadequately captured, and; whether the innovative nature of the technology offers benefits that are inadequately captured in the QALY measure. With ICERs of above £30,000 stronger arguments with respect to these three factors are required in order for a positive recommendation to be made.³ In addition, in 2009 NICE provided supplementary advice to its Appraisal Committees indicating that higher weightings for QALYs gained in “End-of-Life” circumstances may be appropriate. Interventions that fulfil the specified criteria may receive positive recommendations even if they have incremental cost-effectiveness ratios (ICERs) that are higher (less favourable) than the range in which an intervention may usually be considered cost-effective.⁴¹ Therefore, while NICE appears to generally rely upon a decision rule based upon a cost-effectiveness threshold, there does appear to be scope for distributional judgements to be made based upon other factors.

NICE’s use of a cost-effectiveness threshold implicitly means that it is assumed that recommending interventions with an ICER of less than £20,000 to £30,000 will result in a re-allocation of resources that will increase the total number of QALYs obtained using the health care budget. To achieve this, interventions that are stopped in order to finance newly recommended interventions (assuming that there is a limited health care budget and thus re-allocation is required) should have ICERs of greater than £20,000 to £30,000. However, in practice the interventions that are displaced – and thus the true opportunity cost of implementing the new intervention – are unknown.⁴² The approach of using a cost-effectiveness threshold decision rule without specifically considering disinvestment and opportunity cost has been criticised, with commentators arguing that the basis of such economic evaluation is not an economic one.⁴³⁻⁴⁷ However, NICE attempts to apply Extra-Welfarist resource allocation theory in a practical way, and it may be regarded as impractical to attempt to assess the ICERs of all health care interventions and programs funded by the NHS – which is essentially what would be required if the absolutely optimal allocation of resources were to be identified.

This highly developed quantitative approach to assessing the cost-effectiveness of new interventions means that NICE furnishes examples of analysis in which it is possible to

demonstrate quantitatively the consequences of choosing one approach to correcting for crossover rather than another. Examples given in this thesis have therefore been taken from NICE. Nevertheless, the implications of treatment crossover are similarly important when different approaches to assessing cost-effectiveness are taken.

2.4 Implications for this thesis

It is apparent that, whether the approach to determining the value of a new intervention is CBA under a Welfarist perspective, or CUA under an Extra-Welfarist perspective, the need for accurate estimation of the survival impact of new cancer therapies is unaffected. Both approaches require accurate estimation of the health effect of the new intervention being appraised. An important part of the treatment effect relates to survival. No matter how extensions in life are valued, the extension in life itself is definitive and separate from other aspects of the total benefit. Even if different people value life years differently, or if individuals value life years differently at different times in life, an accurate estimate of the number of life years gained by an intervention is required before any value can be applied to this. For this thesis to be important it matters, irrespective of the approach to valuation and decision making taken, that survival is an important argument in the decision function. This is likely to be generally the case.

It is therefore clear that accurately estimating the survival benefit of a novel drug compared to an existing standard is of paramount importance in economic evaluation. This is made more difficult in the presence of treatment crossover. Williams (1997) states that a CBA can only be successfully undertaken if (amongst other requirements) it is possible to separate one service from another in a sensible way, there is the possibility of choice between them, and if it is possible to estimate and value the outcomes and costs associated with each service.⁴⁸ These requirements also hold for a CUA and a CEA, irrespective of the theoretical framework perspective. As stated in Chapter 1, in the instance of treatment crossover an RCT no longer separates comparators and does not allow a simple analysis of comparative effectiveness. In these circumstances an ITT analysis of the RCT will not provide evidence that is relevant for the decision problem and suitable crossover adjustment methods are required.

2.5 Conclusions

Chapters 1 and 2, which form Part 1 of this thesis, have set the scene for the remainder of the thesis by describing the treatment crossover problem and why it is important both

theoretically and practically. In particular, Chapter 2 has explained why treatment crossover represents an important problem for health economists and decision-makers irrespective of the specific theoretical perspective taken to economic evaluation. Simple and more complex methods for addressing the treatment crossover problem have been introduced, illustrating that the research contained in this thesis is needed due to limitations associated with existing research.

In Part 2 of this thesis (Chapter 3) I move on to consider the importance of the treatment crossover problem by reviewing a sample of technology appraisals completed by NICE. This serves to illustrate the frequency with which the crossover problem arises, and also allows methods that are commonly used to address the crossover problem to be identified, and what impact the application of these methods has on the results of economic evaluations. The rest of the thesis is dedicated to identifying all relevant methods for addressing treatment crossover and assessing their usefulness in an economic modelling context (Part 3), evaluating their performance in a range of scenarios (Part 4), and preparing recommendations on best practice analytical approaches to address the treatment crossover problem given an economic evaluation context (Part 5).

Chapter 3

Treatment crossover in a practical, health technology assessment context

3.1 Chapter overview

The purpose of this chapter is to explore the prevalence of treatment crossover issues in oncology trials, to examine how these are usually dealt with in a health technology assessment context and, where possible, to assess the impact of treatment crossover adjustment methods on the results of economic evaluations. This is achieved by reviewing technology appraisals (TAs) conducted by the National Institute for Health and Clinical Excellence (NICE). This is important, and novel, because previously such systematic evidence on the impact of treatment crossover in HTAs has not been available.

Section 3.2 introduces the review and Section 3.3 presents the search strategy used. Section 3.4 summarises the review findings before Section 3.5 details the prevalence of crossover identified by the review. Section 3.6 describes methods used to adjust for crossover, providing examples of TAs in which each method was used. Some of the simpler, “naive” methods are evaluated because they are not novel and do not require a great deal of discussion. However, the more complex methods (such as the RPSFTM and IPCW methods) that provide the main focus of this thesis are not reviewed in detail here. These methods are considered in much more detail in Chapter 4, which presents a systematic review of methods that may be used to address the treatment crossover problem. Section 3.7 considers the potential impact of treatment crossover on recommendations made by NICE. In Section 3.8 the review is updated for two TAs that involved the use of complex treatment crossover adjustment methods, that were completed after the review cut-off date. Given the relative lack of use of such methods, it was helpful to complete this update in order to demonstrate the recent drive toward more complex methods. The review findings are discussed in Section 3.9, before conclusions and the implications of these for the remainder of the thesis are outlined in Section 3.10. Evidence tables and more details on each appraisal reviewed are presented in Appendices 1, 2 and 3.

3.2 Introduction

NICE appraisals were chosen as the source of the reviewed HTAs as these were deemed to provide a good representation of technologies that have been assessed for use in health care systems in recent years, and due to the ease with which detailed appraisal documents can be accessed. All appraisals that involved assessments of interventions for advanced and/or metastatic cancer were analysed in order to identify those that were affected by treatment crossover. The methods used to address treatment crossover were reviewed, and where possible the impact of the methods on the results of the economic evaluation were reported. There were three primary aims associated with this task. The first aim was to assess how often treatment crossover affected pivotal clinical trials of new cancer interventions. The second aim was to assess the consistency of methods used in different appraisals and their potential biases. The third aim was to demonstrate the impact that the application of methods to address treatment crossover can have on cost-effectiveness results. The overarching objective associated with these aims was to demonstrate the importance of the treatment crossover problem – associated with how often crossover occurred, what impact it may have had on treatment recommendations, and how inconsistently and inappropriately it had been addressed.

3.3 Search strategy

All NICE TAs completed by December 2009 were screened for inclusion. The following inclusion and exclusion criteria were applied:

Inclusion criteria

- Appraisals must be complete (with guidance issued)
- Appraisals must assess an intervention for advanced and/or metastatic cancer, or for all stages of cancer

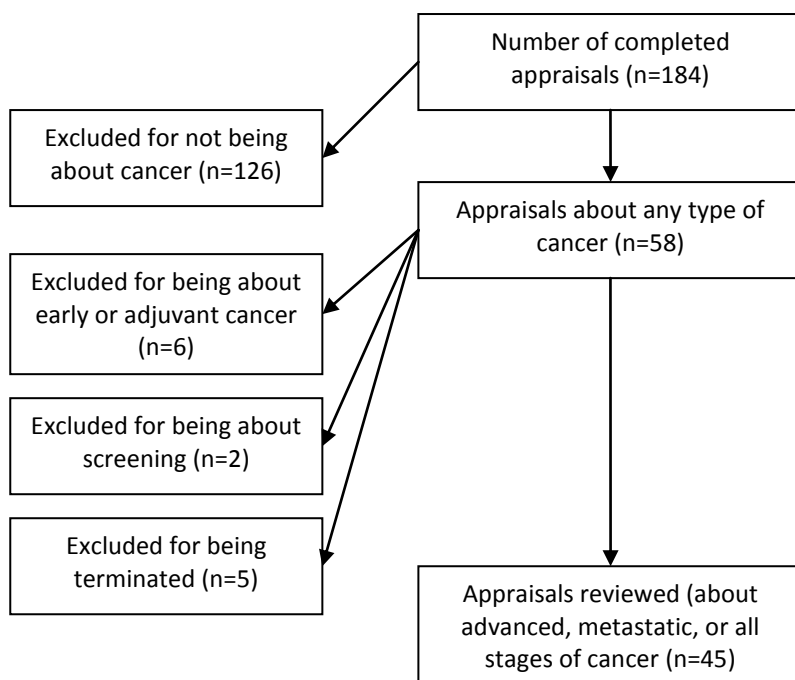
Exclusion criteria

- Appraisals which do not assess an intervention in the cancer setting
- Appraisals which assess an intervention for early or adjuvant cancer
- Appraisals which assess a screening intervention
- Appraisals which were terminated (thus no economic analysis was undertaken)

184 NICE TAs had been completed, 45 of which met inclusion criteria and were reviewed. For included appraisals, the assessment reports developed by the independent assessment group or Evidence Review Group (ERG) provided the primary source of evidence, but all other

relevant documents available on the NICE website were also reviewed. These included manufacturer submissions, final appraisal determinations (FADs), appeal documents, Decision Support Unit (DSU) reports, and documents containing updated analyses. On a number of occasions (particularly in earlier appraisals) the review of the analysis conducted was restricted by commercial-in-confidence data being removed from reports, a lack of availability of the manufacturer submission, or a lack of detail in the reporting of the methods used. Linked to this, due to the absence of any scenario or sensitivity analysis it was often not possible to identify the impact of the method used to adjust for treatment crossover on the cost-effectiveness results. Figure 3.1 shows the results of the search.

Figure 3.1: NICE technology appraisals search results



The included appraisals are listed in table 3.1.

Table 3.1: NICE technology appraisals included in the review

Ref	Title	Date Issued
TA3	Ovarian cancer - taxanes (replaced by TA55)	May 2000
TA6	Breast cancer - taxanes (replaced by TA30)	Jun 2000
TA23	Brain cancer - temozolomide	Apr 2001
TA25	Pancreatic cancer - gemcitabine	May 2001
TA26	Lung cancer - docetaxel, paclitaxel, gemcitabine and vinorelbine (updated by and incorporated into CG24 Lung cancer)	Jun 2001
TA28	Ovarian cancer - topotecan (replaced by TA91)	Jul 2001
TA29	Leukaemia (lymphocytic) - fludarabine (replaced by TA119)	Sep 2001
TA30	Breast cancer - taxanes (review)(replaced by CG81)	Sep 2001
TA34	Breast cancer - trastuzumab	Mar 2002

TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (replaced by TA93)	Mar 2002
TA37	Lymphoma (follicular non-Hodgkin's) - rituximab (replaced by TA137)	Mar 2002
TA45	Ovarian cancer (advanced) - pegylated liposomal doxorubicin hydrochloride (replaced by TA91)	Jul 2002
TA50	Leukaemia (chronic myeloid) - imatinib (replaced by TA70)	Oct 2002
TA54	Breast cancer - vinorelbine (replaced by CG81)	Dec 2002
TA55	Ovarian cancer - paclitaxel (review)	Jan 2003
TA62	Breast cancer - capecitabine (replaced by CG81)	May 2003
TA61	Colorectal cancer - capecitabine and tegafur uracil	May 2003
TA65	Non-Hodgkin's lymphoma - rituximab	Sep 2003
TA70	Leukaemia (chronic myeloid) - imatinib	Oct 2003
TA86	Gastro-intestinal stromal tumours (GIST) - imatinib	Oct 2004
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review)	May 2005
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (review)	Aug 2005
TA101	Prostate cancer (hormone-refractory) - docetaxel	Jun 2006
TA105	Colorectal cancer - laparoscopic surgery (review)	Aug 2006
TA110	Follicular lymphoma - rituximab	Sep 2006
TA116	Breast cancer - gemcitabine	Jan 2007
TA118	Colorectal cancer (metastatic) - bevacizumab & cetuximab	Jan 2007
TA119	Leukaemia (lymphocytic) - fludarabine	Feb 2007
TA121	Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide	Jun 2007
TA124	Lung cancer (non-small-cell) - pemetrexed	Aug 2007
TA129	Multiple myeloma - bortezomib	Oct 2007
TA135	Mesothelioma - pemetrexed disodium	Jan 2008
TA137	Lymphoma (follicular non-Hodgkin's) - rituximab	Feb 2008
TA145	Head and neck cancer - cetuximab	Jun 2008
TA162	Lung cancer (non-small-cell) - erlotinib	Nov 2008
TA169	Renal cell carcinoma - sunitinib	Mar 2009
TA171	Multiple myeloma - lenalidomide	Jun 2009
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab	Jun 2009
TA174	Leukaemia (chronic lymphocytic, first line) - rituximab	Jul 2009
TA178	Renal cell carcinoma	Aug 2009
TA176	Colorectal cancer (first line) - cetuximab	Aug 2009
TA179	Gastrointestinal stromal tumours (GIST) - sunitinib	Sep 2009
TA181	Lung cancer (non-small cell, first line treatment) - pemetrexed	Sep 2009
TA183	Cervical cancer (recurrent) - topotecan	Oct 2009
TA184	Lung cancer (small-cell) - topotecan	Nov 2009

Data on the survival analysis methods used as well as any treatment crossover issues were extracted from the 45 included appraisals. The review was focused primarily on treatment crossover issues, in particular:

- i. Whether treatment crossover was identified as being an issue
- ii. Whether any adjustments were made to the analysis to address treatment crossover

- iii. What impact adjustments made to address treatment crossover had on cost-effectiveness results

Alongside these concerns, the review also examined survival analysis techniques used in the economic evaluation sections of the TAs, independent of whether treatment crossover was an issue. The purpose of this was to provide evidence on commonly used extrapolation techniques in the context of economic analyses. Findings on these are not reported in this chapter, but are discussed in Chapter 5 which addresses the importance of extrapolation in the context of economic evaluation, with particular reference to the outputs provided by methods that may be used to address treatment crossover. This section of the review focussed on the following questions:

- i. Were mean survival times or medians used in the economic analysis?
- ii. How was mean survival estimated (for example, through a restricted mean approach or through extrapolation)?
- iii. How was extrapolation performed?
- iv. How were methods used for the survival analysis rationalised or justified?

Comprehensive details on both extrapolation and crossover adjustment techniques included in each appraisal are given in Appendix 1. The key details from each appraisal are summarised more briefly in Appendix 2. Using tick-boxes Appendix 3 summarises for each appraisal whether treatment crossover was an issue, whether and how it was addressed, and the positivity or otherwise of associated recommendations made by NICE.

3.4 Summary of review findings

Methods used to address treatment crossover in NICE TAs have become more advanced over time. In earlier appraisals issues associated with treatment crossover were often not acknowledged at all (although there were exceptions to this – for instance crossover was recognised as an important issue in TA28 (topotecan for ovarian cancer),⁴⁹ TA33 (irinotecan, oxaliplatin and raltitrexed for advanced colorectal cancer),⁵⁰ and TA34 (trastuzumab for breast cancer),⁵¹ which were completed in 2001 and 2002). More complex methods to address treatment crossover have been used in recent times – as illustrated by the use of the Rank Preserving Structural Failure Time Model (RPSFTM) in TA179 (sunitinib for GIST, completed in September 2009).⁵² A partial update to the review, presented in Section 3.8, further demonstrates the tendency towards the use of complex methods to address treatment crossover in recent appraisals. Both RPSFTM and IPCW methods were used in TA215

(pazopanib for the first line treatment of metastatic renal cell carcinoma (RCC)) and TA219 (everolimus for the second-line treatment of advanced RCC).^{53;54}

A number of alternative methods have been used to account for treatment crossover. However those that have been used most regularly – censoring or excluding crossover patients (which were used in TAs 28, 34, 70, 86, 129, 169, 178, 179)^{49;51;52;55-60} – are very likely to be associated with selection bias. Where more complex methods such as RPSFTM and IPCW have been used, the key questions regarding their suitability have not been addressed in detail – chosen methods have not been well justified and potential biases have not been investigated. For instance, in TA179 (sunitinib for GIST), it was recognised that an important weakness of the RPSFTM method is its reliance on a “common treatment effect” assumption (whereby the treatment effect received is the same no matter when the treatment is taken (relative to the time for which the treatment is taken)), but the implications of this assumption and potential alternative methodologies were not considered.⁶¹ Similarly, in TA215 and TA219 the “no unmeasured confounders” assumption (that is, that data are available on all key prognostic covariates) was identified as a key weakness of the IPCW method, resulting in an RPSFTM analysis being preferred.^{53;54} However the plausibility of this assumption was not investigated, or placed in context compared to the assumptions associated with the RPSFTM. The “no unobserved confounders” and “common treatment effect” assumptions are discussed in more detail in Chapter 4, particularly in Sections 4.10.1 and 4.10.2.

The review demonstrates that treatment crossover issues are prevalent in HTAs of cancer treatments. Further, the review shows that adjusting survival estimates in the presence of treatment crossover can have a significant effect on incremental cost-effectiveness ratios (ICERs) that may affect treatment recommendations – for example ICERs were more than halved when crossover was addressed in TA171 (which used external data to adjust for crossover),⁶² TA179 (which used the RPSFTM method),⁶³ and TA86 (in which crossover patients were excluded from the analysis).⁵⁷ Given the range of methods used to adjust for treatment crossover there are important inconsistencies between NICE appraisals of advanced and/or metastatic cancer treatments, which may have led to inconsistent and sub-optimal resource allocation decisions. There is a risk that inconsistent treatment recommendations have been made in the past, and may continue to be made in the future. There is a clear need for methodological guidance describing how treatment crossover can be adjusted for in a more systematic way, allowing more robust and consistent evaluations. Providing such guidance is a key objective of this thesis.

3.5 The prevalence of crossover

The prevalence of treatment crossover in the reviewed TAs was identified through an analysis of the details provided on the protocols of the key clinical trials. Treatment crossover was identified as certainly being an issue in 25 (55.6%) of the TAs. These are listed in Table 3.2. In these TAs treatment crossover occurred in at least one of the key clinical trials of the novel intervention(s) being appraised.

Table 3.2: NICE technology appraisals that involved treatment crossover

Ref	Title	Crossover
TA3	Ovarian cancer - taxanes (replaced by TA55)	✓
TA28	Ovarian cancer - topotecan (replaced by TA91)	✓
TA34	Breast cancer - trastuzumab	✓
TA55	Ovarian cancer - paclitaxel (review)	✓
TA70	Leukaemia (chronic myeloid) - imatinib	✓
TA86	Gastrointestinal stromal tumours (GIST) - imatinib	✓
TA179	Gastrointestinal stromal tumours - sunitinib	✓
TA6	Breast cancer - taxanes (replaced by TA30)	✓
TA30	Breast cancer - taxanes (review)(replaced by CG81)	✓
TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (replaced by TA93)	✓
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review)	✓
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (review)	✓
TA101	Prostate cancer (hormone-refractory) - docetaxel	✓
TA118	Colorectal cancer (metastatic) - bevacizumab & cetuximab	✓
TA119	Leukaemia (lymphocytic) - fludarabine	✓
TA121	Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide	✓
TA124	Lung cancer (non-small-cell) - pemetrexed	✓
TA129	Multiple myeloma - bortezomib	✓
TA169	Renal cell carcinoma - sunitinib	✓
TA178	Renal cell carcinoma	✓
TA171	Multiple myeloma - lenalidomide	✓
TA176	Colorectal cancer (first line) - cetuximab	✓
TA116	Breast cancer - gemcitabine	✓
TA162	Lung cancer (non-small-cell) - erlotinib	✓
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab	✓

The review demonstrated that the treatment crossover issue is often complicated further by post-progression treatment other than crossover treatment – in fact, in 15 of the 25 TAs affected by treatment crossover issues with other post-progression treatments were also identified. As stated in Chapter 1, the focus of this thesis is specifically on treatment crossover, rather than other post-study treatments. However, in some instances post-study

treatments were considered in tandem with treatment crossover in the reviewed TAs, and thus steps taken to adjust for crossover together with other post-study treatments are reported in this chapter.

3.6 Methods used to adjust for crossover

Of the 25 appraisals in which treatment crossover was identified as an issue, some adjustments were made to the economic analysis to account for this in 18 cases. However, in general the methods used could be described as “naïve” (defined based upon the complexity of the statistical approach taken to adjust specifically for treatment switching) and open to bias. A summary of the methods used to account for treatment crossover are presented in Table 3.3.

Table 3.3: Methods used to account for crossover in NICE technology appraisals

Method	TAs which use method
“Naïve” methods	
Censored patients	6 (TAs 28, 86, 129, 169, 178, 179)
Excluded patients	5 (TAs 34, 70, 86, 169, 178)
Included costs of post-progression treatments	4 (TAs 101, 116, 118, 121)
Modelled based on PFS, not OS	2 (TAs 6, 33)
Used sequencing models	2 (TAs 93, 176)
Applied the same risk of death upon disease progression	1 (TA 118)
Assumed equal OS for the two treatment groups	1 (TA 119)
More complex methods	
Rank preserving structural failure time model (RPSFTM)	1 (TA 179)
Adjusted survival estimates using a case-mix approach	1 (TA 34)
Used external data	1 (TA 171)

Often the methods listed in Table 3.3 above were not used specifically to account for treatment crossover – for example the sequencing models used in TAs 93 and 176 were perceived to provide the best fit to the disease pathway.^{64;65} Frequently methods such as basing models on progression-free survival (PFS) rather than overall survival (OS), assuming the same risk of death upon disease progression, or assuming equal OS for the two treatment groups were used primarily due to a lack of OS data, rather than specifically because of treatment crossover. However, each of the above methods in some way addresses the crossover problem and therefore will be considered below.

3.6.1 Naïve Methods

3.6.1.1 Censoring and excluding patient groups

In several TAs censoring and exclusion techniques were used to address the treatment crossover problem. As noted in Chapter 1, if censoring or exclusion of trial participants is related to prognosis or endpoints (such as PFS and OS) in any way, selection bias will result. Censoring allows information to be used for each patient up until the point at which the event which triggers the censoring occurs, and was the most commonly used approach for adjusting for treatment crossover, being used in 6 of the reviewed TAs. Completely excluding data from patients who crossed over was the second most common method, being used in 5 TAs.

It seems likely that exclusion will result in more bias than censoring, because exclusion involves using none of the data from the excluded group in the analysis, whereas censoring techniques at least involve using data from the censored groups up until the censoring time-point (usually the point at which treatment crossover occurs). Indeed, one of NICE's Appraisal Committees expressed a preference for an analysis based on a censoring approach rather than an exclusion approach.⁶⁶ However there is not a clear consensus on this – in TA169 the Appraisal Committee preferred an exclusion approach to a censoring approach, demonstrating inconsistent preferences across TAs.⁶⁰

In the TA of topotecan for ovarian cancer (TA28) the key clinical trial was a crossover trial whereby patients with stable disease after 6 courses of treatment crossed over to the alternative treatment (topotecan if the patient began on the control treatment (paclitaxel), and paclitaxel if the patient started on topotecan).⁴⁹ A censoring approach was used in an attempt to correct for the crossover. It seems likely that both PFS and OS estimates of survival will have been biased due to this censoring. Prolonged stable disease is highly likely to be linked both to PFS and OS, and therefore censoring patients based on treatment crossover which is directly related to prolonged stable disease is highly likely to cause bias in survival estimates and associated cost-effectiveness results.

In TA34 (trastuzumab for breast cancer) crossover was a key issue in the clinical trial, as a large majority of patients in the control group switched on to trastuzumab following disease progression. In response to this the manufacturer initially presented an economic evaluation which only used data from patients who did not cross over – crossover patients were excluded from the analysis. This resulted in a substantial change in the survival gain estimated for trastuzumab, with the median gain increasing by 10.9 months compared to the ITT analysis (17.9 month median gain in the adjusted analysis compared to 7.0 months in the ITT analysis).⁵¹ The assessment group considered that this approach was very likely to result in biased survival estimates, since the exclusion analysis was based on very low patient numbers

(approximately 35).⁵¹ Due to this the manufacturer recalculated their survival estimates based on a casemix extrapolation approach,⁶⁷ which will be discussed in more detail in Section 3.6.2.2 of this chapter.

In the key clinical trial that informed TA70 (imatinib for chronic myeloid leukaemia) 58% of control group patients had switched from the control treatment onto imatinib treatment after 18 months.⁵⁵ Initially the manufacturer and the assessment group based their economic analyses on ITT trial data but given the extent to which treatment crossover had occurred, the Appraisal Committee requested that an analysis be undertaken using per protocol data rather than ITT data. Essentially this involved excluding crossover patients. This resulted in a substantial reduction in the ICER (from £87,000 per QALY gained to £60,000 per QALY gained for imatinib compared to hydroxyurea).⁵⁵

In TA86, which assessed the use of imatinib for GIST, several different approaches were used to address the treatment crossover problem. The manufacturer chose to exclude patients who crossed over, which involved excluding 51% of control group patients.⁵⁶ In contrast, the assessment group tested the results of the economic model when an ITT approach was taken, irrespective of crossover.⁵⁶ Further analysis by the NICE Decision Support Unit (DSU) was commissioned, and yet another approach was taken as the DSU chose to censor patients as and when they crossed over treatments.⁶⁶

These different approaches led to significant differences in the results of the economic evaluation, with the manufacturer's approach generating an ICER of around £14,000 per QALY gained, the assessment group's approach (including other changes to model parameters) resulting in an ICER of approximately £30,000 per QALY gained and the DSU's approach leading to an ICER of approximately £32,000 per QALY gained.^{56;66} As would be expected due to the dilution of the treatment effect typically caused by treatment crossover, the assessment group's approach of including all patients irrespective of crossover resulted in a higher ICER than the manufacturer's approach of excluding any patient that crossed over.

The Appraisal Committee stated in the FAD that the approach taken by the DSU was least likely to result in bias.⁶⁶ This is because the ITT approach would be expected to result in an overestimate of survival for control group patients, whereas the exclusion method wholly ignores any data from patients who crossed over at some point – even unconfounded data prior to crossover is ignored. The censoring at the time of crossover approach may be seen to improve on this because data are used prior to crossover, allowing for a more complete

analysis of the dataset, whereas any beneficial effect of the treatment crossover is excluded from the analysis. However, as previously stated, censoring crossover patients is also likely to result in bias. If patients who crossover have relatively good prognoses their censoring is likely to result in an underestimate of OS for the control group, whereas the opposite is true if crossover patients have poorer prognoses and are in more “need” of further treatment. Either way, each of the methods used to estimate OS for the control group in TA86 are likely to have been subject to bias, leading to potentially inaccurate cost-effectiveness results.

TA129 (bortezomib for multiple myeloma) also involved problems associated with treatment crossover. In the key clinical trial patients in the control group were allowed to crossover on to bortezomib after an 8.3 month interim analysis or upon disease progression.⁵⁸ Due to this, only data up to 8.3 months were used to contribute to the extrapolation conducted for the economic model, with the remaining data being censored.⁵⁸ However, by this time point, 44% of control group patients had already crossed over. Although the problems caused by crossover were addressed to some extent in this TA (in that only data up to 8.3 months were used) the remaining confounding of the treatment effect estimate caused by the treatment crossover that occurred prior to 8.3 months was not considered.

In the recent renal cell carcinoma (RCC) TAs (TA169 and TA178) various approaches were used in an attempt to control for treatment crossover. Either censoring, exclusion or both approaches were used for sunitinib, bevacizumab and sorafenib. For sunitinib, excluding all patients who received any second line treatment resulted in a hazard ratio (HR) of 0.65, whereas censoring those patients gave an HR of 0.81, compared to the ITT HR of 0.82.⁵⁹ NICE commissioned the DSU to perform an analysis, and the DSU stated that the censoring approach was preferable in comparison to the ITT and exclusion methods, but that the censoring approach itself was also open to selection bias, because the receipt of second-line treatment was likely to be associated with prognosis and performance status.⁵⁹

In the exclusion analysis conducted by the manufacturer of sunitinib, patients who received any second-line treatment (including crossover treatment) were excluded in the control group, but they were not excluded in the intervention group.⁵⁹ 49% of patients in the intervention group received second-line treatment, and including these patients in the intervention group whilst similar patients were excluded from the control group seems inconsistent and open to bias. However, an analysis was completed whereby these patients were excluded from both the control and intervention groups and the resulting OS estimates for sunitinib were so high that they were deemed unrealistic by the Appraisal Committee.⁶⁰ Therefore, whilst excluding

control group patients who received second-line treatments led to an improvement in the HR for sunitinib (suggesting that patients in the control group who received second-line treatment had relatively good OS), excluding intervention group patients who received second-line treatments also led to an improvement in the HR for sunitinib (suggesting that patients in the sunitinib group who received second-line treatment had a relatively poor OS).

This may be due to chance and small patient numbers, but is indicative of second-line treatment choices being dependent upon a range of factors that were related to the treatment arm to which patients were initially randomised. Certainly the analyses demonstrate that the impact of treatment crossover and other post-progression treatments on the estimate of the treatment effect is highly uncertain and simply censoring or excluding certain patient groups without further consideration is unlikely to be a robust approach. It also shows that censoring or excluding various groups and comparing results can help demonstrate what the impact of post-progression treatments may have been, which may inform a choice of which type of censoring or exclusion (if any) may be most acceptable. In TA169 the NICE Appraisal Committee decided that the approach of excluding control group patients who received any second-line treatment, but not excluding intervention group patients who received any second-line treatment, was the most acceptable approach – even though in theory this appears particularly open to bias.^{59;60;68}

The manufacturers of bevacizumab used an approach whereby all patients who received second-line treatments were censored in their submission for TA178. This reduced the OS HR from 0.75 (for the ITT analysis) to 0.61.⁵⁹ However, in this case the DSU stated that further censoring of alternative groups should have been undertaken in order to allow the likelihood of informative censoring to be assessed.⁵⁹ In the absence of this the DSU stated that the approach taken by the manufacturer could not be supported, since the OS HR had fallen below the PFS HR of 0.63. Owing to this, the DSU tested an analysis whereby the OS HR was equal to the PFS HR.⁵⁹

In TA179 (sunitinib for GIST) censoring was used as an approach to control for treatment crossover as a secondary analysis. The primary analysis in the TA used the RPSFTM method, which will be discussed in Section 3.6.2.1. However to supplement this the Appraisal Committee asked for a censoring analysis to be completed as this method had been used commonly in previous TAs.⁶⁹ Eighty-four percent of control group patients crossed over upon disease progression, and only 15 patients did not crossover. Hence when crossover patients were censored very few OS events were observed in the control group. This resulted in

control group OS times that were longer than those estimated for the experimental group and due to the perceived lack of face-validity of these, the Appraisal Committee did not take this analysis into account when making their decision.⁶⁹ Although based on very few patients, the censoring analysis in this case would seem to suggest that only the control group patients with the very best prognosis were not switched onto the new intervention upon disease progression. Again this indicates that the decision of whether a patient is offered crossover treatment is a complex one, and that this decision is likely to be influenced by expected survival.

Several examples of how censoring and exclusion approaches have been used in an attempt to control for treatment crossover in economic evaluations included within NICE TAs have been identified. However, clearly there are problems with these approaches, resulting from informative censoring and exclusion, confounding, and selection bias. Performing several analyses of censoring and excluding different patient groups may allow more understanding to be gained with respect to the likely biases associated with censoring or excluding different groups, as was the case in TA169.⁵⁹ However, given the strong likelihood of a link between treatment crossover and survival (as seems to have been demonstrated by the analyses presented in some of the TAs discussed above), the censoring and exclusion methods are likely to be inherently associated with bias.

3.6.1.2 Including costs of post-progression treatments

In four of the TAs included in the review, the costs of post-progression and/or crossover treatments were included in the economic evaluation in order to account for the occurrence of post-progression treatment and/or crossover. In terms of producing an accurate economic evaluation of the key clinical trial, this technique may be considered reasonable. However, the usefulness of the technique for use in NICE TAs given the decision problem is much less clear. As discussed in Chapter 1, the aim of an economic evaluation of a new intervention is typically to compare a state of the world where the new intervention exists to one in which it does not, and simply analysing the trial data as observed and allocating costs to crossover patients does not satisfy this aim – it trades internal consistency for external validity. The NICE TAs (TA101, TA116, TA121, TA118) that included costs of post-progression treatments did so as an alternative to attempting to adjust survival estimates due to the perceived difficulties associated with this.⁷⁰⁻⁷³ They also typically did so as a means for modelling “other” post-study treatments, rather than specifically to address treatment crossover.⁷¹⁻⁷³ Where the post-study treatments received represent a realistic treatment pathway that is relevant for the decision problem this represents a suitable modelling approach – however it is not suitable for

addressing treatment crossover. This approach was only considered specifically to address treatment crossover in one TA (TA101).⁷⁰

3.6.1.3 Modelling based only on PFS

In two of the reviewed TAs in which treatment crossover was an issue, the economic modelling focussed only on PFS, rather than OS. This approach was generally taken due to a lack of OS data or due to confounding factors, such as treatment crossover.

Excluding OS from the economic evaluation involves quantifying only the impact of the new treatment on PFS. However this does not mean that there is an implicit assumption that the new intervention only affects PFS. Rather, the method implies an assumption that the absolute QALY gain associated with the extension of PFS is exactly equivalent to the absolute QALY gain if OS had also been modelled, thus assuming that post-progression survival is identical in the two treatment groups.

The initial conclusion may be that this method is likely to underestimate the cost-effectiveness of the new intervention: If the new intervention increases the duration of PFS, and if there is a link between PFS and OS, or if the new intervention has any independent effect on OS, then modelling based only upon PFS will underestimate cost-effectiveness. However, this is not necessarily the case. Modelling based only upon PFS essentially assumes that upon disease progression a patient dies – no more costs are incurred and no more QALYs are obtained. This is important because additional QALYs (and costs) are accrued after disease progression, and indeed if the absolute effect of the new treatment on OS is smaller than that on PFS then modelling based only on PFS may overestimate the cost-effectiveness of the new intervention. When considering this, it is important to note the important distinction between absolute and relative effects. In TA178 the NICE DSU suggests that it is most likely that the relative effect of a new treatment will be lower for OS than PFS,⁵⁹ but this does not necessarily mean that the absolute difference in OS will also be lower. Because OS is a longer time period than PFS an absolute difference in OS that is the same (or greater) than the absolute difference in PFS can be achieved with a worse (higher) hazard ratio. Therefore it is clear that economic analyses that only include PFS could lead to underestimation or overestimation of the cost-effectiveness of the new intervention.

In TA6 (taxanes for breast cancer) the economic evaluation of the first-line indication of paclitaxel appeared to be based on PFS only.^{74;75} Treatment crossover as well as the use of other post-progression treatments and other confounding factors appeared to have led to the

decision to conduct the economic evaluation based upon PFS. In TA33 (irinotecan, oxaliplatin and raltitrexed for advanced colorectal cancer) a similar approach was taken, with the assessment group stating that treatment crossover, post-study treatments, and difficulties associated with attributing OS gains to different treatments led to the decision to conduct the economic evaluation based upon PFS.⁵⁰

It is clear that an economic evaluation that only includes PFS is likely to lead to inaccurate results. Using this approach as a method for adjusting for treatment crossover is only likely to be reasonable where post progression survival is expected to be identical in both treatment arms – meaning that the incremental QALYs gained in the PFS period are identical to those that would be expected to be gained over a lifetime. Even if this were the case there would be some inaccuracies in the PFS model due to potential cost differences and due to discounting. It is essentially a clinical question as to whether this assumption is reasonable in specific cases, but it certainly seems a strong assumption to make, and one which should generally be considered inappropriate.

3.6.1.4 Applying the same risk of death upon disease progression

An extension to the method of basing the economic analysis only on PFS is to model OS, but to assume that once a patient has experienced disease progression, the risk of death is the same whether the patient is in the control group or the intervention group. Using this technique will mean that the absolute difference in OS for the two treatments will be similar to the absolute difference in PFS.

Essentially, this method assumes that the treatment effect of the new intervention is of limited duration – it lasts only until disease progression. After that there is no additional gain to having been treated with the new intervention. On the other hand, the risk of death is not greater in the new intervention group, and thus the PFS gain associated with the new treatment is assumed to lead to an OS gain. This assumption is potentially flexible. For example, at a conservative extreme it could be assumed that a new intervention that increases PFS has no impact on OS – that is, the risk of death upon progression in the intervention group is higher than in the control group to the extent that OS is identical between the two groups. Alternatively, it could be assumed that the treatment effect is maintained into the post-progression period, with the most liberal assumption being that the treatment effect is maintained for an entire lifetime. Another option could be to assume that the treatment effect disappears after (for example) two years. The assumption made about the duration of

treatment effect would ideally be based upon clinical data, or expert clinical knowledge including a consideration of the biological nature of the intervention and the disease itself.

In TA118 (bevacizumab and cetuximab for metastatic colorectal cancer) the manufacturer of bevacizumab chose to use the same risk of death for the new intervention and the control group once disease progression had occurred.⁷³ Rather than conventional treatment crossover (whereby control group patients switch onto the new intervention) being the reason for this, the primary motivation was that in the two key clinical trials patients in the bevacizumab arm were allowed to continue receiving bevacizumab after disease progression.⁷³ This was off-licence, and the manufacturer expected that this would increase OS. Therefore, to complete an economic evaluation relevant for the drug licence, the manufacturer attempted to adjust for this continued treatment with bevacizumab by assuming the same risk of death upon progression for both the bevacizumab group and the control group.

The assessment group considered whether the assumption of equal risk of death following disease progression would be likely to bias the economic model in favour of the intervention or the control treatment.⁷³ Based upon one key clinical trial, the assessment group noted that the assumption seemed reasonable, since the absolute OS benefit was similar to the absolute PFS benefit. However, in another relevant trial the PFS benefit appeared to be greater than the OS benefit, and thus the assessment group considered that the equal risk of death upon progression assumption would be likely to cause bias in favour of the new intervention.

The results of the assessment group's analysis are interesting. Theoretically, the manufacturer would appear to have a reasonable concern about the applicability of their OS data, considering that the intervention was allowed to be given off-licence after disease progression. Assuming an equal risk of death following disease progression may be a relatively naïve method for accounting for this but it also seems to represent a reasonably conservative approach since it does not extend the treatment effect beyond disease progression, and there may be no reason to expect that the risk of death would actually be higher in the intervention group following disease progression. However, the analyses undertaken by the assessment group seemed to demonstrate that the risk of death after progression was actually higher in the intervention group than in the control group, despite continued treatment with bevacizumab.

This may not be an entirely expected result, but is a particularly important one which demonstrates the dangers associated with seemingly reasonable (even apparently

conservative) assumptions regarding OS when trial data are confounded by crossover. It may seem counter-intuitive to assume that the risk of death is higher in the intervention group following disease progression when the intervention has been shown to confer a PFS advantage, but it may actually be so. This also shows the importance of making good use of all available data (and clinical opinion) to inform modelling assumptions, and decisions on how best to adjust for treatment crossover.

3.6.1.5 Assumed equal OS for the two treatment groups

As discussed in Section 3.6.1.4, an extreme conservative assumption that could be made in the presence of OS data confounded by treatment crossover could be that the new intervention does not confer any OS benefit, even if a PFS benefit has been demonstrated. In some cases this may be a useful analysis for decision makers, even if it is not likely to be accurate. This analysis may present a type of “worst-case” analysis for the new treatment, providing a “maximum” ICER associated with the intervention (assuming all other assumptions in the model – for example utility scores – were acceptable). If this “maximum” ICER were acceptable then the decision-maker (in NICE’s case the Appraisal Committee) may recommend the intervention with a greater degree of confidence. However, in the more likely scenario, whereby this analysis resulted in an ICER that was not acceptable, the analysis would be less useful to the decision-maker without several other sensitivity analyses demonstrating the ICER for alternative OS estimates.

This approach was taken in TA119 (fludarabine for lymphocytic leukaemia). The manufacturer assumed that OS was equal in the control and intervention groups in their economic model because OS data were sparse and confounded by treatment crossover and other post-study treatments.⁷⁶ Although this analysis would seem to be conservative, the assessment group noted a number of problems with it.⁷⁶ Firstly, the manufacturer initially used an approach which ensured that *median* OS was similar between treatment groups in their model, rather than *mean* OS. A similar median does not guarantee a similar mean, and thus important OS differences could still occur in the model. Secondly, the OS data that was available showed that the mortality rate was marginally (non-significantly) higher in the intervention group than in the control group. Therefore, there was some indication that OS may actually have been lower in the intervention arm, despite a PFS benefit. Thirdly, setting OS to be equal meant that in the economic model patients in the control group spent longer in the progressed state than patients in the intervention group. Because this state was costly, and because the utility associated with the state was low, spending additional time in this state had the potential to worsen the cost-effectiveness case for the control treatment. Similarly, an analysis that

modelled an OS advantage for the new intervention may actually have led to a higher (less favourable) ICER. This demonstrates that care should be taken with “simple”, seemingly conservative analyses to adjust for treatment crossover, because they may lead to bias in unexpected directions.

3.6.1.6 Using sequencing models

In some circumstances, it may be the case that treatment pathways are well defined, and data are available demonstrating the effectiveness of treatments given at different points in a pathway. This lends itself to a treatment sequencing economic evaluation whereby post-progression treatments are explicitly modelled as part of a treatment pathway. This represents a method for addressing the treatment crossover problem because typically crossover only occurs after disease progression, and a sequencing model would only incorporate information from the trial of the new intervention up until the point at which the next treatment in the pathway would be administered, which would often be upon disease progression. Data on survival beyond this point would generally be taken from another source and the period confounded by treatment crossover would not be used within the economic model. However, problems with the analysis would still occur if treatment crossover occurred in the trial of the final-stage treatment, and often data on the effectiveness of interventions at different stages of the disease pathway are hard to come by.

Two of the reviewed TAs specifically modelled treatment sequences in part to address the treatment crossover problem. Both TAs considered treatments for advanced colorectal cancer and used the same sequencing trial (TA93: Irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer; TA176: Cetuximab for colorectal cancer).^{64,65} In TA93 the assessment group modelled first line PFS, second line PFS and OS using data from a trial in which patients were randomised to a planned sequence of chemotherapy.⁶⁴ This approach was taken because treatment crossover and post-progression treatments occurred regularly in the key trials, but also because there was a sequencing trial available that included the key comparators. Indeed, the assessment group specifically noted that in situations where treatment crossover is an issue and sequencing trials are not available, using PFS as the key outcome measure in the economic analysis is the only way to avoid confounding (although the group do also note the problems associated with relying on PFS – notably that it is a surrogate outcome).⁷⁷ It was also noted that the sequencing approach did not fully avoid the problems associated with treatment crossover because the trial that was used to estimate OS after the final disease progression involved a degree of post-study treatment that may have included treatment crossover.⁷⁷

3.6.2 More Complex Methods

The above discussion demonstrates how relatively simple methods have most often been used in attempts to adjust for treatment crossover in NICE TAs. In a small number of TAs more complex statistical methods have been used.

3.6.2.1 The Rank Preserving Structural Failure Time Model (RPSFTM)

Treatment crossover was a serious issue in TA179 (sunitinib for GIST), as 84% of patients in the control group of the pivotal clinical trial switched onto sunitinib upon disease progression.⁶³ When the survival data were analysed under an ITT approach, the manufacturer estimated an ICER of approximately £77,000 per QALY gained and an HR of 0.876 for OS.⁶³ However, recognising the confounding present in the control group OS data, the manufacturer implemented the RPSFTM method developed by Robins and Tsiatis^{22;78} to adjust the estimate of the treatment effect on OS. This resulted in an OS HR of 0.505, associated with an ICER of £27,000 per QALY gained.⁶³ This reduction in the ICER was potentially large enough to alter the treatment recommendation, as NICE typically recommends interventions that have an ICER of less than £20,000-£30,000 per QALY gained.³

The assessment group discussed the advantages and disadvantages associated with the RPSFTM in their report. In particular, they noted that the method is based upon the ITT population and therefore the randomisation of the trial is preserved, which they perceived to be an important advantage of the method.⁶¹ On the other hand, the assessment group noted that a weakness of the method is the “common treatment effect” assumed – thus the validity of the approach was questioned.⁶¹ However, the key implication of the “common treatment effect” assumption was not discussed by the assessment group; that is, if crossover patients are expected to receive a lower treatment effect than patients initially randomised to the experimental group (perhaps due to a lower capacity to benefit after disease progression) the RPSFTM might be expected to over-estimate the crossover treatment effect and hence would overcompensate for this and underestimate the ICER. These key issues are considered much further in this thesis – in Chapter 4, in which the RPSFTM method is described in detail (see section 4.10.2), and in subsequent chapters in which the potential biases associated with the method are investigated.

The assessment group considered the use of the RPSFTM method with the help of an external statistician with expertise in the area (Ian White).⁶³ They compared the RPSFTM to “naive”

approaches for addressing treatment crossover, and concluded that the RPSFTM approach was preferable and appeared to be the correct method to use. The group noted that censoring crossover patients would be expected to result in selection bias, while an ITT analysis would estimate a diluted treatment effect. The group were concerned that they could not guarantee that the RPSFTM had been implemented correctly because they did not have access to the individual patient-level data, and advice from Ian White revealed that the manufacturer had calculated the confidence intervals associated with their RPSFTM analysis incorrectly. The assessment group stated that the RPSFTM method had not to their knowledge been used in cost-effectiveness analyses before and that it represented the only method available that can correct for time-dependent treatment changes in survival data while respecting the randomisation.⁶³ Strictly speaking this is not true – randomisation-respecting derivatives of the RPSFTM exist that are described in Chapter 4 – however this demonstrates how little these methods have been used and how novel they are in the context of health economic evaluation.

3.6.2.2 Adjusted survival estimates based on case-mix of patients that switched

As discussed in Section 3.6.1.1, treatment crossover was a key issue in TA34 (trastuzumab for breast cancer). Initially the manufacturer provided an analysis which involved excluding all patients who crossed over but the assessment group considered that this approach was likely to result in bias.⁵¹ In response to this, the manufacturer recalculated their survival estimates based on a weighted case-mix extrapolation approach. Unfortunately, this analysis appears to have been included in a revised submission by the manufacturer as it is not discussed in the Assessment Report and only limited details are provided in the FAD document.⁶⁷

Based upon the discussion in the FAD it appears that a group of patients were selected from the trial and survival time estimates for these patients were weighted to take into account the case-mix of patients who crossed over onto trastuzumab treatment. It is not clear whether this involved placing particular weight on non-crossover patients who had case-mix attributes most similar to the crossover patients, or whether some other adjustment method was used to re-estimate survival times for crossover patients. Thus, it is not possible to provide a full critique of this method. It seems possible that if there were a reasonable number of patients in the control group who did not crossover, survival for particular patients in this group could be weighted in such a way as to attempt to minimise selection effects – this would be a similar approach to the IPCW method described by Robins and Finkelstein.²⁴ However it is not clear how the weighting process was carried out in TA34. In addition, given the very low number of patients left in the control group of the trial after crossover patients had been excluded, basing

survival estimates on the few that remained with a certain case-mix seems likely to have resulted in very high levels of uncertainty and potential bias.

The casemix approach used in TA34 led to a mean survival advantage of approximately 10 months for trastuzumab – which was close to mid-way between the ITT estimate and the estimate based on the exclusion approach. This is likely to have given the Appraisal Committee some confidence in the estimate, and they used this as a lower bound of the likely true OS gain due to evidence from other (non-controlled) trials that suggested a similar or higher benefit.⁶⁷

3.6.2.3 Using external data

Treatment crossover was an important issue in TA171 (lenalidomide for multiple myeloma). Around 50% of control group patients in the key clinical trial crossed over onto lenalidomide, with 75% of that crossover occurring after disease progression.⁷⁹ To address the crossover problem the manufacturer used patient-level data from previous trials that included similar (but not identical) control group arms that were not confounded by crossover.⁶² The manufacturer provided analyses to demonstrate that the OS that could be expected for the control group treatment (dexamethasone) used in their novel lenalidomide trial was similar to that observed in the external trials (which used dexamethasone as well as some other standard treatments as control).⁶² In addition, the manufacturer produced an analysis to demonstrate that there was not evidence of an OS improvement over time.⁶² This was important because the external trial datasets were dated, with patients enrolled between 1980 and 1997. Based upon these analyses, the manufacturer rationalised the use of the external trial datasets for inferring what control group survival in the novel lenalidomide trial would have been, had treatment crossover not occurred.

The manufacturer fitted parametric survival models to the external trial data in order to derive an equation for OS that included a range of patient characteristic variables.⁶² The values of these variables were then set to reflect the patient characteristics observed in the lenalidomide trial, and hence survival times that would have been observed in the external trial had the patient characteristics in the control arm matched those in the novel lenalidomide trial were estimated. The manufacturer did not use this estimate of OS directly in the economic model, because PFS and post-progression survival (PPS) were modelled as distinct states, with PFS estimated based only on the novel lenalidomide trial (this in itself is questionable, since 25% of crossover occurred before disease progression). In the economic model the manufacturer used a “calibration factor” applied to PPS such that the median OS

estimated from the external trial dataset adjusted for the lenalidomide trial patient characteristics equalled the median OS estimated by the model, as a function of PFS plus PPS.⁶²

The assessment group noted some problems with the manufacturer's analysis.⁶² Firstly, they noted that mean OS rather than median OS should have been used to calibrate the estimated OS in the control arm of the lenalidomide trial with the external data. The manufacturer defended their use of the median, stating that estimating means was problematic in the presence of censored data.⁸⁰ However the assessment group re-asserted the importance of the use of means in economic evaluation, and pointed out that the external trial data involved relatively little censoring (94% were said to have died), dismissing the manufacturer's argument.⁸⁰ The assessment group showed that calibrating to the mean rather than the median increased ICERs by approximately £8,000 (to over £30,000) for two of the subgroups analysed.⁸⁰

A second problem highlighted by the assessment group was that there were likely to be important patient characteristics not reported in both the novel lenalidomide trials and the external trials which could not be included in the OS equations.⁶² Hence it may not have been possible to fully adjust the external trial survival estimates to reflect the lenalidomide trial patient population. The analysis is essentially reliant on a "no unmeasured confounders" assumption, and the lack of analysis to identify any important variables missing from either the lenalidomide or external trial datasets represents an important oversight on the part of the manufacturer.

Finally, the assessment group noted that alternative data sources suggested improvements in survival in the relevant patient group between 1995 and 2006, thus suggesting the dated external trials may indeed represent an underestimate of present-day control group OS.⁶²

An additional issue which was not mentioned by the assessment group but discussed by the Appraisal Committee surrounded the clinical validity of the manufacturer's analysis. There were two lenalidomide trials relevant to the appraisal, and the application of the manufacturer's analyses to these trials led to control group OS estimates that were approximately half those observed in the trials themselves. These details were marked as "commercial-in-confidence" in the TA documents, but were reported in a subsequent published paper.⁸¹ Therefore, based upon the manufacturer's analysis, the impact of approximately 50% of control group patients crossing over was to cause the mean OS for the

control group as a whole to approximately double. For this to be the case, the experimental treatment would have to more than double life expectancy for crossover patients.

In the key lenalidomide clinical trial the gain in PFS for lenalidomide was large: 13.4 months compared to 4.6 months in the control arm (2.9 times longer for lenalidomide).⁸¹ Therefore a similar relative effect on OS could potentially lead to the OS estimates derived by the manufacturer. However, this would assume that the relative effect of lenalidomide on OS is the same (if not higher) than for PFS, and that receiving lenalidomide after disease progression leads to the same (if not higher) impact on OS as is the case when it is given before disease progression. The Appraisal Committee noted that the manufacturer's approach led to an improvement in OS predicted by the economic model which was out of proportion given the improvement seen in PFS.⁷⁹

The manufacturer's method led to an ICER of £25,000 per QALY gained. When an ITT analysis was run, the ICER was £79,000.⁶² Therefore, addressing the crossover problem was extremely important and could potentially have affected treatment recommendations made. The methodology used by the manufacturer seems reasonable, but there are clear limitations to it – to the extent that the assessment group suggested that a simple analysis in which crossover patients were censored may have represented a preferable approach.⁶² Along with the limitations of the method specific to TA171 – that is the calibration to the median, and the age of the external datasets – there are several more general limitations that may restrict its practicality. In particular, often relevant external datasets are not available or do not exist. Also, the method requires that all important prognostic variables are available from both the novel clinical trial, and the external trials – otherwise different trial populations cannot be adjusted appropriately for comparison. This “no unmeasured confounders” assumption is testable to an extent. When variables are measured in one trial but not another, analyses can be conducted to determine whether these are prognostic. However it is not possible to determine the importance of variables that are excluded from both data sources, although light might be shed on this by clinical experts.

3.7 Potential impact of treatment crossover on NICE's decisions

From the above analysis it is clear that treatment crossover is a common issue in NICE TAs of interventions for advanced and/or metastatic cancer, and the issue has generally not been dealt with adequately. Of the 25 TAs in which treatment crossover was identified as being an issue, attempts were made to adjust survival estimates in 18. The economic evaluation results

in the remaining 7 TAs are likely to be biased, and given that treatment crossover typically moves the estimate of the treatment effect towards the null, the cost-effectiveness of these interventions is likely to have been underestimated. In 4 of these TAs the intervention was recommended for use in the NHS (TA3: Taxanes for ovarian cancer; TA30: Taxanes for breast cancer; TA55: Paclitaxel for ovarian cancer; TA91: Paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan for advanced ovarian cancer).⁸²⁻⁸⁵ In these TAs it may be the case that adjusting for treatment crossover would not have altered the final appraisal determinations made by the Appraisal Committee. However, in 2 of the TAs in which crossover was an issue but was not addressed, the intervention was not recommended for use (TA124: Pemetrexed for non-small-cell Lung Cancer; TA172: Cetuximab for Head and neck Cancer),^{86;87} and in one TA the intervention received a restricted recommendation that was subject to a costing agreement (TA162: Erlotinib for non-small-cell Lung Cancer).⁸⁸

In TA124, treatment crossover was from the new treatment, pemetrexed, to the control treatment, docetaxel, rather than from the control treatment to the new treatment. Therefore in this instance the treatment crossover present would not be expected to bias the estimate of the treatment effect against the new treatment. In addition, clinical data seemed to suggest very little if any difference in either PFS or OS for the two treatments, and the ICER was extremely high (approximately £450,000 per QALY gained).⁸⁹ Therefore it seems unlikely that the failure to address the treatment crossover issue would have impacted upon the adoption decision.

In TA172 treatment crossover was not mentioned in the TA documents, but another publication reporting the results of the key clinical trial showed that 6% of the control group received the new treatment (cetuximab) after disease progression.⁹⁰ Cetuximab was not recommended by NICE partially because of a high ICER (around £100,000 per QALY gained) and partially because the Appraisal Committee felt that the treatment did not fulfil the “End-of-Life” criteria set out by NICE, which allow some interventions to be recommended by NICE even if they have an ICER of substantially greater than £30,000 per QALY.^{41,87} One of the criteria that must be satisfied is that the new treatment must be expected to offer an OS advantage of 3 months or greater and the Appraisal Committee believed that cetuximab did not satisfy this.⁸⁷ Although only a small proportion of control group patients switched treatments in the clinical trial, there is some potential that this could have impacted upon the Appraisal Committee’s view. The Committee considered that cetuximab was likely to offer an OS advantage of around 2.2-2.7 months – just below the 3-months required by the “End-of-Life” criteria. It is possible, though unlikely, that adjusting OS estimates to take into account

treatment crossover could have resulted in a mean OS gain of 3 months or more and a reduced ICER. The ICER would be unlikely to fall to levels of around £30,000 per QALY gained simply by adjusting for 6% crossover, but there remains a slight possibility that the combination of an increased OS benefit and a reduced ICER may have altered the Appraisal Committee's decision.

In TA162 erlotinib was recommended as a second-line treatment option as an alternative to docetaxel under an equal costing agreement with the manufacturer.⁸⁸ There was a significant amount of debate in the TA concerning the manufacturer's assumption that OS was equivalent for erlotinib and docetaxel, which led to the restrictive recommendation made. Crossover was not considered in the appraisal, but two patients (less than 1%) in the placebo arm of the key trial inadvertently received erlotinib but were included in the placebo arm for the statistical analysis, and 7% of placebo patients compared to 2% of Erlotinib patients went on to receive epidermal growth factor receptor (EGFR) inhibitor therapy after withdrawal from the study.^{88;91,92} This is not strictly treatment crossover as considered in this thesis, but there could be grounds to adjust the analysis. The proportions were low, but any additional analyses that may have bolstered the OS argument for erlotinib may have increased the confidence of the Appraisal Committee in the analysis, and could potentially have led to a less restrictive recommendation.

In the 18 TAs in which some attempt was made to address the treatment crossover issue, generally naïve methods were used and these are open to substantial bias. Therefore it is difficult to determine whether the resulting economic evaluations under-estimated or over-estimated the cost-effectiveness of the new interventions that they analysed. Potentially, flawed evaluations could have led to positive recommendations due to an over-estimated survival gain and an under-estimated ICER, or to negative or restricted recommendations due to an under-estimated survival gain and an over-estimated ICER.

In the 18 TAs, 22 recommendations were made (as several were multiple technology appraisals in which multiple recommendations were made). Six of these were positive recommendations (in TAs 6, 33, 70, 93, 116 and 169);^{55;93-97} Five were recommendations for subgroups (in TAs 28, 33, 34, 101 and 121);^{67;94;98-100} One was a positive recommendation dependent on a treatment response restriction (TA86);⁵⁷ One was a positive recommendation dependent on a rebate scheme (TA129);¹⁰¹ Three were positive recommendations based upon costing agreements (in TAs 171, 176 and 179);^{69;102;103} and six were negative recommendations (in TAs 6, 33, 93, 118, 119 and 178).^{93-95;104-106} The impact that addressing the crossover issue can have on cost-effectiveness results is highlighted by consideration of TA171 (ITT-based ICER

of approximately £79,000 per QALY gained was reduced to approximately £25,000 per QALY gained based on the external data approach),⁶² TA179 (ITT-based ICER of approximately £77,000 per QALY gained was reduced to approximately £27,000 per QALY gained using the RPSFTM method);⁶³ and TA86 (ITT-based ICER of approximately £30,000 per QALY gained was reduced to approximately £14,000 per QALY gained by excluding crossover patients).⁵⁷ Therefore attempting to account for treatment crossover can have very important impacts on the ICER, and may well have influenced the recommendations made in the TAs. It is reasonable to assume that if adjustments had not been made to account for treatment crossover in TA171, TA179 and TA86 the restrictive recommendations that were made may have instead been negative recommendations.

3.8 Review update

The review of NICE TAs presented in this chapter had a cut-off date of December 2009. Since this time complex crossover adjustment methods have been used in at least two NICE TAs. While I have not systematically updated the review for all cancer-related TAs completed since December 2009, it is useful to make note of these appraisals, particularly given that there is little evidence of the use of more complex crossover adjustment methods in the review. These appraisals demonstrate the ongoing issue of treatment crossover within HTA, and also show that there has been a recent tendency to move towards the more complex adjustment methods.

In TA215 (pazopanib for the first-line treatment of metastatic RCC) pazopanib was recommended subject to certain restrictions, including a 12.5% discount on the price of the drug.⁵³ Within the pivotal RCT patients in the control group could be offered open-label pazopanib once their disease had progressed, provided their Eastern Cooperative Oncology Group (ECOG) performance status was less than or equal to 2. At the time of the final analysis of the trial, 40 of 78 control group patients (51%) had crossed over and the OS HR was 1.03, compared to the PFS HR of 0.40.⁵³ Given that the offer of treatment crossover within the trial was associated with performance status, crossover was not random and simply censoring or excluding crossover patients would be expected to result in selection bias. The manufacturer implemented the RPSFTM and IPCW methods to adjust for the crossover, and these resulted in OS HRs of 0.50 and 0.64 respectively. In addition, an analysis that excluded crossover patients was presented, which gave a HR of 0.30. The manufacturer considered the RPSFTM method to be the most appropriate purely because it had been regarded as an acceptable approach in TA179 (sunitinib for the treatment of GIST).¹⁰⁷ The NICE Appraisal Committee stated that it

was appropriate to adjust for crossover using statistical modelling techniques and accepted the use of the RPSFTM method – noting that the independent academic ERG had suggested that the approach was reasonable.⁵³ Adjusting for treatment crossover had a substantial impact on the ICER, with an ITT analysis resulting in an ICER of approximately £322,000 per QALY gained compared to best supportive care, the IPCW analysis resulting in an ICER of approximately £49,000 per QALY gained, and the chosen RPSFTM method resulting in an ICER of approximately £33,000 per QALY gained.¹⁰⁸ Because the Appraisal Committee accepted that pazopanib fulfilled the criterion of an “End-of-Life” treatment an ICER of over £30,000 per QALY gained could be considered acceptable. Therefore, adjusting for treatment crossover in this appraisal may have significantly altered the Appraisal Committee’s decision.

The ERG deemed the RPSFTM method to be appropriate because it is a randomisation-based method which maintains the validity of between-group comparisons. However, they noted that care should be taken when assessing the results of relatively new methods (such as the RPSFTM and the IPCW) because there is currently no consensus on which is the best approach to use.¹⁰⁸ The choice of crossover adjustment method was further complicated by the manufacturer’s use of a weighted version of the RPSFTM, rather than the standard version of the method described by Robins and Tsiatis (1991). The weighted version of the RPSFTM involved the same basic approach to adjustment (which is described in detail in Section 4.10.2 of Chapter 4), but the log-rank estimator, used to identify the adjusted treatment effect, is weighted. The manufacturer stated that they took this approach because an unweighted analysis gave multiple possible adjusted treatment effects, and because the power of an unweighted log rank test was poor because the extent to which crossover occurred meant that patients in the control group were more likely to be on the experimental treatment after a certain time-point than patients initially randomised to the experimental treatment.¹⁰⁷ However, the ERG were unsure of the value of this weighted approach given that it had not been published in peer reviewed journals, and also because it did not allow baseline covariates to be taken into account (unlike a standard, unweighted application of the RPSFTM). The ERG suggested that the unweighted version of the RPSFTM may actually have been preferable.¹⁰⁸ While the weighted and unweighted versions of the RPSFTM gave different estimates of the treatment effect (HRs of 0.50 and 0.31 respectively), it was unclear to what extent this was due to the weighting of the method, because, unlike the unweighted version, the weighted version did not adjust for baseline patient characteristics. In the ITT analysis, adjusting for baseline patient characteristics reduced the HR from 1.03 to 0.86,¹⁰⁷ and so it appears possible that the difference in the estimates resulting from the two types of RPSFTM analyses may be

due to the inclusion or otherwise of baseline covariates, rather than the weighting of the method.

Despite the manufacturer conducting an IPCW analysis, relatively little attention was paid to this in the appraisal. The manufacturer noted that the IPCW analysis was limited by a lack of post disease progression data that hampered the ability to model the treatment crossover and survival process, while also noting that the RPSFTM approach was preferable due to being randomisation based and therefore not reliant on the “no unmeasured confounders” assumption.¹⁰⁷ The fact that there were problems with data availability in the appraisal demonstrates that the IPCW method may sometimes not be practical. However, it is surprising that neither the manufacturer nor the ERG discussed one of the key limitations associated with the RPSFTM approach – that is its “common treatment effect” assumption. Instead both the manufacturer and the ERG consider the main weakness of the RPSFTM to be that it involves recensoring, which means that the treatment effect estimate is weighted towards the early trial follow-up period, which may not be representative of the effect over the entire follow-up period.^{107;108} While this is a legitimate concern that will be analysed in detail in later chapters of this thesis (in particular, see Section 4.10.2 of Chapter 4), the “common treatment effect” assumption is also likely to be very important. TA215 therefore highlights that while complex crossover adjustment methods were applied, there remains some lack of appreciation of the key limitations of the alternative approaches. The appraisal also demonstrates that the more complex methods for adjusting for treatment crossover – the RPSFTM and IPCW – can provide substantially different point estimates of the treatment effect. This reflects the fact that these methods are very different and therefore it is important to consider which is likely to be most appropriate in different circumstances.

In TA219 (everolimus for the second-line treatment of advanced RCC) everolimus was not recommended.⁵⁴ The decision was subject to an appeal, but this was not upheld.¹⁰⁹ In the RCT that was pivotal to the appraisal, patients randomised to the control group were permitted to cross over onto everolimus upon disease progression. At the latest reported data cut-off 81% of control group patients had crossed over – leaving only 32 control group patients that did not cross over (15 of whom had not experienced disease progression).^{54;110} An ITT analysis gave an OS HR of 0.87, which was associated with an ICER (including a patient access scheme) of approximately £91,000 per QALY gained. Initially, the manufacturer provided an IPCW analysis to adjust for the crossover, which resulted in an HR of 0.55 and an ICER of approximately £52,000 per QALY gained.⁵⁴ The manufacturer explained in their submission that they chose to present an IPCW analysis rather than an RPSFTM analysis for a number of reasons – in

particular, the IPCW analysis provides an HR which is convenient for economic modelling; it is non-parametric and so does not require assumptions about the distribution of the survival data; and it does not “borrow” information from crossed over patients to inform the estimate of the treatment effect.¹¹⁰ Borrowing information from crossed over patients refers to the fact that the RPSFTM method uses data both from patients initially randomised to the experimental group and from crossover patients to estimate an average treatment effect, whereas the treatment effect estimated using the IPCW method is associated only with patients randomised to the experimental group.

In their review of the manufacturer’s analysis, the ERG sought the advice of Ian White, a leading expert in the area of treatment crossover adjustment methods. He disagreed with some of the rationale offered by the manufacturer in favour of the IPCW approach – in particular he noted that the RPSFTM method was originally developed using a non-parametric framework, and hence neither the RPSFTM or the IPCW require parametric assumptions.¹¹¹ Also, White stated that it was possible to obtain HRs from an RPSFTM analysis and thus the provision of an HR from an IPCW analysis should not be considered an important advantage.¹¹¹ Although it is true that the IPCW analysis censors crossover patients and so does not use the treatment effect received by these patients in its estimation process, unlike the RPSFTM, the ERG and White stated that this was at the expense of the “no unmeasured confounders” assumption. This was deemed to be a key weakness of the IPCW analysis, and it was suggested that an RPSFTM analysis should also be provided.¹¹¹

In response to this the manufacturer provided an RPSFTM analysis, which resulted in an ICER of approximately £53,000 per QALY gained.⁵⁴ The Appraisal Committee stated their preference for the RPSFTM analysis rather than the IPCW analysis, primarily because it avoids the “no unmeasured confounders” assumption.⁵⁴ However, as in TA215, the “common treatment effect” assumption required by the RPSFTM method was not discussed in any detail. Also, while the “no unmeasured confounders” assumption was regarded as a key weakness of the IPCW method the viability of the assumption was not investigated.

TA219 also demonstrated important issues associated with combining treatment crossover adjustment analyses with extrapolation within an economic model. The manufacturer initially argued that the IPCW analysis was preferable to an RPSFTM analysis due to the provision of an HR which could be used within the economic model. The inverse of the HR was applied to the mortality transition probabilities for the everolimus treatment group to derive transition probabilities for the control group. However, the ERG identified that the HR had been used

incorrectly since it should be applied to rates, rather than probabilities.¹¹¹ In addition, the HR was applied to transition probabilities from stable disease and progressive disease health states in the economic model, without taking into account that patients progressed more quickly in the control group and the mortality probability from the progressed state was higher than from the stable disease state. Due to this, the ERG showed that the manufacturer's application of the IPCW HR overestimated the effect of everolimus on OS.¹¹¹ Correcting for these errors, the ICER increased from approximately £53,000 per QALY gained to £65,000 per QALY gained.

Similarly, problems with the manufacturer's extrapolation were identified in their RPSFTM analysis. The manufacturer extrapolated the RPSFTM-adjusted Kaplan-Meier survival curve by applying the finally observed transition probability from one model cycle to the next to future time-points, rather than by fitting a parametric model to the complete survival curve.¹¹² The ERG conducted exploratory analyses around this and demonstrated that the manufacturer's ICER was likely to represent an underestimate. Depending upon different modelling assumptions, the ERG estimated that the RPSFTM-based ICER was likely to be approximately £58,000 - £76,000 per QALY gained, compared to the manufacturer's estimates of £50,000 – £53,000 per QALY gained.^{54;112}

During the course of TA219 several different analyses and patient access schemes were presented by the manufacturer. Despite the Appraisal Committee's acceptance that the drug met "End-of-Life" criteria and therefore a higher ICER could potentially be considered acceptable, everolimus was rejected because the ICER was too high and the uncertainty that surrounded it was substantial. Methodological uncertainty around the appropriate use of novel methods to adjust for treatment crossover added to the overall uncertainty that led to the Committee's decision.⁵⁴ The appraisal clearly demonstrates the potential importance of crossover adjustment methods for decision making, as well as the current uncertainty surrounding which methods are likely to be most appropriate. Importantly, the appraisal also shows the complexities associated with combining crossover adjustment methods with extrapolation methods for use in economic models, which demonstrates the importance of Chapter 5 of this thesis.

3.9 Discussion

There were three primary aims associated with the review reported in this chapter. Firstly, to determine the extent to which treatment crossover was an issue in appraisals of new cancer

treatments. Secondly, to identify which methods have been used to address treatment crossover in an economic evaluation context and finally, to demonstrate what impact these methods could have on cost-effectiveness results and related treatment recommendations.

It is clear that treatment crossover often occurs in clinical trials of new cancer treatments – of 45 TAs reviewed, treatment crossover occurred in 25 (55.6%) of these. The proportion of patients that crossed over differed substantially across these 25 appraisals – on two occasions less than 10% crossed over (TA172, TA162,^{90;91} hence the bias caused by the crossover might be expected to be low. However usually the crossover proportion was much higher (for example, the proportion of control group patients that crossed over was higher than 40% in TA34, TA86, TA129, TA171 and was 84% in TA179)^{51;57;58;63} and would therefore be expected to have a substantial impact on survival time and cost-effectiveness estimates. Although the exact impact of the crossover on ICERs and treatment recommendations is often hard to tease out from the appraisal documents, in a number of examples it was possible to determine what the ICER would have been with and without adjustments for treatment crossover. Frequently the differences were large, with ICERs being more than halved when crossover was addressed in TA171, TA179 and TA86.^{57;62;63} It is clear that these differences may have impacted upon treatment recommendations made by NICE.

An array of methods have been used to adjust for treatment crossover in NICE TAs, demonstrating a lack of consistency between appraisals, and also a lack of clarity over which methods are appropriate. Typically naive censoring or exclusion approaches have been used, which are clearly associated with selection bias. In many cases, treatment crossover was not addressed at all. This might be because analysts decided that the crossover was unlikely to be important, but this was not stated in the appraisal documents reviewed. Modelling approaches based only upon PFS, or assuming an equal risk of death after progression, were used in a small number of TAs. These were generally used due to a lack of OS data rather than due specifically to treatment crossover; however, they can be seen as methods for conducting the economic analysis avoiding the use of OS data that is confounded by crossover. These methods are limited by important assumptions and do not attempt to make use of trial data available on OS. The focus of this thesis is on making use of OS data when it is confounded by crossover, and thus these methods will not be considered in further detail. However, if the analyses shown in the forthcoming chapters demonstrate that in some circumstances OS data cannot be adjusted appropriately in the presence of treatment crossover, other more simplistic modelling-based approaches may seem reasonable.

Occasionally more complex methods have been used in NICE TAs, such as Robins and Tsiatis's RPSFTM.^{22;63;63;78} The analysis conducted in TA171 showed that it might not always be necessary to adjust confounded trial data to account for treatment crossover, rather external datasets might be available that could be used instead.⁶² This is a potentially valid approach, but is heavily reliant on data availability and the existence of suitable external datasets – it is unlikely to represent a method that is generalisable. An analysis of more recent NICE TAs has demonstrated the tendency towards more complex treatment crossover adjustment methods such as RPSFTM and IPCW.^{113;114} These TAs have also clearly demonstrated the uncertainty around which methods are appropriate for adjusting for treatment crossover in an economic evaluation context, and also the potentially important lack of understanding of what these methods entail. For example, in both TA215 and TA219 the weakness of the IPCW method due to its “no unmeasured confounders” assumption was highlighted, whereas the “common treatment effect” assumption made by the RPSFTM method was not discussed in any detail.^{107;111} Hence, while the RPSFTM method appeared to be preferred in these appraisals, the advantages and disadvantages associated with each method did not appear to have been fully taken into account. Once again, this demonstrates the importance of this thesis as more information and guidance on these methods is required.

3.10 Conclusions

This chapter has highlighted that the treatment crossover problem is common in HTAs of cancer treatments, and that addressing the problem can lead to very large – potentially decision-changing – reductions in ICERs. It has also highlighted that there is no consensus regarding appropriate methods for adjusting for treatment crossover, and that there are important inconsistencies between NICE appraisals which may have led to inconsistent resource allocation decisions. This demonstrates the importance of the treatment crossover problem and completes Part 2 of this thesis. The remainder of this thesis seeks to address this problem, first, in Part 3 (Chapters 4 and 5), by identifying and reviewing all existing crossover adjustment methods.

Chapter 4

Systematic review of statistical methods for adjusting survival estimates in the presence of treatment crossover

4.1 Chapter overview

Part 2 (Chapter 3) of this thesis demonstrated the importance of the treatment crossover problem in an HTA context. Together with Chapter 5, the present chapter comprises Part 3 of the thesis – focussing on identifying potential solutions for the treatment crossover problem, given an HTA context. This chapter conducts a systematic review to identify potentially relevant methods, and Chapter 5 adds to this by considering methods in a specific economic evaluation context. The present chapter determines which methods are taken forward for further analysis and assessment via simulation and real-world data studies reported in Part 4 (Chapters 6 and 7) of the thesis. Given that simulation studies can be computationally intensive – especially when complex data are simulated and several methods are applied – it is practically important to only include the most relevant methods. Hence an important part of this chapter involves deciding which methods will be taken forward for further analysis. Previous studies of crossover adjustment methods that have considered an HTA context have not included a systematic review of potentially appropriate methods,²¹ so this chapter represents a novel and important addition to previous research.

The literature search presented is broad in scope, because methods that may be suitable for addressing treatment crossover have not always been developed specifically for this purpose. For example, Robins and Tsiatis's Rank Preserving Structural Failure Time Model (RPSFTM) was primarily developed for addressing non-compliance issues rather than treatment crossover problems.^{22;22;22} Therefore the search terms used are broad. However, the review is specific because identified methods are only included if they are directly applicable to the treatment crossover context. In addition, only methods that are potentially generalisable are included. Case-specific methods, such as that used in the NICE TA of lenalidomide for multiple myeloma, are not included.⁶² This is not to say that these methods are not potentially useful for addressing the treatment crossover problem, but rather that they are only implementable in certain restrictive circumstances (that is, where patient-level data are available from relevant external trials in which treatment crossover did not occur). Such methods will be considered in Chapter 8, alongside the results of my evaluation of more generalisable methods.

4.2 Introduction

In this chapter a systematic search and review of the medical and statistical literature is undertaken to identify relevant methods for adjusting survival estimates in the presence of treatment crossover. Section 4.3 describes the methods used to review the identified papers, while Section 4.4 discusses the methods used to conduct the search systematically. Section 4.5 introduces the search strategy and Sections 4.6, 4.7 and 4.8 describe the methods used for quality assessment, data extraction and data synthesis. Section 4.9 presents the search results and Section 4.10 provides a narrative synthesis and review of all identified methods, categorised by broad methodological group. A decision is made as to whether each method will be included or excluded from further analyses in Part 4 (Chapters 6 and 7) of this thesis. The inclusion/exclusion decision is made based upon my own assessment of the theoretical and practical properties of the methods. Section 4.11 summarises the outcomes of the review. Section 4.12 introduces a novel method that was not identified by the review, but which may provide a useful approach given the treatment crossover mechanism often observed in oncology trials. Section 4.13 provides discussion, conclusions, and considers implications for the remaining parts of the thesis.

4.3 Methods for the review

It has been noted in the health services research literature that systematically reviewing methodologies presents significantly different problems to systematically reviewing clinical interventions.^{115;116} There is no “gold standard” or “standard of care” that the different methods can be compared against and therefore the choice between rival methods depends upon other considerations, such as theoretical suitability and potential bias.

In addition, systematic reviews of the clinical effectiveness of interventions typically compare the results of different studies taking into account study quality, based on issues such as randomisation, blinding, population, allocation, protocol and reporting. Edwards *et al* (1998), and others, state that in the same way, reviews of methods must be considered according to an intellectual framework, allowing their validity to be assessed.^{116;117} Hutton and Ashcroft (1998) and Edwards *et al* (1998) suggest that reviews of statistical methods are likely to be driven by issues such as their practical applicability, reliability and mathematical properties.^{115;116} Theoretical argument is also important because the assumptions made by different methods affect their suitability. Hutton and Ashcroft (1998) state that reviews of methods should be driven by the nature of the arguments that the review will make, and the

uses it will have.¹¹⁵ Given the aims of the review presented in this chapter it is essential to extract information and review the methods identified based upon their suitability (both theoretically and practically) for adjusting survival estimates obtained from RCTs in the presence of treatment crossover, given the economic analysts decision problem. Therefore, based upon the ideas of Hutton and Ashcroft and Edwards *et al* I developed a framework through which alternative methods can be compared, driven by the aims of the review. The framework, which was also used for the extraction of information from identified papers, is presented in Table 4.1.

Table 4.1: Framework for the review of identified methods

Factor	Considerations
Origin	Was the method developed specifically in the survival analysis context? If not, what was the original context and how has the method been adapted? Does the method represent an extension to another method?
Theoretical Suitability	How does the method work? What are the key assumptions? What are the theoretical advantages and disadvantages associated with the method? What are the potential biases associated with the method? Why might the method not be appropriate? How does the method compare to others identified (What are the similarities and differences of the method compared to others identified)?
Application	Is there a worked example in the survival setting? Is the example relevant? What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach)?
Other Issues	Are there any other relevant characteristics associated with the method?

The framework focuses upon three key factors – methodological origin, theoretical suitability, and application – a review of which allows alternative methods to be assessed for inclusion in the analyses conducted in subsequent chapters of this thesis. The origin of the method refers to whether it was originally developed specifically to address the treatment crossover problem, whether it was developed for some other purpose but has been adapted to address treatment crossover, or whether it represents an extension to another method. This may influence whether or not the method is suitable to be taken forward to subsequent sections of this thesis, because if one method extends or supersedes another, it may be unnecessary to include both in further analyses.

Theoretical suitability considers the framework, assumptions, limitations and potential biases associated with the method, as well as its applicability to the treatment crossover context. This is particularly important, as methods that have clear and fundamental flaws will be excluded from analyses in subsequent chapters. All methods are likely to have important limitations but often they will still be useful in certain circumstances and so these remain relevant. However some methods may have flaws that mean they are theoretically incorrect or are inappropriate for addressing the treatment crossover problem. These methods will be excluded from further analyses in this thesis.

Application refers to whether the method has been tested in an applied or simulated setting, and if so whether it was perceived to have performed well. In the interests of practicality the review will not include papers that only describe the application of a method – instead the focus will be on methodological papers that develop novel methods. Hence the review is limited with regard to the extent to which applications of the methods will be picked up. However, often papers that develop new methods include simulated studies demonstrating the performance of the method, or demonstrate the application of the method to a real-world dataset. Where this is the case it is of interest, because it may be revealed that a method works better in some circumstances and poorly in others. Occasionally this may justify the exclusion of the method from further analysis in this thesis.

A methods extraction table was completed for each identified paper using the framework specified in Table 4.1, and these are presented in Appendix 4. The main text of this chapter provides a narrative synthesis of the review focussing upon the key points associated with each identified method.

4.4 Methods for the search

Systematically searching for methodological literature requires a different approach to that employed when systematically searching for literature on clinical interventions. When searching for literature on clinical interventions each paper that reports a new trial of the intervention in question is helpful and relevant for the review. However, as noted by Hutton and Ashcroft (1998), methodological reviews are based upon theoretical arguments, and reviewing the same argument multiple times is not necessary.¹¹⁵ This view is echoed by Edwards *et al* (1998) who accept that it is important to attempt to identify all relevant methods, rather than to pursue each individual paper that discusses a particular identified method.¹¹⁶ Hence, once a method has been identified, it is unnecessary to exhaustively review

all other papers that discuss or apply that method. Such papers, and papers that effectively “duplicate” a method already identified without contributing anything novel, are excluded.

An additional problem associated with searching for methodological literature is that conducting a search prior to having a full understanding of the relevant literature may result in the use of inadequate search terms, and as a result key papers may be missed. Even if a very large range of potentially relevant search terms were used some may be missed without a better knowledge of the relevant literature, and such an approach would be likely to lead to unmanageable results, with very large numbers of irrelevant papers found. Also, often methodological terms may not be listed as keywords, and so searches based upon these (even when appropriate keywords have been identified) may not identify all relevant papers.¹¹⁶

To ensure that I completed an effective systematic methodological search I discussed these issues with an Information Specialist at the University of Sheffield (Suzy Paisley, Head of the Information Resources Group within Health Economics and Decision Science). We decided that an iterative searching approach represented the approach most suited to my needs. To this end, I used a “pearl growing” technique,^{118;119} which has previously been used in methodological reviews in the health economics arena in circumstances where keyword based searches would be likely to yield a large number of irrelevant references.¹²⁰⁻¹²² The approach involves an iterative searching technique similar to the “explosion” technique described by Hutton and Ashcroft (1998).¹¹⁵ Firstly, key papers (the “pearls”) are identified through a very specific search based upon keywords identified in relevant papers already known to the searcher. Identified references are then reviewed and new keywords are identified such that subsequent searches can be run in an attempt to find more relevant papers. In combination with this, the references listed in the “pearls” are checked, and citation searches are run for each “pearl” to identify additional papers. Edwards *et al* (1998) note that searches for methodological topics often lead to a very large yield of theoretical articles, but that the marginal returns associated with reviewing additional papers diminish very quickly after a certain point.¹¹⁶ Therefore, I undertook two iterations of this search process (which will be described in more detail below). As an additional check that no relevant methods had been missed by the search, the list of identified methods were checked with Ian White (Medical Research Council (MRC) Biostatistics Unit) a widely-published expert in the area of statistical methods for addressing treatment crossover.

4.5 Search strategy

4.5.1 Initial search

First an initial search based on specific terms related to the topic area, appearing in the titles, abstracts or keywords of relevant papers already identified was conducted. For example, keywords included in the title of Robins and Tsiatis's (1991) paper introducing the RPSFTM method were "non-compliance" and "failure time".^{22;22} I have identified other useful papers through general reading in the topic area and the terms included in these also informed the search – for example White *et al's* (1999) paper discusses randomisation and "treatment changes".¹²³ Branson and Whitehead (2002) discuss estimating the treatment effect when patients "switch treatment".²³ Mittlbock and Whitehead (1998) consider analysing immediate versus "delayed therapy".¹²⁴

The terms searched for are shown in Table 4.2.

Table 4.2: Initial search terms

Search	Search terms
1	Title=(Treatment crossover OR Switch* treatment OR Treatment switch*)
2	Topic=method* OR approach OR model*
3	Topic=survival OR time-to-event OR failure time
4	#3 AND #2 AND #1
5	Title=(Non-compliance OR Noncompliance OR Non-adherence OR Nonadherence OR Non-randomized OR Nonrandomized OR Post-randomization OR Compliance OR adherence OR randomized)
6	Title=(adjust* OR correct* OR impact* OR amend*)
7	#6 AND #5 AND #3 AND #2
8	Title=(treatment comparisons AND (adjust* OR correct* OR impact* OR amend*)) OR (Treatment actually received) OR (delayed therapy) OR (delayed treatment) OR (treatment changes AND (adjust* OR correct* OR impact* OR amend*)) OR (non-compliance AND (adjust* OR correct* OR impact* OR amend*)) OR (noncompliance AND (adjust* OR correct* OR impact* OR amend*))
9	#8 AND #3 AND #2
10	#9 OR #7 OR #4

Note: ISI Web of Knowledge All Databases; Timespan=All years;

The above search strategy can be simplified by considering it as three separate searches that are then combined. The search specified by (4) identifies papers that have "treatment crossover" or derivatives of "treatment switching" in the title, as well as terms such as "methods", "approach" or "model" and "survival", "time-to-event" or "failure time" in the title, abstract, or key words. Although a search restricted to the appearances of all relevant terms in titles only would be more specific, terms such as "survival" and "methods" may only appear in abstracts rather than titles, and therefore the search is not restricted to titles only.

In summary, search (4) is likely to identify relevant papers that specifically deal with methods for adjusting for treatment crossover in a survival context.

The search specified by (7) is designed to supplement search (4) by picking up relevant papers that do not include the terms “treatment crossover” or “treatment switching” in the title, but which describe treatment crossover in another way (such as “non-compliance”). Search (9) further supplements these searches by picking up additional terms in titles that are directly relevant to treatment crossover, given terms included in the titles of already identified papers (such as “delayed therapy”, from Mittlbock and Whitehead’s 1998 paper).¹²⁴

The ISI Web of Knowledge was used for the search, including all databases. The review was limited to peer-reviewed publications and therefore grey literature including unpublished or ongoing research was excluded.

4.5.2 Inclusions / Exclusions

The initial search was planned to be specific, however due to the subject area a wide range of search terms were required. Therefore a substantial number of citations were anticipated, many of which would not be relevant. The titles and abstracts of identified papers were screened and relevant full-text papers were obtained. The following inclusion and exclusion criteria were applied:

Inclusion criteria

- Relevant methodological papers
- Methods papers directly applicable to the treatment crossover context
- Methods papers which are potentially generalisable

Exclusion criteria

- Papers which only apply a method, rather than describe, discuss or develop the method
- The method is not developed for use in survival analysis
- The method does not allow survival estimates to be adjusted to take into account treatment crossover
- Conference abstracts, with insufficient description of the method
- Naive methods: excluding or censoring crossover patients, or including treatment as a time-varying covariate, since these are well-known and subject to selection bias

4.5.3 Secondary searches

In order to ensure the topic searches were exhaustive, a secondary search was run based upon additional keywords or terms that appeared in the titles or abstracts of relevant papers identified by the initial search but were not included in the initial search terms. Added to this were terms included in papers referenced by identified papers that were relevant but had not been identified by the initial search. The initial search was then run again including these key terms, and the newly identified papers were sifted.

4.5.4 Citations and references

Following completion of the secondary search, it was hoped that most relevant “pearl” papers would have been identified. However, to verify this, the references and citations of the “pearl” papers were reviewed. First, citation searches for each of the “pearl” papers were conducted, and the results were screened to identify any methodological papers that may have developed new methods, or may have further developed already identified methods. Then, the references given in each of the “pearl” papers were screened in order to identify the origins of methods, or alternative methods. It was anticipated that this would provide a complete set of “pearl” papers.

4.6 Quality assessment

No appropriate, published, quality assessment criteria exists for this type of methodological review. Therefore, the framework for the review presented in Table 4.1 was used as a *de facto* quality assessment tool. This relates primarily to “theoretical suitability” – if a method was found to be fundamentally flawed it was excluded from detailed review in the narrative synthesis, and from any further analysis in this thesis. Such flaws included situations where a method had been shown to be mathematically incorrect. A critical appraisal of each method is included in the narrative synthesis.

4.7 Data extraction

Data extraction was based upon the review framework presented in Table 4.1. Details on each element included in the framework were extracted from each paper where relevant information was reported. The evidence tables for each method are presented in Appendix 4.

4.8 Data synthesis

The extracted details on each crossover adjustment method are synthesised in a narrative review. The narrative review discusses the key points identified from the evidence tables presented in Appendix 4. Identified papers were categorised into groups of over-arching methods, and each of these methods are described in detail based on the review framework presented in Table 4.1. For each method, papers which offer extensions and the contribution of these in a treatment crossover context are described.

4.9 Search results

4.9.1 Initial and secondary searches

The initial search identified 560 citations, while the expanded secondary search identified 2,381 citations. The citations generated from the specific searches are presented in Table 4.3 – the additional search terms used in the secondary search are highlighted in italics.

Table 4.3: Initial (29/09/10) and secondary (01/12/10) search results

Search	Search Terms	Results (Initial search)	Results (secondary search)
1	Title=(Treatment crossover OR Switch* treatment OR Treatment switch*)	1,695	1,731
2	Topic=method* OR approach OR model*	>100,000	>100,000
3	Topic=survival OR time-to-event OR failure time OR death	>100,000	>100,000
4	#3 AND #2 AND #1	35	38
5	Title=(Non-compliance OR Noncompliance OR Non-adherence OR Nonadherence OR Non-randomized OR Nonrandomized OR Post-randomization OR Compliance OR adherence OR randomized OR <i>second-line OR second line OR crossover OR cross-over OR switch* OR efficacy OR confound* OR time-vary* OR deviation*</i>)	>100,000	>100,000
6	Title=(adjust* OR correct* OR impact* OR amend* OR <i>account* OR estimat* OR identif* OR analys* OR causal effect OR causal inference</i>)	>100,000	>100,000
7	#6 AND #5 AND #3 AND #2	250	2,053
8	Title=(treatment comparisons AND (adjust* OR correct* OR impact* OR amend* OR <i>account* OR estimat* OR identif* OR analys*</i>) OR (Treatment actually received) OR (delayed therapy) OR (delayed treatment) OR (treatment changes AND (adjust* OR correct* OR impact* OR amend* OR <i>account* OR estimat* OR identif* OR analys*</i>)) OR (non-compliance AND (adjust* OR correct* OR impact* OR amend* OR <i>account* OR estimat* OR identif* OR analys*</i>)) OR (noncompliance AND (adjust* OR correct* OR impact* OR amend* OR <i>account* OR estimat* OR identif* OR analys*</i>))	3,399	3,884

9	#8 AND #3 AND #2	287	324
10	#9 OR #7 OR #4	560	2,381

Note: ISI Web of Knowledge All Databases; Timespan=All years;

Upon screening, 536 of the citations identified by the initial search were excluded because they were not relevant based upon their title and abstract. Several papers were excluded because they did not report on relevant methodologies, rather they had been identified due to the broad nature of the search terms used. Twenty-four papers were selected for review and the full papers were obtained. Of these a further 15 were excluded due to not being relevant owing to reasons bulleted above. Of the 2,381 citations identified by the secondary search, 1,821 had not been identified by the initial search. The titles of these were sifted and the majority were excluded. One-hundred-and-six abstracts were retrieved and of these 48 references appeared potentially relevant, 2 of which were letters referring to other identified papers, and 5 of which were abstracts. Thus based on the secondary search 41 full papers were obtained for review. Of these a further 34 were excluded based upon the exclusion criteria outlined above. Excluded papers from these and subsequent searches, and rationale for their exclusion, are listed in Appendix 5.

4.9.2 References and citation search

The initial and secondary searches led to 16 papers being identified which included novel methods relevant for adjusting time-to-event estimates in the presence of treatment crossover. In total, 317 papers had cited at least one of the 16 “pearls”. The titles and abstracts of these papers were screened and 30 full papers were obtained as they appeared to be potentially relevant. Of these, 5 papers included relevant methods, and 25 were excluded.

Upon completion of the citation search, 21 relevant papers had been identified, and the references included in these papers were then screened in a final attempt to identify any remaining relevant methodologies. Thirty-one new, potentially relevant papers were identified, and full papers were obtained. Of these, several were relevant but discussed or introduced methods that had already been identified by previous searches, or the same method was discussed in more than one paper. In total, three papers identified by the references search were included in the review.

4.9.3 Expert advice

As a check on whether all relevant methodological papers had been identified, I contacted Ian White (MRC Biostatistics Unit) to discuss the methods that had been identified. He confirmed

that he knew of no other relevant methods, but did suggest two papers relevant for inclusion that had not been identified by the search. These were both papers that described the application of identified methods (specifically, the RPSFTM and Marginal Structural Models (MSM)) in the statistical package STATA (White *et al* 2002; Sterne and Tilling 2002).^{125,126} These papers had cited key papers identified by my search, but were published in the STATA Journal which has only been included in the Web of Knowledge database since 2005, and hence they were not picked up. The identification of these papers was useful, but it is not of great concern that articles published in the STATA Journal prior to 2005 were not identified by my search, since these are highly likely to report on methods for applying previously existing methods in STATA, rather than the development of novel methods. These papers were included in the review because they extended previous research on the RPSFTM and MSM methods in that they presented computer programming code demonstrating how these methods could be implemented in practice – hence they describe *how* to apply these methods, rather than simply reporting an *application* of these methods (which would have resulted in exclusion under the first bulleted exclusion criterion in Section 4.5.2).

After the identification of the paper written by Sterne and Tilling (two researchers based at the University of Bristol, UK, who were clearly well aware of available methods for addressing treatment non-compliance) I contacted the authors as a final check on whether my search had identified all relevant methods. They could think of no other relevant methods. They suggested one additional paper (Young *et al* 2010) that did not introduce any new methods, but made interesting comments on MSM and Structural Nested Models (SNM) that were identified by my search.¹²⁷ This paper was included in the review.

After my initial contact with Ian White, he informed me that a potentially relevant paper had been published subsequent to the completion of my search. The paper (Howe *et al*, 2011¹²⁸) was added to the review.

4.9.4 Summary of search results

In total 28 papers were identified that developed methods relevant for adjusting survival estimates to take into account treatment crossover. The literature search process is illustrated in Figure 4.1 and the identified papers are listed in Table 4.4.

Figure 4.1: Summary of literature search

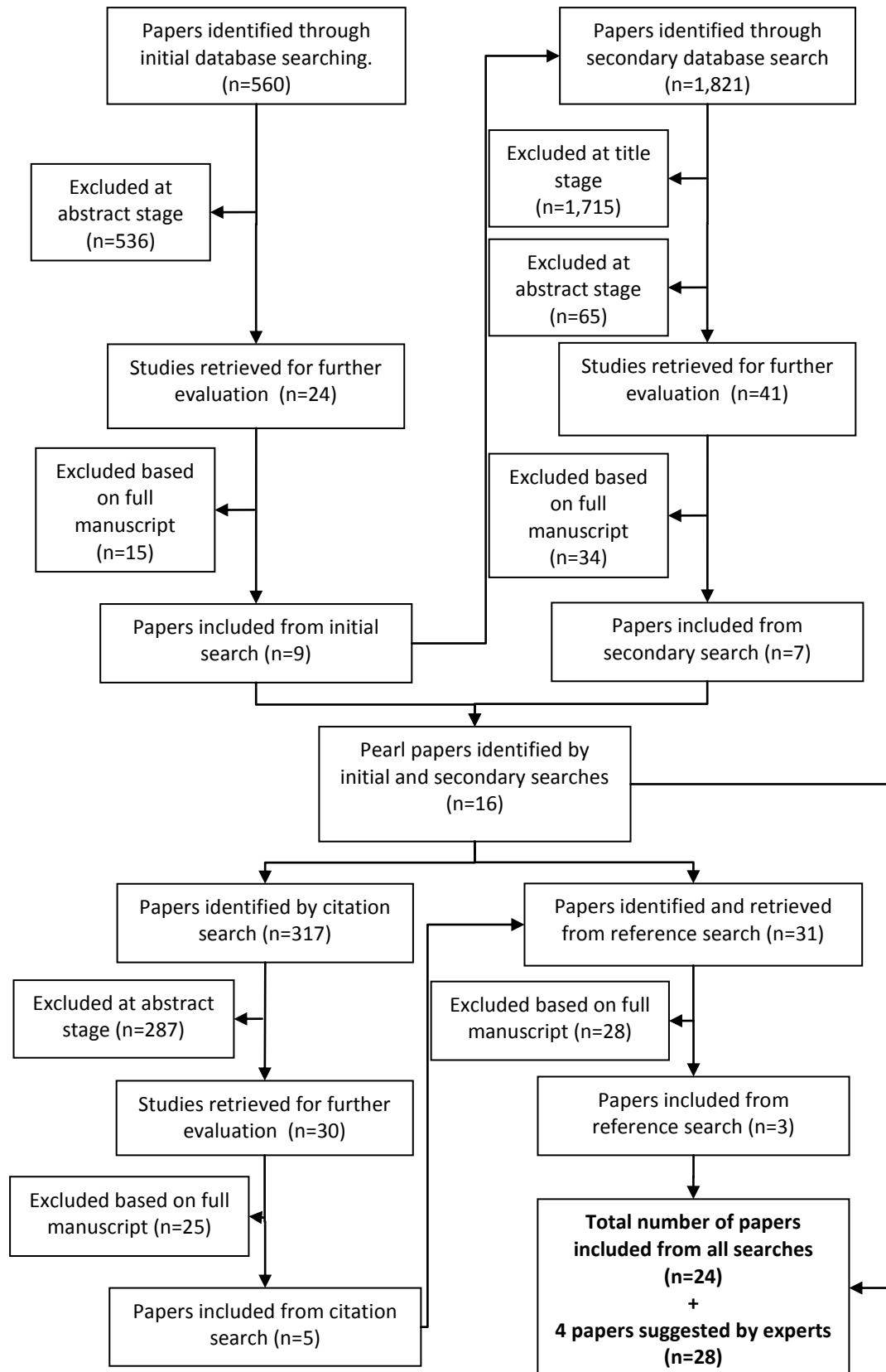


Table 4.4: Papers included in the systematic review

Number	Reference
1	Robins JM and Tsiatis AA (1991) ^{22,22}
2	Robins JM and Greenland S (1994) ¹²⁹
3	Law MG and Kaldor JM (1996) ¹³⁰
4	Mittlbock M and Whitehead J (1998) ¹²⁴
5	White IR, Babiker AG, Walker S and Darbyshire JH (1999) ¹²³
6	Robins JM and Finkelstein DM (2000) ²⁴
7	Branson M and Whitehead J (2002) ²³
8	Walker AS, White IR and Babiker AG (2004) ¹³¹
9	Tanaka Y, Matsuyama Y and Ohashi Y (2008) ¹³²
10	Mark SD and Robins JM (1993) ¹³³
11	Mark SD, Robins JM (1993) ¹³⁴
12	Hernan MA, Brumback B, Robins JM (2001) ¹³⁵
13	Nagelkerke N, Fidler V, Bernsen R, Borgdorff M (2000) ¹³⁶
14	Loeys T, Vansteelandt S, Goetghebeur E (2001) ¹³⁷
15	Yamaguchi T and Ohashi Y (2004) ¹³⁸
16	Huang X, Cormier JN, Pisters PWT (2006) ¹³⁹
17	Shao J, Chang M, Chow SC (2005) ¹⁴⁰
18	Hsu C-H, Taylor JMG, Murray S, Commenges D (2006) ¹⁴¹
19	Lee MLT, Chang M, Whitmore GA (2008) ¹⁴²
20	Murray S, Tsiatis AA (1996) ¹⁴³
21	Robins JM (1998) ¹⁴⁴
22	Robins JM (1999) ¹⁴⁵
23	Hsu CH and Taylor JMG (2010) ¹⁴⁶
24	Witteman JCM, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MAJ <i>et al.</i> (1998) ¹⁴⁷
25	White IR, Walker S, Babiker AG (2002) ¹²⁵
26	Sterne JAC, Tilling K (2002) ¹²⁶
27	Young JG, Hernan MA, Picciotto S, Robins JM (2010) ¹²⁷
28	Howe CJ, Cole SR, Chmiel JS, Munoz A (2011) ¹²⁸

4.10 Results: Narrative synthesis of identified methods

In this section the methods identified are grouped and described, with an emphasis on their theoretical and practical (applied) suitability. The commentary included here focuses on methods that are of most relevance to the treatment crossover problem and that will be taken forward for further analyses in this thesis. Methods that are identified but are not deemed relevant for further analysis are described only briefly and the rationale for their exclusion is given. More details on these (and the included methods) are available in methods extraction tables which follow the framework set out in Table 4.1 – these details are presented in Appendix 4. Not all 28 included papers are discussed in detail, as often they cover very similar methods or only include small extensions to previously discussed methods. Hence several papers are referred to only briefly, with the emphasis of the review on the methods identified, rather than individual papers.

For three of the included papers^{23;130;140} letters were found that had been written to the publishing journal to comment upon the methods introduced.^{148;149} These letters are not formally included in the review, but their points are noted in the commentary on the relevant methods. In addition, the research reported by Yamaguchi and Ohashi (2004) had two parts; Part 1 introduced the methods,¹³⁸ and Part 2 demonstrated their application.¹⁵⁰ Only Part 1 is formally included in the review, but the findings of Part 2 are noted in the relevant commentary.

The search and review identified that there are three broad methodological approaches that are relevant for addressing the treatment crossover problem in the context of RCTs and the economic evaluation decision problem. These are Structural Nested Models (SNM); Rank Preserving Structural Failure Time Models (RPSFTM) which are actually a type of SNM but which are considered separately because they are importantly distinct due to their use of the randomisation assumption, and; Marginal Structural Models (MSM). Various other methods were also identified but were not suitable for further analysis in this thesis – these are not discussed in detail. Table 4.5 groups the methodological categories and their extensions alongside the relevant papers identified by the search, demonstrating how the systematic search has been translated into the narrative synthesis presented in this section, which follows the structure of Table 4.5. The narrative synthesis describes and summarises each group of methods, highlighting the key points from the evidence tables provided in Appendix 4. The three primary methodological categories are described with reference to their origins, their theoretical characteristics and their practical applicability, in line with the review framework outlined in Table 4.1.

Table 4.5: Methodological categories, sub-categories and papers identified

Method category	Method sub-category	Brief description	Papers
Structural Nested Models (SNM) (see section 4.10.1)	Original SNM	Observational-based method for estimating causal effects of time-dependent treatments using accelerated failure time models and g-estimation	Robins (1998) Wittman <i>et al</i> (1998) Sterne and Tilling (2002) Young <i>et al</i> (2010) Mark and Robins (1993a)
	Two-stage SNM	Extension of the SNM that allows application in an RCT context	Robins and Greenland (1994) Yamaguchi and Ohashi (2004)
Rank Preserving Structural Failure Time Model (RPSFTM) (see section 4.10.2)	Original RPSFTM	Randomisation-based version of the SNM	Robins and Tsiatis (1991) Mark and Robins (1993) White <i>et al</i> (1999) White <i>et al</i> (2002)
	Multiparameter RPSFTM	Extension that allows the “common treatment effect” assumption to be relaxed	Mark and Robins (1993b) Robins and Greenland (1994) White <i>et al</i> (1999) Yamaguchi and Ohashi (2004)

section 4.10.2)	Parametric randomisation-based methods	Fully parametric extension of the RPSFTM	Mittlbock and Whitehead (1998) Walker <i>et al</i> (2004)
	Iterative Parameter Estimation	Extension with parametric estimation procedure	Branson and Whitehead (2002) Shao <i>et al</i> (2005)
	Cluster RCTs	Extension for application in cluster RCTs	Loeys <i>et al</i> (2001)
Marginal Structural Models (MSM) (see section 4.10.3)	Original MSM	Observational-based proportional hazards model method developed as an alternative to SNM, using inverse probability weighting	Robins (1999) Hernan <i>et al</i> (2001) Yamaguchi and Ohashi (2004) Huang <i>et al</i> (2006) Young <i>et al</i> (2010)
	Inverse Probability of Censoring Weights	A type of MSM applicable to an RCT	Robins and Finkelstein (2000) Hernan <i>et al</i> (2001) Howe <i>et al</i> (2011)
Other methods (see section 4.10.4)	Auxiliary Variables (AV)	AVs used to to recover lost information on censored observations – superseded by IPCW	Murray and Tsiatis (1996) Hsu <i>et al</i> (2006) Hsu and Taylor (2010)
	Intensity Score	Similar to the IPCW approach, and has not been used in practice	Tanaka <i>et al</i> (2008)
	Adjusted Hazard Ratio method	Adjusted HRs estimated by splitting patients into several groups based upon whether or not they cross over in the future – fundamentally flawed	Law and Kaldor (1996)
	Adjusted Treatment Received	Attempt to model unobserved confounders. Performs poorly in survival context	Nagelkerke <i>et al</i> (2000)
	Threshold Regression Mixture Model	Models the disease process, and may be used in a crossover context. However not directly applicable as currently developed and not used in practice	Lee <i>et al</i> (2008)

4.10.1 Structural Nested Models

4.10.1.1 Origins

Structural Nested Models (SNMs) were initially developed by Robins. The description of them here is based mainly on his 1998 paper.¹⁴⁴ The context of the approach is not specific to treatment crossover, rather it was developed for the estimation of causal effects in the analysis of time-to-event data in observational datasets. However, the method is also potentially useful in a treatment crossover context.

4.10.1.2 Theoretical characteristics

Structural nested failure time models are causal models which estimate the effect of a time-dependent treatment or exposure on a survival time outcome in the presence of time-dependent confounding covariates.¹⁴⁴ The models work by mapping a subject's observed

failure time to the failure time that would have occurred if, possibly contrary to fact, treatment had been withheld, taking into account observed treatment, a patient's confounder history, and an unknown causal parameter. The causal parameter can be identified if the treatment at any time point is randomly assigned conditional on past treatment and confounder history. This requires the assumption that there are no unknown or unobserved confounders – referred to as the “no unmeasured confounders” assumption.¹⁴⁴

When treatment is initially received at different time-points, randomisation at baseline – or controlling for baseline covariates – does not suffice if an estimate of the treatment effect is required, because time-dependent confounders are likely to exist. Time-dependent confounders exist when there are covariates which vary over time (possibly as a consequence of treatment), and these are both prognostic (for example, they predict the risk of death) and they predict exposure to the treatment in question.^{126;134} In the presence of such confounders an analysis that estimates the treatment effect controlling only for baseline covariates will be subject to bias, whereas an analysis that incorporates time-dependent covariates will also result in bias because part of the treatment effect is likely to act through the time-dependent covariates, and this part of the effect would be controlled for in the analysis.¹²⁶

For example, in a clinical trial of a new cancer intervention, the treatment might reduce carcinoembryonic antigen levels over time, and this may lead to a reduction in the risk of death. If treatment crossover is allowed in the trial and patients with a high antigen score are more likely to cross over then some allowance for this needs to be made if the treatment effect in crossover patients is to be estimated. However, including antigen score as a time-dependent covariate would control for the treatment effect that occurs through the antigen, and so bias would result.

In the context of treatment crossover in cancer trials it seems particularly likely that time-dependent confounders will exist because control group patients are usually only permitted to cross over once their disease has progressed, by which time prognostic covariates that were similar between randomised groups at baseline may have diverged, and because clinicians are likely to decide which control group patients will cross over based upon such prognostic covariates. It is this that leads “naive” methods (such as simple censoring and exclusion of crossover patients) to result in selection bias.

SNMs attempt to avoid the bias associated with time-dependent confounders by using information on past treatment and confounder history and by making the “no unmeasured

confounders” assumption. In the context of treatment crossover this assumption means that if there is a reason why some control patients are allowed to cross over and others are not, this reason must be explained by covariates included in the SNM. If this is the case, g-estimation applied to an SNM can be used to provide robust semi-parametric estimators of the causal parameter.¹⁴⁴

Robins states that a specific example of an SNM is a version of the accelerated failure time model of Cox and Oakes (1984).¹⁵¹ The equation given by Robins (1998) is as follows:

$$U = \int_0^T \exp[\psi A_i(t)] dt \quad [1]^{144}$$

Where U is the counterfactual survival time for each patient, which is a known function of observed survival time (T), observed treatment (A, where A is a binary time-dependent variable equal to 1 or 0 over time), and an unknown parameter ψ_0 . If observed treatment was zero – i.e. the treatment of interest was not given – then $U=T$. In addition, if the unknown parameter $\psi_0=0$, then U will always equal T. If ψ_0 does not equal zero, then for a constantly treated patient $T = e^{-\psi_0}U$, hence the patient’s untreated survival time is expanded or contracted by the factor $e^{-\psi_0}$. If $\psi_0 > 0$ treatment is harmful and shortens survival, whereas the opposite is true if $\psi_0 < 0$.¹⁴⁴ $e^{-\psi_0}$ represents the acceleration factor (AF) – if this is less than 1 the treatment is harmful, whereas it is beneficial when it is greater than 1.

Robins (1998) demonstrates that an estimate of ψ_0 , ψ can be estimated using g-estimation, making use of the assumption that there are no unobserved confounders. For each potential value of ψ the counterfactual survival time can be estimated for each patient because all parameters making up its function are known (observed survival time, observed treatment, and ψ). Given this, a grid-search can be undertaken to identify an unbiased estimate of the true value of ψ_0 . This is achieved by estimating the counterfactual survival time for each potential value of ψ and then performing a g-test. The g-test identifies a g-estimate, $\hat{\psi}$, which leads to the treatment received at any time point being random conditional on past treatment and confounder history – essentially this is the value of ψ which leads to counterfactual survival times (the survival time had treatment been withheld) being independent of treatment received. The confidence interval for ψ_0 is given by the values of ψ that result in the g-test not being rejected at the 0.05 level.¹⁴⁴

The model used for the g-test, as specified by Robins (1998), is a time-dependent Cox proportional hazards model for the hazard of treatment change:

$$\lambda_0(t) \exp[\alpha'W(t)] \quad [2]^{144}$$

Where $W(t)$ is a known vector valued function of treatment history and covariate history up until time t , α is an unknown parameter vector, and $\lambda_0(t)$ is an unspecified baseline hazard function. To conduct the g-test the term $\theta Q(t, \psi)$ is added to $\alpha'W(t)$ in the above model, where $Q(t, \psi)$ is a function of treatment and covariate history up until time t and the estimated counterfactual survival time for a given value of ψ . It is the value of ψ that results in a Cox partial likelihood score test (g-test) statistic of zero for the hypothesis $\theta = 0$ in this model that provides a consistent and asymptotically normal estimator of ψ_0 , given the “no unobserved confounders” assumptions holds, the Cox model of the hazard of treatment change is correct, and the SNM is correct.¹⁴⁴ With this value of ψ the treatment received is independent of counterfactual survival times and thus $\hat{\psi}$ represents the causal treatment effect.

Censoring presents a problem for SNMs because it means that counterfactual survival times can only be estimated for a subset of patients (those that were not censored). Robins (1998) presents details of how censoring (both informative and random) can be incorporated within an SNM.¹⁴⁴ In a carefully planned RCT, administrative censoring (that is, censoring at the end of follow-up) is likely (and some censoring due to loss-to-follow up may occur), but this should be uninformative (that is, random and independent of the counterfactual survival time). In this case, censoring for each individual i , (C_i), can be dealt with relatively simply in an SNM, as demonstrated by Wittelman *et al* (1998). In the g-test model the counterfactual survival time for a given value of ψ ($U_i(\psi)$) is replaced by $\Delta_i(\psi)$, a function of $U_i(\psi)$ and C_i , that is observed for all subjects.¹⁴⁷ For example, as defined by Wittelman *et al* (1998), let

$$\Delta_i(\psi) = 1 \text{ if } U_i(\psi) < C_i(\psi) \text{ and}$$

$$\Delta_i(\psi) = 0 \text{ if } U_i(\psi) \geq C_i(\psi)$$

where $C_i(\psi) = C_i$ if $\psi \geq 0$ and $C_i(\psi) = C_i \exp(\psi)$ if $\psi < 0$.¹⁴⁷ With $C_i(\psi)$ defined in this way when a subject is censored ($T_i > C_i$), $U_i(\psi)$ will always be greater than $C_i(\psi)$ and thus $\Delta_i(\psi)$ will always be observed, and will equal zero (thus patients are “ ψ -censored”) for all patients who had censored survival times.¹⁴⁷ It may also be zero for some patients who had observed survival times, depending upon the size of the treatment effect ψ and the individuals

treatment histories. This allows the g-test to be conducted, but with counterfactual survival time indicated by $\Delta_i(\psi)$ taking into account censoring, instead of simply being $U_i(\psi)$. This is reasonable because for treatment received to be independent of U_i conditional on past treatment and covariate history, it must also be independent of Δ_i because this is a function only of U_i and C_i , and C_i is essentially a baseline covariate (since it represents the end of follow-up time).¹⁴⁷

More complex censoring due to competing risks (for example, causes of death unrelated to the disease in question) can also be handled within an SNM, by combining an IPCW approach with the SNM. This is described by Robins (1998),¹⁴⁴ but will not be explained here, as the IPCW approach is addressed later in this chapter.

4.10.1.3 Practical applicability

The original SNM is designed for estimating causal treatment effects in observational studies – it is not specifically designed for treatment crossover and does not make use of the randomisation element of an RCT in any way. The method could not be applied simply to an RCT dataset, because it would be inappropriate to attempt to model the treatment process (that is, the treatment received by patients over time) when patients are randomised to treatment groups – patients randomised to the control group would not receive the intervention (until crossover was allowed) irrespective of their covariates, and similarly patients in the intervention group would remain in that group (though they may discontinue treatment) irrespective of their covariates. Attempting to model the treatment process based upon observed covariates would be counter-intuitive. Even if this were possible, if the method were applied to the whole trial population in order to estimate a causal effect of the treatment the result would be an average effect based upon all patients who took the treatment. It is arguable how useful this is from the health economist's decision problem perspective, because what is desired is an estimate of the treatment effect in patients initially randomised to the intervention group – not an average effect incorporating this group and patients who received this treatment later on, once their disease had progressed.

However, these issues do not mean that the SNM method cannot be useful in an RCT context, given the decision problem faced in an economic evaluation. In the context of an RCT, the control group after the point at which treatment crossover becomes possible could be treated as an observational dataset. The SNM method could then be applied to this dataset to estimate the treatment effect specific to control group (crossover) patients. Given the resulting treatment effect estimate (in terms of an acceleration factor – working on the time

scale) the survival times of control group crossover patients could be adjusted to estimate counterfactual survival times had crossover not occurred. This is similar to the approach taken by Robins and Greenland (1994) and Yamaguchi and Ohashi (2004).^{129;138}

The benefits of the SNM approach as applied to an RCT are that it allows a treatment effect specific to crossover patients to be estimated, and in turn this leads to an estimate of the treatment effect in the intervention group that meets the requirement for the economic evaluation decision problem. Linked to this, the method does not require that the treatment effect is the same in patients initially randomised to the experimental group and crossover patients. Also, because the approach makes use of all available data on covariates it may result in relatively certain estimates of the treatment effect (that is, with relatively (compared to a randomised analysis) narrow confidence intervals).

However, the approach also suffers from important limitations. Most important is the “no unmeasured confounders” assumption. For the method to be unbiased the covariates included in the model must cover all variables that may impact upon treatment received and survival. This assumption cannot be tested using the observed data.^{129;138} In reality it is likely that no matter how many covariates are measured the treatment process and survival will not be fully explained, and thus some bias may be expected when using methods that rely on “no unmeasured confounders”. However, if all important covariates are included this bias may be small. Potential problems associated with the “no unmeasured confounders” assumption were highlighted by Yamaguchi and Ohashi (2004) who attempted to apply an SNM in an RCT context and had problems with the availability of important prognostic covariate data.¹³⁸

Robins and Greenland (1994) recommend that SNMs are richly parameterised as this increases the likelihood of the “no unmeasured confounders” assumption holding, and also has been shown to reduce bias when the proportional hazards model is misspecified. A further advantage is that it increases efficiency when the model is specified correctly.¹²⁹ Importantly, they note that an advantage of an alternative approach – the IPCW – is that it does not require the correct specification of an SNM for the second-line (crossover) treatment effect, although the IPCW method is associated with other potential disadvantages (such as the inefficiency associated with loss of information – which will be discussed later).¹²⁹

The SNM method is also limited for the purpose of adjusting for treatment crossover in an RCT due to its observational data origins. Observational datasets are usually much larger than RCT datasets, and so more data are usually available. When fewer data are available SNMs may

become less stable and confidence intervals may actually be wider than analyses that make use of the randomisation assumption. This is particularly important when it is considered that the method would only be applied to the control group in an RCT setting – usually meaning that the study size is halved (assuming 1:1 randomisation). For an SNM to be applied, a substantial amount of data need to be collected in the clinical trial, and also it must be possible to define for each patient when they became “at risk” of treatment crossover so that the SNM is only applied to the relevant “observational” dataset.

If either almost all or very few control group patients crossover, estimating the treatment effect and the treatment process based upon observed covariates may be inaccurate and open to bias. In practice the SNM approach may only be suitable for adjusting for treatment crossover in trials where the control group was relatively large, and where a reasonable number of control group patients both did and did not crossover. The SNM method is suitable for inclusion in the further analyses included in this thesis, and one aim of the simulation study presented in Chapter 6 is to determine scenarios in which the SNM method performs well.

4.10.2 Rank Preserving Structural Failure Time Models (RPSFTM)

4.10.2.1 Origins

In 1991 Robins and Tsiatis developed the RPSFTM specifically for estimating causal effects in the presence of non-compliance in an RCT setting.^{22,22} Their method was developed to deal with situations in which patients did not take the prescribed amount of a treatment, but could also be used in cases where patients took treatments that were not assigned to them. Thus, the method is relevant for dealing with treatment crossover. Robins and Tsiatis’s approach is simply an application of SNM methodology to a randomised trial in which only one treatment effect is measured, and where inference is based purely on randomisation.

4.10.2.2 Theoretical characteristics

Like the SNM approach described above, the RPSFTM uses a counterfactual framework to estimate the causal effect of the treatment in question. However rather than modelling the treatment process using treatment and prognostic covariate history to identify the causal treatment effect, the RPSFTM identifies the treatment effect using only the randomisation of the trial, observed survival and observed treatment history. It is assumed that if two patients have the same observed event time and neither have received treatment, those two patients would also have the same event time if they both received treatment. This assumption is linked to the associated assumptions that the treatment effect is equal for all patients no matter when the treatment is received (the “common treatment effect” assumption), and –

most importantly – that the randomisation of the trial means that there are no differences between the treatment groups, apart from treatment allocated.^{22;22}

The method splits the observed event time (T_i) for each patient into two, that is the event time when the patient is on the control treatment (T_{Ai}), and the event time when the patient is on the intervention treatment (T_{Bi}). For patients who are randomised to the intervention treatment, and who do not switch onto the control treatment (that is, when compliance is full in the treatment group), T_{Ai} is equal to zero. For patients randomised to the control group who do not switch onto the intervention (i.e. compliance is full in the control group) T_{Bi} is equal to zero. However, for patients who switch treatments (for whom compliance is imperfect) both T_{Ai} and T_{Bi} will be greater than zero.

The RPSFTM method relates T_i to the counterfactual event time (U_i) with the following causal model:

$$U_i = T_{Ai} + e^{\psi_0} T_{Bi} \quad [3]$$

As stated in Section 4.10.1.2 $e^{-\psi_0}$ represents the acceleration factor associated with the intervention. By defining a binary process $X_i(t)$ which equals 1 when a patient is on the intervention treatment, and equals zero when the patient is on control treatment, the causal model can be rewritten as:

$$U_i = \int_0^{T_i} \exp[\psi X_i(t)] dt \quad [4]$$

Which is identical to the SNM introduced above (see equation [1]). The value of ψ is estimated using a grid search. U_i is estimated using the causal model for each value of ψ , and the true value of ψ is that for which $U(\psi)$ is independent of randomised groups. Because the estimation procedure is no longer reliant on a vector of various time-dependent and time independent covariates the RPSFTM described by Robins and Tsiatis (1991) may be regarded as a simplification of the SNM which is made possible because the RPSFTM was developed specifically for use with data from RCTs. As described by Mark and Robins (1993), a log-rank or Wilcoxon test can be used for the RPSFTM g-test in a non-parametric setting, testing the hypothesis that the baseline survival curves are identical in the two treatment groups, or a Wald test could be used for parametric models.¹³³ The point estimate of ψ is that for which the test (z) statistic equals zero.

White *et al* (1999) demonstrate that censoring is problematic for the RPSFTM.¹²³ They define C_i as the administrative censoring time which corresponds to the end of follow-up and which is known for all patients, and show that the censoring time for $U_i(\psi)$ is given by:

$$D_i(\psi) = \int_0^{C_i} \exp[\psi X_i(t)] dt \quad [5]$$

However, this demonstrates that $D_i(\psi)$ depends upon X_i (treatment received), which may depend upon prognostic factors. This is because a positive or negative treatment effect may increase or decrease the probability that the survival time of an individual is censored, and, where treatment crossover occurs, treatment received is likely to be associated with prognosis. In turn, this means that D_i may depend on prognostic factors, and thus the censoring of $U_i(\psi)$ is informative.¹²³ White *et al* suggest possible bias from this be avoided by breaking the dependence between censoring time and X_i and by recensoring $U_i(\psi)$ at the earliest possible censoring time given the treatment effect ψ .¹²³ For example, if treatment group could have always been 0, or always been 1, then the minimum censoring time for counterfactual event times is:

$$D_i^*(\psi) = \min(C_i, C_i \exp \psi) \quad [6]$$

Given this, the censoring time of the counterfactual event time is now independent of X_i and in the estimation procedure the counterfactual event time $U_i(\psi)$ is replaced by the censoring time of the counterfactual event time $D_i^*(\psi)$ if $D_i^*(\psi) < U_i(\psi)$. This means that for patients that had an unfavourable treatment history (for example, control group patients that did not crossover onto a beneficial new treatment; or control group patients that did crossover onto a detrimental new treatment) and who experienced the event of interest (such as death) close to their administrative censoring time, the event may not have been observed if they had received more favourable treatment and therefore these patients have their survival times recensored and their events are no longer observed.¹²³ Recensoring is applied to all patients in groups in which crossover occurred. As for the SNM, informative censoring due to competing risks can be controlled for within an RPSFTM analysis using an IPCW approach.

The RPSFTM developed by Robins and Tsiatis^{22;22} makes a number of important assumptions – in particular, the time at which an individual would fail (that is, the survival time) if never treated is not related to the treatment arm to which the individual is assigned (the

randomisation assumption); the treatment effect is equal no matter when treatment is received (the “common treatment effect” assumption); if an individual fails before another individual on one treatment regime, he will also fail before that other individual on all other treatment regimens and if two individuals had identical observed failure times and treatment histories, they would also have identical failure times if treatment had been withheld. Finally, the accelerated failure time model must be correct, which cannot be tested.

Perhaps the most important of these assumptions are that randomisation has worked perfectly, and the assumption of a “common treatment effect”. The randomisation assumption should be reasonable in the context of an RCT, but may be called into question if sample sizes are particularly small. The “common treatment effect” assumption is more problematic and is a legitimate concern when considering the use of the RPSFTM to adjust for treatment crossover in an RCT and economic evaluation context. The RPSFTM makes the assumption that there is not a difference in the treatment effect (relative to the time for which the treatment is taken) in patients initially randomised to the intervention compared to control group patients who cross over. Therefore, effectively it estimates an “average” treatment effect taking into account these two sets of patients. Unlike the observational SNM it would not result in different treatment effects being estimated for intervention group patients and crossover patients.

Owing to this, if crossover patients actually receive a lower treatment effect than intervention group patients, the treatment effect (in terms of an acceleration factor) estimated by the RPSFTM will be an underestimate of the effect actually received in the intervention group – and thus an underestimate of what is really required given the economic evaluation decision problem. However, while the acceleration factor will be an underestimate, the control group counterfactual survival times are also likely to be underestimated, because the acceleration factor applied to these patients to derive counterfactual survival times will be larger than the acceleration factor actually experienced by these patients. Hence if the experimental group is compared to the control group by comparing the observed experimental group survival times to the control group counterfactual survival times the effect of the experimental treatment will be over-estimated. Conversely, if crossover patients happen to receive a higher treatment effect than patients in the intervention group, the RPSFTM acceleration factor will be an overestimate of the effect received in the intervention group, and control group counterfactual survival times will be over-estimated. Either way, unless the “common treatment effect” assumption holds, the RPSFTM treatment effect will not deliver an unbiased estimate of the treatment effect required to address the economic evaluation decision

problem. It seems unlikely that the treatment effect will remain constant no matter when treatment is given – a patient with highly progressed disease seems likely to have a lower capacity to benefit than a patient with disease that has not progressed – and therefore the “common treatment effect” assumption may be biologically implausible. It is also clear that the method for characterising the treatment effect in the economic model is very important – use of the RPSFTM acceleration factor or the control group counterfactual survival times is likely to lead to bias in opposite directions.

An additional limitation associated with the RPSFTM is the recensoring method used to avoid bias through informative censoring. Recensoring involves “throwing away” information, and the larger the treatment effect the more information will be discarded. Therefore in this sense the RPSFTM is inefficient, and information at the tail of the survival curve that may be important for extrapolation purposes is excluded. This is a problem particularly from the perspective of a health economist, because usually extrapolation is required to estimate mean survival times (as discussed in Chapter 1).

However, a key advantage of the RPSFTM method is that it is a randomisation-based effect estimator (RBEE) as classified by White (1999).¹²³ By design, its significance level is equal to the significance level of an ITT test, and it is equal to an ITT effect estimator if there are no treatment changes or if the treatment effect ψ is zero. Hence the z-statistic value when ψ is zero reflects the usual ITT logrank test comparing survival times between the two groups.¹²³ Whenever the ITT analysis is not significant, the RPSFTM confidence intervals will include 0. The method may therefore inspire more confidence than an observational-based approach (such as an SNM, or IPCW) that does not use the randomisation assumption and which is based upon the “no unmeasured confounders” assumption. Linked to this, the confidence intervals associated with the RPSFTM treatment effect may be relatively large because the p-value from the ITT analysis is maintained. This may be regarded as a weakness of the approach, if an observational method leads to narrower confidence intervals.

4.10.2.3 Practical applicability

Practically, there are alternative ways to implement the RPSFTM method. If equation [3] is directly implemented, patients in the experimental group would be classed as “crossing over” onto control treatment at the time that they discontinue experimental treatment. Also, control group patients who switch onto the experimental treatment would “switch back” to the control treatment once they discontinue the experimental treatment. This “on treatment” approach would result in a “causal” treatment effect being estimated for the experimental

treatment – that is, the effect while on treatment, and it would be assumed that the treatment effect is lost as soon as treatment is discontinued. In the context of an economic evaluation this is problematic, because the decision problem requires that a state of the world in which the new treatment exists is compared to a state of the world in which the treatment does not exist, irrespective of whether patients in the “new treatment” world have to discontinue treatment at some point. If patients in the experimental group have to discontinue treatment in the clinical trial, it is likely that patients will also have to discontinue treatment in the real-world. Therefore, what is required from the analysis of the trial is an estimate of the treatment effect associated with being in the intervention group compared to the control group, rather than the causal treatment effect measured only while treatment is being received. This could be achieved under the “on treatment” approach by estimating counterfactual survival times for control group patients using the RPSFTM, and then comparing these counterfactual survival times to observed experimental group survival times – an “on treatment – observed” approach. An alternative would be to simply apply the treatment effect estimated by the “on treatment” approach only while patients are receiving treatment within the economic model. However, the “on treatment” and “on treatment – observed” methods of application remain limited due to two other important factors.

Firstly, the use of the RPSFTM (and IPE) method is made problematic if the comparator treatment used in the RCT is active. The RPSFTM counterfactual survival model requires that patients are either “on treatment” or “off treatment”. If patients in the control group receive an active treatment followed by supportive care upon treatment failure the “off treatment” category represents more than one type of treatment and the counterfactual survival model is not appropriate – in particular, survival in the control group is likely to be “diluted”, as the active comparator treatment is combined with post-treatment care. Secondly, the RPSFTM counterfactual survival model assumes that the treatment effect is only received while a patient is “on treatment” – it disappears as soon as treatment is discontinued. The clinical plausibility of this assumption should be considered.

If a continuing treatment effect is expected or the comparator is an active therapy the RPSFTM (or IPE) methods could be applied on a “treatment group” basis – where patients in the experimental group are always considered to be “on treatment” and patients that switch remain “on treatment” from the time of switch until death. This analysis ignores treatment discontinuation times and requires there to be a common treatment effect associated with the sequence of treatments received by patients randomised to the experimental group and the sequence of treatments received by switchers after the point of switch. Any benefits

associated with post study treatments will be attributed to the experimental treatment, though similarly any benefits from post-study treatments received by control group non-switchers would be attributed to the control group. If the post study treatments received in all groups represent realistic treatment pathways this approach may appropriately address the economic evaluation decision problem – particularly if the costs of the post-study treatments are also incorporated within the economic model.

In a “treatment group” analysis groups still have to be combined – time spent “on” treatment is any time after experimental treatment initiation, ignoring subsequent treatment discontinuation; and time spent “off” treatment is all time prior to experimental treatment initiation, including time after control group treatment discontinuation if crossover does not occur. Hence such analyses may dilute the effectiveness associated with the experimental treatment and the active comparator treatment as for both groups time spent after discontinuation of the treatments is attributed to those treatments. However, while not providing a causal effect of the novel treatment, this approach may provide a reasonable estimate of the treatment effect of being randomised to the experimental treatment group rather than the control group, which is useful for economic evaluation. Rather than diluting only the comparator group, as would be the case in an “on treatment” analysis, the “treatment group” approach dilutes both groups and may be preferable in situations where the comparator treatment is active and importantly different to post-study treatments, where post-study treatments are similar and have similar effectiveness in crossover patients and patients initially randomised to the experimental group, and particularly where some of the benefit associated with the experimental treatment is maintained beyond treatment discontinuation.

The “on treatment – observed” and “treatment group” approaches both provide estimates of the treatment effect associated with the observed experimental group. However, these methods differ in two important ways. Firstly, the “on treatment – observed” approach assumes the treatment effect disappears as soon as treatment is discontinued, whereas the “treatment group” approach does not. Secondly, the “on treatment” approach is likely to generate a larger treatment effect than the “treatment group” approach because the initial analysis compares full treatment to no treatment, rather than observed treatment to no treatment (this may not be the case, however, if there was a continuing benefit of treatment after discontinuation which would not be attributed to the experimental treatment by an “on treatment” analysis). Usually the counterfactual survival times generated by the “on treatment” approach, which are then used within the “on treatment – observed” approach,

will be subject to more recensoring than those produced by the “treatment group” approach (because recensoring is associated with the size of the treatment effect). Thus the potential limitations associated with recensoring may be more important for the “on treatment – observed” approach. Hence, the choice between “treatment group” and “on treatment – observed” analyses should be informed by the clinical plausibility of assuming (or not) a continuing treatment effect, and a consideration of the importance of recensoring. In reality, the “truth” may be somewhere between the two options – there may be some continuation of the treatment effect, but this may wane over time. Hence it is likely to be useful to present both methods alongside clinical justification for their relative plausibility.

The “treatment group” approach presents its own problems, because the RPSFTM treatment effect then becomes an average treatment effect of that observed over a lifetime in the experimental group (irrespective of when treatment discontinuation occurred) and that observed from the point of crossover until death in crossover patients. This is likely to “dilute” the acceleration factor in the intervention group relative to the control group if the majority of the treatment benefit is accrued while it is being taken, because the control group only begin to receive the treatment at a later time point and the time between their discontinuation of treatment and death is likely to be shorter. Hence calculating an “average” treatment effect (acceleration factor) across all patients that received the intervention may actually mean that the treatment effect relevant for the economic evaluation is over-estimated as the acceleration factor may actually be larger in crossover patients than in intervention group patients – which would lead to under-estimated counterfactual control group survival times.

Also, the “treatment group” approach requires that the post-study treatments received in the randomised groups are representative of a realistic treatment pathway – in which case if subsequent benefits received due to post-study therapy differ depending upon previous treatment these benefits can appropriately be attributed to the experimental treatment group in the economic evaluation (provided the costs of these are also included). If, on the other hand, post-study treatments received are not representative of realistic treatment pathways and these and their effectiveness differ between randomised groups then a “treatment group” analysis may attribute benefits or dis-benefits to the experimental treatment group inappropriately, and would lead to inaccurate cost-effectiveness results. Hence it is very important to assess the post-study treatments received and to obtain clinical expert opinion on whether these represent realistic treatment pathways.

Yet another alternative to the “on treatment”, “on treatment – observed” and “treatment group” approaches may be to apply the RPSFTM method on an “ever treated” basis, whereby if a patient was randomised to the intervention group, or if they crossed over onto the experimental treatment at any time-point, they are assumed to be in the “treated” group from time zero until death. However this seems somewhat artificial, as a treatment effect would be associated with crossover patients before the point of crossover, and implies the assumption that treatment duration is similar between experimental and control groups. If – as seems likely – treatment duration is shorter in crossover patients than in the intervention group the average treatment effect across these groups is likely to be an underestimate of what is required for the economic evaluation.

Despite these issues, the RPSFTM method has been shown to produce very low levels of bias when its assumptions are satisfied,²¹ and it has been used in NICE TAs. Therefore it is important to include this method in the analysis contained in subsequent chapters of this thesis. It is particularly important to understand the implications for the performance of the method when its assumptions about the treatment effect are not satisfied – this forms an important question that is addressed in the simulation study presented in Chapter 6.

4.10.2.4 Extensions

Various authors have attempted to address some of the problematic issues associated with the RPSFTM, and their methods are discussed below.

- Equal treatment effect assumption

Robins and Tsiatis^{22;144} acknowledged that a key limitation of their RPSFTM approach was the non-interaction (common treatment effect) assumption. They state that their model could be generalised to allow the effect of treatment to depend on time-dependent covariates using a multi-parameter model. However, when analysts have attempted to apply a multi-parameter version of the RPSFTM the results have not been successful as meaningful point estimates for causal effects are difficult to determine.^{123;129;133;134;138} Hence, at present this limitation has not been resolved.

- Parametric randomisation-based methods

Walker *et al*¹³¹ note that the recensoring method previously described by White *et al*¹²³ entails a loss of information because recensoring can lead to information from late in follow-up being omitted. They state that this is a particular problem if there is a treatment by time interaction. Motivated by this, they developed a bivariate parametric frailty model for time to treatment

change and time to trial endpoint which avoids informative censoring and the need for recensoring.¹³¹ Essentially, the method is a fully parametric extension of Robins and Tsiatis's approach. However, from a practical perspective this approach has proven unsuccessful. Walker *et al* (2004) acknowledge that their approach requires a number of models to be correctly specified,¹³¹ and this cannot be tested. Morden *et al* (2011) found that the method was of limited practical use because it often failed to converge and often led to substantial overestimates of the treatment effect.²¹ In fact, it generally produced more bias than naive approaches (ITT, censor or exclude crossover patients). Therefore, this method is not included in further analyses in this thesis. For information, further details of the method are presented in Appendix 4.

Mittlbock and Whitehead (1998) also developed a parametric approach to adjusting survival estimates in the presence of treatment crossover; this involved estimating counterfactual survival times by modelling the dependence between the observed survival time and the time of switching to the new treatment.¹²⁴ The authors developed their method primarily because they were concerned that treatment switching causes an important loss of power in trial analysis. They note that adjustment methods such as the RPSFTM involve maintaining the same significance level as the unadjusted ITT analysis, and thus no attempt is made to regain lost power. However, the authors admit that their parametric model is not put forward as a comprehensive alternative to an unadjusted ITT analysis, and that it is highly speculative and requires further development – they state that the main message of their paper is that limited information about the absolute effect of a new treatment will be forthcoming from a trial that effectively tests immediate versus delayed therapy.¹²⁴ In addition, the method produced questionable results when applied to a real-world dataset. Hence, this method is not considered further here. Again, for information, further details of the method are presented in Appendix 4.

- Iterative Parameter Estimation

Branson and Whitehead (2002) extended the RPSFTM method using parametric methods, developing a novel iterative parameter estimation (IPE) procedure.^{22,23} The same type of accelerated failure time model is used, but a parametric failure time model is fitted to the original, unadjusted ITT data to obtain an initial estimate of ψ . The observed failure times of crossover patients are then re-estimated using $e^{-\psi}$ and the counterfactual survival time model presented in equation [3], and the treatment groups are then compared again using a parametric failure time model. This will give an updated estimate of ψ , and the process of re-estimating the observed survival times of crossover patients is repeated. This iterative process

is continued until the new estimate for $e^{-\psi}$ is very close to the previous estimate (the authors suggest within 10^{-5} of the previous estimate but offer no particular rationale for this), at which point the process is said to have converged.²³

Branson and Whitehead (2002) take administrative censoring into account by replacing the transformed (counterfactual) survival time with the administrative censoring time if the transformed survival time is greater than the administrative censoring time. In this way, recensoring is included in the method but it is restricted to crossover patients and is only applied if the new treatment is detrimental compared to the control – only in this case could the counterfactual survival time be longer than the administrative censoring time. This is different from the recensoring approach developed by White *et al* (1999) in their extension of the RPSFTM method, in which recensoring is applied whether or not the new treatment is detrimental (although recensoring is not applied in randomised groups in which no crossover occurs).^{123;125}

Branson and Whitehead (2002) state that their recensoring approach is beneficial because it minimises the associated loss of information.²³ However, White (2006) demonstrates that Branson and Whitehead's recensoring is insufficient because recensoring is required whether or not the new treatment is detrimental.¹⁴⁹ Informative bias is likely even if the new treatment is effective, because counterfactual survival times are associated with patient prognosis, as demonstrated by equation [5]. Patients who benefit from crossing over and have their survival time censored may have died if they had not crossed over. Because crossover is associated with prognosis, counterfactual censoring times are therefore also associated with prognosis and hence recensoring at the minimum possible censoring time ($C_i \exp \psi$) is required. White demonstrated that applying the IPE method without full recensoring may result in important bias.¹⁴⁹ Therefore, if the IPE method is to be applied, full recensoring should be applied.

Branson and Whitehead (2002) state standard errors and confidence intervals for the treatment effect could be taken from the final regression in the IPE process, or bootstrapping could be used.²³ However, the authors state that the standard errors from the final regression will be underestimates because the covariance matrix from the final iteration of the IPE algorithm does not take into account the fact that control arm patients have had their survival times transformed by the algorithm – therefore bootstrapping may be preferable. The authors also recommend that the IPE estimate and confidence interval is accompanied by the ITT p-value because as for the RPSFTM method the significance level remains the same as that associated with the ITT analysis.²³

The IPE procedure makes similar assumptions to the RPSFTM method – for example the randomisation assumption is made, as is the “common treatment effect” assumption. An additional assumption made by Branson and Whitehead (2002) is that survival times follow a parametric distribution, and thus they state that it is important to identify suitable parametric models using tools such as log-cumulative hazard plots. Hence similar advantages and disadvantages exist for the IPE method as compared to the RPSFTM method, with the addition of possible disadvantages if a suitable parametric distribution cannot be identified.

Branson and Whitehead (2002) and Morden *et al* (2011) conducted simulation studies to assess the performance of the IPE method, and found that the method produces very low levels of bias when the assumptions made by the method (in particular, the “common treatment effect” assumption) hold.^{21;23} Morden *et al* (2011) found that the method’s performance was very similar to the RPSFTM, which is to be expected as the methods are similar other than in their estimation procedure. Given the success of the method it is included in further analyses in this thesis. Again, it is particularly important to assess the performance of the method when its assumptions are not satisfied.

Shao *et al* (2005) attempted to extend the IPE method.¹⁴⁰ They claimed that the IPE and RPSFTM methods failed to account for the fact that the treatment crossover decision is often associated with prognosis. As noted by White (2006) in his response to Shao *et al*’s paper, this is a confusing statement because in fact the RPSFTM and IPE methods make no assumptions around prognosis because instead their estimation is based upon the randomisation of the trial.¹⁴⁹ It seems that the point Shao *et al* (2005) are attempting to address is that the treatment switching decision is specific to individual patient-level characteristics and prognosis, and that the effect of the treatment once switching has occurred will depend upon these variables. Hence, it seems that the true objective of Shao *et al*’s method is to allow treatment effect to vary between patients depending upon certain prognostic characteristics – in particular whether they were initially randomised to the treatment or whether they switched on to the treatment after being randomised to the control group.

Shao *et al* (2005) develop two methods.¹⁴⁰ One is based on the IPE method, and the other takes a semi-parametric approach using a Cox proportional hazards model, whereby counterfactual hazard rates are considered rather than counterfactual event times. In both methods, an extra term is incorporated within the models to allow the treatment effect to differ for crossover patients compared to patients initially randomised to the experimental

group. This is potentially useful, because the “common treatment effect” assumption is a key weakness of the RPSFTM and IPE methods. However, both of the methods developed by Shao *et al* involve estimating the treatment effect using conditional likelihood functions, where the results are conditional on switching (crossover) time. White (2006) identifies this to be a serious problem, because the conditional likelihood approach is only valid if the treatment crossover observed is random and independent of prognosis.¹⁴⁹ This is the same problem associated with the naive method of censoring crossover patients from a standard ITT analysis, which is very likely to lead to selection bias because typically crossover patients have a different prognosis (either better or worse) to non-crossover patients. Hence, the Shao *et al* methods are very likely to be biased. They are therefore not considered any further in this thesis. For information, further details on the methods are available in Appendix 4.

- Cluster randomised trials

Loeys *et al* (2001) state that the RPSFTM method ignores clustering in the randomisation process, which creates the risk that the method may ignore correlation between survival outcomes.¹³⁷ This may occur if the health experience of earlier cluster members impacts upon the compliance of later cluster members. The authors consider an extension to the RPSFTM method to allow for cluster randomisation, involving the use of a cluster-specific frailty term. Typically pharmaceutical trials of new cancer interventions are not cluster randomised and therefore Loeys *et al*'s method represents an unnecessary extension to the RPSFTM approach for the context investigated in this thesis. Therefore it is not considered any further, although details on the method are available in Appendix 4.

4.10.3 Marginal Structural Models and Inverse Probability of Censoring Weights

4.10.3.1 Origins

Robins (1999) and Hernan *et al* (2001) originally introduced marginal structural proportional hazards models (MSMs) as a proportional hazards alternative to SNMs.^{135;145} MSMs are built on similar assumptions as SNMs – for example, correct functional forms of models are assumed, and it is assumed that data have been obtained on all time-independent and time-dependent covariates that predict subsequent treatment and mortality (the “no unmeasured confounders” assumption is made).^{135;145} However MSMs use a Cox proportional hazards model weighted by inverse probability of treatment weights (IPTW), rather than an accelerated failure time model estimated using g-estimation. Like SNMs, MSMs were originally developed for use with observational data, so their estimation does not rely on the randomisation assumption.

A standard MSM estimates an average treatment effect across all patients who took the treatment. In the context of treatment crossover that would include both experimental group patients and crossover patients. This is problematic from the perspective of the economic evaluation decision problem, as we are interested in the treatment effect specific to patients initially randomised to the experimental treatment. In addition, as for the SNM method it is problematic to apply the MSM in an RCT context. The MSM relies on being able to model the treatment process, yet when patients are randomised to treatment groups attempting to model the probability of treatment received based upon observed covariates is counter-intuitive. A similar approach to that taken for the SNM could be applied, whereby the MSM would be applied only to patients in the control group after the time-point at which treatment crossover becomes possible. This would result in a treatment effect estimate specific to crossover patients. However, the SNM approach uses accelerated failure time models and produces an acceleration factor which works on the time-scale, allowing observed survival times of crossover patients to be “shrunk” in order to arrive at an estimated counterfactual dataset. Conversely, the MSM produces a hazard ratio, which works on the hazard scale rather than the time scale. Therefore survival times of crossover patients cannot be shrunk in the same way and there is no obvious way to estimate the counterfactual dataset. For these reasons, neither a standard MSM nor a “two-stage” MSM represent a suitable approach for addressing the treatment crossover problem in the context of an RCT, given the economic evaluation decision problem.

However, the inverse probability of censoring weights (IPCW) method is a type of MSM which is directly relevant for the context considered in this thesis, as it attempts to estimate the treatment effect specifically for the experimental group. The method can be used to address any type of informative censoring and represents a method for improving upon the naive censoring approach. Instead of simply censoring patients the covariates of censored patients are taken into account in an attempt to remove selection bias. In the context of treatment crossover, the method involves artificially regarding subjects as dependently censored at the time crossover occurs.

4.10.3.2 Theoretical characteristics

Robins and Finkelstein (2000) used an IPCW method to adjust survival estimates to account for treatment crossover and stated that censoring based upon crossover must be regarded as dependent or informative because crossover is likely to be linked in some way to prognosis.²⁴ If there are data on all time-dependent prognostic factors for mortality that independently predict censoring (crossover), then the dependence between the censoring and failure can be

corrected for by replacing the Kaplan-Meier estimator, log-rank test, and Cox partial likelihood estimator of the hazard ratio by their IPCW versions.²⁴

In the IPCW Kaplan-Meier the contribution of a subject at risk at time t is weighted by the inverse of an estimate of the conditional probability of having remained uncensored until time t . Robins and Finkelstein (2000) used “stabilised” weights, as these are shown to be more efficient.²⁴ Unstabilised weights are simply the inverse of the conditional probability of having remained uncensored until time t conditional on baseline and time-dependent covariates, whereas stabilised weights are the conditional probability of having remained uncensored until time t given baseline covariates, divided by the conditional probability of having remained uncensored until time t given baseline and time-dependent covariates. The stabilised weight will be equal to 1 for all t if the history of the included prognostic factors for failure do not impact upon the hazard of censoring at t – thus there would be no informative censoring and treatment crossover would be random.²⁴

Formally, the stabilised weights applied to each individual for time interval (t), as specified by Hernan *et al* (2001) are:¹³⁵

$$\widehat{W}(t) = \prod_{k=0}^t \frac{\Pr[C(k)=0|\bar{C}(k-1)=0,\bar{A}(k-1),V,T>k]}{\Pr[C(k)=0|\bar{C}(k-1)=0,\bar{A}(k-1),\bar{L}(k),T>k]} \quad [7]$$

Where $C(k)$ is an indicator function demonstrating whether or not informative censoring (crossover) had occurred at the end of interval k , and $\bar{C}(k-1)$ denotes censoring history up to the end of the previous interval ($k-1$). $\bar{A}(k-1)$ denotes an individual’s treatment history up until the end of the previous interval ($k-1$), and V is an array of an individual’s baseline covariates. $\bar{L}(k)$ denotes the history of an individual’s time-dependent covariates measured at or prior to the beginning of interval k , including baseline values. Hence the numerator of [7] represents the probability of an individual remaining uncensored (not crossed over) at the end of interval k given that that individual was uncensored at the end of the previous interval ($k-1$), conditional on baseline characteristics and past treatment history. The denominator represents that same probability conditional on baseline characteristics, time-dependent characteristics and past treatment history. When the cause of informative censoring is treatment crossover, past treatment history is removed from the model because as soon as crossover occurs the individual is censored.

The probabilities required to estimate the stabilised weight can be estimated by fitting logistic regression models with informative censoring (crossover) as the dependent variable, and with baseline covariates included as independent variables for calculating the numerator of the weight, and baseline and time-dependent covariates included as independent variables for the denominator.¹³⁵

Howe *et al* (2011) showed that the survival function corrected for selection bias in the presence of informative censoring using the IPCW method can be estimated using the following equations, adapted from those presented by Robins and Finkelstein (2000):^{24;128}

$$\hat{\lambda}(t_j) = \frac{\sum_{i \in D_j} \widehat{W}_i(t_j)}{\sum_{i \in R_j} \widehat{W}_i(t_j)} \quad [8]$$

$$\hat{S}(t) = \prod_{t_j \leq t} [1 - \hat{\lambda}(t_j)] \quad [9]$$

Where t_j is the time corresponding to the j th visit at which an event was observed to have occurred, R_j is the subset of the patient group for whom the minimum of their observed survival or censoring time is greater than or equal to t_j , and D_j is the subset of R_j who experience the event at t_j . $\widehat{W}_i(t_j)$ is the estimated weight for individual i at t_j , $\hat{\lambda}(t_j)$ is the estimated hazard at t_j , and $\hat{S}(t)$ is the estimated survival at time t . If crossover is random all weights will equal 1, and equations [8] and [9] will reduce to the standard Kaplan-Meier estimator.

The IPCW adjusted Cox hazard ratio (HR) can be estimated by fitting a time-dependent Cox model to the dataset in which crossover patients are artificially censored. The model includes baseline covariates and uses the time-varying stabilised weights for each patient and each time interval. An adjusted HR is obtained, but in their discussions of MSMs Robins (1999) and Hernan *et al* (2001) both state that because weights are used in the estimation process standard error estimates will be incorrect, and thus robust variance estimators – which will provide a conservative confidence interval – must be calculated.^{135;145} Alternatively, bootstrapping could be used.

A slight complication regarding the calculation of an MSM or IPCW adjusted HR was noted by Hernan *et al* (2001).¹³⁵ While using the IPCW approach an ordinary time-dependent Cox model can be weighted to estimate consistent causal treatment effects, few software programs

actually allow weights that vary by time for each individual within a Cox model. To avoid this problem, they explain that a weighted pooled logistic regression model can be fitted, treating each person-month as an observation and allowing for a time-dependent intercept using a spline-based approach.¹³⁵ This could cause problems, as noted by Young *et al* (2010), because the logistic models will only be equivalent to the Cox model when the hazard in any single observational period is small – Young *et al* call this the “rare disease” assumption.¹²⁷ However provided the period between observations is relatively short (as is usually the case in RCTs) this should not be a critical issue.

The IPCW approach shares some of the same advantages associated with the SNM method. It makes use of a large amount of covariate data in order to inform estimates of survival adjusting for informative censoring (crossover) and owing to this the method is potentially more powerful than a standard ITT or randomisation-based analysis. However, as noted for the SNM method, this may not be the case if there is much uncertainty in the data and/or the trial dataset is relatively small.

An additional advantage of the IPCW approach that differentiates it from the SNM approach is that it more closely resembles standard, commonly used survival analysis methods – because it results in an adjusted HR obtained using a Cox model.^{135;138;145} Reflecting this, the IPCW approach has been used relatively often in the literature. Several papers reporting applications of the IPCW method were excluded from the review presented in this chapter, and several included papers discussed either an MSM or IPCW approach.^{24;135;138;139;145}

Similarly, the IPCW method has a number of limitations in common with the SNM approach. In particular, the untestable “no unmeasured confounders” assumption is critical, and therefore data requirements may restrict the practicality of the method. Also, as for the SNM method, models must be correctly specified.²⁴ In addition, the IPCW approach cannot work if there are any covariates which ensure (that is, the probability equals 1) treatment crossover will or will not occur.^{135;138;145} In defence of the problems associated with the “no unmeasured confounders” and model specific assumptions made by MSMs and SNMs, Hernan *et al* (2001) point out that when estimating the effect of a time-independent treatment using standard methods and observational data the same assumptions are made – there must be no unmeasured confounders, noninformative censoring and no model misspecification.¹³⁵ However, as Robins (1999) states, this is why it is dangerous to draw causal inferences from observational datasets.¹⁴⁵

A further potential disadvantage of the IPCW approach compared to the SNM approach is that it ignores potentially useful information from patients who cross over and thus are censored from the analysis. However, from the economic evaluation perspective this may actually be perceived as an advantage, as it means that the IPCW method specifically addresses the decision problem – that is, it specifically estimates the treatment effect in patients randomised to the experimental treatment.

4.10.3.3 Practical applicability

The IPCW method has a strong theoretical basis and has been shown to be unbiased when its assumptions hold.²⁴ It has been used in an economic evaluation context,^{113;114} and is therefore relevant for further analysis in this thesis. However applications of the method have highlighted problems associated with data availability,^{138;150} and simulation studies have shown that generating data that satisfies the modelling assumptions made by MSMs is a complex process, and their assumptions are quite restrictive. For instance, Young *et al* (2010) developed a data generating mechanism that allowed them to satisfy the assumptions made by SNMs and MSMs, but in order to satisfy both models an exponential distribution for counterfactual survival times was required. When instead they used a Weibull model the data generating mechanism no longer satisfied the assumptions made by an MSM, and the bias associated with the MSM increased to approximately 20%.¹²⁷ In the context of survival analysis of cancer patients this is important, because there is no guarantee that survival data will follow the required distribution. Hence it is important to assess the performance of the IPCW method (and other methods) when data are not generated precisely as required by the method's assumptions.

In addition, Howe *et al* (2011) conducted a simulation study to demonstrate that the IPCW method may result in biased estimates of survival when sample sizes are very small ($n=50$), selection bias is very strong (with a sample size of 500), and if there exist unmeasured confounders.¹²⁸ This demonstrated that even if all important covariates are included in an analysis, the IPCW method may still result in bias under certain scenarios. Howe *et al* state that four conditions are required for the method to be unbiased: all common predictors must be appropriately measured and accounted for in the analysis; there must be a sufficient number of participants under follow-up at all relevant times – among those at risk and those under follow-up there must be a nonzero probability of not being censored for every combination of values observed for the common predictor histories at each time point (the actual number of participants required to satisfy this will be case-specific); the common predictors cannot be deterministic or nearly deterministic in relation to both the outcome of

interest and the artificial censoring mechanism among patients over time, and; models must be correctly specified.¹²⁸ Therefore in a real-world cancer RCT context the “no unmeasured confounders” assumption is not the only concern – problems may occur if very high proportions cross over (indicating almost deterministic covariates), particularly if the sample size is relatively small. Howe *et al*'s simulations were relatively simplistic – indicators of censoring were constant over time, the hazard of censoring was constant over time, and the percentage of patients censored was fixed at 60%. In a more complex case, with time-dependent confounding and/or altering risks of censoring over time and/or high crossover proportions, the IPCW method may be even more likely to result in important biases.

When implementing the IPCW method the covariates included in the analysis need to be determined. For SNMs Robins and Greenland (1994) recommend that models are richly parameterised in order to increase the likelihood that “no unmeasured confounders” assumption will hold, and this is likely to hold true for the IPCW approach.¹²⁹ Robins and Finkelstein (2000) demonstrated a method for retaining only covariates that are significant prognostic factors for mortality, using a time-dependent Cox proportional hazards model for failure. They decided to keep covariates that were significant at the $p=0.12$ level (though offered no particular rationale for choosing this significance level), but also constructed models to test the relationship between included and excluded variables, to identify whether any of the excluded variables were predictors of future values of the retained variables which would therefore be relevant for inclusion in the IPCW model. This represents one method that an analyst may use to determine which covariates should be included in an IPCW analysis, but Robins and Finkelstein note that other approaches might also be considered.²⁴ In reality variables for which data are collected in RCTs may often be chosen because they are potentially prognostic or informative in some way, and thus it is likely that most of these would be relevant for inclusion in an IPCW model.

4.10.4 Other excluded methods

The review identified several other methods that, based upon their titles and abstracts, appeared to be potentially useful methods for addressing the treatment crossover problem. These methods were included in the review, but were deemed not relevant for further analyses within this thesis. Brief details of these are given below. Further information on the methods are available in Appendix 4.

4.10.4.1 Auxiliary Variables

The Auxiliary Variables (AV) approach for dealing with treatment crossover is closely related to the IPCW approach. The method treats censored data as missing and associates other information collected about patients with event times using auxiliary variables. These are used to help recover some of the information lost due to censoring.^{141;143;146}

Robins and Finkelstein (2000) note that a key problem with an AV approach is that when there are a range of auxiliary variables, conditional modelling is required for the estimated event time and the process by which this is affected by the auxiliary variables.²⁴ If the models used to capture these relationships are misspecified the resulting treatment effect estimates can be biased and, for example, inconsistent estimates of the survival curve can be produced even when censoring is independent and a standard Kaplan-Meier would have been consistent. Robins and Finkelstein (2000) state that this problem can be avoided by creating a pseudo population using the IPCW approach rather than an AV approach and that the IPCW approach requires fewer modelling assumptions.²⁴ Thus, it appears that the IPCW approach supersedes AV approaches and thus these methods are not considered any further in this thesis.

4.10.4.2 Intensity Score

Extending from the MSM methodological branch is the Intensity Score approach, developed by Tanaka *et al* (2008).¹³² The approach is quite different in that it uses a parametric AFT model (more similar to a SNM-type approach), but it involves weighting the AFT by the intensity score, which is derived from the propensity score, which is similar to the IPCW. The intensity score reflects the cumulative differences over time between treatment actually received and treatment predicted by prior observed medical history – thus the method relies upon the “no unmeasured confounders” assumption. There appear to be no instances of Tanaka *et al*'s method being applied by other authors and given that it is similar to other methods discussed here, it is not considered further in this thesis.

4.10.4.3 Adjusted Hazard Ratio Methods

Law and Kaldor (1996) developed an approach for estimating an adjusted hazard ratio specifically within the context of treatment switching.¹³⁰ Their method works by splitting patients into four groups depending upon their initially randomised group and whether or not they switch treatments. Hazards are assumed to be proportional, and a Cox model is fitted with a time-varying covariate for switching time. White noted a key limitation of this approach – patients are grouped based on future events, that is, before switching occurs.¹⁴⁸ Once patients are split into groups they are assumed to have a certain hazard function before they switch, however crossover patients cannot die before they switch treatment as otherwise they

would not be in their allocated group – hence in reality they have a hazard of zero up until the point at which they switch treatment. White demonstrates that this leads to bias,¹⁵⁰ and in the simulation study conducted by Morden *et al* (2011)¹ the method produced substantial amounts of bias compared to other methods – therefore I do not consider this method further.

4.10.4.4 Adjusted Treatment Received Method

Nagelkerke *et al* (2000) developed a method for adjusting estimates of the treatment effect to take into account non-compliance that is quite separate from the other methods identified by the review.¹³⁶ The authors consider a situation in which the relationship between compliance and thus treatment received and outcome is influenced by unobserved confounders. They show that these confounders are usually unknown and standard methods to correct for bias cannot be used. However, they suggest that a variable (E) can be used in lieu of the real confounder(s) (C) to adjust for confounding in a multivariate analysis in order to estimate the true causal effect. The authors only extended their approach to a survival setting as a secondary analysis and did not deal specifically with the issue of treatment crossover. The authors conducted a simulation study and found that in a survival setting their method performed poorly,¹³⁶ and therefore I do not consider this method any further.

4.10.4.5 Threshold Regression Mixture Models

A further method identified by our search was the Threshold Regression Mixture Model approach developed by Lee *et al* (2008).¹⁴² The approach involves the use of a first hitting time model to model the disease progression process, and four treatment regimens are specified such that the treatment effect can be estimated in each one. Thus the treatment effect is allowed to differ depending upon when and if treatment switching occurred. However, the method is essentially an alternative method for modelling survival data in general and treatment regimens specified by the authors were not those that would primarily be of interest when considering treatment crossover. The method would require substantial amendment if it were to be taken forward here. Given that this method is very new, and has so far only been cited by the original authors in two follow-up papers, it is not considered further. Ideally it would be accepted for standard survival analysis before it is considered for dealing with treatment crossover.

4.11 Summary

The narrative synthesis has described in detail the origins, theoretical suitability and practical applicability of three identified methodological categories relevant for this thesis – SNM;

RPSFTM, and; MSM/IPCW. It has demonstrated that these methods may further be categorised based upon their key assumptions and theoretical basis – in particular, the key distinction between the methods is whether they are observational-based or randomisation-based. The observational-based methods (SNM and MSM) require the “no unmeasured confounders” assumption and are prone to bias if this does not hold. The randomisation-based methods are less reliant on covariate data but require randomisation to have worked adequately, and the practical workable versions of these methods rely upon the “common treatment effect” assumption. Therefore, the three methodological categories can be further classified as:

- Observational-based methods
 - Structural nested models with g-estimation
 - Marginal structural models (IPCW)
- Randomisation-based approaches
 - RPSFTM (and the IPE algorithm)

These methods are taken forward for further analyses in subsequent chapters of this thesis.

4.12 A novel two-stage method

When undertaking my review an alternative method for adjusting for treatment crossover occurred to me that was not identified by my search, and which has not been used in previous HTAs. The method is linked to the two-stage process required in order to apply a SNM with g-estimation to an RCT dataset. The SNM attempts to adjust for time-dependent confounding but is heavily data reliant and requires g-estimation. A simple alternative would be to take a similar two-stage approach (that is, using a disease event such as disease progression as a secondary baseline and treating data for patients in the control group as an observational dataset after this point) and to fit a simple parametric accelerated failure time model to this dataset in order to estimate the treatment effect associated with the experimental treatment in crossover patients. Fitting such a model including covariates measured at the secondary baseline (including a covariate indicating treatment switch) would be expected to produce a reasonable estimate of the treatment effect received by patients who crossed over, provided the model fits the data, there are “no unmeasured confounders” at the point of the secondary baseline and provided crossover occurs soon after the secondary baseline. The resulting acceleration factor associated with switching could then be used to “shrink” survival times in switching patients to derive a counterfactual survival dataset upon which standard survival analysis could be undertaken.

Although the assumptions required in order for such an approach to be taken are restrictive, they fit the treatment crossover mechanism often observed in oncology trials well; for instance, given that new treatments are often licensed based upon data on PFS, crossover is usually only permitted after the point of disease progression. If crossover happened a long time after this secondary baseline time-dependent confounding could become important, but in the context of a clinical trial, where crossover is allowed after disease progression, it seems likely that if crossover is to occur it will happen soon after the point of disease progression. The approach relies upon the “no unmeasured confounders” assumption, but only at the point of disease progression, and it does not require the “common treatment effect” assumption. It is limited in that it can only be applied in certain circumstances (such that there is an appropriate secondary baseline), but in practice this may often be the case. The method is further limited by the fact that unless all switching occurs immediately at the secondary baseline time-point it will be prone to time-dependent confounding – however this may often be the case (or may be close to being the case) and so the resultant bias may not be great. Because this two-stage method involves the estimation of counterfactual survival times recensoring is required.

4.13 Discussion and conclusions

The primary aims associated with this chapter were to identify statistical methods for adjusting survival estimates in the presence of treatment crossover, to review these and to determine which are appropriate for further analyses and evaluation in this thesis. These aims have been met. A broad, iterative search procedure was undertaken in order to identify novel methods. It is clear that research in this area has focussed upon two methodological strands – structural nested models and marginal structural models. Most of the methods identified fit into one of these broad categories. Within these strands there are important distinctions between methods, which are particularly important given the context of this thesis – given that the problem being investigated is treatment crossover in RCTs rather than in observational datasets.

The origins of both SNMs and MSMs lie in observational settings, and therefore these rely on covariate data and the “no unmeasured confounders” assumption. In rich datasets with a large number of covariates measured over time this may be reasonable, but typically RCT datasets are smaller than observational datasets and less information is available. Therefore, methods such as the RPSFTM and IPE algorithm that use the power of randomisation in their

adjustment mechanism – termed here as “randomisation-based approaches” – offer potential advantages. The RPSFTM and IPE methods are effectively randomisation-based versions of the SNM approach. No such version of the MSM approach exists. The IPCW approach does compare groups as randomised, but relies upon observational data to adjust survival estimates and so it is not referred to as a randomisation-based approach here.

All of the identified approaches have important limitations. Whilst the data requirements of the observational-based methods are very important in the context of an RCT, a key advantage is that assumptions regarding the commonality of the treatment effect are not required. This represents a key limitation associated with the randomisation-based methods – and this problem has not been overcome in the literature despite the attempted use of multiparameter models. Given these limitations it is clear that none of the methods identified will successfully adjust for crossover in an unbiased way across all realistic scenarios. However, given that the treatment crossover problem must be addressed, it remains important to determine which methods are likely to result in least bias in different scenarios. Also given the limitations associated with the methods identified by my review, it is important that I have identified an additional two-stage method to adjusting for treatment crossover, described in Section 4.12. This is designed based upon the particular crossover mechanism often observed in oncology RCTs, and this method is taken forward for further analyses in subsequent chapters of this thesis.

It should be noted that the weighted version of the RPSFTM used in NICE technology appraisal 215 (pazopanib for the first line treatment of metastatic RCC),¹⁰⁷ as discussed in Chapter 3, has not been considered in this chapter. This is because the method has not been described in a peer reviewed published paper and so was not identified by the review. Given this, it is not possible to adequately review this method, and the method cannot be taken forward for further analysis in this thesis. The limited details presented in the NICE appraisal documents seem to suggest that the treatment effect estimated is likely to be similar to that resulting from the unweighted RPSFTM, hence not including this extension should not devalue the findings of this thesis.

Only 1 of the 25 NICE TAs reviewed in Chapter 3, in which treatment crossover was an issue, used one of the potentially appropriate methods identified by the review presented in this chapter (the RPSFTM method was used in TA179⁶³). However, both of the more recent TAs included in Section 3.8 of Chapter 3 used potentially relevant methods (RPSFTM and IPCW methods were used in TA215 and TA219^{53;54}). In general though, this demonstrates the lack of

use of potentially appropriate methods for adjusting for treatment crossover in HTAs, and that potentially appropriate methods and variations of methods exist that have not been used at all.

SNMs with g-estimation, IPCW, RPSFTM and the IPE algorithm will be taken forward for further assessment and evaluation in Part 4 (Chapters 6 and 7) of this thesis. First though, Chapter 5 considers the implications that these methods have for extrapolation and incorporation into economic models. This helps towards achieving the primary objective of this thesis – that is, to determine how best to address the treatment crossover problem in the context of economic evaluation.

Chapter 5

Survival analysis and economic evaluation – implications for treatment crossover methods

5.1 Chapter overview

In Chapter 4 a systematic search and review of the literature was undertaken in order to identify statistical methods that are potentially suitable for addressing the treatment crossover problem within an economic evaluation context. This involved assessing the theoretical characteristics of identified methods in order to determine their capacity to produce adjusted estimates of the treatment effect that satisfy the decision problem faced by resource allocators – the treatment effect must represent a comparison of a world in which the novel treatment does not exist to a world in which the novel treatment does exist. However, as stated in Chapter 1, trial-based estimates of the survival effects of a new treatment are usually insufficient for economic evaluation – there is a requirement to assess the effect of a new treatment on the entire disease population over the duration of a lifetime. Due to the censoring typically observed in clinical trials, extrapolation is usually required. This is the case whether or not treatment crossover is present – in the presence of crossover the treatment effect must be adjusted *and* extrapolated. The present chapter complements Chapter 4 by specifically considering how crossover adjustment methods may be combined with extrapolation methods for use in an economic model. Combined, these chapters complete Part 3 of this thesis, identifying potential solutions to the crossover problem in an economic evaluation context.

Various parametric modelling options are available to perform extrapolation, and much of the research focus on survival modelling surrounds the choice of parametric model and the fitting method used.^{11;16;17;19;25;152-154} Different crossover adjustment methods result in different types of survival-related outputs, which impact upon how extrapolation can be undertaken. For example, the RPSFTM provides the analyst with an adjusted acceleration factor and a counterfactual dataset that has been recensored, while the IPCW provides an adjusted hazard ratio and a weighted Kaplan-Meier (WKM) curve. It is therefore important to consider the amenability of different crossover adjustment methods to extrapolation – a method is required that is successful in adjusting for treatment crossover, *and* that is amenable to suitable extrapolation for use in an economic model. Previous research that has considered

crossover adjustment methods in an economic evaluation context has not addressed the issue of extrapolation.²¹ Therefore, this chapter provides a novel and important addition to the literature.

Sections 5.2, 5.3, 5.4 and 5.5 of this chapter present an overview of how survival modelling is generally conducted in the context of economic evaluation by drawing upon evidence from NICE technology appraisals (TAs) conducted in the cancer disease area. Lengthier versions of these sections formed the basis of the recently published NICE DSU Technical Support Document (TSD) on “Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data”¹⁵ (an abridged version of this TSD is currently at the “revise and resubmit” stage with the Medical Decision Making journal). In Section 5.6 it is considered whether “best practice” extrapolation can be undertaken in combination with the application of treatment crossover adjustment methods. Section 5.7 provides discussion and Section 5.8 makes conclusions and considers implications for the remainder of the thesis.

Extrapolation of survival data is a very broad topic, with a number of research groups currently working in the area. The focus of this thesis is treatment crossover and therefore I do not review alternative extrapolation methods in exhaustive detail. In keeping with this, detailed formulations and theoretical characteristics of different parametric models are not discussed. The hazard and survivor functions and brief details of a range of parametric models are presented in Appendix 6. More details on these are available from several standard statistics text-books, such as Collett (2003).¹²

5.2 Introduction

In Chapter 3 a search and review of NICE TAs was described, with an emphasis on methods used to adjust for treatment crossover. In Section 3.3 it was noted that this search was also used to examine survival modelling techniques used in the economic evaluation sections of the TAs, independent of whether treatment crossover was an issue. The purpose of this was to provide evidence on commonly used extrapolation techniques in the context of economic analyses. This section of the review focussed on the following questions:

- i. Were mean survival times or medians used in the economic analysis?
- ii. How was mean survival estimated (for example, through a restricted mean approach or through extrapolation)?
- iii. How was extrapolation performed?
- iv. How were methods used for the survival analysis rationalised or justified?

The full results of this review were used to develop the NICE DSU TSD on survival modelling. However, given the focus of this thesis questions (ii) and (iii) are concentrated upon in this chapter. In this chapter commonly used methods for estimating the entire survival distribution (which may be summarised using a “mean” survival measure) are identified and it is then considered how these can be combined with treatment crossover analyses.

5.3 Search strategy

The search strategy and identified NICE TAs are described in Chapter 3; 45 TAs were identified and included in the review – these are listed in Table 3.1.

5.4 Modelling methods

Table 5.1 presents the approaches used in the NICE TAs to estimate mean survival (or the entire survival distribution). Five broad approaches were identified: 1) restricted means analysis; 2) parametric modelling; 3) Proportional Hazards (PH) modelling; 4) external data modelling; and 5) other “hybrid” methods. In 17 (38%) TAs extrapolation was not performed, with the survival analysis based purely on the observed trial data (restricted means analysis). Appropriately, this was generally only the case when there was relatively little censoring in the survival data from the trial. It is important to note that survival data may appear to be relatively complete even in the presence of substantial censoring – if the last observed data-point represents an event rather than a censored observation the Kaplan-Meier curve will fall to zero. A restricted means approach may be taken in these circumstances, but where numbers at risk at the tail of the survival curve are very low this will be subject to substantial uncertainty and extrapolating the data may in some circumstances be more appropriate.

Table 5.1: Methods used to estimate mean survival in NICE technology appraisals

Method for Estimating Mean	Number of TAs (%)
Restricted Means	17 (38%)
Parametric Models	32(71%)
Weibull	23 (51%)
Exponential	20 (44%)
Gompertz	6 (13%)
Log-logistic	9 (20%)
Log normal	6 (13%)
Gamma	2 (4%)
Piecewise modelling	1 (2%)

Method for Estimating Mean	Number of TAs (%)
Proportional Hazards modelling	19 (42%)
External data	4 (9%)
Other “hybrid” methods	2 (4%)
Kaplan-Meier - Exponential method	1 (2%)
Gelber method	1 (2%)

5.4.1 Parametric modelling

Thirty-two of the 45 TAs reviewed (71%) used parametric extrapolation techniques in order to produce estimates of survival. The most popular parametric models were the Weibull and exponential; the Weibull was used in 23 TAs (72% of those that involved extrapolation), and the exponential in 20 (63% of those that involved extrapolation). Other models (such as the Gompertz, log normal, log-logistic and generalised gamma) were used considerably less often (see Table 5.1).

The methods used to fit the parametric models varied. Usually the manufacturer had access to patient-level data and thus could fit parametric models using these, whereas the independent academic groups that review the manufacturers’ analyses typically had to use a computer digitisation program in order to reproduce published Kaplan-Meier curves so that parametric models could be fitted. This data availability issue has been important in the past but in the future it is likely to be less so, due to novel research that allows patient-level data to be accurately replicated from a published Kaplan-Meier curve.¹⁵⁵ It was most common for all survival data-points to be used when fitting parametric models to trial data. However, in some TAs, models were fitted using a restricted dataset. For example, in TA86 (imatinib for gastro-intestinal stromal tumours (GIST)) and TA121 (carmustine implants and temozolomide for glioma) the manufacturer and the independent academic group fitted exponential parametric models using trial data only up to certain specified time-points.^{56;72} The final observed survival events were excluded from the model-fitting process because these were argued to be highly uncertain due to high levels of censoring. The robustness of this is questionable because excluding data-points means that the level of uncertainty is increased further. Data should only be excluded if it can clearly be shown that implausible outliers exist, based upon external data or clinical expert opinion.

A variation on the approach of restricting the data to a certain time-point when fitting parametric models was used in TA169 (sunitinib for renal cell carcinoma) and TA179 (sunitinib for GIST).^{63;156} In both cases, the independent academic group stated that they approved of the manufacturer’s analysis whereby a Weibull model was fitted to the survival data using only

one data-point per month. This was perceived to allow the fitted models to follow the Kaplan-Meier curve more closely from a visual perspective. However this implicitly places disproportionate weight to those segments of the Kaplan-Meier where there are fewer data-points, and does not place sufficient weight on areas where a large number of data-points were observed. Furthermore, this requires single data-points to be chosen for inclusion in the analysis; the choice of the included points is likely to be arbitrary, whilst excluding other data-points leads to greater uncertainty. This is therefore a potentially biased technique, and is directly at odds with the method of excluding data from the right-hand-side of the Kaplan-Meier curve from the analysis – the latter places no weight on the events observed at the right-hand-side of the Kaplan-Meier, whereas the former implicitly places a high weight on these events. Both methods are likely to bias the resulting survival estimates and should be avoided.

5.4.2 Proportional Hazards modelling

Some use of PH modelling was evident in 19 (59%) of the 32 TAs that involved extrapolation of survival data. This relies upon the proportional hazards assumption (whereby the HR remains constant over time). An alternative involves independently fitting models to each treatment group, avoiding proportionality assumptions. In some instances, PH modelling was tested as a structural uncertainty sensitivity analysis (with independent model fitting forming the base case), whilst in other TAs it was the only method for extrapolation used.

PH modelling was most often used when multiple comparators were included in the evaluation, and where patient-level data were not available for all comparators (for example, TA70 (imatinib for leukaemia), TA91 (paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan for ovarian cancer) and TA93 (irinotecan, oxaliplatin and raltitrexed for colorectal cancer)).^{64;157;158} However, there was also some use of PH modelling when single comparators were included in the economic model, and where patient-level data from the pivotal RCT were available (for example, TA137 (rituximab for lymphoma) and TA174 (rituximab for leukaemia)).^{159;160}

Of the TAs that used PH modelling, few made explicit assumptions about the duration of treatment effect. This implies that the HR observed in the trial lasts for the entire duration of the economic model – typically a lifetime. This may be clinically implausible. Alternatively one could assume that the treatment effect is maintained until the end of the trial follow-up reflecting the available evidence (this approach was taken in TA65¹⁶¹), or could select a cut-off point related to disease progression or time (for example, TA70¹⁵⁷). Both of these options are

arbitrary in nature and require explicit justification. Within these TAs, such justification was not evident.

5.4.3 External data

Within 4 TAs, external registry data were used to inform the extrapolation of survival estimates, due to a lack of long-term survival data within the relevant trials. This typically involved applying death rates from external data to the post-trial time period. Usually it was assumed that the risk of death in this period was the same whether the patient was initially randomised to the intervention or the control treatment, or PH modelling was used. The former approach was taken in TA110 (rituximab for follicular lymphoma).¹⁶² This implies an overall survival (OS) benefit correlated with the PFS gain. Within the rituximab appraisal the validity of this surrogate relationship had not been proven, hence the independent academic group conducted sensitivity analysis assuming that none of the PFS gain translated to OS.¹⁶²

In TA129 (bortezomib for multiple myeloma) PH modelling was used, whereby external data were used to estimate long-term survival for the control group and a HR was applied to estimate survival in the experimental group, assuming that the treatment effect declined over time but was maintained for 3 years.¹⁶³ The independent academic group highlighted this as a key weakness of the manufacturer's analysis because the assumptions around the duration and decline of the treatment effect were not rationalised.¹⁶⁴

5.4.4 Other "hybrid" methods

Most TAs fitted simple parametric models to estimate mean survival. However some novel approaches were used, most notably by the independent academic group who reviewed the manufacturer's analyses in TA181 (Pemetrexed for lung cancer), which bears similarities to a method previously suggested by Gelber *et al.*¹⁶⁵ The independent academic group stated that the extrapolation techniques used by the manufacturer (exponential and Weibull models) provided poorly fitting survival curves. To investigate this further they examined the cumulative hazard function and noted that standard parametric models were not compatible with the trial data, due to hazard rates and ratios that changed over time.¹⁶⁶ However, they also observed that for each group at some point following the end of treatment the cumulative hazard function assumed a steady linear increase that was indicative of a constant risk of death per unit of time. Consequently, the independent academic group elected to model survival using the observed Kaplan-Meier curve itself for short-term survival, and supplemented this with an exponential distribution for long-term survival. Although this is unlikely to be suitable in all circumstances (often the long-term survival prognosis will not be

accurately represented by an exponential distribution), it demonstrates the value of closely examining the hazard rates observed in the trial.

5.5 Model selection process algorithm

Based upon the review of NICE TAs and previous research, in the NICE DSU TSD I presented a survival model selection process algorithm (see Figure 5.1).¹⁵ Whilst other researchers have previously made recommendations on the use of specific survival modelling techniques,^{11;16;17;19} such an algorithm has not before been published. Because research on extrapolation methods (particularly methods more complex than the “standard” parametric models) is currently ongoing, the algorithm is likely to evolve over time as novel research is completed. However, in the interim it provides an important means to improve the quality and consistency of survival modelling incorporated within economic evaluations.

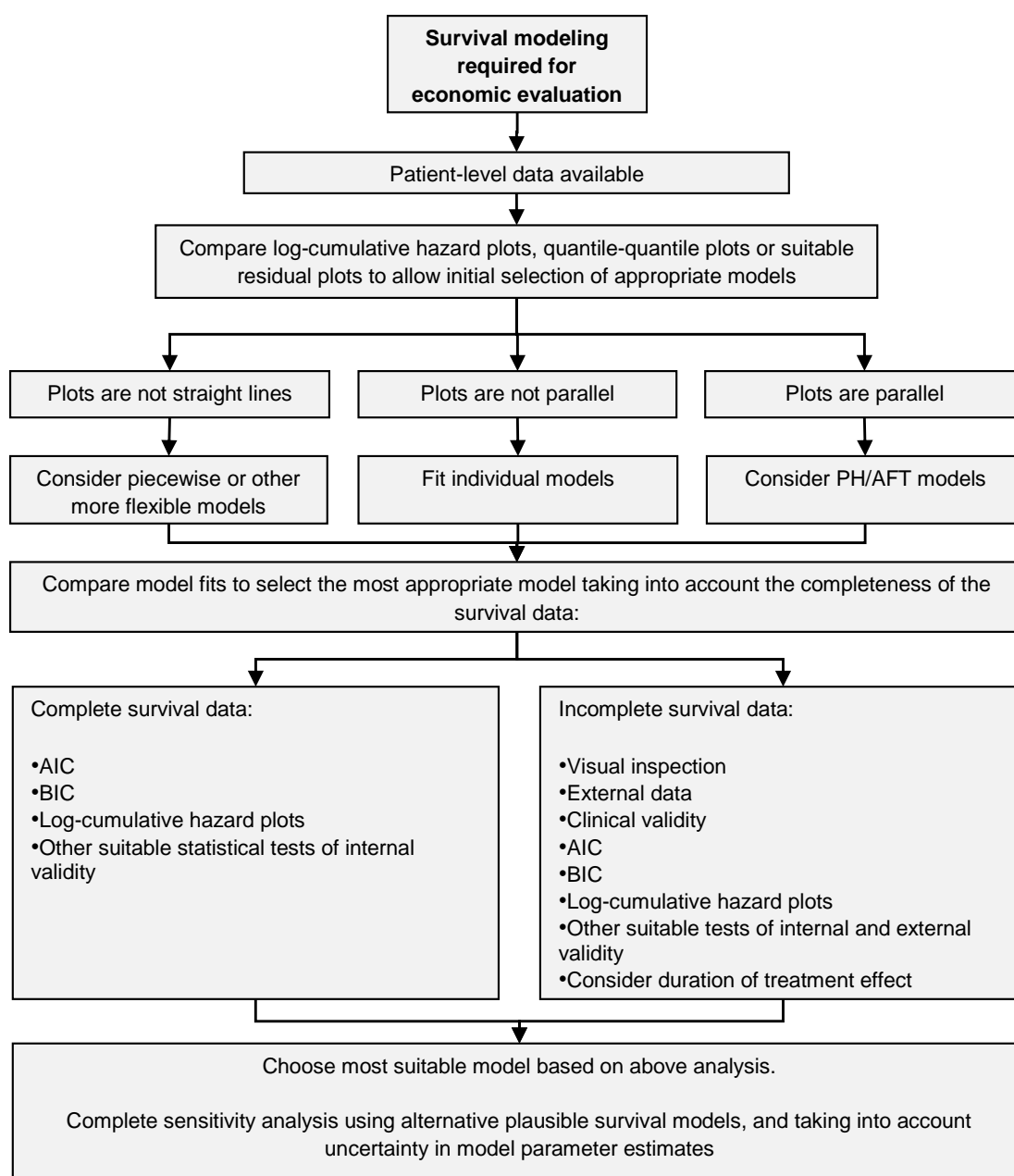
Three key underlying principles influence the algorithm:

- The analyst should demonstrate that all “standard” parametric models (exponential, Weibull, Gompertz, log-logistic, log normal) have been considered and compared, in order to make evident that the model choice has not been arbitrary. If standard models appear unsuitable, the use of more flexible parametric models (such as the generalised gamma and generalised F models), piecewise modelling and other novel survival modelling methods such as those demonstrated by Royston and Parmar and Jackson *et al* should be considered.^{19;167}
- The fit of alternative models should be assessed systematically, including consideration of the fit to the observed data (internal validity) and the plausibility of the extrapolated portion of the curves (external validity).
- In general, complete trial data should be used when fitting models. Data points should not be excluded unless it can be clearly demonstrated that certain points are erroneous outliers. This does not preclude sensitivity analysis investigating the influence of highly uncertain longer term data, as suggested by Connock *et al*.¹⁷

The model selection process algorithm presents a step-by-step process through which appropriate survival models can be identified. First, models that are appropriate for modelling the observed data are identified based upon an analysis of the hazards. This includes a consideration of the proportional hazards assumption. Appropriate models are then compared based upon their statistical fit to the observed data and models that provide a plausible extrapolation are identified by drawing upon external data and clinical expert

opinion. The duration and the trend of the treatment effect should be explicitly considered at this stage. In the model selection decision the weight given to the comparative fit of alternative models to the observed data depends upon the extent to which extrapolation is required. If very little extrapolation is required the fit to the observed data is of most importance. On the other hand, if substantial extrapolation is required the plausibility of the extrapolated portion of alternative models is of greater importance than the fit to the observed data. Finally, alternative plausible models (based upon internal and external plausibility) should be considered in sensitivity analysis.

Figure 5.1: Survival model selection process algorithm



5.6 Extrapolation and treatment crossover adjustment methods

In this section, bearing in mind the algorithm suggested within the NICE DSU TSD, I assess whether it is possible to take a similar approach to survival model selection when treatment crossover adjustment methods have been applied. The implications for extrapolation are considered for each of the treatment crossover adjustment methods identified as commonly used in Chapter 3 (that is, generally naive methods), or as important for further assessment in Chapter 4.

5.6.1 Naive Approaches

5.6.1.1 Intention to treat

An ITT analysis does not attempt to adjust for treatment crossover, but was often the only analysis undertaken in NICE TAs in which treatment crossover occurred. Hence it is included in this and subsequent chapters as a potential analysis that could be chosen even in the presence of crossover. The survival model selection process algorithm suggested by the NICE DSU TSD would generally be expected to be followed using an ITT analysis in the absence of treatment crossover. This approach is therefore amenable to the algorithm, but is subject to bias in the presence of treatment crossover.

5.6.1.2 Per protocol – excluding and censoring switchers

Excluding or censoring crossover patients represent the most commonly used methods for addressing the treatment crossover problem in NICE TAs, as shown in Chapter 3. Neither of these approaches affect the ability of an analyst to extrapolate survival data using the survival model selection process algorithm. However, if treatment crossover is associated with prognosis the associated extrapolations would be subject to selection bias – even if they fit the adjusted dataset well.

5.6.1.3 Treatment as a time-dependent covariate

This naive approach to adjusting for treatment crossover is essentially a simplified version of a marginal structural model (MSM) as described in Chapter 4, in which the average treatment effect in all patients who ever took the treatment is estimated. However patient characteristics at the time of treatment initiation are not controlled for appropriately. This method was not observed in any of the reviewed NICE TAs, but was included as an option in the simulation study undertaken by Morden *et al* (2011)²¹ – hence it is included here. Like the per-protocol approaches, this method is prone to selection bias and is particularly open to time-dependent confounding.

This approach is more problematic than other naive approaches with regard to following the survival model selection process algorithm. Adjusting for crossover in this way would require a proportional treatment effect modelling approach to extrapolation (proportional hazards for a proportional hazards model, or a constant acceleration factor for an accelerated failure time model), because a treatment received indicator is included in the survival model. This removes some of the flexibility associated with the survival model selection process algorithm, because models could not simply be fitted independently to treatment groups. Because this crossover adjustment approach is subject to bias an inappropriate extrapolation is likely, and this is compounded if (based upon an analysis of observed hazards and failure times) the proportional hazards and/or constant acceleration factor assumptions appear not to hold.

5.6.1.4 Crossover as a time-dependent indicator variable

This method differs very slightly from the “treatment as a time-dependent covariate” method, because in this method survival effects are estimated separately for crossover patients and for experimental group patients due to the crossover indicator being included in the model. This method was not observed in any of the reviewed NICE TAs and was not considered by Morden *et al* (2011). However it is similar to the “treatment as a time-dependent covariate” method with the potential advantage that it can produce different treatment effect estimates for patients randomised to the experimental group and crossover patients – which may make it appealing to analysts. Hence it is included here and in subsequent chapters. The method is expected to be prone to selection bias and time-dependent confounding.

Like the “treatment as a time-dependent covariate” method, the treatment group is included in the survival model and therefore a proportional treatment effect modelling approach must be taken for extrapolation purposes – hence this method restricts the flexibility associated with the survival model selection process algorithm.

5.6.2 More complex approaches

5.6.2.1 Structural Nested Models

The observational SNM method described in Chapter 4 uses g-estimation to provide an estimate of the treatment effect in the form of an acceleration factor. The method relies upon successfully modelling the treatment process (that is, the probability of crossover) and is typically applied to observational datasets. As discussed in Chapter 4, applying the method to an RCT is more complex. The randomised group and the disease progression indicator are pivotal indicators of the treatment process – patients in the experimental group will receive

the experimental treatment regardless of their covariates at baseline and over time, and patients in the control group cannot receive the experimental treatment until disease progression has occurred regardless of their covariates over time (assuming treatment crossover is only permitted after disease progression). The SNM approach is not designed to work in these circumstances and so the approach that is taken forward for further analysis in Chapter 6 is to treat the period after disease progression in the control group as an observational study (similar to the approaches taken by Robins and Greenland¹²⁹ and Yamaguchi and Ohashi^{138;150}). The SNM method can then be applied to this period in order to estimate the treatment effect specific to crossover patients.

Because the SNM takes the form of an accelerated failure time model, which works on the time scale, the estimated effect (an acceleration factor (AF)) can be used to shrink survival times in crossover patients, in order to arrive at a counterfactual dataset which is adjusted for treatment crossover. This “shrinkage” approach is achieved by multiplying the difference between overall survival and crossover time by the inverse of the AF:

$$\text{counterfactual survival time} = \text{crossover time} + \left(\frac{1}{\text{AF}} * (\text{observed survival} - \text{crossover time}) \right) \quad [10]$$

Once the counterfactual dataset is obtained parametric models could be fitted to it according to the NICE DSU TSD survival model selection process algorithm. A proportional treatment effect modelling approach could also be taken, by fitting a proportional hazards or accelerated failure time model to the counterfactual dataset. However, there would be implications for uncertainty, because the counterfactual dataset represents an amended dataset, rather than what was actually observed in the trial. The coverage of treatment crossover adjustment methods (that is, the proportion of times that the confidence interval estimated by the method contains the true value) will be assessed and discussed further in subsequent chapters of this thesis.

5.6.2.2 Inverse Probability of Censoring Weights

The IPCW approach provides an estimate of the HR, adjusted for treatment crossover, and a weighted Kaplan-Meier (WKM) curve. The provision of an adjusted HR is amenable to a PH modelling extrapolation approach, but fitting parametric models independently to treatment arms would be more problematic. Whereas standard Kaplan-Meier curves are based on the underlying dataset, the WKM does not represent the underlying data, rather it represents a weighted version of this – the contributions that each patient makes to the WKM at each time

point are weighted according to their baseline and time-dependent covariate values. Therefore the IPCW method does not provide an adjusted dataset to which parametric survival models could be fitted. A potential solution to this problem could be to fit parametric survival models to trial data “created” to represent the WKM survivor function. There are numerous ways in which such a dataset may be created, including novel methods recently published by Guyot *et al* (2012).¹⁵⁵

Even a PH modelling approach to extrapolation is less straightforward when combined with an IPCW analysis than is usually the case. Usually this approach would involve fitting a parametric survival model to the survival dataset, with treatment group included as a covariate, thus providing survival curves for the control and intervention groups. An alternative (but equivalent) approach would be to fit a parametric survival model to the survival dataset with treatment group included as a covariate, and to apply the resulting hazard ratio to the control group hazard function in order to obtain the experimental group hazard function and survival curve. However, when the IPCW method is used there are problems with these approaches. The method incorporates a semi-parametric Cox model, rather than a parametric model and therefore does not provide survival extrapolations for the experimental group or the control group. Hence, a PH approach to extrapolation is not straightforward – an alternative method is required if such an approach is to be taken.

Since – in the treatment crossover context considered in this thesis – survival in the experimental group is not confounded and does not require adjustment, an independent parametric model could be fitted to that group to represent the experimental group survival curve. The survivor function and hazard function associated with that model could then be calculated, and the hazard function could then be multiplied by the inverse of the IPCW HR to obtain the control group hazard function. From this, the control group survivor function and survival curve could be derived. This “survivor function” approach is not without limitations – notably the IPCW HR is associated with the weighted Cox model, rather than the parametric model fitted to the experimental group, and therefore it is theoretically incorrect to apply the IPCW HR to the independently fitted parametric model. However this approach does represent a way in which a PH modelling approach could be taken for extrapolation purposes in combination with an IPCW analysis. The bias associated with applying the IPCW HR to the independently fitted parametric model may be relatively small (certainly in relation to the bias caused by treatment crossover), as in most cases the Cox model and alternative parametric models can be expected to result in similar estimates of the HR. Substantial bias would only be expected if the parametric model provided a particularly poor fit to the observed data.

5.6.2.3 RPSFTM

The RPSFTM method provides an estimate of the treatment effect adjusted for treatment crossover in the form of an acceleration factor (AF). It also provides counterfactual survival times – that is, survival times that would have been observed if nobody had received treatment. Given this, theoretically, any of the extrapolation approaches specified by the survival model selection process algorithm could be undertaken. However, there are likely to be issues with several of these.

A proportional hazards modelling extrapolation approach could be applied by fitting proportional hazards models to the counterfactual dataset. However, the counterfactual dataset is subject to recensoring, which results in a loss of information and potentially less robust survival extrapolations because longer-term data-points are recensored at an earlier time (with the size of the impact of recensoring dependent upon the size of the estimated treatment effect). An alternative approach would be to take a “survivor function” approach similar to that described for the IPCW method above. A parametric model could be fitted independently to the uncensored experimental group data (this group does not require recensoring because no crossover occurred in it). Because the RPSFTM provides an AF, the control group survivor function could then be derived by dividing the time associated with each experimental group survivor function probability by the RPSFTM AF. A similar limitation exists for this approach as for the IPCW survivor function approach – that is, the RPSFTM AF is applied to a parametric model to which it is not related.

An independently fitted modelling approach could be taken in combination with the RPSFTM method by simply fitting independent parametric models to separate treatment groups based upon the counterfactual dataset. However, this too would be subject to the limitations associated with recensoring and loss of information. While recensoring may ensure an unbiased treatment effect estimator (provided the other model assumptions hold), it may cause problems for extrapolation because extrapolations are based upon data with a shorter follow-up time.

An alternative approach to extrapolation in combination with the RPSFTM method is to “shrink” the survival times of crossover patients in a similar way as described by equation [10] for the SNM approach. The inverse of the RPSFTM AF could be used to shrink survival times in crossover patients in an approach that may not incorporate full recensoring. While the AF would be estimated under full recensoring, the counterfactual dataset would only involve

recensoring of crossover patients, not all patients in the control group (as is the case in the fully recensored RPSFTM counterfactual dataset). Whilst this removes some of the information loss associated with recensoring, it creates further potential for bias due to the possible association between counterfactual censoring times and prognostic factors (as discussed in Section 4.10.2 of Chapter 4).¹²³

5.6.2.4 IPE algorithm

While the IPE algorithm method uses the same counterfactual survival model as the RPSFTM, its parametric estimation procedure means that it provides slightly different outputs, which have slightly different implications for extrapolation. In addition to an AF adjusted for treatment crossover, the IPE method provides the parameter values of the final parametric model used to estimate the adjusted treatment effect. Hence the IPE method allows a more direct extrapolation approach, as the final parametric model can be used for extrapolation purposes. The final parametric model could be used in either an independently fitted parametric model extrapolation process – by using the IPE final parametric model for the control group and a parametric model independently fitted to the unadjusted experimental group data – or in a PH type approach. However, because the IPE method uses an accelerated failure time model framework, a constant acceleration factor assumption would be made rather than the proportional hazards assumption. Because the final parametric model fitted under the IPE algorithm is fitted to fully recensored data, using this model for extrapolation purposes is prone to error due to loss of information.

As for the SNM and RPSFTM methods, because the IPE method uses an accelerated failure time model a “shrinkage” approach would also be feasible, whereby a counterfactual dataset is created using the IPE AF. Alternative extrapolation approaches as specified by the NICE DSU TSD survival model selection process algorithm could then be applied. However, again, the limitation with this approach is that it is not subject to full recensoring, and there are likely to be further implications for uncertainty analyses (as previously touched upon in Section 5.6.2.1).

5.6.2.5 Two-stage method

Like the RPSFTM and IPE methods, the simple two-stage method introduced in Section 4.12 of Chapter 4 creates a counterfactual dataset for the control group. A proportional hazards modelling extrapolation approach could be applied by fitting proportional hazards models to the counterfactual dataset, or independent parametric models could be fitted to separate treatment groups. However, the counterfactual dataset is subject to recensoring, which

results in a loss of information and potentially less robust survival extrapolations because longer-term data-points are recensored at an earlier time (with the size of the impact of recensoring dependent upon the size of the estimated treatment effect). It would not be logical to take a “survivor function” approach alongside this two-stage method, because the treatment effect initially estimated is for crossover patients, not for the experimental group.

5.7 Discussion

In NICE TAs standard parametric models such as the exponential and Weibull models are typically used to extrapolate survival data. It is common either for models to be fitted independently to each treatment arm – making no structural assumption about the treatment effect over time (other than that the survival curves are from the same family of parametric distribution) – or a PH modelling approach is taken where the proportional hazards assumption is made. The work undertaken for this chapter has been used to produce a NICE DSU Technical Support Document, advising upon how extrapolation should be approached where it is being undertaken alongside a clinical trial from which patient-level data are available.¹⁵ Analysts must consider the assumptions made by different parametric distributions, must analyse hazards observed in the clinical trial and the proportionality of these, and must consider internal validity in combination with external data or clinical expert opinion when selecting the most appropriate parametric model. More complex and flexible models such as the generalised gamma, piecewise models and spline-based models such as those described by Royston and Parmar¹⁶⁷ have been used very rarely in practice but should be considered.¹⁵

When treatment crossover adjustment methods are used the analyst’s ability to follow the survival model selection process algorithm advised by the NICE DSU TSD is somewhat compromised. Importantly, extrapolating the results of any of the methods will result in bias if the crossover adjustment method itself is biased. For example, while naive censoring and exclusion approaches are amenable to the full survival model selection process algorithm, they are highly prone to selection bias and so the results of their extrapolations are likely also to be biased. Similarly, if the “common treatment effect” assumption does not hold the RPSFTM and IPE methods will lead to biased extrapolations, with the same being true for the IPCW and SNM methods if the “no unmeasured confounders” assumption does not hold.

In addition to this, the crossover adjustment methods create practical problems for undertaking extrapolation. If it is desired to fit parametric models independently to treatment

arms after applying the IPCW method, a “new” dataset that represents the weighted Kaplan-Meier must be created, because the method only produces a WKM rather than an adjusted dataset. The IPCW method may seem more amenable to a proportional hazards modelling approach to extrapolation, but this cannot be undertaken as usual as the method is not fully parametric and does not provide an adjusted control group survival extrapolation – instead the inverse of the IPCW HR must be applied to the experimental group hazard function in order to extrapolate control group survival – an approach that is subject to its own theoretical limitations.

The SNM method causes fewer problems with regard to following the NICE DSU TSD survival model selection process algorithm, owing to the fact that in an RCT treatment crossover context a two-stage approach is required. This involves creating a counterfactual dataset to which any extrapolation approach could be applied. Similar is true for the simpler two-stage method and the RPSFTM method, although the counterfactual dataset created is subject to full recensoring – while this ensures an unbiased estimate of the treatment effect (assuming other methodological assumptions hold) this may increase bias associated with the extrapolation. The IPE algorithm’s different estimation procedure leads to slightly different outputs but the NICE DSU TSD survival model selection process could still be followed, although again this may be subject to extrapolation bias caused by recensoring.

In addition to the treatment crossover method-specific implications for extrapolation, standard extrapolation issues should also be considered when assessing how appropriate different treatment crossover adjustment methods are. For example, the IPCW method uses a weighted Cox model and therefore is a proportional hazards approach. If such an approach is to be used, the proportional hazards assumption should be justified, as stated by the NICE DSU TSD.¹⁵ Similar is true for the RPSFTM and IPE methods, which are based upon an accelerated failure time model and assume that there is a constant AF over time. Justifying these approaches is more problematic when treatment crossover adjustment is required; usually log-cumulative hazard plots or quantile-quantile plots could be used, but it would be inappropriate to produce these plots using ITT data because these would be subject to confounding due to crossover. For the RPSFTM and SNM methods these plots could be produced for the counterfactual dataset once adjustments had been made for treatment crossover, but the results of such an exercise would be subject to bias if the method used to produce the counterfactual dataset was biased – for example if the RPSFTM method was used to create the counterfactual dataset (under the assumption of a constant AF), and then a quantile-quantile

plot was produced, the plot would be likely to suggest that there is a constant AF – however that may be due to the use of the RPSFTM, rather than a true constant AF.

Given this, it might be difficult to satisfactorily justify a proportional hazards or constant AF assumption based upon the trial data. Therefore, it is even more important to attempt to justify such assumptions based upon external data, clinical expert opinion, or biological plausibility – the importance of which is touched upon by the NICE DSU TSD.¹⁵ When considering the plausibility of assumptions about the treatment effect, one advantage of the IPCW method compared to the RPSFTM and IPE methods is that while it assumes proportional hazards between the experimental and control groups, it makes no assumption about the treatment effect received by crossover patients. A key weakness of the RPSFTM and IPE methods is that they assume a constant AF across patients initially randomised to the experimental group, and patients who cross over onto the experimental treatment from the control group. The “two-stage” SNM method makes no “common treatment effect” or proportional hazards assumptions regarding the randomised experimental and control groups, but does assume that there is a constant AF received by control group crossover patients compared to control group patients that do not cross over.

The analysis included in this chapter has primarily focussed on survival modelling when patient-level data are available, because such data are required in order to be able to apply any of the treatment crossover adjustment methods. However, in an economic evaluation it may be required that external comparators (i.e. those that were not included in the pivotal clinical trial) are incorporated. Often this is achieved by conducting a comparison of hazard ratios across different trials, making use of common comparators (as described in Section 5.4.2). This may appear to be most amenable to an IPCW analysis, which provides an adjusted HR. However, adjusted HRs can also be obtained from an analysis of the counterfactual dataset under the RPSFTM or SNM methods, or using the final parametric model obtained from the IPE algorithm.

5.8 Conclusions

Together with Chapter 4, the present chapter comprises Part 3 of this thesis, the aim of which was to identify potential solutions to the treatment crossover problem given an economic evaluation context. Usually survival data need to be extrapolated for use in an economic model and it was important to consider how amenable each crossover adjustment method is to the extrapolation techniques that are typically observed in economic models – this has not

been previously considered in the literature. I have found that although complexities arise with regard to the performance of extrapolation following the application of crossover adjustment methods, SNMs, RPSFTM, the IPE algorithm and IPCW represent potentially appropriate methods for addressing treatment crossover in an economic evaluation context.

Given these findings I now move on to Part 4 of the thesis, which assesses how the identified treatment crossover adjustment methods perform in practice. First, in Chapter 6, a simulation study is presented. This demonstrates the performance of the identified crossover adjustment methods across a range of simulated scenarios. Chapter 7 then presents a real-world application of treatment crossover adjustment methods. The findings of Part 4 are drawn upon in Part 5 (Chapter 8) to formulate recommendations on the use of crossover adjustment methods in the context of economic evaluation – fulfilling the primary objective of this thesis.

Chapter 6

Simulation study of methods for adjusting survival estimates to take into account treatment crossover

6.1 Chapter overview

Chapters 4 and 5 have identified existing methods that are most relevant for adjusting survival estimates in the presence of treatment crossover, given an economic evaluation context. This chapter assesses the performance of the identified methods in a simulation study. The scenarios considered within the study aim to ensure that realistic circumstances reflecting the characteristics of real-world cancer clinical trials are tested. This leads to crossover adjustment methods being applied to complex scenarios in which their key assumptions do not hold. The simulation study is novel in two important ways. Firstly, although some of the crossover adjustment methods have previously been compared in a simulation study undertaken in the context of economic evaluation,²¹ never before have the complete range of methods considered in this chapter been compared in such a study. Secondly, the comparative sensitivities of crossover adjustment methods to a range of key methodological assumptions have not before been tested in realistic scenarios designed to reflect cancer RCTs.

The aim of the chapter is to provide evidence on the relative performance (in terms of bias, coverage and mean squared error) of alternative crossover adjustment methods in a range of scenarios, in order to inform recommendations on the use of these methods made in Part 5 of this thesis. The more complex, potentially appropriate methods identified in Chapter 4 are included alongside the commonly used naive methods discussed in Chapters 3 and 5. In Section 6.2 the rationale for a simulation study is briefly introduced. In Section 6.3 a previously completed simulation study in this area is discussed, which helps rationalise the design of the novel simulation study that is presented in Section 6.4. The simulation study is complex and the methods for generating the data are described in detail, in order that the results can be accurately interpreted and understood. Section 6.4 also describes the simulated scenarios, and the approaches used for applying the crossover adjustment methods. Section 6.5 presents a brief proviso on what can be expected of the simulation study results. The results are then given in Section 6.6. The large number of scenarios considered means that there are many results and for practicality these are grouped by method and scenario type. The grouping of results is explained at the start of Section 6.6. Limitations of the study are discussed in Section 6.7 and conclusions are made in Section 6.8.

6.2 Introduction

A simulation study is required to investigate the treatment crossover problem because to understand how well alternative methods work the “truth” must be known – that is, what would have happened if crossover had not occurred – so that the size of the bias (and coverage and mean squared error) associated with each method can be assessed. In real-world datasets the survival times that would have been experienced by crossover patients had they not crossed over are unknown. However, in a simulation study this truth is known and hence methods can be compared. Given that in this thesis I aim to make recommendations upon which crossover methods are most appropriate in a range of different circumstances, the challenge of the simulation study is to simulate data as realistically as possible, and to assess as many relevant scenarios as possible.

6.3 Relevant previous simulation study

As discussed in Section 1.2.6 Morden *et al* (2011) presented a simulation study of methods for dealing with treatment switching in RCTs.²¹ I was a co-author on the paper. Although the study was useful, there is great value in extending it. Lessons can be learned from the previous simulation study to ensure that the study described here is as informative and efficient as possible. In particular, lessons can be learned from the performance of the methods included in the previous study, and the sensitivity of results to the different scenarios included in the study.

Morden *et al* (2011) did not include observational-based approaches such as IPCW or SNMs in their analysis. Instead naive approaches (such as censoring, exclusion and ITT analyses) were compared to a limited subset of more complex methods. The more complex methods considered were:

- Adjusted Cox Model (Law and Kaldor, 1996)¹³⁰
- Causal proportional hazards estimator (Loeys and Goetghebeur, 2003)¹⁶⁸
- Rank Preserving Structural Failure Time Model (RPSFTM) (Robins and Tsiatis, 1991)^{22;22}
- Iterative Parameter Estimation (IPE) (Branson and Whitehead, 2002)²³
- Parametric Randomisation-based Method (Walker *et al*, 2004)¹³¹

The authors show that Law and Kaldor’s Adjusted Cox Model and Loeys and Goetghebeur’s causal proportional hazards estimator were outperformed across their simulations by the

RPSFTM and IPE methods.²¹ Law and Kaldor's method was excluded from the review presented in Chapter 4 because the method conditions on future events and is therefore highly likely to be biased. Similarly Loeys and Goetghebeur's method was excluded because it is designed for "all-or-nothing" compliance, which is not the case with treatment crossover as defined in this thesis. In addition, Morden *et al* found that Walker *et al*'s parametric method performed very poorly and often failed to converge.²¹ Therefore this method is also excluded from the simulation study presented in this chapter.

Morden *et al* tested estimation within the RPSFTM approach using the log-rank test, Cox-test, exponential test, and Weibull test and found that each gave very similar results. However, some estimation problems occurred in a small proportion of simulations when using the exponential, Weibull and Cox tests, and so the authors suggest that the log-rank test may be most appropriate for use. Hence, in the interests of efficiency only the log-rank test version of the RPSFTM is included in the simulation study presented here. Morden *et al* (2011) found that in each scenario tested bias was relatively small for the RPSFTM, although the treatment effect was slightly over-estimated, suggesting that the method was over-adjusting for treatment crossover. Across the scenarios tested the IPE algorithm performed best, producing the least bias.²¹

Morden *et al* (2011) simulated independent datasets using a Weibull model in which the true treatment effect on survival was known. A baseline prognostic covariate ("good" or "poor" prognosis) which influenced the probability of crossover was incorporated, but no time-dependent covariates or effects were included. It was assumed that the treatment effect was constant over time, and was equal in crossover patients and patients initially randomised to the experimental group – therefore, the "common treatment effect" assumption was assumed to hold. The bias, mean squared error (MSE) and coverage of each method was analysed across 16 scenarios. Weibull parameters were chosen to reflect a disease population that had a decreasing mortality rate over time, of whom 90% would have died after 3 years of follow-up (which was assumed to be the administrative censoring time). The authors tested scenarios which varied the prognosis of crossover patients, the difference in survival between prognosis groups, the probability of crossing over (dependent on prognosis group) and the treatment effect (HRs of 0.9 and 0.7 were tested).

In their results, Morden *et al* (2011) found that the bias associated with the methods (particularly the naive methods) was related to the proportion of control group patients that crossed over. Hence testing the bias associated with alternative methods for different

crossover proportions is important. Morden *et al* (2011) also found that their results were sensitive to the treatment effect, with the bias of naive methods typically increasing with an increasing treatment effect, but bias actually falling for some of the more complex methods (for example, the IPE algorithm). Hence it is important to test different levels of treatment effect in the novel simulation study presented in this chapter. Conversely, the authors found that the difference in expected survival between “good” and “poor” prognosis patients was not important, and therefore testing scenarios around this parameter is not a priority.

6.4 Novel simulation study

6.4.1 Introduction

There were two main limitations associated with the simulation study undertaken by Morden *et al* (2011). Firstly a “common treatment effect” was assumed – which satisfied the key assumption associated with randomisation-based methods like the RPSFTM and IPE algorithm. This assumption may be unrealistic, since crossover patients who receive the experimental treatment at a more advanced stage of disease progression may have a lower capacity to benefit. Considering this, it is important to consider how well alternative methods perform when the treatment effect is allowed to vary by group and over time. Secondly, Morden *et al* (2011) did not include any more complex observational-based approaches, such as SNMs with g-estimation or IPCW. It is important to consider the relative performance of these methods compared to the randomisation-based methods and the naive methods. The novel simulation study presented here aims to address these issues by including a time-dependent covariate in the data generating mechanism, by applying different treatment effects to crossover patients, and by including the complex observational-based approaches.

The inclusion of a time-dependent covariate is of pivotal importance in order to make the simulation study as realistic as possible. Prognosis-related crossover explains why naive methods for dealing with treatment crossover are likely to be biased. While crossover may be linked to baseline prognostic covariates, it is likely that prognostic time-varying covariates will also impact upon the probability of crossover, and these could be impacted upon by initial treatment. For example, a disease marker such as carcinoembryonic antigen (henceforth referred to as “antigen”), levels of which have been found to be raised in individuals with a range of cancers – such as colorectal, gastric, pancreatic, lung and breast cancer –, may be used to reflect the recurrence or spread of cancer. Equally, the Eastern Cooperative Oncology Group (ECOG) Performance Status score may be used over time to assess the progression of disease. If these indicators are influenced by treatment, influence whether and when

treatment crossover occurs, and have an impact on survival, then they represent time-dependent confounders and not accounting for these in a statistical analysis would lead to selection bias. However including these factors as time-varying covariates in a standard analysis would also lead to bias, because part of the treatment effect acts through them.¹²⁶ As discussed in Chapter 4, SNM and MSM (such as IPCW) methods attempt to appropriately adjust for time-dependent confounders.

Incorporating a prognostic time-dependent covariate makes the simulation study substantially more complex than that carried out by Morden *et al.* Given this, care was taken over the number of scenarios simulated, as running each scenario 1000 times (as in the Morden *et al.* study) and applying each identified method is computationally intensive.

6.4.2 Simulation Study Design

In this section a detailed protocol for the simulation study is presented, following the structure recommended by Burton *et al.* (2006).¹⁶⁹ In Section 6.4.2.1 the aims and objectives of the simulation study are specified. In Section 6.4.2.2 the data generating mechanism used within the simulation study is described. In Section 6.4.2.3 values of parameters used within the data generating mechanism are specified for the “base case” simulation scenario, Scenario 1. The parameter values were chosen in order to ensure the simulation of realistic survival times (as described in Section 6.4.2.3), treatment effects (as described in Section 6.4.2.4) and crossover proportions (as described in Section 6.4.2.5). In Section 6.4.2.6 the range of scenarios simulated are described. Sections 6.4.2.7 and 6.4.2.8 specify the performance measures that are used to assess the crossover adjustment methods and the key estimates from each scenario that are stored for analysis and interpretation. Sections 6.4.2.9 and 6.4.2.10 summarise the crossover adjustment methods that are included in the simulation study, and the approaches used to apply these.

6.4.2.1 Aims and Objectives

The objective of the simulation study is to assist in determining which crossover adjustment methods are most appropriate for use in economic evaluations. The study should help determine which methods work best in a variety of different scenarios. A key aim is to assess the sensitivities of alternative methods to departures from their theoretical assumptions – the following questions are of particular interest:

1. Do methods such as the RPSFTM and IPE still outperform naive approaches when the “common treatment effect” assumption does not hold?

2. How sensitive are the results of methods based on the “no unmeasured confounders” assumption to circumstances when certain prognostic covariates are not included in their estimation? This reflects a situation where data are not available on all important covariates, or where there is a risk that there may be some unmeasured confounders.
3. Observational-based methods rely on creating a weighted pseudo population based upon observed data from patients who do not switch treatments. A key question is how inaccurate these methods become when there are very few patients who do not cross over and at what stage do randomisation-based methods become more efficient and precise? This requires various scenarios to be run with different crossover proportions.
4. In which scenarios do randomisation-based methods perform best, and in which scenarios do observational-based methods perform best, when the “no unmeasured confounders” assumption holds? Building on this, which randomisation-based methods perform best, and which observational-based methods perform best? This is particularly important as the objective of the simulation study is to determine which methods work best in different scenarios – a key aim is to provide evidence on when an observational-based approach (and which specific approach) and when a randomisation-based approach (and which specific approach) is likely to be most appropriate.

The simulation study is designed to address these questions.

6.4.2.2 Methods for generating simulated datasets

Generating survival data influenced by a time-dependent confounder is not straightforward. In order to perform this task as efficiently as possible I collaborated with experts at the University of Leicester (Professor Keith Abrams, Dr Paul Lambert, and Mr Michael Crowther). Professor Abrams and Dr Lambert were members of an expert advisory group who agreed to help and advise me during my NIHR Doctoral Research Fellowship, and they recommended that some work completed by Mr Crowther could be adapted to generate the data that I required.

Mr Crowther’s work involves the simulation of complex survival data and allows survival times and a time-dependent covariate to be generated simultaneously in a joint model.¹⁷⁰ In order to generate the data for the simulation study presented here, a two-stage Weibull model was used to simultaneously generate the time-dependent covariate (referred to as “antigen”) and

survival times. The simulation study was conducted using STATA software, version 11.0.¹⁷¹ The STATA program used for to generate initial survival times – written by Mr Crowther, but amended as part of collaborative discussions between myself, Mr Crowther, Professor Abrams and Dr Lambert – is presented in Appendix 7. Once these initial survival times and antigen values had been generated, treatment crossover was applied, survival times were amended accordingly, and crossover adjustment methods were applied using the STATA program presented in Appendix 8 – written solely by me.

Within the initial data-generating joint model, the longitudinal model for the antigen value for the i th patient at time t is:

$$\text{antigen}_i(t) = \beta_{0i} + \beta_1 \log(t) + \beta_2 \log(t) \text{trt}_i + \beta_4 \text{badprog}_i \quad [11]$$

Where,

$$\beta_{0i} \sim N(\beta_0, \sigma_0^2)$$

β_{0i} is the random intercept, β_1 the slope for a patient in the control group, $\beta_1 + \beta_2$ the slope for a patient in the experimental treatment group (all assuming the same “badprog” status). β_4 is the change in the intercept for a patient with a poor (referred to as “badprog”) prognosis compared to a patient without a poor prognosis, trt_i is a binary covariate that equals 1 when the patient is in the experimental group and 0 otherwise, and badprog_i is a binary covariate that equals 1 when a patient has poor prognosis at baseline, and 0 otherwise. Hence the value of the antigen over time depends upon treatment group and the baseline prognosis group of the patient.

Based on methods described in detail by Bender *et al*, the antigen level is incorporated into the survival simulating process based upon the Weibull model hazard function:¹⁷²

$$h(t) = \lambda \gamma t^{\gamma-1} \exp(X\beta) \quad [12]$$

where,

$$X\beta = \delta_1 * \text{trt}_i + (\eta * \log(t)) * \text{trt}_i + \delta_2 \text{badprog}_i + \alpha(\text{antigen}(t)) \quad [13]$$

Given this, the survivor function can be calculated and survival times for the control group and the experimental group can be estimated. For the experimental group the full hazard function is:

$$h(t) = \lambda \gamma t^{\gamma-1} \exp(\delta_1 + \eta \log(t) + \delta_2 \text{badprog}_i + \alpha(\beta_{0i} + \beta_1 \log(t) + \beta_2 \log(t) + \beta_4 \text{badprog}_i)) \quad [14]$$

Where δ_1 is the baseline log hazard ratio, η is the time-dependent change in the treatment effect, δ_2 is the impact of poor prognosis, and α is the coefficient of the antigen level. The β 's are defined as for equation [11].

Survival is simulated with the antigen level having a linear relationship with log(time) rather than time, because otherwise the survival equations become analytically intractable. Owing to the logs included in equation [14], this is equivalent to:

$$h(t) = \lambda \gamma t^{\gamma-1+\alpha(\beta_1+\beta_2)+\eta} \exp(\delta_1 + \delta_2 \text{badprog}_i + \alpha(\beta_{0i} + \beta_4 \text{badprog}_i)) \quad [15]$$

The integral of which (between 0 and t) is the cumulative hazard function:

$$\int h(t) dt = \left[\frac{\lambda \gamma t^{\gamma+\alpha(\beta_1+\beta_2)+\eta}}{\gamma+\alpha(\beta_1+\beta_2)+\eta} \exp(\delta_1 + \delta_2 \text{badprog}_i + \alpha(\beta_{0i} + \beta_4 \text{badprog}_i)) \right]_0^T \quad [16]$$

The cumulative hazard function therefore is:

$$H(t) = \frac{\lambda \gamma}{(\gamma+\alpha(\beta_1+\beta_2)+\eta)} \exp(\delta_1 + \delta_2 \text{badprog}_i + \alpha(\beta_{0i} + \beta_4 \text{badprog}_i)) (t^{\gamma+\alpha(\beta_1+\beta_2)+\eta}) \quad [17]$$

This is equivalent to the $-\log$ of the survivor function, and therefore the survivor functions are:

Survivor function in experimental group with bad prognosis:

$$S(t) = \exp\left(\frac{-\lambda \gamma}{\gamma+\alpha(\beta_1+\beta_2)+\eta} \exp(\delta_1 + \delta_2 + \alpha(\beta_{0i} + \beta_4)) (t^{\gamma+\alpha(\beta_1+\beta_2)+\eta})\right) \quad [18]$$

Survivor function in experimental group with good prognosis:

$$S(t) = \exp\left(\frac{-\lambda \gamma}{\gamma+\alpha(\beta_1+\beta_2)+\eta} \exp(\delta_1 + \alpha(\beta_{0i})) (t^{\gamma+\alpha(\beta_1+\beta_2)+\eta})\right) \quad [19]$$

Similar steps can be taken to define the survivor functions in the control group, which are:

Survivor function in control group with bad prognosis:

$$S(t) = \exp\left(\frac{-\lambda\gamma}{\gamma+\alpha(\beta_1)} \exp(\delta_2 + \alpha(\beta_{0i} + \beta_4))(t^{\gamma+\alpha(\beta_1)})\right) \quad [20]$$

Survivor function in control group with good prognosis:

$$S(t) = \exp\left(\frac{-\lambda\gamma}{\gamma+\alpha(\beta_1)} \exp(\alpha(\beta_{0i}))(t^{\gamma+\alpha(\beta_1)})\right) \quad [21]$$

Through this model treatment and baseline prognosis is allowed to impact upon the antigen level over time. Treatment group, baseline prognosis and antigen level all also impact upon survival probabilities over time, and antigen level and time impact upon the treatment effect over time. Hence the model allows the kind of association between variables that would be expected in reality. The relationship between antigen level, time and the treatment effect is demonstrated using the following equations.

The Acceleration Factor (AF) is (as demonstrated by Collett (2003)):¹²

$$S_E = S_C \left(\frac{t}{\varphi}\right) \quad [22]$$

Where S_E is the survivor function in the experimental group, S_C is the survivor function in the control group, and φ is the AF. Given the survivor functions presented in equations [18-21] the AF over time can be estimated. First consider equation [22] with the full survivor functions included (for “bad prognosis” patients, but note that later prognosis cancels and so the treatment effect is not affected by baseline prognosis group):

$$\exp\left(\frac{-\lambda\gamma}{\gamma+\alpha(\beta_1+\beta_2)+\eta} \exp(\delta_1 + \delta_2 + \alpha(\beta_{0i} + \beta_4))(t^{\gamma+\alpha(\beta_1+\beta_2)+\eta})\right) = \exp\left(\frac{-\lambda\gamma}{\gamma+\alpha\beta_1} \exp(\delta_2 + \alpha(\beta_{0i} + \beta_4))\left(\left(\frac{t}{\varphi}\right)^{\gamma+\alpha\beta_1}\right)\right) \quad [23]$$

Cancelling, this becomes:

$$\frac{1}{\gamma+\alpha\beta_1+\alpha\beta_2+\eta} \exp(\delta_1)(t^{\gamma+\alpha\beta_1+\alpha\beta_2+\eta}) = \frac{1}{\gamma+\alpha\beta_1} \left(\left(\frac{t}{\varphi}\right)^{\gamma+\alpha\beta_1}\right) \quad [24]$$

Multiplying by the denominators, the following is obtained:

$$(\gamma + \alpha\beta_1)\exp(\delta_1)t^{\alpha\beta_2+\eta} = (\gamma + \alpha\beta_1 + \alpha\beta_2 + \eta)\varphi^{-\gamma-\alpha\beta_1} \quad [25]$$

Rearranging equation [25] gives us:

$$\frac{(\gamma + \alpha\beta_1)\exp(\delta_1)t^{\alpha\beta_2 + \eta}}{(\gamma + \alpha\beta_1 + \alpha\beta_2 + \eta)} = \varphi^{-(\gamma + \alpha\beta_1)} \quad [26]$$

We then take logs to obtain:

$$\log\left(\frac{(\gamma + \alpha\beta_1)\exp(\delta_1)t^{\alpha\beta_2 + \eta}}{(\gamma + \alpha\beta_1 + \alpha\beta_2 + \eta)}\right) = -(\gamma + \alpha\beta_1)\log(\varphi) \quad [27]$$

And thus the AF over time equals:

$$\varphi = \exp\left[\frac{-\log\left(\frac{(\gamma + \alpha\beta_1)\exp(\delta_1)t^{\alpha\beta_2 + \eta}}{(\gamma + \alpha\beta_1 + \alpha\beta_2 + \eta)}\right)}{(\gamma + \alpha\beta_1)}\right] \quad [28]$$

The 't' term in this equation illustrates the time-dependency of the treatment effect, and the β terms illustrate the antigen-dependency.

6.4.2.3 Parameter values – survival times

Values for the parameters introduced in Section 6.4.2.2 were selected in order to ensure that the simulated data suitably represented the type of dataset that the study was designed to replicate. In particular, simulated survival times, treatment effects, and crossover proportions had to be reasonable.

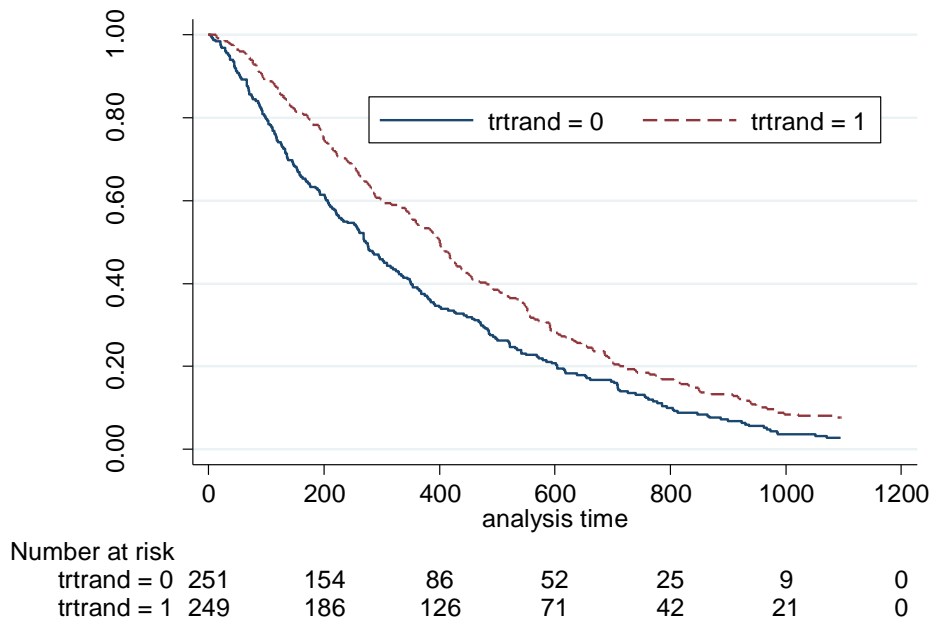
To ensure that the survival times produced by the data generating mechanism represented the survival times seen in RCTs of metastatic cancer treatments variations of parameter values were tested. In the "base case" (Scenario 1) analysis the parameter values for the Weibull survival model and the longitudinal antigen model were:

$$\beta_0 = 20, \sigma_0^2 = 1, \beta_1 = 15, \beta_2 = -4, \beta_4 = 5, \delta_1 = -0.7, \delta_2 = 0.5, \alpha = 0.02, \lambda = 0.0005, \gamma = 0.9, \eta = 15, \omega = 0.50^1$$

The Kaplan-Meier curves that are produced by the simulation model (in the absence of treatment crossover) using these parameter values are presented in Figure 6.1. Note that trtrand=0 represents the control group, and trtrand=1 represents the experimental group.

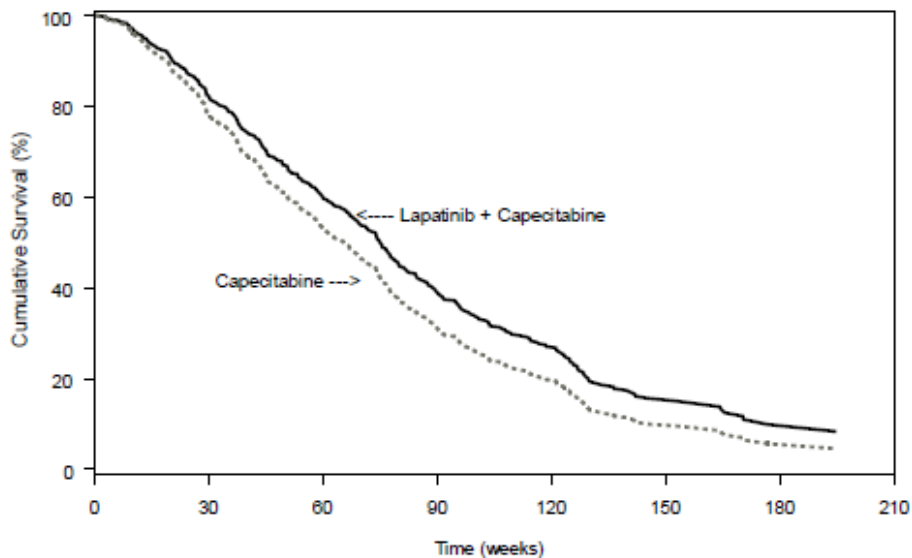
¹ Note, the parameter ω was not introduced in Section 6.4.2.2. It is defined in Section 6.4.2.4

Figure 6.1: Overall Survival Kaplan-Meier from simulated dataset Scenario 1: Zero crossover



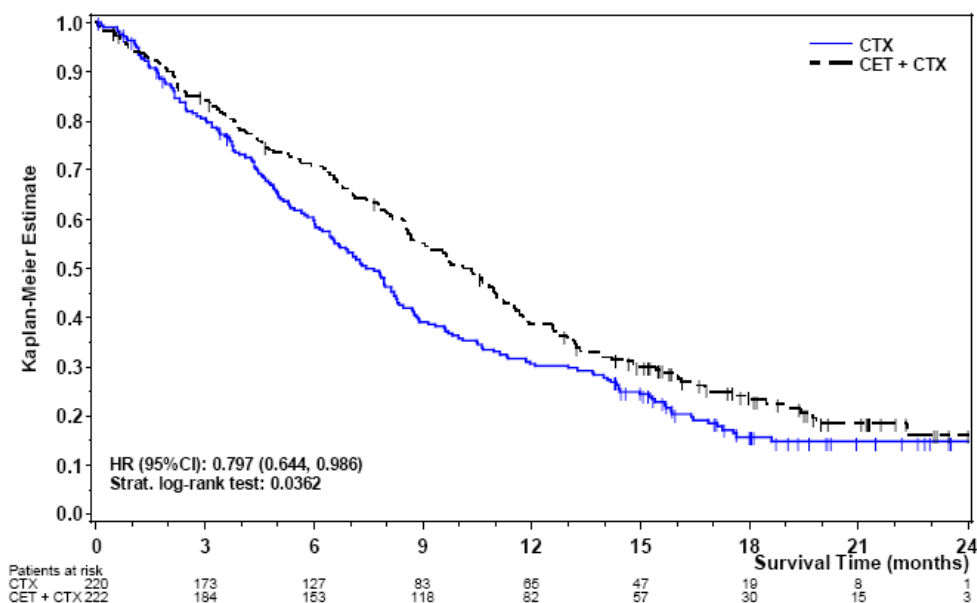
The shape of these Kaplan-Meier curves seem reasonable when compared to Kaplan-Meier's for OS taken from two recent NICE technology appraisals in metastatic cancer in which treatment crossover was seen to be an issue, shown in Figures 6.2 and 6.3. Note that the timescale is days in Figure 6.1, weeks in Figure 6.2, and months in Figure 6.3.

Figure 6.2: Overall Survival Kaplan-Meier - lapatinib+capecitabine for metastatic breast cancer



Note: This is Figure 2 in GSK's submission to NICE, August 2009.¹⁷³

Figure 6.3: Overall Survival Kaplan-Meier - cetuximab for metastatic squamous cell carcinoma of the head and neck



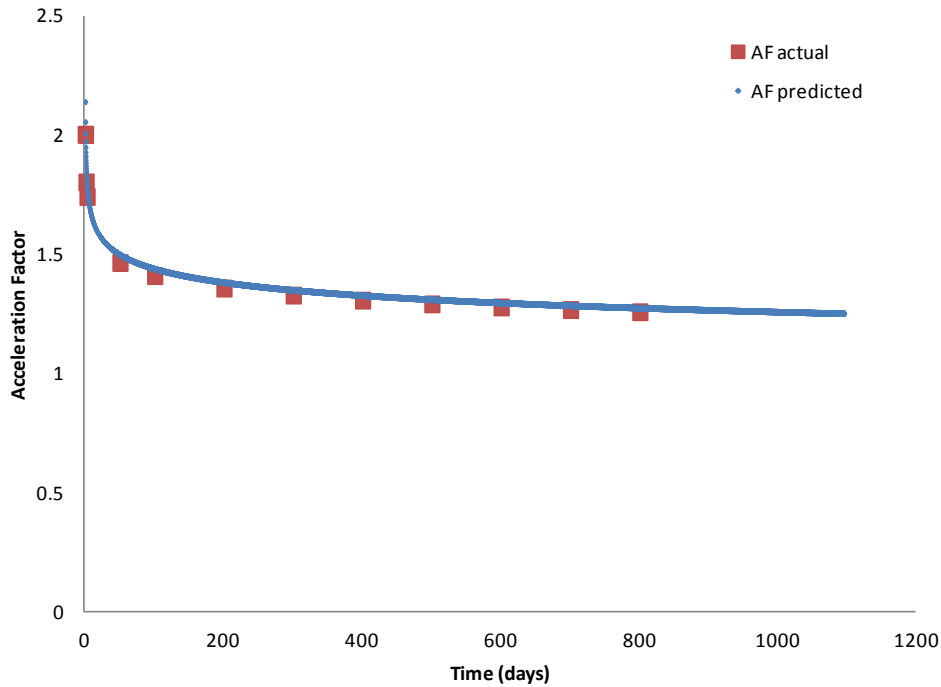
Note: This is Figure B2 from: Merck Serono’s submission to NICE, 2009.¹⁷⁴

6.4.2.4 Parameter values – treatment effect

In tandem with ensuring suitable survival times were obtained, values for model parameters were selected to produce a reasonable treatment effect over time. The aim was to simulate a treatment effect that reduced the progression of the antigen value over time (and thus had an additional effect through the antigen). In combination with this, the overall treatment effect was assumed to reduce over time as disease steadily progresses and capacity to benefit is assumed to reduce. To achieve this suitable values for α and η were chosen. It was necessary to include parameter η because otherwise a treatment that reduced the progression of the antigen level would have a treatment effect that increased over time, which I assumed to be unrealistic.

Using the parameter values stated in Section 6.4.2.3, the treatment effect illustrated in Figure 6.4 is obtained. As desired, the treatment effect reduces over time. “Actual” AF values are calculated empirically from the application of the survivor functions presented in equations [18-21], whereas “predicted” AF values are those estimated using the formula presented in equation [28]. This demonstrates that the formula for the time-dependent AF is accurate. Importantly, Figure 6.4 illustrates that the treatment effect reduces over time but always remains above 1, which means that the treatment effect never becomes negative.

Figure 6.4: Simulated acceleration factor over time



The decrease in the treatment effect over time is kinked – it reduces rapidly at first before the slope reduces in gradient. This is because survival is simulated with the antigen level having a linear relationship with $\log(\text{time})$ rather than time. This may be regarded as a limitation of the approach, but is not of high importance – a treatment effect that reduces over time is still modelled, as desired.

The treatment effect applied to crossover patients in the simulation study was not calculated using the time-dependent AF equation, since this equation estimates the treatment effect at a certain time-point assuming the patient had been receiving the treatment since baseline – which is not the case for crossover patients. Instead, the baseline treatment effect was applied to crossover patients, but was multiplied by a factor (ω) such that the effect received by crossover patients was lower than the average effect received by experimental group patients. This is necessary to ensure that the treatment effect applied to crossover patients was not higher than the average effect received by patients in the experimental group, which would appear counterintuitive. The magnitude of ω was varied to test scenarios in which the treatment effect received by crossover patients was different to the average effect received by experimental group patients to a greater and lesser extent. A two-step approach was taken to calculate plausible values for ω . First, the simulation program was run with a very large sample size (1,000,000 simulated patients) in the absence of treatment crossover, and the average AF associated with the experimental treatment was calculated. Second, the value of ω applied to the baseline treatment effect that would be required to result in crossover

patients receiving a treatment effect that is 15% lower than the average effect received by the experimental group was calculated. A 15% reduction was regarded to represent a reasonably substantial decrease in the treatment effect, but this figure was chosen fairly arbitrarily and therefore values of ω were varied to investigate the impact of altering the size of the treatment effect received by crossover patients – in addition to the 15% decrement, decrements of 0% and 25% were considered.

The approach for estimating the treatment effect applied to crossover patients means that while the treatment effect received by these patients is less than that received by those in the experimental group, it is equal for all crossover patients and does not reduce over time. This may be regarded as a limitation of the study, but it represents a reasonable simplifying assumption. Because all crossover patients switch treatment either at or very close to disease progression, they may have a similar capacity to benefit from the treatment even if (for example) one patient received the treatment 100 days after randomisation and another received it 300 days after randomisation. An alternative approach would have been to estimate the effect received by crossover patients using equation [28], but as stated previously this equation is suitable only for patients that have received the treatment since baseline. In addition, using equation [28] would mean that a lower treatment effect is applied to a patient who experiences a long progression free survival period before switching treatments than a patient who progresses and crosses over more quickly; which in itself may be deemed implausible.

Essentially, the mechanism of the treatment effect in crossover patients is open to debate and is likely to differ in different diseases and patients. The approach taken here allows the bias in crossover methods when the “common treatment effect” assumption does not hold to be tested, and allows control over the extent to which the treatment effect differs in these two groups. Hence the approach is sufficient to address the research questions focussed upon in this thesis and the aims specified in Section 6.4.2.1. It may be of interest to test alternative scenarios in which the treatment effect in crossover patients is estimated in different ways. However, it is important to note that this would not be expected to alter the bias associated with the complex methods such as RPSFTM and IPCW – the randomisation-based methods do not attempt to model the treatment effect process and so the important factor that determines their bias is the extent to which the average treatment effect differs between crossover and experimental group patients – not how this treatment effect difference is estimated. The IPCW approach censors crossover patients and so the treatment effect

received (and how it is estimated) by these patients does not influence the bias with which the method estimates the treatment effect received in the experimental group.

6.4.2.5 Parameter values – the crossover mechanism

While simultaneously ensuring that suitable survival times and treatment effects were simulated, I developed a crossover mechanism that produced the desired crossover proportions in simulated datasets. In the base case scenario the probability of treatment crossover was allowed to depend upon the antigen value at the time of disease progression (split into three categories – referred to as “antigen group at progression”) and the time of progression itself (also split into three categories – referred to as “time to progression group”). Crossover was not allowed prior to disease progression, to reflect the treatment crossover typically seen in metastatic cancer trials. In addition, crossover was only allowed to occur at one of the three consultations immediately following disease progression (including the consultation at which progression was first observed), and the probability of crossover declined in each of these consultations. Consultations were assumed to occur every 21 days (also in line with metastatic cancer trials) and hence the earliest that crossover could occur was 21 days after randomisation, and the latest that crossover could occur was 42 days after the first consultation at which disease progression was observed. This reflects that in a clinical trial setting if a patient has not crossed over onto the experimental treatment soon after experiencing disease progression, they are very unlikely to cross over. In addition, the requirement that crossover could only occur at one of three consultations following disease progression allowed a practical simplification of the simulation program.

The probability of crossover was calculated for each control group patient using a logistic function (in keeping with the definition of treatment crossover in this thesis, crossover could not occur from the experimental group onto the control treatment). In the base case the probability of crossover increased if the antigen value was low at the time of disease progression, and if time-to-progression was high. This was assumed to reflect a clinical decision in which patients with high antigen values (indicative of more progressed disease) may be clinically inappropriate for crossover, whereas patients with a long time-to-progression were likely to have a good prognosis. Hence in the base case scenario crossover was linked to good prognosis. This is complicated by the fact that a patient with low antigen levels at progression is likely to have progressed quickly (because the antigen value increases over time), and so the treatment crossover mechanism may be regarded as not being “clear-cut”. However this is likely to reflect reality, and the impact of this complicating factor was tested in

further scenarios which modelled a clearer link between prognosis and crossover probability (this is discussed in Section 6.4.2.6).

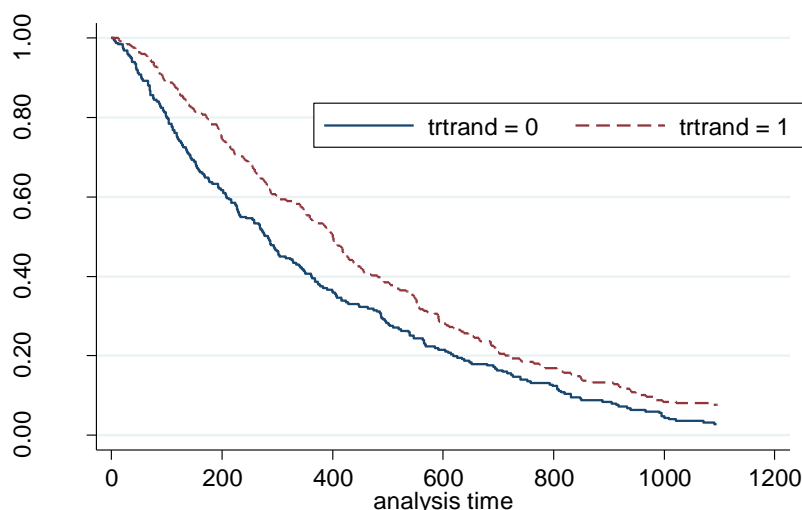
The probability of crossover in the different progression and antigen groups at the three consultations following disease progression for the base case scenario are presented in Table 6.1. Higher group numbers represent higher values for that group (that is, “time to progression group” 0 are the control group patients that had time-to-progression times in the lowest 33.3% of the control group). Note however that these groups only refer to patients who became “at-risk” of crossover – that is, those control group patients that survived for longer than 21 days. Hence the lowest 33% represent the lowest third of the at-risk group, not the control group as a whole.

Table 6.1: Probability of treatment crossover by prognostic groups and consultation

Consultation 1		Antigen group at progression		
		0	1	2
Time progression group	to 0	0.30	0.20	0.10
	1	0.51	0.41	0.30
	2	0.88	0.75	0.60
Consultation 2		Antigen group at progression		
		0	1	2
Time progression group	to 0	0.23	0.15	0.07
	1	0.42	0.32	0.23
	2	0.84	0.68	0.51
Consultation 3		Antigen group at progression		
		0	1	2
Time progression group	to 0	0.14	0.09	0.04
	1	0.29	0.22	0.15
	2	0.75	0.55	0.38

When crossover is incorporated into Scenario 1 using the crossover probabilities presented in Table 6.1, 156 (62%) control group patients switch treatments. The resulting Kaplan-Meier curve is as shown in Figure 6.5. As desired, the control group Kaplan-Meier curve moves towards the experimental group Kaplan-Meier curve (as can be seen by comparing Figure 6.5 with Figure 6.1). The ITT HR in this instance increases to 0.743 from 0.719, demonstrating the effect of the crossover.

Figure 6.5: Overall Survival Kaplan-Meier from simulated dataset Scenario 1: With crossover



Number at risk							
trtrand = 0	251	155	91	54	31	12	0
trtrand = 1	249	186	126	71	42	21	0

6.4.2.6 Scenarios to be investigated

The simulation study was designed taking into account information supplied by companies that are part of the Pharmaceutical Oncology Initiative (POI), a group who provided a real-world clinical trial dataset for use in Chapter 7 of this thesis, and who engaged with the research. The POI is made up of representatives from pharmaceutical companies that are part of the Association of the British Pharmaceutical Industry (ABPI) that have a particular interest in oncology products, and have considerable experience in clinical trials of oncology treatments. The POI were asked for advice regarding relevant scenarios to include in the study, in order to reflect the typical metastatic oncology clinical trials funded by industry.

The simulated data generating mechanism has several variables for which values must be assumed. These are listed in Table 6.2. Values for each variable in Scenario 1 are quoted, as are alternative values for different scenarios. Values for each parameter are specified for each scenario in Appendix 9.

Table 6.2: Simulated scenarios – Parameter values and alternatives tested

Variable	Value (Scenario 1)	Alternative Values
Sample size	500	-
Number of prognosis groups (prog)	2	-
Probability of good prognosis	0.5	-
Probability of poor prognosis	0.5	-
Maximum follow-up time	3 years (1095 days)	-
Baseline effect of being in the "bad prognosis" group	Log hazard ratio = 0.5	-
Survival time distribution	Weibull parameters when time-dependent effect included:	Alter scale parameter to 0.001 to represent a more severe disease (and

	<p>Shape parameter 0.9 (mortality decreasing over time) Scale parameter 0.0005</p> <p>Weibull parameters when time-dependent effect not included: Shape parameter 0.9 (mortality decreasing over time) Scale parameter 0.004</p>	<p>hence less censoring) in scenarios with time-dependent effect</p> <p>Alter scale parameter to 0.007 to represent a more severe disease (and hence less censoring) in scenarios without time-dependent effect</p>
Progression free survival	Overall survival time multiplied by a value from a beta distribution with shape parameters (5,5) – this implies the assumption that PFS is approximately half of OS. This is not an important assumption – PFS is only included because I model a situation where crossover cannot occur before disease progression	-
Baseline treatment effect (note this is not the true treatment effect as this does not take into account the effect of the treatment that occurs through the time-dependent confounder, antigen level, or the time-dependent part of the treatment effect, η)	<p>Baseline log hazard ratio in scenarios that include an additional time-dependent effect = -0.7</p> <p>Log hazard ratio in scenarios that do not include an additional time-dependent effect = -0.3</p>	<p>Alter log hazard ratio to -1.1 to represent a larger treatment effect in scenarios with time-dependent effect</p> <p>Alter log hazard ratio to -0.7 to represent a larger treatment effect in scenarios without time-dependent effect</p>
Antigen intercept	Calculated using a normal distribution with mean of 20 and standard deviation of 1	-
Antigen value progression over time	As demonstrated by formula [11]. $\beta_2 = -4$ to represent that the antigen value increases more slowly in the experimental group, and $\beta_4 = 5$ so that bad prognosis patients start with higher levels of the antigen	-
Impact of antigen value on overall survival	As demonstrated by formulas [12-21]. Increased antigen value increases the risk of death. The strength of this relationship depends on the variable α , which equals -0.02 in Scenario 1	Remove impact of the antigen value by setting $\alpha=0$
Impact of antigen value on treatment effect	As demonstrated by formulas [22-28]. Because treatment reduces the progression of the antigen value and increased antigen values increase the risk of death, the treatment has an additional effect through the antigen. The strength of this relationship depends on the variable α , which equals -0.02 in Scenario 1	<p>Remove impact of antigen value by setting $\alpha=0$</p> <p>Model larger time-dependent effect by applying additional decrement multiplier to crossover patients</p>
Time-dependent portion of treatment effect, η	$\eta = 0.15$ to generate a reduction in the treatment effect over time, in scenarios where a time-dependent treatment effect is assumed	Remove impact of η by setting $\eta = 0$ in scenarios where the treatment effect is not time-dependent
Assumed frequency of consultations	One every 3 weeks (21 days)	-
Probability of switching treatment over time	As shown in Table 6.1. This results in a crossover proportion of approximately 63% in Scenario 1	Test a high crossover scenario where all probabilities are increased – to an extent where approximately 90% of patients that survive longer than 21 days crossover
Prognosis of crossover patients	As shown in Table 6.1. This makes crossover more likely in good prognosis patients, via a mechanism that takes into account both time to progression	<p>Make crossover more likely in poor prognosis patients.</p> <p>Test scenarios where crossover is</p>

	and antigen value at progression	based on a simpler mechanism (only based on the antigen value)
Treatment effect in crossover patients	Equal to baseline treatment effect multiplied by ω . Set ω such that treatment effect received by crossover patients is 85% of the average effect received by experimental group patients in base scenarios.	Alter ω such that treatment effect received by crossover patients equals 75% - 78% of the average effect received by experimental group patients. In scenarios where the treatment effect is not time-dependent, set ω to 1 – such that the treatment effect received by crossover patients is the baseline effect received by the experimental group.

Based on the variables and alternative values presented in Table 6.2, 72 scenarios were run. For the scenarios that included a time-dependent treatment effect the impacts of altering disease severity, the treatment effect, the crossover proportion, and the treatment effect decrement applied to crossover patients were tested. Applying these four alterations to Scenario 1 meant that a total of 16 scenarios were required. In addition, I wished to evaluate a range of scenarios in which the “common treatment effect” assumption held, and thus scenarios that did not incorporate a time-dependent treatment effect (referred to as “zero TDC scenarios”) were included. In these disease severity, treatment effect, and the crossover proportion were altered, which required 8 scenarios. This combined to 24 scenarios.

Given that the more complex observational-based methods rely upon being able to accurately model the treatment crossover risk, different treatment crossover mechanisms were tested to assess the sensitivity of the methods to more complex scenarios. In the base case the probability of treatment crossover increased with “time to progression group”, but decreased with “antigen group at progression”. Because the antigen value increases over time patients with a long time to progression are likely to have high antigen values, and hence the crossover decision is not clear-cut in its relationship to prognosis. For this reason all 24 of the base scenarios were tested again in simulations in which the treatment crossover decision was only based upon “antigen group at progression”. In Scenarios 25-48 it was assumed that treatment crossover was more likely if the antigen value was low at progression (hence patients who progressed quickly and were therefore more likely to have poor prognosis were more likely to crossover). In Scenarios 49-72 Scenarios 25-48 were repeated, but under the assumption that crossover was more likely if the antigen value was high at progression (hence patients who progressed slowly and were therefore more likely to have good prognosis were more likely to crossover).

The sensitivity of observational-based crossover adjustment methods to the “no unmeasured confounders” assumption was not addressed through scenarios, instead different versions of the SNM and IPCW methods were applied in which it was assumed that data on certain prognostic variables were and were not available. This will be described further in Sections 6.4.2.9 and 6.4.2.10.

In total 72 scenarios were run. In each scenario a sample size of 500 patients was simulated, to reflect study sizes generally found in oncology clinical trials, and matching the sample size used by Morden *et al* (2011).²¹ One-thousand simulations were run for each scenario in accordance with the approach taken by Morden *et al*.

6.4.2.7 Performance measures

The time-dependency of the simulated treatment effect means that it is not possible to produce a single “true” HR or acceleration factor that the results of the crossover adjustment methods can be compared to. Hence, an alternative “truth” was required, in order that the crossover adjustment methods could be assessed and compared. By integrating the survivor functions given in equations 18 to 21 the true mean survival time can be calculated, and therefore restricted mean survival at 1095 days (the administrative censoring time in the simulated dataset) was used as the “truth” upon which performance measures were based. The *restricted* mean was convenient to use as it avoided potential additional biases associated with having to extrapolate survival times and placed the focus of the simulation study on crossover adjustment methods, rather than extrapolation methods.

However, the use of restricted mean survival as the performance measure meant that this had to be calculated for the control group for each of the crossover adjustment methods, which is more complex than simply estimating the treatment effect in terms of a HR or AF. In some circumstances extrapolation-type techniques were required (although little extrapolation was actually undertaken, owing to the fact that a restricted mean at 1095 days was used), and so the amenability of the crossover adjustment methods to extrapolation – as discussed in Chapter 5 – was important. The methods used to obtain estimates of mean survival at 1095 days for each of the crossover adjustment methods are described in Section 6.4.2.10.

This use of restricted mean estimation as the primary performance measure is particularly relevant given the economic evaluation context of this thesis. From an economic modeller’s perspective the ideal approach for adjusting for crossover would provide an adjusted control group dataset or survival curve that can be used (or extrapolated for use) in the economic

model. The ideal method would produce counterfactual survival datasets and survival curves that are similar to the true unconfounded datasets and survival curves, and would thereby minimise the difference between true mean survival and that estimated using the crossover adjustment method.

The performance of the crossover adjustment methods was evaluated according to the bias in their estimate of control group restricted mean survival at 1095 days. Bias (δ) was measured by the difference between the true restricted mean (β) and the restricted mean estimated by the crossover methods ($\bar{\beta}$):

$$\delta = \bar{\beta} - \beta \quad [29]$$

The mean squared error (MSE) was also calculated to provide information on the bias associated with each method in combination with the variation of estimates. This is useful in this study because some of the methods are observational-based whereas others are randomisation-based, and these two different approaches are likely to result in important differences in standard errors (SE) of estimates. The MSE was calculated as:

$$\text{MSE} = (\bar{\beta} - \beta)^2 + (\text{SE}(\hat{\beta}))^2 \quad [30]$$

Where $\text{SE}(\hat{\beta})$ is the standard error associated with the mean restricted mean estimated by each crossover adjustment method over the 1,000 simulations run for each scenario.

The coverage of each method was also calculated, defined as the proportion of times the 95% confidence intervals of the restricted mean estimated by each method contained the true restricted mean. A good method should result in coverage of approximately 95%, reflecting that 95% of confidence intervals contain the true value. However, when a method is biased it is unlikely that it will have good coverage. The proportion of times that each method resulted in an estimate of the treatment effect (that is, the proportion of times they converged) was recorded, to illustrate the reliability of the methods. Where methods did not converge the bias and coverage performance measures were calculated based upon simulations in which convergence did occur. Scenarios in which convergence of certain methods did not occur were further explored in order to identify potential reasons.

6.4.2.8 Simulation outputs to be stored and summarised

For each simulation the restricted mean survival estimated for the control group for each crossover adjustment method was recorded, as well as the associated confidence intervals. In addition the proportion of patients that crossed over, and the proportion of patients that were censored, were recorded. These proportions could not be directly controlled in the simulation program, and recording these aids the analysis and interpretation of results. The treatment effect estimate (in terms of a HR or an AF and their confidence intervals) associated with each method were also recorded. Although these are inherently biased because they are not time-dependent whereas the true treatment effect is time-dependent (in the majority of scenarios), the “average” true HR and AF *can* be estimated for each scenario. Comparing these to the estimates from each of the crossover adjustment methods represents a useful secondary analysis for assessing the performance of the methods. The “average” true HR and AF for each scenario was estimated by generating the scenario data for a large number of patients (1,000,000) without applying treatment crossover, and applying Cox (for a HR) and Weibull (for an AF) models to this data.

6.4.2.9 Methods included in the simulation study

The methods included in the simulation study are listed in Table 6.3. Naive methods are included so that their relative bias can be assessed compared to the more complex methods identified in Chapter 4.

Table 6.3: Crossover methods for inclusion in simulation study

Naive Methods	
Intention to treat analysis	Accelerated Failure Time (AFT) model
	Proportional Hazards (PH) model
Per protocol analysis	Exclude crossover patients
	Censor at crossover
Treatment as a time-dependent covariate	PH model
	AFT model
Treatment crossover as an indicator variable	PH model
	AFT model
Complex Methods	
Structural Nested Models	Two-stage method, observational SNM with <i>g</i> -estimation to estimate treatment effect in crossover patients, followed by standard randomisation-based ITT analysis on adjusted “counterfactual” data
	SNM excluding bad prognosis covariate in order to test sensitivity to “no unmeasured confounders” assumption
IPCW	IPCW approach using logistic models for the probability of censoring, and using stabilised weights as recommended by Fewell ¹⁷⁵
	Standard IPCW approach, but with antigen covariates excluded from models in order to test sensitivity to “no unmeasured confounders” assumption
Randomisation-based complex approaches	RPSFTM with log-rank test
	RPSFTM with log-rank test including baseline covariates in

	estimation
	IPE algorithm using Weibull model
	IPE algorithm using Exponential model
	IPE algorithm using Weibull model including baseline covariates in estimation
	IPE algorithm using Exponential model including baseline covariates in estimation
Other Methods	
Two-stage Weibull	Two-stage method in which a Weibull model is used to estimate the treatment effect in crossover patients, using covariates measured at baseline and at disease progression, described in Section 6.4.2.10

The complex methods listed in Table 6.3 cover the main methodological strands identified from the review presented in Chapter 4. The naive methods represent simple methods that have often been used to analyse data in which treatment crossover has occurred – as demonstrated by the review of NICE appraisals reported in Chapter 3. Usually proportional hazards (PH) models (for example, a Cox model) are used to analyse survival data, but since SNMs and the RPSFTM/IPE methods take the form of accelerated failure time (AFT) models it is appropriate to compare the more complex methods to naive methods in the form of AFT as well as PH models to allow a secondary comparison of treatment effects. One “other” method is included, the “two-stage Weibull”, which is a version of the two-stage method introduced in Section 4.12 of Chapter 4. This is described further in Section 6.4.2.10.

6.4.2.10 Applying the methods

In this section I describe precisely how each method was applied and how restricted mean survival was estimated.

- *Intention to treat*

The restricted mean associated with an ITT analysis was calculated by applying STATA’s “stci, rmean” command which computes the mean survival time restricted to the longest follow-up time (that is, 1095 days) for the specified patient group. For the ITT analysis this was applied to the control group data confounded by treatment crossover.

- *Per protocol – exclude switchers*

The restricted mean associated with a per-protocol analysis where crossover patients were excluded was calculated by excluding all crossover patients and using STATA’s stci command.

- *Per protocol – censor switchers*

The restricted mean associated with a per-protocol analysis where crossover patients were censored was calculated by censoring all crossover patients at the time at which they crossed over, and using STATA's `stci` command on this adjusted dataset.

- *Treatment as a time-dependent covariate*

Four versions of this method were applied. Treatment was included as a time-dependent covariate within a Cox model and a Weibull model including and excluding covariates. All methods were expected to be biased, because excluding covariates means that the prognosis of crossover patients is not accounted for, whereas including covariates adjusts for factors through which the treatment has an effect. Restricted means were estimated using “survivor function” approaches – as described in Sections 5.6.2.2 and 5.6.2.3 of Chapter 5. A Weibull model was fitted to the experimental group survival data and the associated survivor function and hazard function were derived. For the treatment as a time-dependent covariate methods that provided a hazard ratio (HR) adjusted for crossover (the Cox model versions of the approach) the experimental group hazard function was multiplied by the inverse of the HR to obtain the control group hazard function, and the control group survivor function was then derived up to 1095 days. For methods that provided an acceleration factor (AF) (the Weibull model versions of the approach) the time associated with each experimental group survivor function probability was divided by the AF adjusted for crossover in order to obtain the survival times associated with the survival probabilities for the control group, and mean survival was estimated up to 1095 days by calculating the area under the survival curve. Confidence intervals (CIs) for the mean survival estimate were calculated by applying the 95% CIs of the estimated treatment effect in the “survivor function” process.

- *Crossover as a time-dependent indicator variable*

Four versions of this method were applied. Crossover was included as a time-dependent indicator variable within a Cox model and a Weibull model including and excluding covariates. This method differed very slightly from the “treatment as a time-dependent covariate” method, because in this method survival effects were estimated separately for crossover patients and for experimental group patients due to the inclusion of the crossover indicator variable. “Survivor function” approaches were used to obtain estimates of restricted mean survival.

- *Structural Nested Models - Observational SNM with g-estimation*

The observational SNM included in the simulation study was applied using the `stgest` command in STATA, in line with the example given by Sterne and Tilling in *The STATA Journal* (2002).¹²⁶ Full details on `stgest` are given in the Sterne and Tilling evidence table in Appendix 4. Generally, the SNM method is applied to observational datasets and in the context of an RCT applying the method is more complex, as discussed in Chapters 4 and 5. The SNM was applied to the control group after disease progression had occurred, to estimate the treatment effect in crossover patients compared to control group patients that did not crossover. The resulting AF was used to “shrink” survival times in crossover patients in order to arrive at a counterfactual dataset which is adjusted for treatment crossover (as described in Section 5.6.2.1).

The restricted mean associated with the adjusted dataset was estimated using STATA's `stci` command, and CIs for the restricted mean estimate were calculated by applying the 95% CIs of the estimated treatment effect in the “shrinkage” process. As will be discussed in Section 6.6.6 of this chapter, this is likely to represent an underestimate of the true confidence interval as it does not take into account the uncertainty in the underlying survival distribution.

In order to test the sensitivity of the observational SNM method to not including all prognostic confounders, two versions of the method were included. In the first version, all covariates (baseline prognosis group, time-to-disease progression group, antigen at baseline group, antigen at progression group, and antigen over time group) were included, and in the second version the baseline prognostic group covariate was excluded.

- *Inverse Probability of Censoring Weights*

IPCW was applied in line with the example given by Fewell *et al* (2004), although unlike Fewell *et al*'s example the IPCW method rather than a full MSM was applied.¹⁷⁵ In addition, weights were only applied to patients in the control group, as the context is an RCT rather than an observational study.

To apply the IPCW method using stabilised weights first the simulated data were split into time intervals and time-dependent covariate values were recorded for each of these. Data were excluded for crossover patients beyond the point of crossover, and OS was censored for these patients. IPCWs were then estimated for each patient and for each time interval. The numerator of each stabilised weight was the cumulative probability of remaining uncensored (that is, not crossing over) from the beginning of follow-up to the end of the interval given only baseline covariates and the number of consultations since randomisation. This was estimated

for all control group patients for all time periods. The denominator of the stabilised weight was the cumulative probability of remaining uncensored (that is, not crossing over) to the end of the interval given baseline and time-dependent covariates, and the number of consultations since randomisation. These weights were only different from 1 for time periods during which patients were at risk of crossover – that is after disease progression had been observed, and before 3 consultations after disease progression had taken place. The probabilities of remaining uncensored (that is, not crossed over) were obtained by fitting pooled logistic models with informative censoring due to treatment crossover as the dependent variable. A Cox proportional hazards model that incorporated baseline (but not time-dependent) covariates was then run, weighted by the stabilised weights, in order to estimate an adjusted IPCW HR.

To test the sensitivity of the IPCW method to the “no unmeasured confounders” assumption two versions of the method were applied. The first included all baseline and time-dependent covariates:

- Baseline prognosis group
- Baseline antigen group (antigen level at baseline split into 3 groups)
- Time-to-disease progression group
- Antigen at disease progression group
- Antigen group (which splits patients into 6 groups based upon their antigen level as this changes over time)
- An interaction term of time-to-disease progression group multiplied by antigen at time of progression group.

The second version excluded all antigen-related covariates.

The IPCW approach allows a weighted Kaplan-Meier (WKM) curve to be obtained, which would provide the optimal restricted mean measure for the method. However, Fewell *et al*'s STATA methodology does not provide this curve, and there are problems with calculating this in the context of a simulation study. As discussed in Chapter 5, for the WKM to be estimated the sum of the weights for all patients at risk, and all patients who experienced the event, for each time point, must be calculated. In the simulation study it was possible that control group patients with the longest follow-up times may cross over and be censored at an earlier date and therefore a new administrative censoring time (that is, the time to which the restricted mean is calculated) would need to be generated for each simulation in order to avoid biased overestimates of mean survival being produced. Therefore, generating the WKM accurately in

a simulation study would be very computationally-intensive and therefore a “survivor function” approach – as described for the “Treatment as a time-dependent covariate” methods, above – using the IPCW HR was taken to estimate restricted mean survival for the control group at 1095 days. This should represent a close approximation of the IPCW WKM. CIs for the mean survival estimate were calculated by applying the 95% CIs of the IPCW HR in the “survivor function” process.

- *RPSFTM with log-rank test*

The RPSFTM included in the simulation study was applied using the *strbee* STATA program developed by White *et al* (2002).¹²⁵ Full details on *stgest* are given in the White *et al* (2002) evidence table in Appendix 4. The method is straight-forward to apply as it relies only upon the randomisation of the (simulated) trial and does not require the incorporation of any covariate information. White *et al* state that censoring that is non-informative on the observed time scale is informative on the counterfactual time scale (as discussed in Chapter 4) and therefore the *strbee* command incorporates recensoring using potential censoring times. *Strbee* also allows baseline covariates to be included in an attempt to marginally increase the accuracy of the approach, and a version of the method was applied including baseline covariates.

The RPSFTM method provides an estimate of the treatment effect adjusted for treatment crossover in the form of an acceleration factor. It also provides counterfactual survival times – i.e. survival times that would have been observed if nobody had received treatment. Given this, there are three possible approaches to calculating the control group restricted mean, as discussed in Chapter 5. The approaches included in the simulation study were:

- “Shrinkage” approach. This is similar to the “shrinkage” approach used for the SNM method. The inverse of the AF was used to shrink survival times in crossover patients and the *stci* command was used to estimate restricted mean survival. This approach does not involve full recensoring as although the AF is estimated using recensoring and survival times of crossover patients are recensored, survival times of other control group patients are not. This creates the potential for bias - when the estimation of the treatment effect is accurate the “shrinkage” approach is likely to lead to higher bias than the “extrapolation” or “survivor function” approaches described below.
- “Extrapolation” approach. Under this approach the recensored counterfactual survival times produced by the *strbee* command were extrapolated out to

1095 days and the area under the extrapolated survival curve (that is, the restricted mean) was estimated. A Weibull model was used to extrapolate. The extrapolated portion was relatively small as mean survival was restricted to 1095 days, but a poor extrapolation could still impact upon the bias associated with this approach.

- “Survivor function” approach. This approach is similar to the “survivor function” approach previously described for the “Treatment as a time-dependent covariate” method, for versions that produce an acceleration factor.

The “extrapolation” and “survivor function” methods differ slightly as the “extrapolation” method extrapolates recensored counterfactual survival times, whereas the “survivor function” approach assumes a constant acceleration factor applied to experimental group survival times that have not been recensored. The “extrapolation” approach suffers from potential inaccuracy associated with the information lost through recensoring, whereas the “survivor function” approach suffers from potential biases associated with extrapolating under the assumption of a constant acceleration factor. All three approaches were included in the study in order to better understand their advantages and disadvantages. For all versions CIs for the restricted mean estimate were calculated by applying the 95% CIs of the estimated treatment effect in the restricted mean estimation process.

- *IPE algorithm*

The IPE algorithm approach was also applied using the *strbee* STATA program.¹²⁵ The method was applied using full recensoring, and included versions with and without covariates. It was also applied using both a Weibull distribution and an exponential distribution in order to examine the sensitivity of the method to the parametric form chosen in the treatment effect estimation process.

In addition to an AF adjusted for treatment crossover, the IPE method provides the parameter values of the final parametric model used to estimate the adjusted treatment effect. These were used to estimate the restricted mean survival at 1095 days associated with the final model. This is similar to the “extrapolation” approach described above for the RPSFTM method. “Shrinkage” and “survivor function” approaches were also applied in order to obtain alternative restricted mean estimates. Importantly when the IPE method was applied using an exponential model substantial differences between the “extrapolation” approach and the “survivor function” approach were expected. Under the “extrapolation” approach the

extrapolation was undertaken using the final exponential model estimated by the IPE approach, whereas under the “survivor function” approach the treatment effect obtained using the exponential version of the IPE method was applied to the experimental group survival times estimated using a Weibull model. The results of these two methods provide information on the impact of choosing alternative parametric models not only for the treatment effect estimation process, but also for the extrapolation procedure.

As for the other crossover methods, CIs for the restricted mean estimate were calculated by applying the 95% CIs of the estimated treatment effect in the restricted mean estimation process. As noted by Morden *et al* (2011) this was likely to provide relatively poor coverage as the confidence intervals associated with the treatment effect from the final IPE iteration are underestimates.²¹

- *Other methods: Two-stage Weibull*

One final approach was included in the simulation study. The “two-stage Weibull” method is not complex in that it does not apply a weighting mechanism or g-estimation in an attempt to control for crossover in the presence of time-dependent confounders. However, given the simulation study data generating mechanism described in Sections 6.4.2.2 – 6.4.2.5, the method was likely to result in reasonably low levels of bias. The method is similar to that described in Section 4.12 of Chapter 4. First a treatment effect is estimated in the control group for the period following disease progression. A Weibull model was used to estimate this treatment effect, including the following covariates:

- Baseline prognosis group
- Baseline antigen group
- Time-to-disease progression group
- Antigen at disease progression group

These are the “baseline” covariates in the secondary dataset that only covers the post-progression period for the control group. The resulting treatment effect was then used as in equation [10] presented in Chapter 5 to shrink survival times in crossover patients. The restricted mean survival of the adjusted dataset produced was then calculated using STATA's `stci` command, and confidence intervals were calculated using the confidence intervals of the treatment effect.

The method was likely to work well in the simulation study for two key reasons. Firstly, crossover was simulated to occur soon after disease progression. This limited the impact of

time-dependent confounders once they had been adjusted for at the disease progression “baseline”. Secondly, the underlying survival data were generated using Weibull models, which may have been to the benefit of Weibull-based methods (although, due to the inclusion of the antigen time-dependent covariate, survival times did not follow a true Weibull distribution).

Despite the fact that the two-stage Weibull approach may perform well in the simulation study “falsely” – that is, because of the way the data were generated – there remains value to including this method. The mechanism used to generate the simulated data reflects a realistic crossover mechanism, whereby crossover can only occur after disease progression, and if it is to occur it will occur soon after progression. In these circumstances a two-stage approach seems reasonable, and can adjust for some of the bias associated with time-dependent confounders without the need for data intensive weighting methods or g-estimation. In addition, the method allows a different treatment effect to be estimated for crossover patients compared to that estimated for the experimental group. However, the method would not be possible if crossover could occur prior to disease progression, and the bias associated with it would increase if crossover could occur at any time point between disease progression and death.

6.5 Proviso

Before reporting the results of the simulation study, it is important to emphasise that there was not an expectation that any of the included methods would operate with close to zero bias across the entire range of scenarios. The randomisation-based methods such as RPSFTM and IPE were expected to produce low bias in scenarios in which there was not a time-dependent confounder causing the “common treatment effect” assumption to be compromised. However, in scenarios where this assumption did not hold substantial bias was expected. It was expected that observational-based methods such as IPCW and SNM may produce lower bias when a time-dependent treatment effect was simulated, as these methods do not require the “common treatment effect” assumption. However, these methods are reliant on being able to appropriately model the treatment crossover and survival process, which may not be possible across all the simulated scenarios (particularly when crossover proportions were very high). In addition, all these methods implicitly assume a constant treatment effect over time in the experimental group (as they produce an adjusted HR or AF), and because a time-dependent treatment effect was simulated in the majority of scenarios some bias associated with this was expected.

However, the fact that none of the included methods were expected to work perfectly across all scenarios does not mean the analysis is unimportant or inappropriate. In practice, in the presence of treatment crossover a choice has to be made as to which method to use to estimate the treatment effect and survival over time. This study seeks to demonstrate which of the existing imperfect methods can be expected to produce least bias across a range of realistic scenarios.

6.6 Results

The performance of each method differed importantly depending upon the scenario investigated. Due to the large number of methods and scenarios assessed it is not helpful to present detailed results for every method and every scenario. Instead I will focus upon groups of methods and groups of scenarios, in order to illustrate key findings.

In Section 6.6.1 details of each of the scenarios are presented, with regard to average treatment effects, true restricted mean survival, crossover proportions and censoring proportions. This gives an overview of the different scenarios that were run, which is useful when analysing how well the alternative methods worked across scenarios.

In Section 6.6.2 the bias associated with the crossover adjustment methods is analysed by method across the simulated scenarios. In Section 6.6.3 the comparative relative bias of the alternative methods is assessed across scenarios, particularly focussing on groups of scenarios defined by whether or not a time-dependent treatment effect was simulated (thus, whether or not the “common treatment effect” assumption held). This comparison of crossover adjustment methods is summarised in Section 6.6.4. In Sections 6.6.5 and 6.6.6 the coverage and MSE of the methods are considered, and the associated implications are discussed. In Section 6.6.7 the estimates of the treatment effect (in terms of a hazard ratio or acceleration factor) associated with each method across scenarios are considered in order to determine whether any anomalies exist in the results (for example, methods that appear to estimate the HR or AF relatively well/poorly, but then lead to poor/good estimates of control group restricted mean survival).

6.6.1 Overview of simulation scenarios

Table 6.4 provides information on each of the scenarios simulated. The true restricted mean unconfounded by treatment crossover is presented, along with the average treatment effect in

terms of a hazard ratio (calculated using a Cox model) and an acceleration factor (calculated using a Weibull model). Where there is a time-dependent treatment effect this reflects only an approximation of the true treatment effect as the proportional hazards/constant acceleration factor assumptions do not hold. In terms of a hazard ratio, the average treatment effect varied between 0.50 and 0.75.

Table 6.4 shows that the crossover proportion varied between 52% and 94% of all control group patients. Scenarios 13-24, 37-48 and 61-72 were designed to result in higher levels of crossover, although these levels are probabilistic and are reliant on other characteristics. This led the level of crossover to vary between otherwise equivalent scenarios with crossover proportions highest in Scenarios 25-48, followed by Scenarios 49-72 and 1-24. Table 6.4 also presents the crossover proportion as a percentage of the control group patients that became “at-risk” of crossover. Control group patients could only cross over if they were alive at their first “consultation” at 21 days. The proportion of patients that died before this point and never became at-risk depended upon disease severity. The proportion of crossover patients as a percentage of patients that became at-risk ranged from 61% to 96%.

This is particularly important to consider for observational-based methods such as IPCW as these are reliant upon modelling the probability of crossover. This can only be achieved by comparing patients who were at-risk of crossing over and becomes increasingly difficult at the extremes – either when almost all patients crossover, or when very few patients crossover. In addition, the IPCW method essentially generates a “pseudo population” based upon uncensored patients, and if there are very few of these patients high weightings are applied which could lead to bias.

Table 6.4 shows that Scenarios 1-24 incorporated a complex crossover probability mechanism in which better prognosis patients were generally more likely to cross over. Scenarios 25-48 and 49-72 incorporated a simpler crossover probability mechanism based only upon antigen level at the time of disease progression. In Scenarios 25-48 patients with a relatively poor prognosis were more likely to cross over. The opposite was true in Scenarios 49-72. It is important to remember that proportions of patients never became at-risk of crossover (3-8% in the low severity scenarios and 6-13% in the high severity scenarios) and therefore even when “poor prognosis” patients were classed as being more likely to crossover these did not represent the *poorest* prognosis patients.

Table 6.4 shows that in Scenarios 1, 2, 5, 6, 13, 14, 17 and 18 the treatment effect received by experimental group patients was dependent upon the antigen level and time and in these scenarios crossover patients received a reduced treatment effect. In Scenarios 9, 10, 11, 12, 21, 22, 23 and 24 the treatment effect received by experimental group patients was time-dependent and related to the antigen level, and an additional decrement (compared to scenarios 1, 2, 5, 6, 13, 14, 17 and 18) was applied to the effect received by crossover patients (in Table 6.4 these are labelled as “Yes+” in the “Time-dependent treatment effect” column). In scenarios 3, 4, 7, 8, 15, 16, 19 and 20 the treatment effect was not time-dependent or related to the antigen – in these scenarios the “common treatment effect” assumption held. This pattern across scenarios was repeated in scenarios 25-48 and 49-72 (i.e. scenarios 25 and 49 are equivalent to scenario 1, except with altered crossover mechanisms, and so on). Further details on the treatment effect decrement applied to crossover patients in each scenario are included (“Treatment effect in crossover patients (AF)”), as well as on the treatment effect as a proportion of that received by the experimental group. This varied between 75% and 100%.

Table 6.4: Overview of simulated scenarios

Scenario	Truth		Average treatment effects		Mean crossover % of total	Mean crossover % of at risk	Mean censoring proportion (%)	Disease severity	Prognosis of crossover patients	Time-dependent treatment effect	Treatment effect in crossover patients (AF)	% of exp group treatment effect
	Restricted mean (Control group)	Restricted mean (Exp group)	HR	AF								
1	372.06	462.27	0.75	1.28	63.37%	65.33%	7.19%	Low	Complex - good	Yes	1.09	85%
2	372.06	579.28	0.52	1.75	61.54%	63.44%	13.42%	Low	Complex - good	Yes	1.48	85%
3	344.47	568.12	0.51	2.15	56.26%	61.07%	19.97%	Low	Complex - good	No	2.15	100%
4	344.47	437.88	0.75	1.39	58.46%	63.45%	11.65%	Low	Complex - good	No	1.39	100%
5	216.96	285.64	0.73	1.32	60.25%	64.04%	0.84%	High	Complex - good	Yes	1.12	85%
6	216.96	381.51	0.50	1.80	58.17%	61.82%	2.74%	High	Complex - good	Yes	1.53	85%
7	201.45	387.21	0.51	2.17	52.48%	60.56%	7.02%	High	Complex - good	No	2.17	100%
8	201.45	271.95	0.75	1.40	54.09%	62.41%	2.80%	High	Complex - good	No	1.40	100%
9	372.06	462.27	0.75	1.28	63.74%	65.71%	6.86%	Low	Complex - good	Yes +	1.00	78%
10	372.06	579.21	0.52	1.75	61.45%	63.36%	12.86%	Low	Complex - good	Yes +	1.31	75%
11	216.96	285.64	0.73	1.32	60.52%	64.32%	0.76%	High	Complex - good	Yes +	1.00	76%
12	216.96	381.51	0.50	1.80	58.02%	61.66%	2.58%	High	Complex - good	Yes +	1.36	75%
13	372.06	462.27	0.75	1.28	88.37%	91.10%	7.25%	Low	Complex - good	Yes	1.09	85%
14	372.06	579.21	0.52	1.75	87.96%	90.68%	13.80%	Low	Complex - good	Yes	1.48	85%
15	344.47	568.12	0.51	2.15	80.99%	87.90%	20.60%	Low	Complex - good	No	2.15	100%
16	344.47	437.88	0.75	1.39	81.29%	88.23%	11.80%	Low	Complex - good	No	1.39	100%
17	216.96	285.64	0.73	1.32	83.30%	88.53%	0.83%	High	Complex - good	Yes	1.12	85%
18	216.96	381.51	0.50	1.80	82.66%	87.85%	2.81%	High	Complex - good	Yes	1.53	85%
19	201.45	387.21	0.51	2.17	74.83%	86.34%	7.24%	High	Complex - good	No	2.17	100%
20	201.45	387.21	0.75	1.40	75.20%	86.77%	2.84%	High	Complex - good	No	1.40	100%
21	372.06	462.27	0.75	1.28	88.39%	91.12%	6.87%	Low	Complex - good	Yes +	1.00	78%
22	372.06	579.21	0.52	1.75	88.02%	90.75%	13.09%	Low	Complex - good	Yes +	1.31	75%
23	216.96	285.64	0.73	1.32	83.30%	88.53%	0.74%	High	Complex - good	Yes +	1.00	76%
24	216.96	381.51	0.50	1.80	82.77%	87.97%	2.65%	High	Complex - good	Yes +	1.36	75%

25	372.06	462.27	0.75	1.28	69.65%	71.80%	7.02%	Low	Simple - poor	Yes	1.09	85%
26	372.06	579.21	0.52	1.75	71.62%	73.84%	13.10%	Low	Simple - poor	Yes	1.48	85%
27	344.47	568.12	0.51	2.15	65.86%	71.48%	19.08%	Low	Simple - poor	No	2.15	100%
28	344.47	437.88	0.75	1.39	64.03%	69.50%	11.15%	Low	Simple - poor	No	1.39	100%
29	216.96	285.64	0.73	1.32	64.89%	68.96%	0.81%	High	Simple - poor	Yes	1.12	85%
30	216.96	381.51	0.50	1.80	66.71%	70.90%	2.60%	High	Simple - poor	Yes	1.53	85%
31	201.45	387.21	0.51	2.17	59.78%	68.98%	6.47%	High	Simple - poor	No	2.17	100%
32	201.45	387.21	0.75	1.40	57.88%	66.79%	2.61%	High	Simple - poor	No	1.40	100%
33	372.06	462.27	0.75	1.28	69.16%	71.30%	6.89%	Low	Simple - poor	Yes +	1.00	78%
34	372.06	579.21	0.52	1.75	71.64%	73.86%	12.52%	Low	Simple - poor	Yes +	1.31	75%
35	216.96	285.64	0.73	1.32	65.02%	69.10%	0.75%	High	Simple - poor	Yes +	1.00	76%
36	216.96	381.51	0.50	1.80	66.95%	71.15%	2.52%	High	Simple - poor	Yes +	1.36	75%
37	372.06	462.27	0.75	1.28	93.31%	96.20%	7.23%	Low	Simple - poor	Yes	1.09	85%
38	372.06	579.21	0.52	1.75	93.41%	96.30%	13.78%	Low	Simple - poor	Yes	1.48	85%
39	344.47	568.12	0.51	2.15	87.06%	94.50%	20.65%	Low	Simple - poor	No	2.15	100%
40	344.47	437.88	0.75	1.39	86.83%	94.25%	11.72%	Low	Simple - poor	No	1.39	100%
41	216.96	285.64	0.73	1.32	89.12%	94.71%	0.84%	High	Simple - poor	Yes	1.12	85%
42	216.96	381.51	0.50	1.80	89.54%	95.16%	2.80%	High	Simple - poor	Yes	1.53	85%
43	201.45	387.21	0.51	2.17	81.17%	93.65%	7.17%	High	Simple - poor	No	2.17	100%
44	201.45	387.21	0.75	1.40	80.77%	93.20%	2.81%	High	Simple - poor	No	1.40	100%
45	372.06	462.27	0.75	1.28	93.31%	96.20%	6.86%	Low	Simple - poor	Yes +	1.00	78%
46	372.06	579.21	0.52	1.75	93.51%	96.40%	12.96%	Low	Simple - poor	Yes +	1.31	75%
47	216.96	285.64	0.73	1.32	88.87%	94.44%	0.75%	High	Simple - poor	Yes +	1.00	76%
48	216.96	381.51	0.50	1.80	89.49%	95.11%	2.62%	High	Simple - poor	Yes +	1.36	75%
49	372.06	462.27	0.75	1.28	66.26%	68.31%	7.15%	Low	Simple -good	Yes	1.09	85%
50	372.06	579.21	0.52	1.75	64.57%	66.57%	13.45%	Low	Simple -good	Yes	1.48	85%
51	344.47	568.12	0.51	2.15	59.10%	64.15%	19.99%	Low	Simple -good	No	2.15	100%
52	344.47	437.88	0.75	1.39	60.78%	65.97%	11.46%	Low	Simple -good	No	1.39	100%

53	216.96	285.64	0.73	1.32	61.53%	65.39%	0.82%	High	Simple -good	Yes	1.12	85%
54	216.96	381.51	0.50	1.80	60.01%	63.78%	2.76%	High	Simple -good	Yes	1.53	85%
55	201.45	387.21	0.51	2.17	53.90%	62.19%	6.96%	High	Simple -good	No	2.17	100%
56	201.45	387.21	0.75	1.40	55.26%	63.76%	2.77%	High	Simple -good	No	1.40	100%
57	372.06	462.27	0.75	1.28	66.15%	68.20%	6.91%	Low	Simple -good	Yes +	1.00	78%
58	372.06	579.21	0.52	1.75	64.45%	66.45%	12.86%	Low	Simple -good	Yes +	1.31	75%
59	216.96	285.64	0.73	1.32	61.95%	65.84%	0.78%	High	Simple -good	Yes +	1.00	76%
60	216.96	381.51	0.50	1.80	60.00%	63.76%	2.58%	High	Simple -good	Yes +	1.36	75%
61	372.06	462.27	0.75	1.28	91.44%	94.27%	7.26%	Low	Simple -good	Yes	1.09	85%
62	372.06	579.21	0.52	1.75	90.87%	93.68%	13.88%	Low	Simple -good	Yes	1.48	85%
63	344.47	568.12	0.51	2.15	84.11%	91.30%	20.81%	Low	Simple -good	No	2.15	100%
64	344.47	437.88	0.75	1.39	84.13%	91.31%	11.77%	Low	Simple -good	No	1.39	100%
65	216.96	285.64	0.73	1.32	86.39%	91.81%	0.85%	High	Simple -good	Yes	1.12	85%
66	216.96	381.51	0.50	1.80	86.25%	91.67%	2.79%	High	Simple -good	Yes	1.53	85%
67	201.45	387.21	0.51	2.17	77.63%	89.57%	7.30%	High	Simple -good	No	2.17	100%
68	201.45	387.21	0.75	1.40	77.78%	89.75%	2.81%	High	Simple -good	No	1.40	100%
69	372.06	462.27	0.75	1.28	91.42%	94.25%	6.87%	Low	Simple -good	Yes +	1.00	78%
70	372.06	579.21	0.52	1.75	91.40%	94.23%	12.98%	Low	Simple -good	Yes +	1.31	75%
71	216.96	285.64	0.73	1.32	86.59%	92.03%	0.75%	High	Simple -good	Yes +	1.00	76%
72	216.96	381.51	0.50	1.80	86.27%	91.69%	2.59%	High	Simple -good	Yes +	1.36	75%

6.6.2 Bias of methods across scenarios

In this section the bias associated with each method or group of methods is analysed across the range of simulated scenarios. First, the bias of the ITT analysis is assessed across all scenarios. Second, naive exclusion and censoring methods are considered. Third, naive methods that incorporated treatment group and crossover indicators as time-dependent covariates are analysed. Fourth, the results of the IPCW method are assessed. Fifth, the RPSFTM and IPE algorithm results are analysed, grouping scenarios according to whether or not the “no unmeasured confounders” assumption held. Sixth, the results of the SNM with g-estimation method are analysed, and finally the results of the simple “two-stage Weibull” method are assessed. Graphs showing bias across scenarios are presented throughout this Section – care should be taken when comparing these because the y-axes use different scales.

- ITT analysis

The ITT analysis resulted in positive bias (over-estimates of mean survival in the control group) across all scenarios, as expected. This is due to the positive effect of the experimental treatment on crossover patients. Figure 6.6 shows that, as expected, this bias increased when the treatment effect was high and when the proportion of control group patients that crossed over was high. A high disease severity also marginally increased bias. The pattern of bias across Scenarios 1-24, 25-48 and 49-72 was similar, and average bias across these scenarios was also similar (12.5%, 12.6% and 13.0% respectively), which suggests that the bias associated with the ITT analysis is not related to the prognosis of patients that crossover. However, because the crossover mechanism used was probabilistic, it was not possible to ensure that the proportion of patients crossing over was identical across the different scenarios, and in Scenarios 25-48 (in which poor prognosis patients were more likely to crossover) a larger proportion of patients crossed over on average compared to Scenarios 1-24 and 49-72 (194 patients crossed over on average in Scenarios 25-48, compared to 178 in Scenarios 1-24 and 184 in Scenarios 49-72). Hence, this may be indicative that if all other circumstances were equal, the ITT bias would be smaller if poor prognosis patients were more likely to crossover. However, this is likely to be marginal, and of less importance than the treatment effect and the proportion of patients that cross over.

It is important to note that in Scenarios 9, 11, 21, 23, 33, 35, 45, 47, 57, 59, 69 and 71 the bias associated with the ITT analysis was very low. This is because these scenarios involved applying a large treatment decrement (approximately 25%) to the treatment effect received by crossover patients, to the extent that the treatment effect received amounted to an acceleration factor only marginally greater than 1. Hence, because crossover patients received

almost zero benefit from crossing over the ITT analysis was unbiased. This is true to a lesser extent in Scenarios 1, 5, 13, 17, 25, 29, 37, 41, 49, 53, 61 and 65, in which a treatment decrement of approximately 15% was applied to crossover patients and the treatment effect was relatively low to begin with (HR of approximately 0.75). Thus the benefit received by crossover patients was again relatively small in these scenarios. This demonstrates that the bias associated with the ITT analysis is likely to be highly associated with the magnitude of the treatment effect, in particular the magnitude of the treatment effect received by crossover patients.

- Naive methods – exclusion and censoring

Exclusion and censoring approaches always performed poorly, with censoring always leading to a positive bias (overestimation of control group survival) and exclusion almost always leading to a negative bias (underestimation of control group survival). The censoring approach never led to less bias than the ITT analysis and the relative biases were often extremely large. This demonstrates that such an approach is very unreliable when there is a relationship between prognosis and crossover. Bias was consistently high across all scenarios, but was most influenced by higher levels of crossover (illustrated in Figure 6.7 by higher levels of bias in Scenarios 13-24, 37-48 and 61-72). Bias was also increased when disease severity was relatively high.

Initially it may appear unexpected that the censoring method led to positive bias across all scenarios. This might be expected in Scenarios 25-48 where poor prognosis patients were more prone to switching (and therefore censoring). However, in scenarios where better prognosis patients were more likely to switch, censoring might be expected to lead to underestimates of survival in the control group. This finding was also apparent in the simulation study undertaken by Morden *et al*,²¹ and appears to occur because when good prognosis patients are more likely to cross over the crossover generally occurs quite late in the study. Therefore several of the long-term observations are censored, and those that remain alive and uncensored at these times tend to be long-term survivors (as to still be in the risk-set they have already survived a long time). Hence there are very few death events at the tail of the survival curve, which leads to an overestimation of mean survival.

The exclusion approach outperformed the ITT analysis in 9.7% of scenarios, but generally the bias associated with the approach was very high (though not as high as the censoring approach). As expected, the bias was generally positive (control group survival was over-estimated) when patients with relatively poor prognosis were more likely to crossover (see

several scenarios between Scenario 25 and Scenario 36). This is due to the exclusion of poor prognosis patients from the control group, leading to an over-estimate of survival in the control group. Conversely, bias was negative when good prognosis patients were more likely to cross over (Scenarios 1-24 and 49-72). However, in Scenarios 37 to 48 bias was negative even though poor prognosis patients were most likely to cross over. It is likely that this is due to the simulated crossover mechanism. In Scenarios 37 to 48 the vast majority of control group patients crossed over, and it is likely that only patients with the very worst prognosis (many of whom never became at-risk of crossover due to dying before 21 days) did not – causing the exclusion of crossover patients to lead to negative bias.

It is notable that the exclusion approach produced relatively low bias when the crossover proportion was relatively low, and patients with relatively poor prognosis were more likely to cross over (see Scenarios 25-36 in Figure 6.7). Again, this is explainable through the simulation mechanism. As discussed above, patients could only cross over if they lived for longer than 21 days – which removes the possibility of crossover for patients with the worst prognosis. Hence, in scenarios where “poor” prognosis patients were more likely to cross over, these patients may actually have been similar to “average” patients, which would explain why their exclusion did not significantly bias the results. The bias associated with the exclusion approach increased with higher crossover proportions, and with higher disease severity. An increase in the size of the treatment effect had minimal impact on bias.

- Naive methods – time-dependent treatment variable, and crossover indicator variable

8 methods were analysed that involved either including treatment received as a time-dependent covariate in the survival model, or including an indicator variable for treatment crossover (as described in the Section 6.4.2.10).

Of the four time-dependent treatment variable methods (TDCM, TDCS, TDCM-Weibull, TDCS-Weibull), the versions that included baseline covariates and a covariate indicating time to PFS (that is, TDCM and TDCM-Weibull) produced less bias than the versions that did not include any other covariates (that is, TDCS and TDCS-Weibull). None of these methods adequately account for time-dependent confounders and thus were expected to produce substantial bias and generally this was demonstrated. The TDCS and TDCS-Weibull methods performed similarly (as shown by Figure 6.8), both always giving substantial positive bias (overestimating control group survival). Bias generally increased with higher treatment effects, and when the crossover proportion was higher. The methods never produced less bias than the ITT analysis.

Figure 6.6: Mean bias (%) across scenarios – ITT analysis

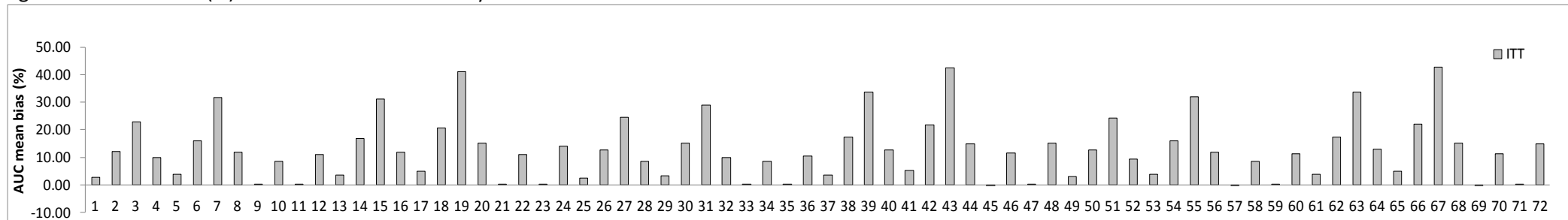


Figure 6.7: Mean bias (%) across scenarios – Exclusion and censoring approaches

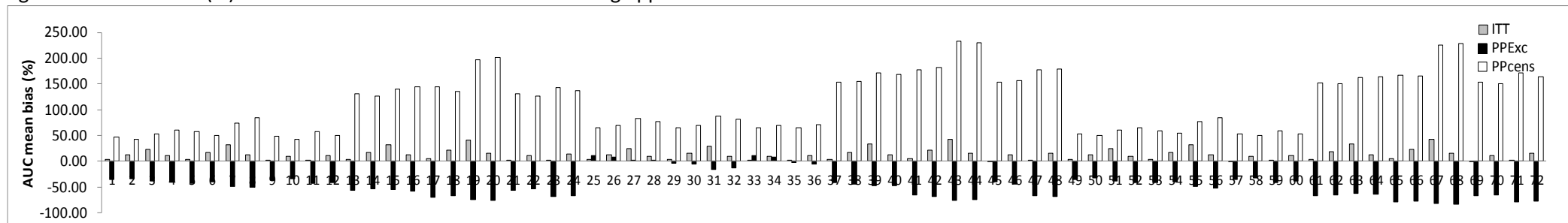
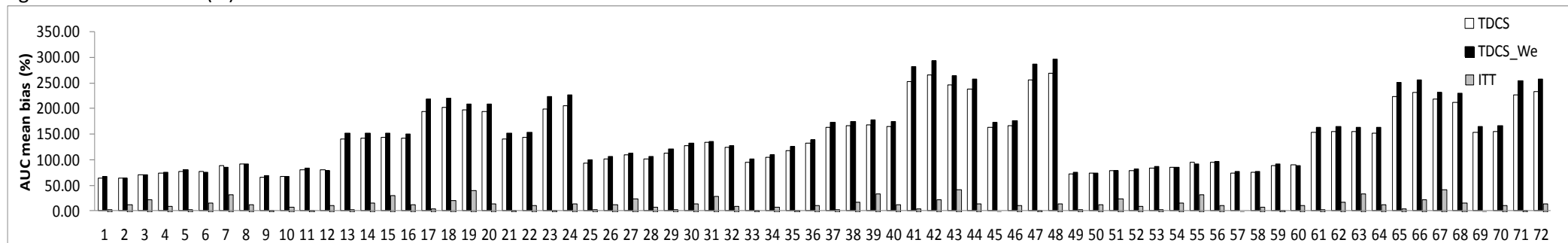


Figure 6.8: Mean bias (%) across scenarios – TDCS and TDCS-Weibull



The TDCM and TDCM-Weibull approaches also performed poorly, but the TDCM approach led to less bias than the ITT analysis in 6.9% of scenarios, and the TDCM-Weibull approach led to less bias than the ITT analysis in 22.2% of scenarios (illustrated by Figure 6.9). The TDCM approach gave a positive bias in all scenarios and produced more bias when the treatment effect was higher, but did not seem to be affected by the crossover proportion. The TDCM-Weibull approach led to negative bias (an underestimate of control group survival) in 41 of the 72 scenarios and its bias was very variable. In general it produced negative bias when crossover patients had relatively good prognosis and bias tended to be less when the treatment effect was relatively low.

As shown by Figures 6.10 and 6.11, the results for the four methods that included treatment crossover as a time-dependent indicator variable were similar to those of the time-dependent treatment variable methods. Comparing Figure 6.8 to Figure 6.10, and Figure 6.9 to Figure 6.11 it is evident that the patterns and magnitude of the bias for the corresponding methods (TDCM compared to XOTDCM, TDCS compared to XOTDCS, TDCM-Weibull compared to XOTDCM-Weibull, and TDCS-Weibull compared to XOTDCS-Weibull) were very similar.

- Inverse probability of censoring weights

The IPCW approach outperformed the ITT analysis in 60% of scenarios (see Figure 6.12). The approach was marginally more successful when antigen covariates were included in the model, but the difference was not substantial. In retrospect, this is not very surprising. The baseline antigen group covariate is directly related to the baseline prognosis group covariate (see equation [11]), and the “antigen group at disease progression” covariate is essentially a time-dependent measure of prognosis which is likely to be highly correlated with the “time to progression group” covariate. Hence, excluding antigen-related covariates from the IPCW estimation may be expected to have a relatively minor impact, providing baseline prognosis and “time to progression group” covariates are included. Importantly though, the IPCW method produced significantly less bias than the naive censoring approach across all scenarios. The IPCW method would reduce to the naive censoring approach if *all* confounders were unmeasured.

Figure 6.9: Mean bias (%) across scenarios – TDCM and TDCM-Weibull

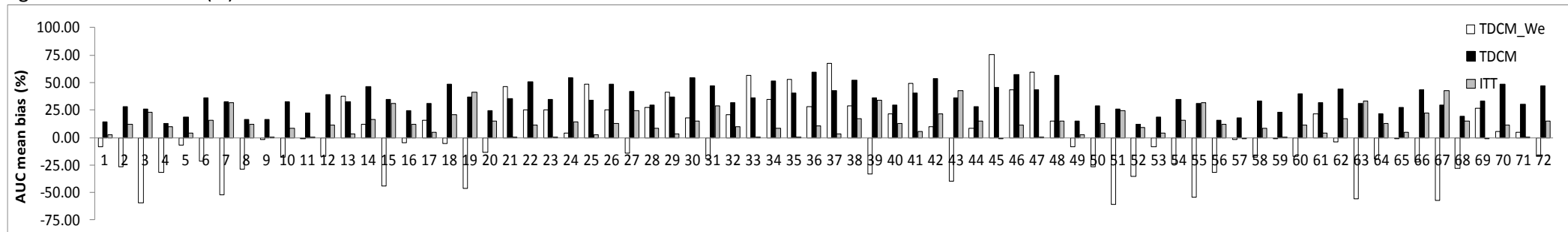


Figure 6.10: Mean bias (%) across scenarios – XOTDCS and XOTDCS-Weibull

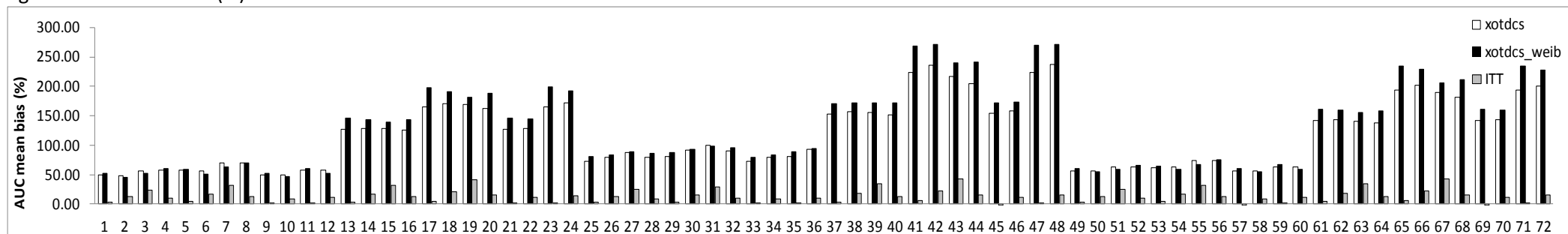
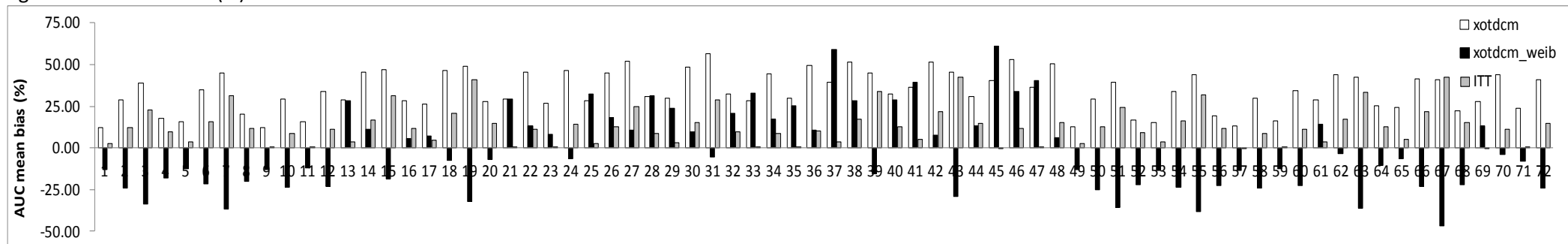


Figure 6.11: Mean bias (%) across scenarios – XOTDCM and XOTDCM-Weibull



In scenarios where the IPCW method worked relatively well it generally led to underestimates of survival in the control group (over-estimates of the treatment effect). In scenarios where the crossover proportion was very high the method produced much higher bias and generally led to substantial underestimates of the treatment effect. This was particularly evident in Scenarios 37 to 48, in which the crossover proportion was very high (approximately 95% of patients who survived more than 21 days).

Excluding Scenarios 37-48 the IPCW method did not seem to be affected by the complexity of the crossover mechanism – similar bias was observed when crossover depended only upon antigen level (Scenarios 25-72) compared to when crossover depended on both antigen level and time to progression (Scenarios 1-24).

Comparing the graph for IPCW bias (Figure 6.12) to Table 6.4, it is clear that the bias associated with the IPCW method is related to the crossover proportion. Very high crossover tends to cause the method to produce more bias. The relationship between the mean percentage bias associated with the IPCW method and the mean proportion of the at-risk population that crossed over is illustrated in Figure 6.15. It is clear that once the proportion of crossover patients increased to approximately 90% and beyond the method became susceptible to very high levels of bias. This is logical because very high crossover proportions leave very few “at-risk” patients who do not crossover (particularly in small datasets) – and it is these patients that form the basis of the IPCW pseudo population. However, Figure 6.15 also shows that under certain circumstances the method can still produce low bias, even with a sample size of 500 and a crossover proportion of 90% or greater.

In the “high-crossover” Scenarios 13-24, 65-67 and 70-72 the IPCW method produced relative mean bias that was generally less than 5%, and was substantially less than that produced in Scenarios 37-48. The first explanation for this is simple – comparatively less at-risk control group patients crossed over in these scenarios compared to the otherwise equivalent Scenarios 37-48 and therefore the IPCW “pseudo” population is based upon a higher number of non-crossover patients. In Scenarios 13-24 there was an average of 22-30 control group patients that did not cross over in each simulation, compared to 9-15 in Scenarios 37-48, and 14-23 in Scenarios 61-72. Hence it might be surmised that a minimum of approximately 20 non-crossover at-risk control group patients are required in order for the IPCW method to be able to work reliably.

However, there is an additional explanation as to why the IPCW method worked much better in Scenarios 13-24, 65-67 and 70-72 than in Scenarios 37-48. In Scenarios 1-24 and 49-72 patients with relatively good prognosis were more likely to cross over. However, as discussed above, the poorest prognosis patients who died before 21 days were not able to cross over – these made up 3-13% of control group patients, depending upon the severity of the simulated disease. Given the crossover mechanism, in scenarios where patients with good prognosis were more likely to cross over the 5-10% of patients who did not cross over were most likely to be those with expected survival times in the 3rd-35th or 13th-42nd percentiles of expected counterfactual control group survival times. In the equivalent scenarios in which poor prognosis patients were more likely to cross over (Scenarios 37-48) the 5-10% of patients who did not crossover were likely to be those with expected survival times in the 68th-100th or 71st-100th percentiles of expected counterfactual control group survival times, depending upon the severity of the simulated disease. These patients were also relatively likely to be administratively censored, thus having no observed survival time. Hence in scenarios where good prognosis patients were more likely to cross over, the weighted “pseudo” population was more likely to be based upon control group patients who had observed survival times that were closer to the average, relative to scenarios where patients with poor prognosis were more likely to cross over. Under these circumstances it appeared that the IPCW method could produce reasonably low bias, even when very high proportions of patients crossed over. On the other hand, basing the “pseudo” population on small numbers of good prognosis patients who may have had censored survival times led to very high levels of bias.

It is important to note that the bias of the IPCW method was not affected by the additional treatment effect decrement applied in Scenarios 9-12, 21-24, 33-36, 45-48, 57-60, 69-72. This is logical because patients are censored at crossover, and thus the treatment effect received by crossover patients is unimportant. This is a particular advantage of the IPCW approach when strong time-dependent treatment effects are likely.

The IPCW method generally produced less bias when the treatment effect was relatively small. In scenarios in which the method worked reasonably well, bias was lower when the treatment effect was lower (bias was lower in Scenarios 1, 4, 5, 8, 9, 25, 28, 29, 32, 33, 49, 52, 56, 57, 64, 65, 68 and 71 than in Scenarios 2, 3, 6, 7, 10, 26, 27, 30, 31, 34, 50, 51, 55, 58, 63, 66, 67 and 72 respectively).

Figure 6.12: Mean bias (%) across scenarios – IPCW

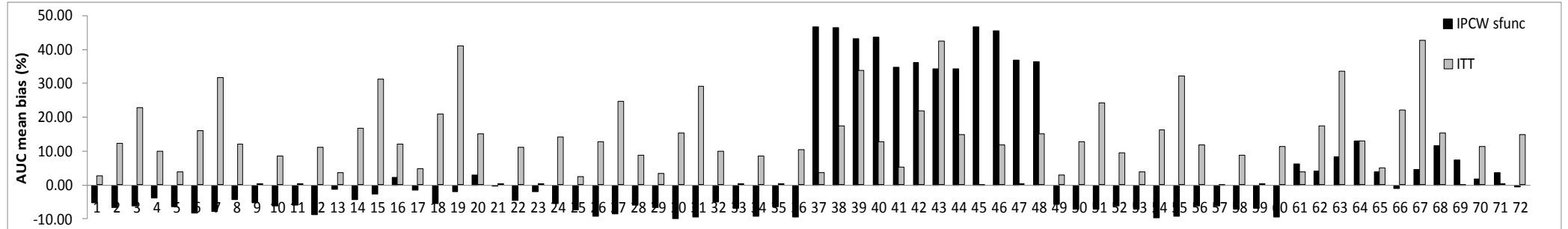


Figure 6.13: Mean bias (%) across scenarios – RPSFTM and IPE Weibull “survivor function” approaches (no covariates)

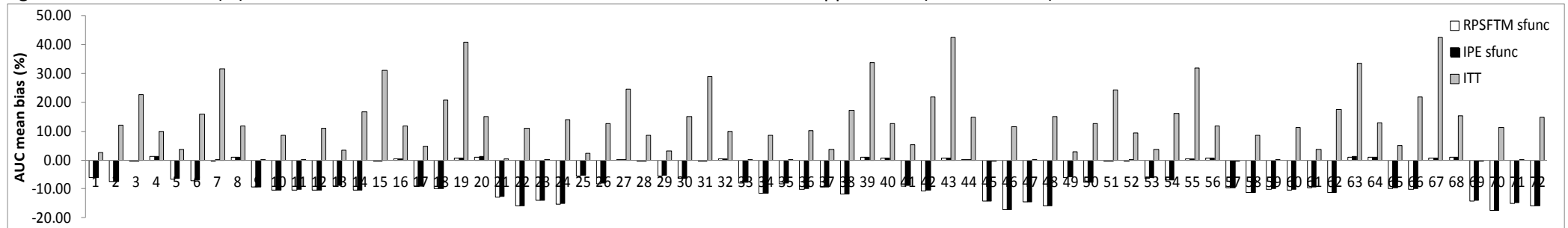


Figure 6.14: Mean bias (%) across scenarios – RPSFTM and IPE Weibull “extrapolation” approaches (no covariates)

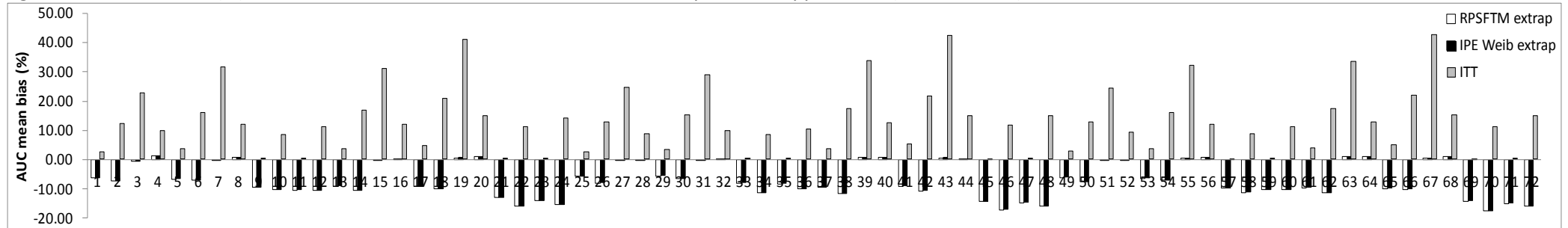
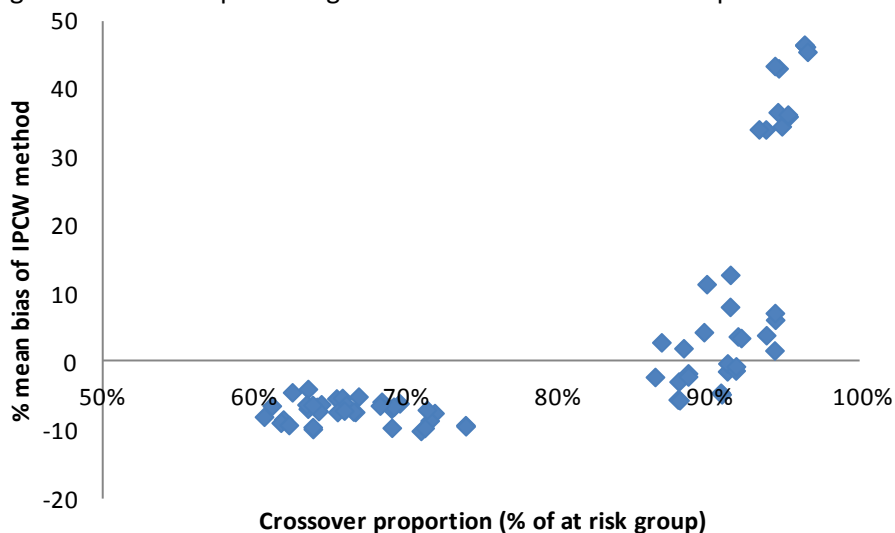


Figure 6.15: Mean percentage bias of the IPCW method compared to crossover proportion



The IPCW method tended to work better when the treatment effect was not time-dependent (bias was lower in Scenarios 3 and 4, 7 and 8, 27 and 28, 31 and 32, 51 and 52, 55 and 56 compared to corresponding Scenarios 1 and 2, 5 and 6, 25 and 26, 29 and 30, 49 and 50, 53 and 54 respectively). There were exceptions to this – bias was lower in Scenario 13 compared to 16; however, this was marginal and may be explained by the slightly higher treatment effect and censoring proportion in Scenario 16 (see Table 6.4). The IPCW method attempts to control for time-dependent confounders, and so if it worked as desired it would not produce more bias when the treatment effect is time-dependent – it should provide a good estimate of the “average” treatment effect, although it remains a proportional hazards-based approach. However, the method is data-intensive and was originally designed for use in observational datasets, which are typically much larger than RCT datasets. Therefore, it is perhaps unsurprising that it fails to work as desired in an RCT context, in which much less data are available, making accurate models more difficult to specify. As noted in Section 4.10.3.3 of Chapter 4, Howe *et al* (2011) demonstrated that the IPCW method was particularly prone to bias when sample sizes were small and selection bias was very strong, even when data were generated very simplistically.¹²⁸ The simulations presented here provide more evidence for this, and show that with more complex survival and crossover mechanisms bias becomes increasingly common with the IPCW method, particularly when covariates are almost deterministic (that is, almost all at-risk patients cross over).

It is difficult to determine whether severity of disease impacts the performance of the IPCW method. Bias was generally marginally lower when the severity of the disease simulated was less (severity was lower in Scenarios 1-4, 9-10, 13-16, 21-22, 25-28, 33-34, 37-40, 45-46, 49-52, 57-58, 61-64 and 69-70 compared to Scenarios 5-8, 11-12, 17-20, 23-24, 29-32, 35-36, 41-44,

47-48, 53-56, 59-60, 65-68 and 71-72 respectively). This may be associated with the “rare disease” assumption noted by Young *et al* (2010) – in practice the IPCW procedure uses a weighted logistic regression model for each time interval and thus it is required that event rates in each interval are low.¹²⁷ However, there were exceptions to this finding – bias was lower in Scenarios 29, 35, 65-66 and 71-72 compared to 25, 33, 61-62 and 69-70 respectively. The results indicate that the crossover proportion and treatment effect have more important effects on the bias associated with the IPCW method than disease severity. However, high proportions of administrative censoring amongst non-crossover control group patients may prevent the method from working well.

- Rank preserving structural failure time models and the Iterative parameter estimation algorithm

The RPSFTM and IPE methods produced very different levels of bias depending upon whether the treatment effect was time-dependent. Hence the results for these methods are analysed separately for these groups of scenarios.

- Scenarios that did not include a time-dependent treatment effect

In the scenarios with no time-dependent treatment effect (“zero TDC scenarios”) the “common treatment effect” assumption held. All variations of the RPSFTM and IPE methods produced less bias than the ITT analysis in all scenarios. This was not reliant on the crossover proportion, or the prognosis of crossover patients. Differences between the alternative versions of the RPSFTM and IPE methods were minor. The “survivor function” and “extrapolation” versions of the methods were most accurate in these scenarios (see Figures 6.13 and 6.14).

The RPSFTM and IPE methods worked marginally less well when the crossover proportion was high (Scenarios 15, 16, 19, 20, 39, 40, 43, 44, 63, 64, 67, 68 compared to 3, 4, 7, 8, 27, 28, 31, 32, 51, 52, 55, 56). In addition, marginally lower bias was produced when the treatment effect was relatively large (for example, Scenarios 3, 7, 15, 19, 63, 67 compared to Scenarios 4, 8, 16, 20, 64 and 68 respectively), but this was not exclusively the case. These methods did not appear to be affected by the complexity of the crossover mechanism, with bias levels remaining similar between Scenarios 1-24 and Scenarios 25-72.

Figure 6.16: Mean bias (%) across scenarios – RPSFTM and IPE “shrinkage” approaches (with covariates)

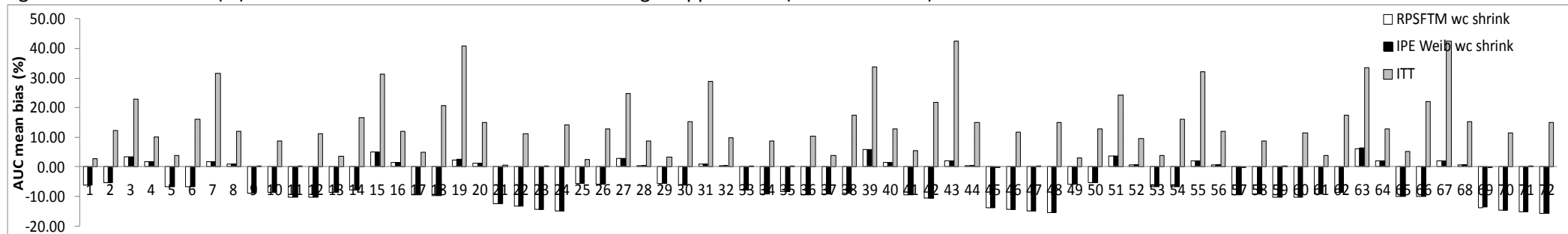


Figure 6.17: Mean bias (%) across scenarios – SNM with g-estimation

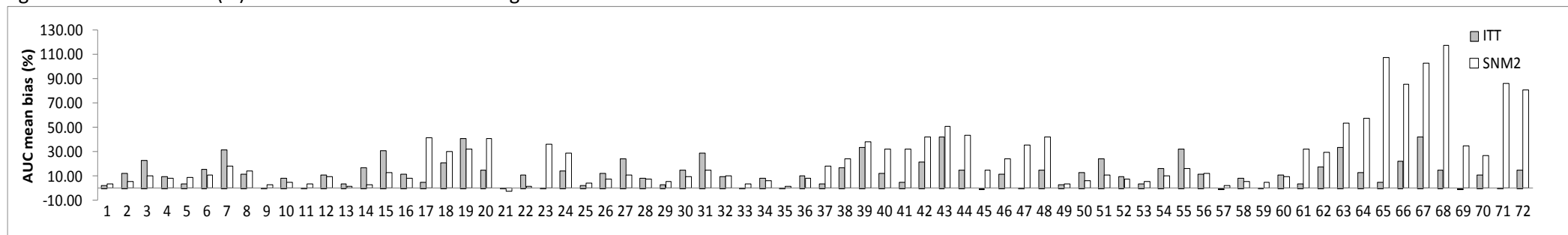
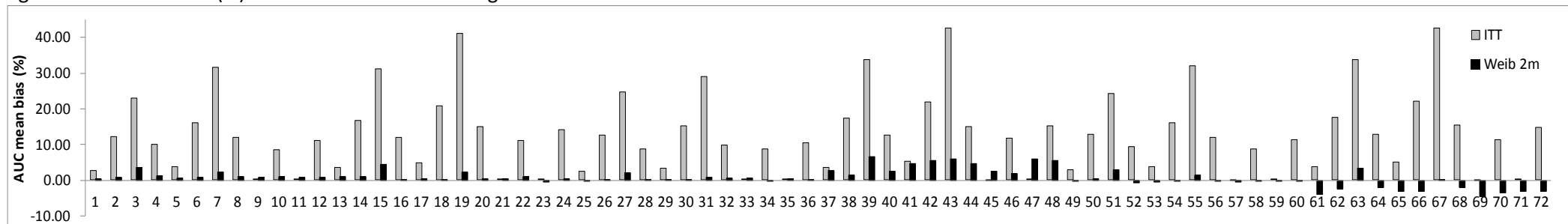


Figure 6.18: Mean bias (%) across scenarios – Two-stage Weibull



While there were important differences between the “shrinkage”, “extrapolation” and “survivor function” approaches applied after the RPSFTM and IPE methods to obtain restricted mean estimates, the differences between the variations of the RPSFTM and IPE methods themselves were less important, with levels of bias remaining similar. This is illustrated by Table 6.5, which reports the mean relative bias associated with each variation of the RPSFTM and IPE methods across the zero TDC scenarios.

Table 6.5: Bias of RPSFTM and IPE approaches in zero TDC scenarios

Method	Shrinkage		Extrapolation		Survivor Function	
	Mean bias (%)	Rank	Relative bias (%)	Rank	Relative bias (%)	Rank
RPSFTM	2.169	4	0.531	1	0.551	4
RPSFTMwc	2.086	2	1.472	5	0.473	1
IPE Weib	2.179	5	0.547	2	0.573	5
IPE Weib wc	2.010	1	1.446	4	0.489	2
IPE Exp	2.188	6	1.177	3	0.594	6
IPE Exp wc	2.103	3	2.196	6	0.503	3

Note: for the “shrinkage” approaches the bias associated with each method was always positive, and thus mean bias is presented in Table 6.5. However, for the “extrapolation” and “survivor function” approaches the bias was sometimes positive and sometimes negative and thus relative bias is presented. “wc” indicates “with covariates”

Table 6.5 demonstrates that the “extrapolation” and “survivor function” approaches generally produced less bias than the “shrinkage” approach in the zero TDC Scenarios. This is likely to be due to the failure of the “shrinkage” approach to perform full recensoring on the adjusted dataset. The “shrinkage” RPSFTM and IPE approaches all perform with similar levels of bias, with the inclusion of covariates in the analysis resulting in slightly lower levels of bias. The “extrapolation” versions of RPSFTM and IPE produced more varied levels of bias, with the RPSFTM and IPE Weibull (without covariates) approaches leading to substantially less bias than the other methods. The “extrapolation” IPE exponential method worked relatively poorly, which was expected because extrapolating with an exponential distribution was unlikely to be appropriate given the survival data generated (using a Weibull model). This demonstrates the importance of identifying appropriate parametric models for extrapolation purposes.

It is interesting that the extrapolation approaches that included covariates worked less well than those that excluded covariates. When covariates were included, they were used in the RPSFTM/IPE estimation procedure, and were also then used (based on mean values of the covariates in the control group) in the production of the survival curve based upon the final parametric model fit (for the IPE approaches), and in the fitting of a Weibull model to the counterfactual dataset (for the RPSFTM approach). This is not entirely surprising, because

using mean values for the covariates will generate bias if the distribution of covariate values are skewed.

The results of the “survivor function” approach offer further insight into this issue. First, it is clear that this approach generally led to relatively small bias in the scenarios in which the treatment effect was not time-dependent. In general, it appears to be a more successful approach than the “extrapolation” approach. However, close inspection reveals that in fact the RPSFTM and IPE Weibull (without covariates) “survivor function” approaches led to slightly higher bias than under the “extrapolation” approach. But, including covariates led to reductions in bias. Under the “survivor function” approach when covariates were included they were used in the RPSFTM/IPE estimation procedure, but were not used in the production of survival curves (a Weibull model with no covariates included was used to model survival in the experimental group, and the RPSFTM/IPE acceleration factor was used to obtain the survival curve for the control group). This suggests that including covariates in the treatment effect estimation procedure marginally improved accuracy.

The IPE exponential “survivor function” approach led to much less bias than the IPE exponential “extrapolation” approach. This was expected, because under the “survivor function” approach the exponential model was used within the IPE estimation procedure, but the resulting acceleration factor was then applied to a Weibull survival model fitted to the experimental group. If the acceleration factor had been applied to an exponential survival model fitted to the experimental group the results would probably have been much worse, as observed for the “extrapolation” approach. The IPE exponential methods were outperformed by the equivalent RPSFTM and IPE Weibull methods as expected (owing to the use of a Weibull in the data-generating mechanism), but this was only marginal. This demonstrates that using the exponential model within the IPE estimation procedure did not significantly reduce the success of the method, even though the underlying data did not resemble an exponential distribution. However, where extrapolation was undertaken it *was* important to select the most appropriate parametric model.

It is likely that the “extrapolation” approach generally produced more bias than the “survivor function” approach due to the loss of information associated with recensoring. Recensoring involves multiplying the potential censoring time (1095 days) by e^{ψ} . So, with an acceleration factor ($e^{-\psi_0}$) of approximately 2.17, equivalent to an e^{ψ} of 0.46 (1/AF) (as in the higher treatment effect scenarios in the simulations), patients are recensored at approximately 500 days rather than 1095 (1095×0.46). This is obviously a substantial difference, and leads to

substantially less data from which to extrapolate survival. Hence when the treatment effect is high (and baseline survival is relatively long – leading to more censored patients) recensoring may be particularly significant and could lead to relatively higher bias in the “extrapolation” RPSFTM and IPE approaches. In these circumstances it is likely to be preferable to take a “survivor function” based approach.

- Scenarios that include a time-dependent treatment effect

The RPSFTM or IPE methods were not expected to work as well when the treatment effect was time-dependent (“TDC scenarios”). In these scenarios, a decrement was applied to the treatment effect received by crossover patients, and thus the “common treatment effect” assumption did not hold. The impact of this is demonstrated in Figures 6.13, 6.14 and 6.16. All scenarios except 3-4, 7-8, 15-16, 19-20, 27-28, 31-32, 39-40, 43-44, 51-52, 55-56, 63-64, and 67-68 included a time-dependent treatment effect, and clearly the bias associated with the RPSFTM and IPE methods was much higher in these scenarios. The results also show, as expected, that when a larger treatment effect decrement was applied to crossover patients the RPSFTM and IPE methods produced even more bias (note the increased bias in Scenarios 9-12, 21-24, 33-36, 45-48, 57-60 and 69-72 compared to the corresponding Scenarios 1-2, 5-6, 13-14, 17-18, 25-26, 29-30, 37-38, 41-42, 49-50, 53-54, 61-62 and 65-66 respectively).

In these scenarios, the RPSFTM and IPE “extrapolation”, “survivor function” and “shrinkage” methods led to lower bias than the ITT analysis in 25-33% of scenarios. The scenarios where these approaches did not improve upon the ITT analysis were those in which a particularly large decrement in the treatment effect was applied to crossover patients (Scenarios 9-12, 21-24, 33-36, 45-48, 57-60, 69-72), and those in which the bias associated with the ITT analysis was very low due to the small treatment benefit received by crossover patients (Scenarios 1, 5, 13, 17, 25, 29, 37, 41, 49, 53, 61, 65). Hence, if the treatment effect received by crossover patients is expected to be around 25% less than the effect received by patients in the experimental group, RPSFTM or IPE methods are likely to be unsuitable, particularly if the treatment effect received is likely to be very small.

In these scenarios the RPSFTM and IPE methods always led to negative bias – that is, they over-adjusted for the treatment crossover effect. Initially, this may appear unexpected – it seems logical that when the treatment effect reduces over time, meaning that crossover patients received a reduced effect, the RPSFTM/IPE methods will underestimate the treatment effect actually received by the experimental group. This is due to the “common treatment effect” assumption, which means that the estimated treatment effect is an average of that

observed across both experimental group and crossover patients. However, the simulation study results show that actually the opposite occurs – the treatment effect obtained using the RPSFTM and IPE methods when crossover patients receive a reduced effect is actually an overestimate of the true treatment effect experienced by experimental group patients. This is likely to be due to the recensoring involved in the treatment effect estimation procedure of the RPSFTM and IPE methods. Recensoring involves basing the treatment effect estimation upon shorter-term data, and where the experimental group treatment effect decreases over time this may lead to an over-estimate of the true treatment effect, even if crossover patients receive a reduced effect. This appears to have been the case across all scenarios. Therefore, it appears that recensoring is a very important issue when a time-dependent treatment effect is suspected, particularly when the average treatment effect is reasonably large (these were equivalent to HRs of approximately 0.50-0.75 in my study). While this recensoring is relevant for both “extrapolation” and “survivor function” applications of the RPSFTM it is worthy of note that in situations where crossover patients receive a lower treatment effect than patients initially randomised to the experimental group, the “extrapolation” approach would always be expected to lead to an over-estimate of the survival advantage of the experimental treatment. Control group counterfactual survival times will be under-estimated because the acceleration factor used to derive counterfactual survival times will be higher than the acceleration factor actually received by crossover patients. Hence while the average treatment effect (in terms of an acceleration factor) estimated by the RPSFTM may be an under-estimate of the AF received by the experimental group, control group counterfactual survival times would be under-estimated. This would only not be the case if the experimental group AF was higher than the AF in crossover patients but the effect actually increased while on treatment, causing recensoring to lead to a serious underestimate of the average AF, which could result in the estimated average AF being lower than the AF received by crossover patients.

In the TDC Scenarios, the RPSFTM and IPE “extrapolation” and “survivor function” approaches led to increased biases when higher proportions crossed over, and when the treatment effect was higher. This is logical as the methods perform poorly in the TDC Scenarios and increased crossover proportions and treatment effects increase the scope for bias.

There was very little difference in the bias produced by the RPSFTM and IPE “survivor function” and “extrapolation” methods across the TDC Scenarios. This suggests that the bias associated with extrapolating from a recensored counterfactual dataset is similar to that associated with applying the estimated treatment effect to an experimental group survival curve that has not been recensored. It is likely that with larger treatment effects – when

recensoring involves a greater loss of information – “extrapolation” approaches will produce higher bias than “survivor function” approaches.

Across the majority of TDC Scenarios the RPSFTM and IPE “shrinkage” approaches led to lower bias than the “extrapolation” and “survivor function” approaches. When the methods estimate a very biased treatment effect this is to be expected. The “shrinkage” approach only applies this treatment effect to crossover patients in order to obtain an estimate of the counterfactual dataset and the corresponding restricted mean, whereas “extrapolation” and “survivor function” approaches apply the biased treatment effect to the entire control group. However the “shrinkage” approach is clearly flawed as a highly biased treatment effect is applied in order to obtain the counterfactual dataset. In fact, if a further analysis of the treatment effect was undertaken on the adjusted dataset the result would be different to that originally obtained using the whole dataset, which clearly demonstrates that the initial estimate of the treatment effect is biased. Therefore this method is difficult to recommend.

- Structural nested models and g-estimation

The SNM with g-estimation approach performed better than the ITT analysis in 38% of scenarios (31% when there was a time-dependent treatment effect and 50% when there was not), and generally performed extremely poorly when the proportion of crossover was very high, and when good prognosis patients were more likely to crossover (see Figure 6.17). The method worked marginally better when the bad prognosis covariate was not included in the model. This may suggest that the method was struggling to cope with the complexity of the dataset (reflected by generally poor performance), or that provided either the bad prognosis covariate or antigen covariates are included, excluding one or the other has little impact on the performance of the method. The latter was also suggested by the results of the IPCW method, discussed above. In several scenarios the method failed to converge in a high proportion of simulations. This occurred particularly when the disease severity was high (for example convergence was 85% in Scenario 8 compared to 100% for Scenario 4, and was 33% in Scenario 5 compared to 99% in Scenario 1), and this was exacerbated when this was combined with a high crossover proportion and a low treatment effect (in such scenarios convergence ranged between 10% and 28%).

The SNM method is similar to the IPCW method in that it is an observational-based approach and so it is heavily reliant on the proportion of control group patients that cross over. If very few control group patients do not cross over (or if very few patients do cross over) the method will be starved of data with which to estimate counterfactual survival times and is likely to

result in biased estimates – this is a particular problem in RCTs, as datasets are typically much smaller than in observational studies. This is clearly shown by Figure 6.17, where bias is generally highest in the high crossover scenarios (Scenarios 13-24, 37-48 and 61-72). However this was not exclusively the case, and the results suggest that the success of the method was reliant on the crossover mechanism – in a way that is different from that observed for the IPCW method. The IPCW method appeared to work better when good prognosis patients were more likely to cross over, whereas the SNM method produced more bias when good prognosis patients were more likely to cross over (Scenarios 61-72).

It is not particularly surprising that the IPCW and SNM methods performed differently across the scenarios. While it was to be expected that neither would perform well with very high crossover proportions, the methods work in very different ways and so could be expected to perform differently across the scenarios. Applied to an RCT, the SNM method involves a two-stage estimation procedure, in which the treatment effect is first estimated for crossover patients, and then is estimated for the experimental group based upon an adjusted dataset. On the other hand, the IPCW method censors crossover patients and the treatment effect they receive plays no part in the estimation procedure. The SNM method requires that a reasonable number of control group patients do not cross over in order for the treatment effect in crossover patients to be estimated accurately. When good prognosis patients are more likely to cross over, crossover patients will more often have administratively censored survival times, making it difficult to obtain reliable estimates of the crossover treatment effect. This may explain why the SNM method produced increased bias when good prognosis patients were more likely to cross over, whereas the opposite was true for the IPCW method.

Occasionally the SNM method produced low bias even when the crossover proportion was very high (for example, Scenarios 13, 14, 21, 22). This is likely to be due to the relatively long control group survival times in these scenarios (see Table 6.4), the crossover mechanism, and the fact that the at-risk crossover proportion was lower in these scenarios compared to equivalent high crossover scenarios (Scenarios 37, 38, 45, 46 and 61, 62, 69, 70). However, these results should be interpreted with care, since the SNM method failed to converge in 4-20% of simulations in these scenarios: in general it is reasonable to conclude that the SNM method is unlikely to be appropriate when crossover proportions are very high.

The SNM method converged in over 90% of simulations in 35 (49%) of the 72 scenarios (Scenarios 1-4, 7, 9-10, 14-16, 22, 25-28, 31, 33-34, 37-40, 43, 45-46, 49-52, 55, 57-58, 62-64). These scenarios tended to simulate a low disease severity and a higher treatment effect. The

method generally performed marginally better in these scenarios, out-performing the ITT analysis in 54% of scenarios – however often substantial bias remained. The method led to positive bias – that is it over-estimated survival in the control group and hence underestimated the treatment effect – in all but one scenario (Scenario 21, where crossover patients received a very small treatment benefit).

The fact that the SNM method seems reliant on disease severity may be due to the `stgest` STATA command.¹²⁶ The command has a “round” option which allows the treatment effect to be estimated allowing counterfactual survival times to be rounded to the nearest unit specified by “round”. In the simulations, a value of “1” was used, which is reasonable given that a day time-scale was used – if a yearly time-scale had been used rounding to the nearest year would not be appropriate. However, often the method failed to converge because estimates of the treatment effect did not allow all counterfactual survival times to be within one day of the value required in order for the method to converge. This appeared to occur more often when survival times were shorter.

- Two-stage Weibull approach

The two-stage Weibull approach consistently performed well and produced less bias than the ITT analysis in 82% of scenarios. Levels of bias only increased significantly in Scenarios 37-48 and 61-72 where crossover proportions were very high (see Figure 6.18). This is to be expected as in these scenarios it is much more difficult to estimate the treatment effect in crossover patients because almost all patients cross over. The scenarios in which the two-stage Weibull method produced higher bias than the ITT analysis were those in which the treatment benefit received by crossover patients was very low (Scenarios 9, 11, 21, 23, 33, 35, 45, 47, 57, 59, 69, 71). In these scenarios the bias associated with the ITT analysis was less than 0.4% and so there was very little scope for alternative approaches to reduce the bias. There was one exception to this – in Scenario 61 the two-stage Weibull approach produced bias of -3.86%, compared to 3.76% associated with the ITT analysis. In this scenario the difference in bias between the two methods was marginal, and the relatively poor performance of the two-stage Weibull method can be explained by the very high proportion of crossover simulated in this scenario, combined with the low ITT bias.

In general, the two-stage Weibull approach led to positive bias when the treatment effect was large, and when poor prognosis patients were more likely to cross over. The method was prone to negative bias when good prognosis patients were more likely to cross over.

6.6.3 Comparative bias of alternative methods across scenarios

In this section the analyses described in Section 6.6.2 are drawn upon to compare alternative methodologies in groups of scenarios. Scenarios are grouped according to whether or not a time-dependent treatment effect was simulated (thus, whether or not the “common treatment effect” assumption held), and if so, the strength of that effect.

Naive methods (censoring, exclusion, treatment as a time-dependent covariate, crossover as a time-dependent indicator variable) are not considered in this section, as Section 6.6.3 demonstrated that these methods were associated with high bias and rarely reduced bias compared to the ITT analysis. In addition, the IPCW and SNM approaches that exclude antigen and bad prognosis covariates are not considered. For simplicity only the RPSFTM and IPE Weibull “survivor function” approaches with covariates are considered, rather than all of the different versions of these methods. RPSFTM and IPE Weibull “shrinkage” approaches (with covariates) are considered when scenarios in which there was a time-dependent treatment effect are addressed.

Relatively little attention is paid to the two-stage Weibull approach in this section, despite it appearing to be the optimal method in a large proportion of scenarios. This is because, in part, it is to be expected that this method works well due to the way in which the simulated data were generated. However, in circumstances where crossover happens as assumed in the simulated scenarios (that is, only after disease progression, and close to the time of progression, and not all patients cross over) this type of method could be useful and should be considered. Key considerations in using this method are:

- If all patients cross over the method is not possible
- If almost all patients cross over the method is likely to be less accurate (as reflected by higher biases in high crossover scenarios)
- The method requires the possibility that data can be split at a specific disease-related time such as disease progression, as this forms a secondary “baseline” at which the values of time-dependent confounders can be taken into account. This would not be possible if crossover could happen before progression.
- If crossover could occur at any time-point after disease progression, the method would be more susceptible to bias due to not appropriately adjusting for time-dependent covariates that impact upon the risk of death.

6.6.3.1 Zero time-dependent treatment effect

Figure 6.19 demonstrates the bias associated with the selected methods as well as the bias associated with the ITT analysis in scenarios that did not include a time-dependent treatment effect (“zero TDC scenarios”). Figure 6.20 presents this data on a truncated axis so that the methods can be compared more accurately (note that bars that have been truncated have arrows indicating this). In these scenarios the “common treatment effect” assumption was satisfied. The results suggest that in scenarios in which the treatment effect is not expected to be time-dependent, RPSFTM and IPE methods are likely to be optimal. Section 6.6.2 suggested that when an estimate of mean survival is required “survivor function” techniques are likely to be preferable to “shrinkage” or “extrapolation” approaches. There is very little difference in the bias associated with RPSFTM and IPE methods, and either are suitable in these scenarios. However, given the parametric nature of the IPE approach it may be more reliant on the survival times representing a suitable Weibull (or exponential) distribution.

Figure 6.21 presents the mean bias associated with the selected methods plotted against mean crossover proportions exclusively for scenarios in which there was not a time-dependent treatment effect. It is clear from this graph and Figures 6.19 and 6.20 that the levels of bias associated with the RPSFTM and IPE “survivor function” approaches were very low in comparison to other methods in these scenarios. These increased marginally when crossover proportions were high, but compared to the IPCW and SNM methods bias remained extremely low. When the crossover proportion was relatively low the IPCW method offered a reasonable alternative to the RPSFTM and IPE methods and generally performed slightly better than the SNM method. However, bias remained substantially higher than for the RPSFTM and IPE methods, and treatment crossover was consistently over-adjusted for (the treatment effect was over-estimated). Both observational-based approaches (IPCW and SNM) became much more prone to bias when the crossover proportion was very high, with the SNM method more sensitive to this than the IPCW method.

Figure 6.19: Bias across zero TDC scenarios – selected methods

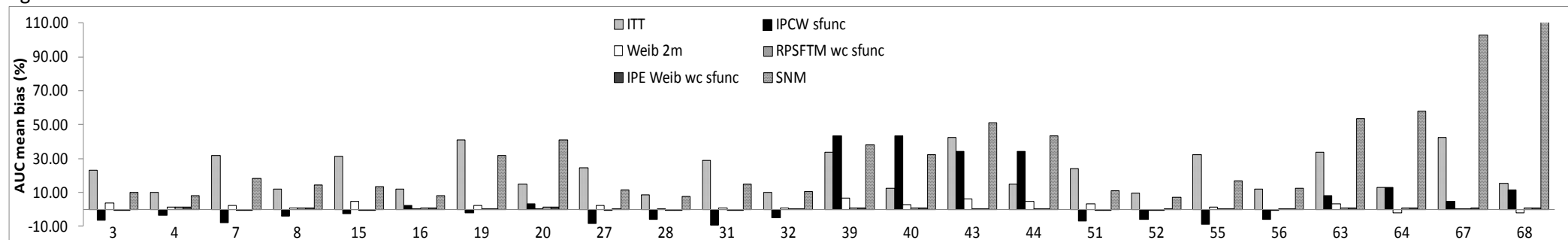


Figure 6.20: Bias across zero TDC scenarios – selected methods with truncated axis

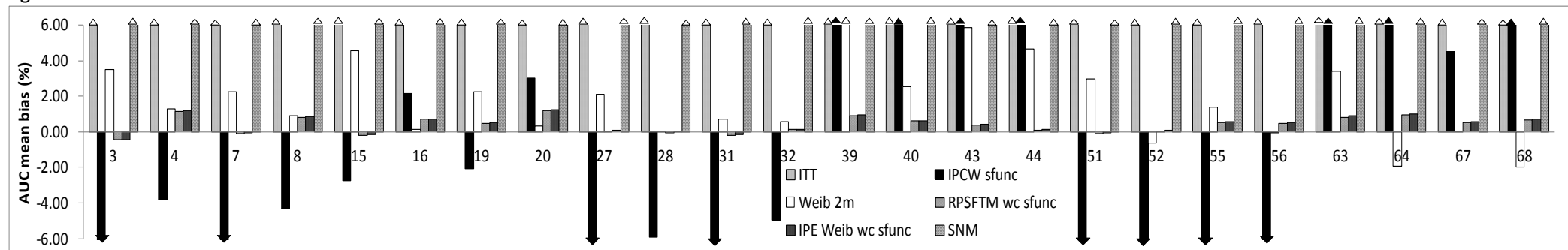
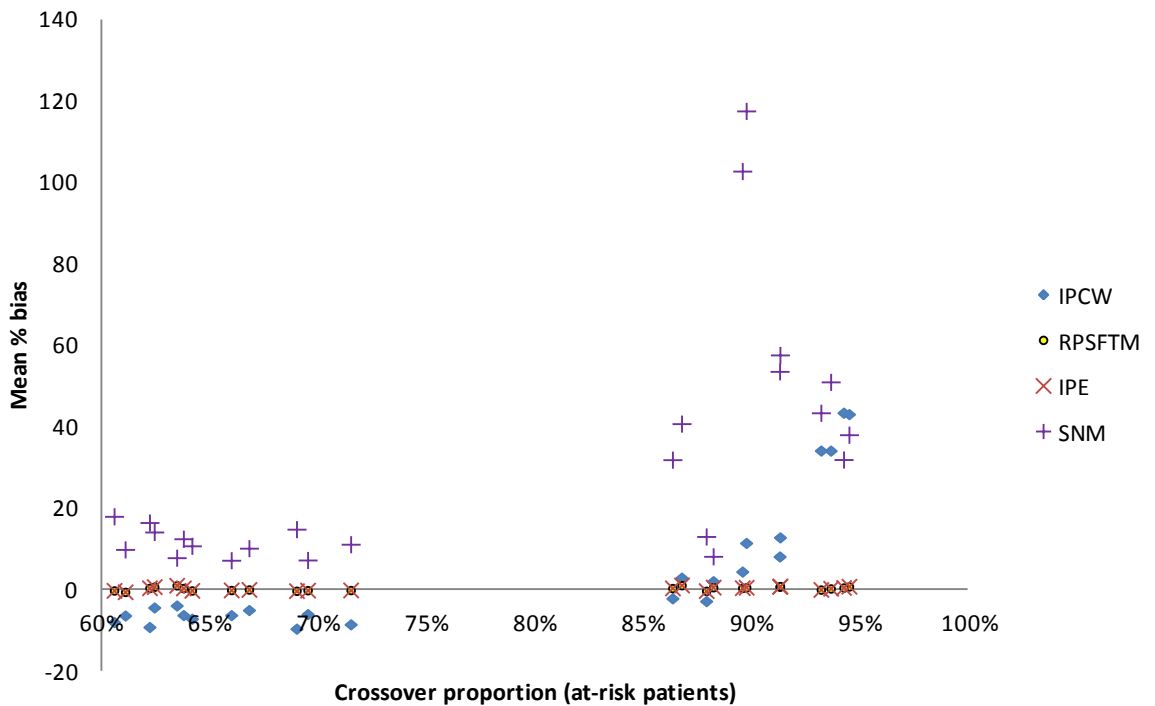


Figure 6.21: Mean bias compared to crossover proportion – zero TDC scenarios



There was not a consistent relationship between bias and the size of the treatment effect for the IPCW method. The RPSFTM and IPE methods and to a lesser extent the SNM method consistently produced marginally higher biases when the treatment effect was lower. However, this has no impact on the optimal methods in scenarios with no time-dependent treatment effect – these remain the RPSFTM and IPE methods. Similarly, disease severity (and therefore the proportion of censoring) did not seriously impact upon the selected methods in these scenarios – the RPSFTM and IPE methods remain optimal. However only censoring ranging from 1% to 21% was tested - higher levels may have led to different results.

6.6.3.2 Time-dependent treatment effect (without additional decrement)

Figure 6.22 illustrates the bias associated with the selected methods and the ITT analysis in scenarios in which there was a time-dependent treatment effect (“TDC scenarios”). In these scenarios the treatment effect received by crossover patients was approximately 15% lower than that received by patients in the experimental group. Scenarios in which an additional decrement was applied to crossover patients (such that the treatment effect they received was approximately 25% lower than that received by patients in the experimental group) are considered in Section 6.6.3.3. Five important observations are immediately apparent:

- I. The selected methods did not always produce less bias than the ITT analysis:
 - IPCW produced less bias than the ITT analysis in 54% of scenarios.

- SNM produced less bias than the ITT analysis in 33% of scenarios.
 - Two-stage Weibull produced less bias than the ITT analysis in 96% of scenarios.
 - RPSFTM “survivor function” produced less bias than the ITT analysis in 50% of scenarios.
 - IPE Weibull “survivor function” produced less bias than the ITT analysis in 50% of scenarios.
 - RPSFTM “shrinkage” produced less bias than the ITT analysis in 50% of scenarios.
 - IPE Weibull “shrinkage” produced less bias than the ITT analysis in 50% of scenarios.
- II. The ITT analysis always overestimated control group survival (underestimated the treatment effect), whereas the selected methods (apart from the SNM) generally underestimated control group survival (overestimated the treatment effect):
 - III. The IPCW method consistently underestimated control group survival except when the treatment crossover proportion was high.
 - IV. The RPSFTM and IPE methods consistently underestimated control group survival across all scenarios.
 - V. The SNM method consistently overestimated control group survival across all scenarios.

Based upon these results it is clear that the two-stage Weibull method was optimal when the treatment effect varied over time. However, as discussed in Section 6.6.3, this result is likely to be driven, at least in part, by the simulation study design. Excluding this method it is more difficult to identify an “optimal” approach. Below the selected methods are compared in more detail, further grouping scenarios by key variables included in the simulation study – crossover proportion, size of the treatment effect, and disease severity.

- High crossover vs low crossover

Figure 6.24 illustrates the mean bias associated with the selected methods plotted against crossover proportion specifically for the scenarios that included a treatment effect decrement of approximately 15% in crossover patients.

Figure 6.22: Bias across TDC scenarios – selected methods

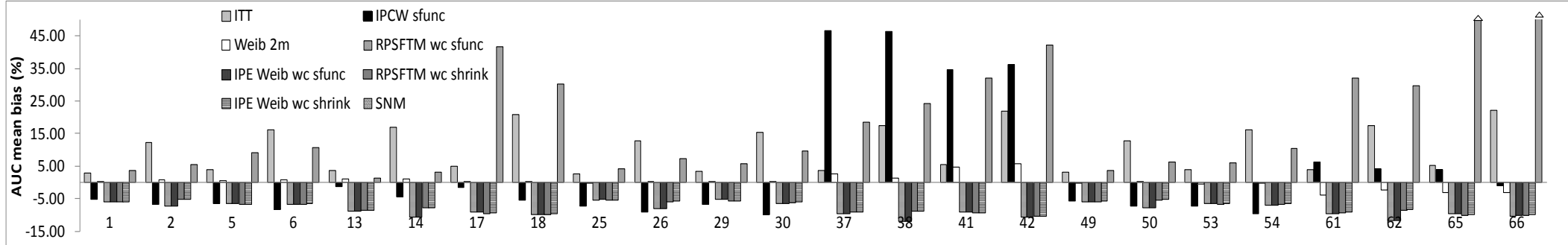


Figure 6.23: Bias across additional TDC scenarios – selected methods

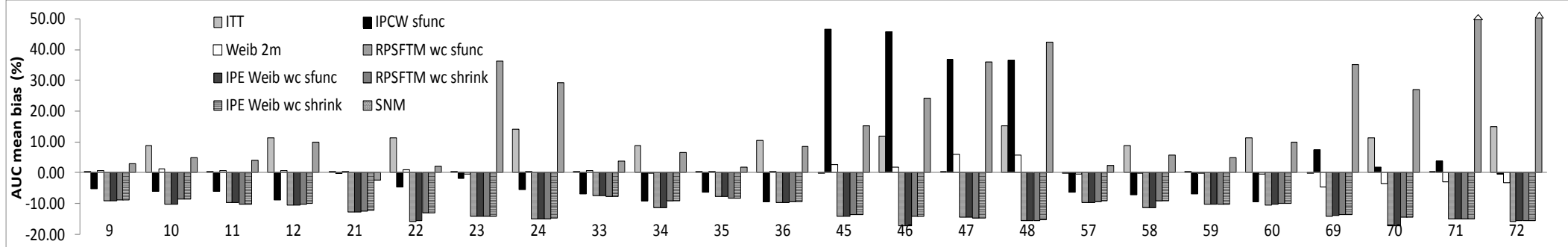
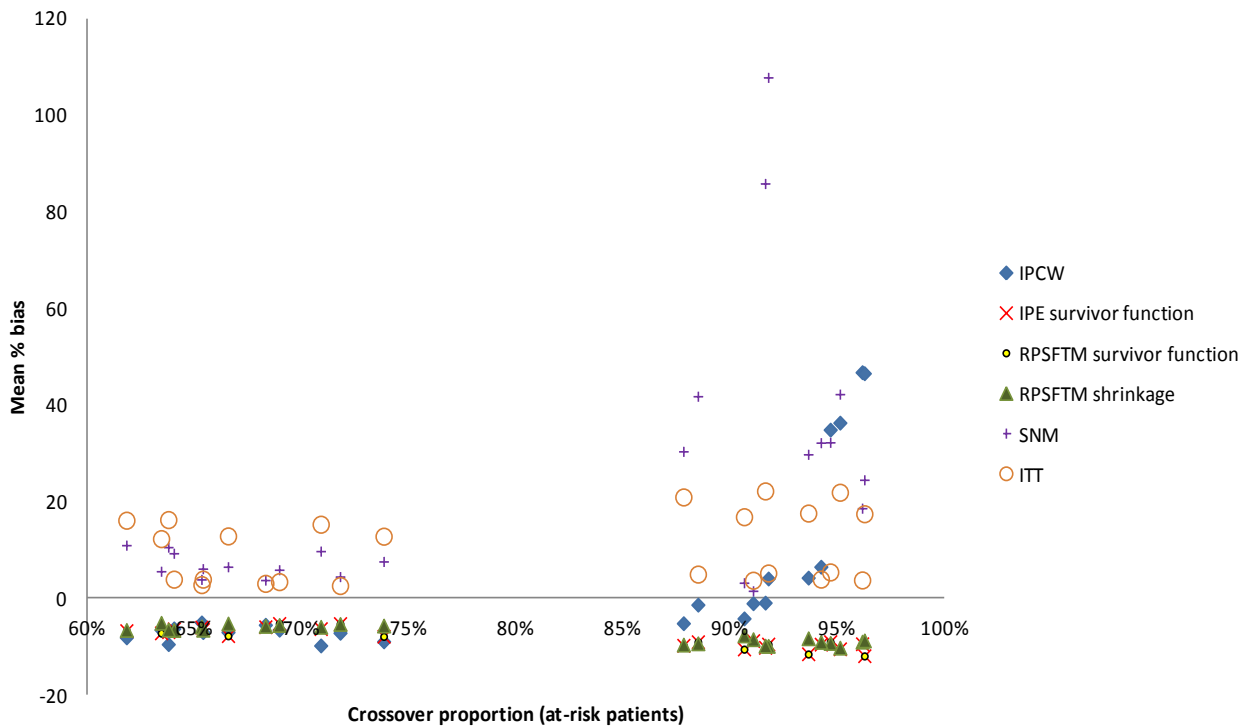


Figure 6.24: Mean bias compared to crossover proportion – TDC scenarios



It is clear from Figure 6.24 that all methods produced relatively high levels of bias in these scenarios, and these increased with the crossover proportion. The SNM method was most affected by the increase in crossover proportion, followed by the IPCW method – although as explained in Section 6.6.3, the IPCW method still produced low levels of bias in scenarios where non-crossover control group patients had an “average” prognosis. In reality though, this might be rare, or difficult to ascertain. The results suggest that in RCTs with a total trial sample size of approximately 500 patients, in which between 60% and 75% of control group patients cross over and receive a treatment effect approximately 15% lower than that received by patients in the experimental group, the IPCW method is likely to produce similar levels of bias compared to RPSFTM or IPE methods – provided the “no unmeasured confounders” assumption holds. However, significant bias of around 10% is likely to remain and an ITT analysis may provide less bias if the treatment effect received by crossover patients is low. The SNM method represents a relatively volatile alternative, and is likely to underestimate the treatment effect. The bias associated with the observational methods was quite stable between crossover proportions of 60 and 75%, suggesting that similar levels of bias might be produced in circumstances where crossover in at-risk patients is less than 60%. However, this was not tested in the simulation study.

Whilst the bias associated with the SNM and IPCW methods increased sharply at crossover proportions of greater than 90%, the increase in relative bias associated with the RPSFTM and IPE methods was modest. Hence these methods are likely to outperform the IPCW method at very high crossover proportions, even if crossover patients receive a treatment effect that is 15% lower than that received by patients in the experimental group. However, again an ITT analysis may produce least bias.

It is important to note that it is the *number* of control group patients who become at-risk of crossover that do not switch treatments that is of most importance, rather than the *proportion*. In the simulated scenarios presented here, 10% of the at-risk control group population was equivalent to approximately 21 patients, and when the IPCW pseudo population was based on numbers lower than this the method generally produced high levels of bias, and the SNM method became very unreliable. The RPSFTM and IPE “shrinkage” approaches generally produced marginally less bias than the other methods across all crossover proportions – apart from when the at-risk patients who did not cross over were of “average” prognosis (when the IPCW method produced least bias). However, the theoretical problems associated with the “shrinkage” approach means that it is difficult to recommend as “optimal”.

- High treatment effect vs low treatment effect

Scenarios 2, 6, 14, 18, 26, 30, 38, 42, 50, 54, 62 and 66 simulated a higher treatment effect than Scenarios 1, 5, 13, 17, 25, 29, 37, 41, 49, 53, 61 and 65. In general, the IPCW approach produced more bias when the treatment effect was higher, although this was not exclusively the case. The RPSFTM and IPE “survivor function” approaches always produced more bias when the treatment effect increased, but this was not always the case for the RPSFTM and IPE “shrinkage” approaches.

Changes in the treatment effect did not alter the ranking of the selected methods with respect to bias. However, the simulations demonstrated that the bias reduction associated with RPSFTM/IPE “shrinkage” approaches compared to “survivor function” approaches reduced with lower treatment effects.

- High severity vs low severity

Scenarios 5, 6, 17, 18, 29, 30, 41, 42, 53, 54, 65 and 66 simulated a higher disease severity than Scenarios 1, 2, 13, 14, 25, 26, 37, 38, 49, 50, 61 and 62, and thus had lower levels of censoring. Changes in disease severity (hence censoring levels) did little to alter the ranking of the

selected methods with respect to bias, and only marginally affected the absolute differences in relative bias between the alternative methods. However it is worthy of note that the IPCW method is likely to be adversely affected if poor prognosis patients are more likely to cross over and disease severity is low, as this is likely to lead to an increased proportion of censoring in control group non-crossover patients. Also, of all the methods the SNM approach was most affected by disease severity – it failed to converge more often and produced more bias when disease severity was high.

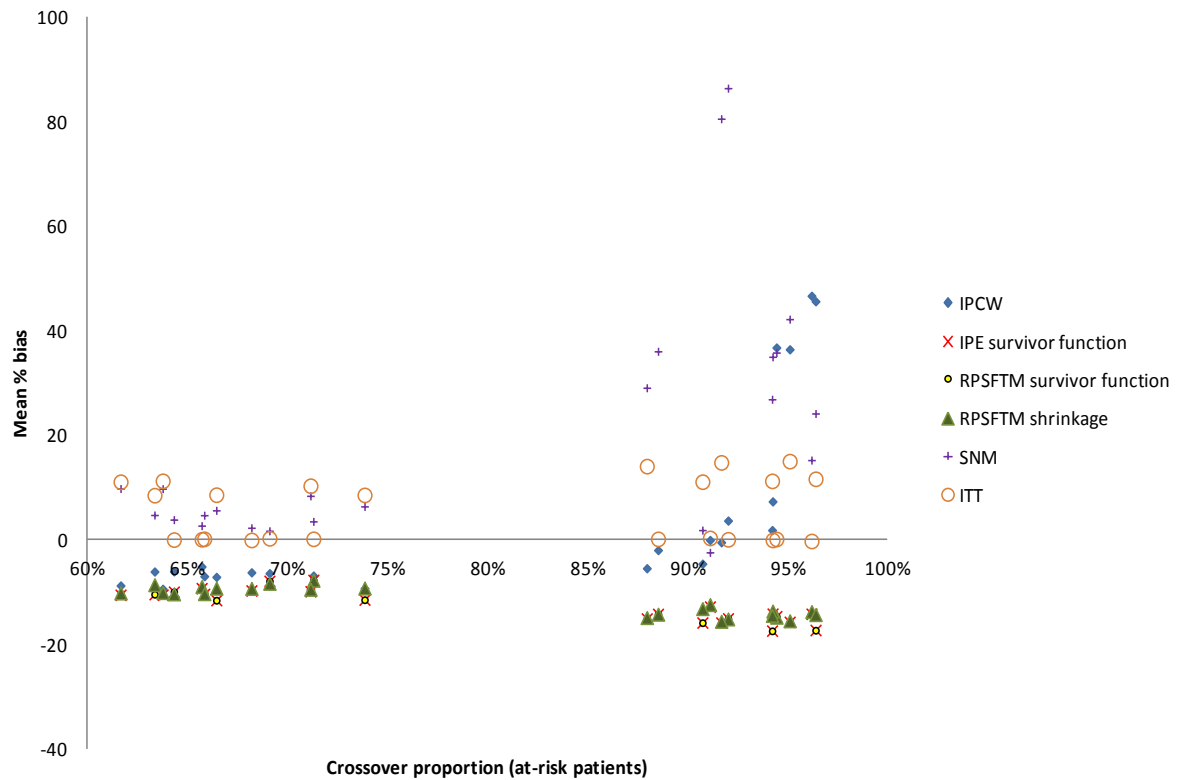
6.6.3.3 Time-dependent treatment effect (with additional decrement)

Figure 6.23 illustrates the bias associated with the selected methods in the 24 scenarios in which a treatment effect decrement of approximately 25% was applied to crossover patients. By comparing Figure 6.22 and Figure 6.23 it can be seen that the pattern of bias associated with the alternative methods was very similar in these scenarios. However, the key difference is that the levels of bias associated with the RPSFTM/IPE “survivor function” and “shrinkage” methods increased and it is more common for these methods not to produce less bias than the ITT analysis. In these scenarios the RPSFTM/IPE methods only led to reduced bias compared to the ITT analysis in 12.5% of scenarios. This is partly due to the increased levels of bias associated with these approaches, but also partly due to the reduced bias associated with the ITT analysis. This occurs because crossover patients received a reduced treatment effect. Owing to this, the RPSFTM/IPE approaches only produced less bias than the ITT analysis when the treatment effect was at its highest – meaning that crossover patients still received a substantial treatment benefit (Scenarios 12, 36, 60).

The IPCW approach was not affected by the additional decrement in the treatment effect received by crossover patients because these patients are censored in an IPCW analysis. However, due to the reduced ITT analysis bias the IPCW method only produced lower bias than the ITT analysis in 42% of scenarios. The SNM approach was also relatively unaffected by the additional decrement in the treatment effect received by crossover patients due to its two-stage procedure – it makes no assumptions about the treatment effect in crossover patients and so performs similarly whatever treatment effect they receive. However, again due to the reduced ITT analysis bias the SNM approach only produced less bias in 29% of scenarios. Similar is true for the two-stage Weibull approach, which led to reduced bias compared to the ITT analysis in 50% of these scenarios.

Figure 6.25 shows the relationship between crossover proportion in at-risk patients and mean relative bias specifically for the scenarios that included a treatment effect decrement of approximately 25% in crossover patients.

Figure 6.25: Mean bias compared to crossover proportion – scenarios with an additional time-dependent treatment effect



In these scenarios the IPCW method produced marginally lower bias than the RPSFTM/IPE methods, providing crossover levels were lower than 90-95%. The SNM approach offered a reasonable – but generally more volatile – alternative when crossover proportions were between 60% and 75%, particularly when disease severity was low. This suggests that when the treatment effect received by crossover patients is expected to be approximately 25% less than the treatment effect received by experimental group patients, the observational methods (IPCW and SNM) are likely to produce less bias than any of the RPSFTM/IPE variants - provided the “no unmeasured confounders” assumption holds, and less than 90% of control group patients cross over (amounting to approximately 20 control group patients *not* crossing over). However, again, the ITT analysis may provide least bias when the treatment effect received by crossover patients is low.

6.6.4 Comparison of methods – Summary

When there is no time-dependent covariate that causes a reduction in the treatment effect over time, therefore meaning that the “common treatment effect” assumption holds, my analyses definitively show that RPSFTM or IPE “survivor function” approaches produce least bias compared to the range of alternative methods. The levels of bias associated with these methods alter with different data characteristics, but they remain optimal.

The results are much more variable when the treatment effect is time-dependent. When the treatment effect received by crossover patients is expected to be approximately 15% less than that received by patients in the experimental group the RPSFTM/IPE approaches and the IPCW approach are likely to produce similar levels of bias, provided the “no unmeasured confounders” assumption holds and less than 90% of at-risk control group patients cross over (leaving approximately 20 control group patients who do *not* cross over). A two-stage SNM method is also likely to produce similar levels of bias in these circumstances (providing disease severity is low), but is more volatile - probably due to the fact that it failed to converge in a significant number of simulations.

When the treatment effect received by crossover patients is expected to be approximately 25% less than that received by patients in the experimental group the IPCW and SNM methods are likely to produce less bias than all RPSFTM/IPE variants, provided the “no unmeasured confounders” assumption holds and less than 90% of at-risk control group patients cross over (leaving approximately 20 control group patients who do *not* cross over). However, in these circumstances the ITT analysis may often produce the least bias if the treatment effect is relatively small. It is likely that in the presence of such a strong time-dependent treatment effect the ITT analysis would be even more likely to produce least bias if the crossover proportion was low. However, I did not test any scenarios with at-risk crossover proportions of less than 60%. If the crossover mechanism is such that a simple two-stage Weibull approach can be applied this is likely to produce the least bias, provided that all time-dependent confounders are measured at the time at which crossover becomes possible (that is, the secondary “baseline”), and that crossover cannot happen substantially after the secondary “baseline”. If it were possible to apply a two-stage approach and crossover could occur at any point after the secondary “baseline” it may be hypothesised that a two-stage SNM may outperform a simple two-stage Weibull, but this was not tested in any scenarios and given the volatility in the SNM results, this is by no means certain.

6.6.5 Coverage of methods

So far the focus of this chapter has been on the bias associated with each method across the range of simulated scenarios. This is because the main aim of the simulation study was to determine which method could be expected to produce least bias. However, in economic analyses uncertainty is important and may impact upon recommendations made by decision-makers. In economic evaluations of treatments for metastatic cancer, survival inputs are likely to be very important in the economic model and therefore it is critical to adequately characterise the uncertainty around these. Hence, it is useful to assess the coverage associated with the crossover adjustment methods – although it is important first to note that generally a poor coverage is expected if a method produces significant bias. Figures 6.26, 6.27 and 6.28 show the coverage of the selected methods across groups of scenarios defined by whether or not the treatment effect was time-dependent.

Figure 6.26 illustrates that the coverage associated with the RPSFTM “survivor function” approach was very good in the “zero TDC Scenarios”, at slightly over 95%. The coverage of the IPE Weibull “survivor function” approach was less good, often between 80% and 90%. This was expected as the confidence intervals around the final iteration of the IPE algorithm were used to generate restricted mean confidence intervals. These represent an underestimate of the true confidence interval, as described in Section 6.4.2.10. However, Morden *et al* (2011) showed that if a bootstrapping approach to estimate confidence intervals for the treatment effect is taken, coverage is satisfactory with the IPE method (in zero TDC scenarios).²¹ Hence, the poor coverage of this method in my simulations is not a concern, provided that when it is used in reality bootstrapping is used to characterise uncertainty.

The coverage of the IPCW method was also good, although slightly lower than the RPSFTM “survivor function” method, which is likely to be due to its decreased accuracy (increased bias). Its coverage was also substantially worse in the scenarios in which the method produced relatively high bias. Given that the IPCW method produced reasonably high levels of bias in the “zero TDC scenarios” (see Figure 6.20), the fact that it provided reasonably good coverage suggests that it produced relatively wide confidence intervals.

It is important to note that the coverage associated with the two-stage Weibull method was poor, often being around 50% to 60%. This is despite the low bias associated with the method and is because the method involved shrinking the survival times of crossover patients to obtain an adjusted dataset before restricted mean survival was estimated. The higher and lower confidence intervals of the treatment effect estimated in crossover patients were used to obtain higher and lower confidence intervals for restricted mean survival. However, this only

takes into account the uncertainty in the treatment effect in crossover patients – it does not take into account the uncertainty in the underlying survival distribution. Similar is true for the SNM approach, which had even lower coverage due to its increased bias. The IPCW, RPSFTM and IPE “survivor function” approaches applied the treatment effect confidence intervals directly into the underlying survivor function, which allowed the full uncertainty to be taken into account. In reality, if a two-stage approach is to be taken, uncertainty around mean survival estimates would need to be taken into account using boot-strapping.

Figures 6.27 and 6.28 demonstrate that in the “TDC Scenarios” the coverage of the RPSFTM and IPE “survivor function” approaches reduced, reflecting the bias associated with these methods in these scenarios. The coverage of the IPCW method was marginally higher than the RPSFTM method in scenarios in which the two methods produced similar levels of bias – hence in these scenarios the IPCW method produced slightly wider confidence intervals. This is interesting because – as discussed throughout Chapter 4 – it may be expected that observational-based methods that make use of a rich dataset have the potential to produce more specific estimates of the treatment effect than randomisation-based methods that produce the same p-value as the ITT analysis. However, my simulations show that in the context of an RCT rather than a much larger observational dataset, this may well not be the case. In these scenarios the coverage of the IPCW method again decreased substantially in scenarios in which its bias was high.

Figures 6.27 and 6.28 also demonstrate that the coverage of the RPSFTM and IPE “shrinkage” approaches was very poor, despite their relatively low bias in the “TDC scenarios”. The reason for this is similar to that for the two-stage Weibull and SNM approaches – using the confidence intervals around the treatment effect to obtain mean survival confidence intervals does not take into account uncertainty in the underlying survival distribution, and hence the full extent of uncertainty is not characterised. This is exacerbated when there are reasonably high levels of bias in the treatment effect estimate. Bootstrapping would be required to increase the coverage of these methods, but levels would still be likely to be sub-optimal due to the bias associated with the treatment effect estimates.

Figure 6.26: Coverage across zero TDC scenarios – selected methods

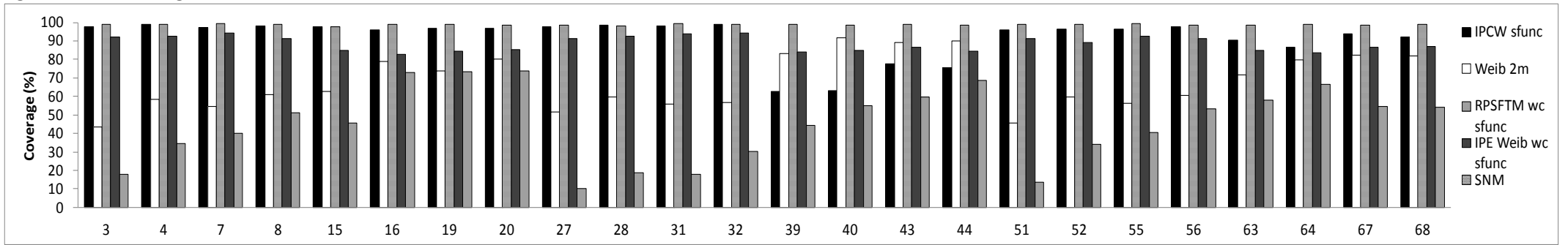


Figure 6.27: Coverage across TDC scenarios – selected methods

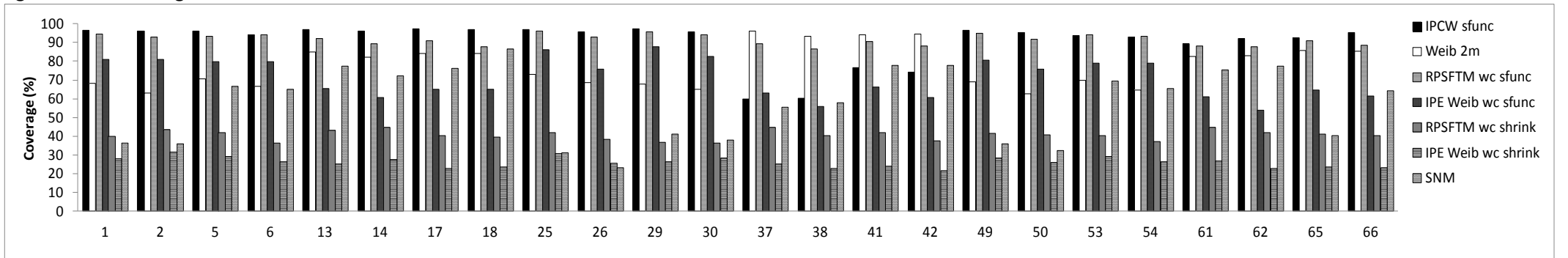
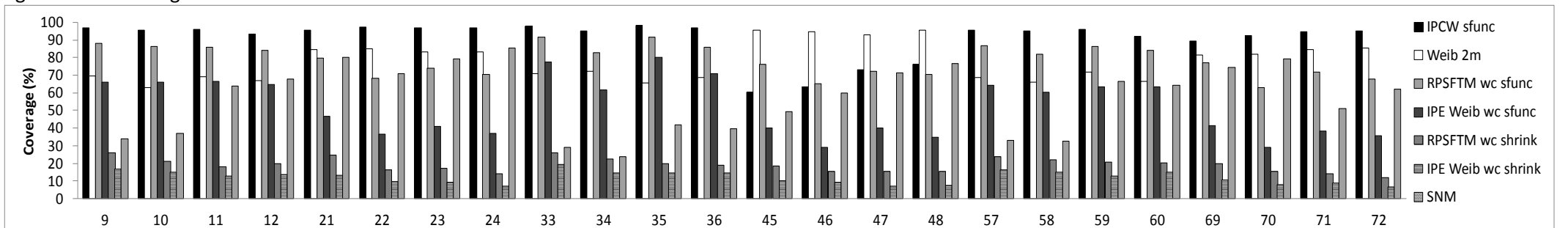


Figure 6.28: Coverage across additional TDC scenarios – selected methods



6.6.6 Mean squared error of methods

Given that a potential advantage of observational-based methods is a higher degree of accuracy than randomisation-based methods, owing to the use of all available covariate data (as discussed in Chapter 4), it is useful to consider the mean squared error (MSE) associated with the selected methods. Generally a lower MSE value represents a more accurate measure, but it is important to consider MSE values while bearing in mind bias and coverage results, as a low MSE may be misleading if bias is relatively high or coverage is relatively low.

Figures 6.29, 6.30 and 6.31 show the MSE associated with the selected methods in the “zero TDC”, “TDC” and “additional TDC” scenarios. In some scenarios the bars representing the SNM and IPCW methods have been truncated due to extremely high MSE values. These are highlighted using arrows at the top of the bars, and occur in the scenarios where treatment crossover proportions were very high, causing highly biased and variable estimates of mean survival.

Figure 6.29 shows that in the “zero TDC scenarios” the RPSFTM and IPE methods produced relatively low MSE. Generally the two-stage Weibull method produced the lowest MSE, but as Section 6.6.5 demonstrates, this should be interpreted with care because the method is associated with low coverage. The IPCW method produced a lower MSE than the SNM method across all scenarios which confirms its better performance in these scenarios, especially considering it generally provides lower bias (see Figure 6.19) *and* improved coverage (see Figure 6.26).

As would be expected given the results already presented in this chapter, in the “TDC Scenarios” the MSE associated with the RPSFTM and IPE methods is higher than in the “zero TDC scenarios”, particularly in the “additional TDC scenarios” (as shown in Figures 6.30 and 6.31). In a selection of the “additional TDC scenarios” the IPCW method (and occasionally the SNM method) produced lower MSE than the RPSFTM and IPE methods. This is a similar pattern to that observed in Section 6.6.3, and demonstrates that relative bias is the key distinguishing factor in the MSE results. This suggests that the standard errors (SE) of the mean restricted mean survival estimates associated with the IPCW, RPSFTM and IPE methods were actually fairly similar (at least in the relatively low crossover scenarios), and further analysis of the results demonstrates that this is the case. However, across the vast majority of scenarios the SE associated with the SNM method were substantially higher, contributing to a higher MSE.

Interestingly, in the scenarios in which the treatment crossover proportion was less than 90%, the SE associated with the IPCW method generally remained approximately the same across the scenarios (given a particular disease severity), whereas the SEs associated with the RPSFTM and IPE methods were higher in the “zero TDC scenarios” than in the “TDC scenarios”. Correspondingly, in the “zero TDC scenarios” the IPCW method produced reasonably low MSE – often very similar to that associated with the RPSFTM and IPE methods – even though the associated bias was relatively high. Conversely, in several “TDC scenarios” in which the RPSFTM, IPE and IPCW methods produced similar levels of bias the RPSFTM and IPE methods produced marginally lower MSE. Further analysis reveals that this is due to the form of the IPCW and RPSFTM/IPE survival models. For example, in Scenarios 1 and 4 the average true HR was 0.75, whereas the average true AF was 1.28 in Scenario 1 and 1.39 in Scenario 4, due to the slightly different survival *times* generated in these scenarios. Similar is true in Scenarios 2 and 3, where the average true HR was 0.51 in Scenario 2 and 0.52 in Scenario 3, whereas the true AFs were 1.75 and 2.15 respectively. The increase in the AFs in the “zero TDC scenarios” (for example, Scenario 4 compared to Scenario 1, and Scenario 3 compared to Scenario 2) provided scope for higher SEs in these scenarios when the RPSFTM and IPE accelerated failure time methods were applied. This was not the case for the IPCW method, which uses a Cox model, and thus relative SEs varied between these methods across these scenarios.

This analysis of the MSEs reveals that the RPSFTM and IPCW methods produced fairly similar SEs, provided the treatment crossover proportion was less than 90%. This is interesting given the different estimation procedures used by each method. It may be surmised that in larger datasets the observational IPCW approach may lead to lower SEs and thus reduced MSE as models may be specified more accurately. On the other hand, longer survival times impact AF estimates relatively more than HR estimates, and therefore with lower disease severity the RPSFTM method might be expected to provide a lower SE and MSE (assuming bias remains the same).

Figure 6.29: MSE across zero TDC scenarios – selected methods

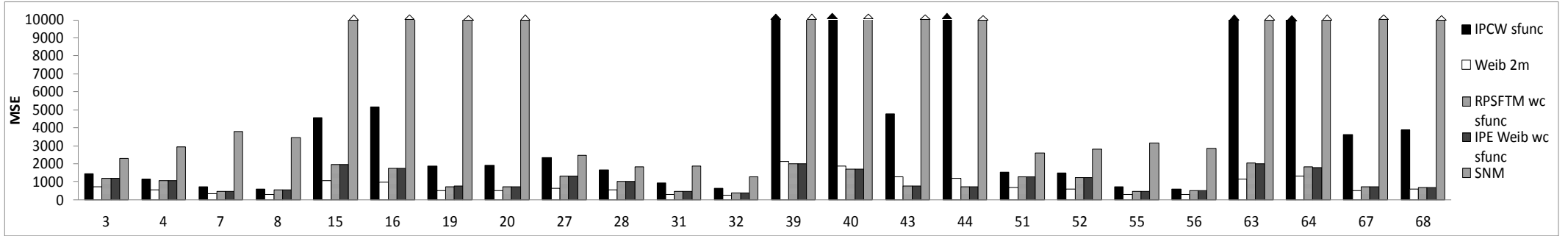


Figure 6.30: MSE across TDC scenarios – selected methods

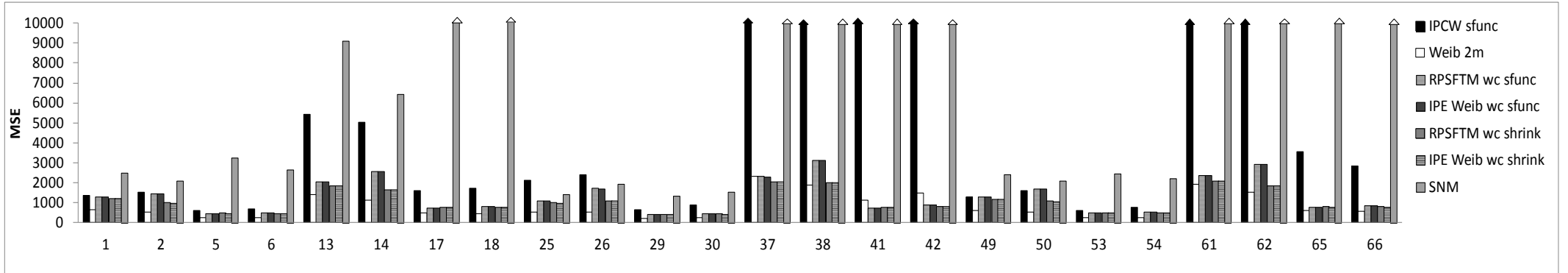
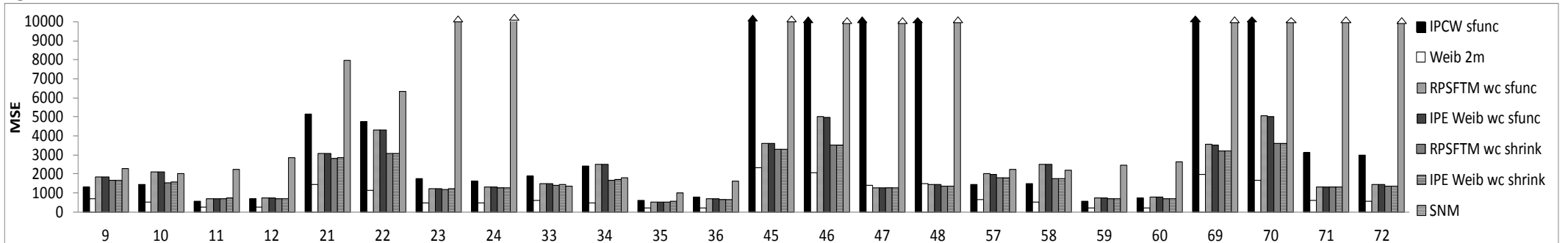


Figure 6.31: MSE across additional TDC scenarios – selected methods



6.6.7 Treatment effect estimates

As a final analysis in this chapter the biases in the treatment effect estimates of the selected methods are considered. This is a secondary analysis because in the majority of scenarios a time-dependent treatment effect was simulated, meaning that no single “true” treatment effect in terms of a hazard ratio or an acceleration factor can be presented. However, as described in Section 6.4.2.8, “average” HRs and AFs can be estimated, and an overview of these results allows any anomalies associated with the restricted means results to be identified (for example, if a method appeared to estimate the average treatment effect well, but produced very biased estimates of restricted mean survival, there may be cause to question the simulation study coding).

Figures 6.32, 6.33 and 6.34 demonstrate the percentage bias for selected methods in the “zero TDC”, “TDC” and “additional TDC” scenarios respectively. For methods that result in a hazard ratio estimated for the experimental group (IPCW, two-stage Weibull, RPSFTM/IPE “shrinkage”, SNM) the bias is measured as a percentage compared to the “true” (average) hazard ratio, and for methods that result in an acceleration factor (RPSFTM and IPE without “shrinkage”) the bias is measured as a percentage compared to the “true” (average) acceleration factor. The figures for the two-stage Weibull, SNM and RPSFTM/IPE “shrinkage” approaches refer to the hazard ratio calculated using a Cox model applied to the adjusted dataset once the survival times of crossover patients had been “shrunk” using initially obtained acceleration factors. The “RPSFTM wc” and “IPE Weibull wc” results refer to the acceleration factors obtained when these methods (including covariates) were applied to the crossover dataset – these are the factors used by the “survivor function” approach to estimate the control group survival curve, and used by the “shrinkage” approach to shrink survival times in crossover patients.

For all approaches and scenarios the results reflect the results presented in Section 6.6.3, with bias following very similar patterns. The only “anomaly” may at first appear to be that the RPSFTM/IPE “shrinkage” approaches give negative bias with regard to the treatment effect, whereas the “survivor function” approaches give positive bias. However, this is logical because the “shrinkage” approaches generate a hazard ratio, whereas the “survivor function” approaches generate an acceleration factor – an over-estimated acceleration factor is equivalent to an under-estimated hazard ratio.

Figure 6.32: Treatment effect bias across zero TDC scenarios – selected methods

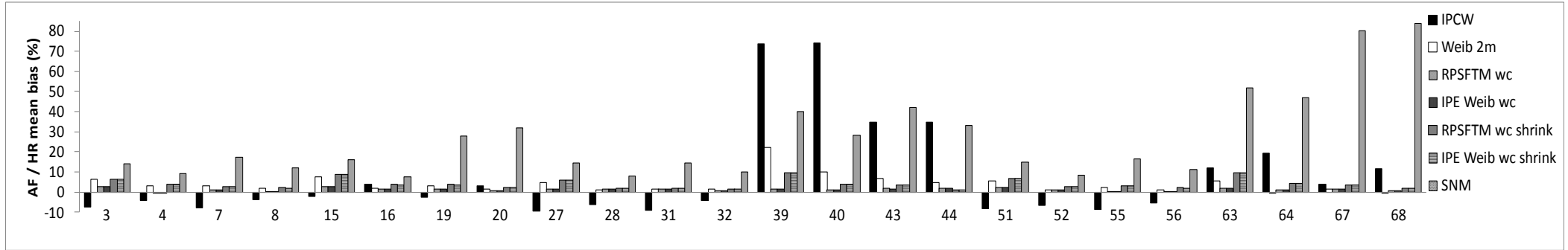


Figure 6.33: Treatment effect bias across TDC scenarios – selected methods

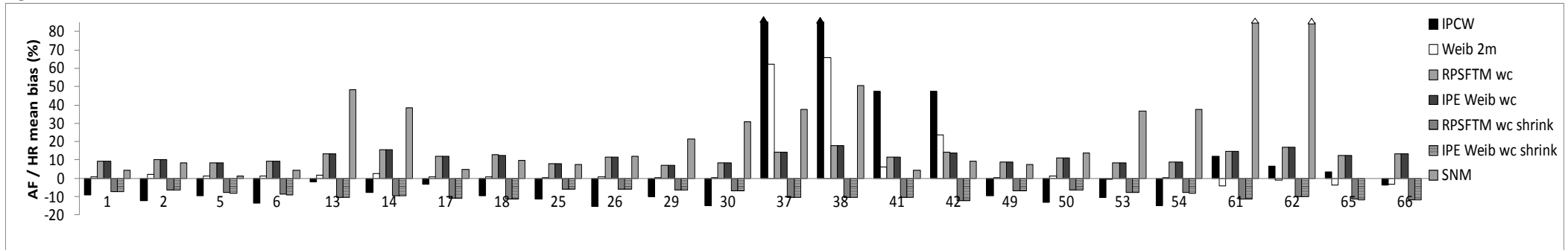
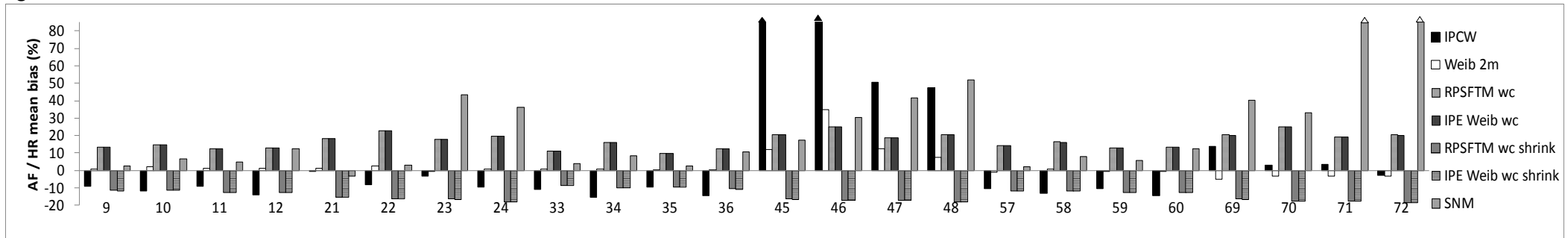


Figure 6.34: Treatment effect bias across additional TDC scenarios – selected methods



6.7 Limitations

Inevitably there are limitations associated with the study presented in this chapter, as with any simulation study. There are also lessons that can be learnt from this study that would benefit future studies. In particular, the study has demonstrated which scenario parameters would be valuable to investigate further.

A limitation associated with any simulation study is that not all scenarios that may be observed in practice can be investigated. I attempted to include all the most important and most relevant scenarios given results of the Morden *et al* (2011) study, realistic cancer trial characteristics and the characteristics of the methods that were being assessed. However, there remain potentially interesting scenarios that were not included. In particular, the proportion of patients that crossed over was an extremely important factor in the results. A range of relatively high crossover proportion scenarios were considered, under the assumption that the crossover adjustment methods would struggle with these most – I assumed that if they worked with a high level of crossover they were likely to also work with a low level of crossover. However, given the high levels of bias associated with the assessed methods in the scenarios that incorporate a time-dependent treatment effect it would be interesting to identify whether levels of bias fall with lower levels of crossover. This is particularly important for the IPCW and SNM approaches, which were particularly sensitive to the crossover proportion. It would be of particular value to understand the relative bias associated with all methods across the full range of crossover proportions (0-100%).

In addition, the study showed that the IPCW approach often produced lower bias than RPSFTM or IPE “survivor function” methods when there was a time-dependent treatment effect (thus, when the “common treatment effect” assumption did not hold). However, only two levels of treatment effect decrements applied to crossover patients were investigated (15% and 25%). It would be interesting to consider a larger range of treatment effect decrements to gain a better understanding of what level of treatment decrement is required for the IPCW method to outperform the RPSFTM/IPE methods.

Also, in all the scenarios tested the assumed sample size was 500 patients. This approximately reflects the common size of metastatic oncology trials, and matches the assumption made by Morden *et al* (2011).²¹ However, given the data requirements of methods such as IPCW and SNMs, this could be an important factor. In fact, if very large proportions of patients crossed over, but the trial was extremely large, there may still be a substantial number of patients

upon which to base the “pseudo” population and so the bias that is created when this number is very small might be avoided. Thus, observational-based methods may produce more accurately specified models, reduced bias and reduced MSE in larger datasets. Given the variable size of cancer trials, it would be valuable to test this assumption further by re-running a sub-set of scenarios with larger (and perhaps smaller) sample sizes.

For practical reasons some crossover adjustment methods that either had performed poorly in previous studies or were not deemed suitable were excluded (for example, the Walker *et al* parametric method). Ideally all possible methods would have been included, but it is unlikely that any of the excluded methods would have proved optimal in any of the scenarios, given that prior evidence suggested they were biased or inferior in less complex scenarios.

A technical limitation of the simulation study was that the IPCW weighted Kaplan-Meier could not be estimated successfully and without bias in the simulation program. However, the IPCW “survivor function” approach is likely to closely resemble results that would have been obtained for the WKM. An additional technical limitation surrounds the use of the *stgest* STATA program to implement the SNM method. The method often failed to converge and this appeared to be due to the “round” option included in the program (as described in Section 6.6.2). This might be argued to be a limitation associated with the STATA program rather than the SNM method. However, it may also demonstrate that it may be difficult to successfully apply a two-stage SNM approach to an RCT dataset because deriving a treatment effect that allows counterfactual survival times to be appropriately close to their required values given the model’s estimation of the hazard of crossover and counterfactual survival may be problematic. Thus this may simply represent a limitation of the two-stage SNM approach in the context of an RCT.

A weakness of the simulation study may also be argued to be the way that the “no unmeasured confounders” assumption was tested. As described in Section 6.6.2, excluding covariates from the estimation procedure for the IPCW and SNM methods had relatively little affect on their results. This was likely to be due to the high correlation between included and excluded covariates. While this made it difficult to assess the importance of the “no unmeasured confounders” assumption, it is also useful because it demonstrates that not *all* covariates may be required by these observational-based methods, provided that the included covariates are sufficient for modelling the treatment crossover process. However, in reality it may be difficult to ascertain whether any independently important covariates are missing. Also, it may be argued that my study did in fact allow a good analysis of the impact of the “no

unmeasured confounders” assumption, because I included a naive censoring approach, which the IPCW method would reduce to if *all* confounders were unmeasured. My results showed that the IPCW method always produced substantially less bias than the naive censoring approach, demonstrating that the inclusion of important confounders is extremely important for observational-based methods.

A general limitation of simulation studies is that the results are likely to always be linked in some way to the chosen data generating process. Attempts were made to limit this by testing different distributions for parametric crossover adjustment methods. Given that a Weibull model was used to generate the underlying survival times, the data generating mechanism may have favoured Weibull-based approaches such as the IPE method. However, the results showed that the IPE method performed similarly well in estimating the treatment effect when it was applied using an exponential model. Also, due to the inclusion of a time-dependent covariate in the data generating model, the resulting survival times no longer followed a true Weibull distribution. Despite this, it may be of value to conduct similar studies using different data generating models.

It may also be of value to re-run the simulations using different methods to estimate the treatment effect received by crossover patients. This was not linked to time, instead the baseline treatment effect was multiplied by a factor to ensure that these patients received a plausible effect. An alternative would be to link this to time and other covariates using formula [28]. However, as discussed in Section 6.4.2.4 this would not be expected to alter the performance of the crossover adjustment methods.

Attempts were made to generate the survival data as realistically as possible, and in such a way that did not satisfy the requirements of any of the crossover adjustment methods. In itself this could be regarded as a limitation, as data were not generated in such a way that satisfies the requirements of methods such as SNMs or the IPCW. Hence these methods could not be expected to produce unbiased results. However, the aim of the simulation study was to demonstrate the performance of these methods in realistic situations – and it is highly likely that in the real-world data will not be generated in a way that satisfies the requirements of these models.

Finally, the simulation study has demonstrated that the application of certain crossover adjustment methods is relatively complex, which could limit the use of these methods by economic modellers. Despite the existence of useful programs written for certain statistics

computer packages, time and/or training would be required in order for economic modellers to be able to appropriately apply the methods.

6.8 Conclusions

The purpose of this chapter was to analyse the comparative performance of alternative crossover adjustment methods in a range of realistic scenarios. There are inevitably limitations associated with any simulation study, but the analysis presented in this chapter extends current knowledge on the practical use of crossover adjustment methods.

The simulation study demonstrates that randomisation-based crossover adjustment approaches such as the RPSFTM and IPE methods produce lower bias than all other methods and provide good coverage in a wide range of scenarios, provided the relative treatment effect received by crossover patients is equal to that received by experimental group patients (that is, the “common treatment effect” assumption holds). However, when the treatment effect is strongly time-dependent, and the “common treatment effect” assumption does not hold, these methods produce high levels of bias and in some circumstances may not be preferable to an intention to treat analysis.

In the presence of such a time-dependent treatment effect existing treatment crossover adjustment methods are limited and are all prone to important bias. Observational-based methods such as the IPCW and SNM require high levels of data availability and are particularly sensitive to bias when the crossover proportion is very high. The SNM method with *g*-estimation as implemented using STATA's *stgest* command was particularly volatile, often failing to converge. When data are available on baseline and time-dependent covariates and when the crossover proportion is less than 90% of at-risk patients, the IPCW method can produce relatively low levels of bias in the presence of a time-dependent treatment effect, and is likely to produce lower bias than an ITT analysis if crossover patients receive a substantial treatment benefit. The SNM method can also produce relatively low levels of bias in these circumstances, particularly when disease severity is low, but is less reliable and generally less accurate than the IPCW method. The simulations suggest that the relatively small size of RCT datasets may cause the observational-based methods to work sub-optimally.

If the treatment crossover mechanism is similar to that simulated by my simulation study (that is, crossover can only occur after disease progression, and if crossover occurs it must happen very soon after disease progression) and data on key prognostic variables are collected upon

disease progression, a simple two-stage Weibull method may be appropriate for adjusting for treatment crossover. In the presence of a time-dependent treatment effect, such a method may outperform more complex methods, although if the treatment effect is not time-dependent RPSFTM/IPE methods are likely to remain optimal.

In the next chapter (Chapter 7), the practical application of the crossover adjustment methods will be assessed in a real-world data study. The present chapter (Chapter 6) and Chapter 7 combine to form Part 4 of this thesis, which aimed to assess the performance of crossover adjustment methods. Following the completion of Part 4, in Part 5 (Chapter 8) recommendations will be made on the use of crossover adjustment methods. These recommendations will draw particularly upon the results reported in this chapter.

Chapter 7

A real-world application of treatment crossover adjustment methods

7.1 Chapter overview

Part 3 of this thesis (Chapters 4 and 5) identified potentially appropriate crossover adjustment methods, in the context of economic evaluation. Chapter 6 tested these methods in a simulation study, and in the present chapter these methods are applied to a real-world dataset confounded by treatment crossover. Chapter 6 and the present chapter combine to form Part 4 of the thesis, which investigates the performance of the adjustment methods identified in Part 3. The present chapter provides an example of the practical use of the methods and also of their potential impact on survival estimates and cost-effectiveness results. This is important because there may be practical limitations associated with methods that affect their performance, that did not arise in the simulation study. The analysis presented in this chapter is novel; although there are examples of applications of alternative crossover adjustment methods applied to RCT datasets in the literature (for example, Yamaguchi and Ohashi,¹⁵⁰ Mark and Robins,¹³⁴ White *et al*,¹²³ Robins and Greenland¹²⁹) these have not included a full range of naive and complex adjustment methods, and also have not considered potential impacts on mean survival estimates (thus incorporating extrapolation within the analysis) and cost-effectiveness results.

7.2 Introduction

This chapter presents a case study of the application of crossover adjustment methods to an RCT confounded by crossover. Section 7.3 provides background on the RCT and the crossover observed in it. Section 7.4 details the crossover adjustment methods that are applied to the confounded dataset. Section 7.5 provides results, first summarising a parametric analysis of the dataset and the implications of this for the application of the crossover adjustment methods, and then presenting the results of the applied adjustment methods. The parametric analysis is required because I seek to demonstrate the impact of the alternative crossover adjustment methods on mean survival and cost-effectiveness results, and therefore extrapolation of the censored RCT survival data is necessary. Hence I determine appropriate models for extrapolation and then combine these with the crossover adjustment methods. In Section 7.6 the potential impact of alternative adjustment methods on cost-effectiveness are

presented, based upon a simple economic evaluation. Section 7.7 considers the limitations associated with the case study and Section 7.8 presents conclusions.

7.3 Background

GSK provided data from the EGF100151 study – a Phase III randomised controlled trial that compared lapatinib plus capecitabine to capecitabine alone in women with advanced breast cancer that had progressed on trastuzumab. Details of the study are reported by Geyer *et al* (2006) and Cameron *et al* (2010).^{176;177} Further updated results from the study are presented in GSK's submissions to NICE.^{173;178} Patients were randomly assigned in a 1:1 ratio to treatment with lapatinib (1,250 mg daily, continuously) plus capecitabine (2,000 mg/m² in two divided doses on days 1 through 14 of a 21-day cycle) or capecitabine alone (2,500 mg/m² in two divided doses on days 1 through 14 of a 21-day cycle). The primary endpoint was time-to-progression (TTP; defined as the time from randomisation to disease progression or death due to breast cancer) as assessed by an independent blinded radiological review using copies of serial radiographs and photographs of visible lesions. Secondary endpoints were progression-free survival (PFS; the time from randomization to disease progression or death due to any cause), overall survival (OS), overall response rate, clinical benefit rate (complete response, partial response, or stable disease at 6 months), and safety as measured by the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE, version 3.0).¹⁷⁷

Accrual to the trial was discontinued and crossover was offered to women receiving monotherapy capecitabine as of April 3, 2006 after an interim analysis had been conducted which demonstrated a HR for TTP of 0.40. By this timepoint 399 women had been recruited (201 randomised to the monotherapy group; 198 randomised to the combination therapy group) and a further 9 were being screened for treatment. These 9 patients were allowed to be included in the trial in the combination treatment arm (resulting in 207 participants in this group). Subsequently, 36 (17.9% of the control group) patients of 39 who remained on capecitabine monotherapy on April 3 2006 switched onto combination therapy. The group of patients who were offered the chance to switch treatments were control group patients who had not yet experienced disease progression, and thus it appeared likely that any patients that switched treatments would do so *before* disease progression. However, this was complicated by the fact that these patients were given the option of crossing over immediately, or at any time between April 3 2006 up to and including the time of disease progression. In practice 10 patients (5.0% of the control group; 27.8% of control group patients that switched treatment) crossed over after they had experienced disease progression.¹⁷³

Given this information, it is possible that treatment crossover would have caused bias in both estimates of TTP (and thus PFS) and OS. Unfortunately, GSK are only aware of the number of crossover patients that had progressed at the time of crossover – for the most part which patients these were and their time of progression could not be identified, because the cut-off date for TTP data was April 3 2006 (the time at which crossover was initially allowed). Of the 36 crossover patients, independent review data specifying time to disease progression is only available for 6 – all of whom experienced disease progression before crossing over. The remainder were censored on or shortly before the TTP cut-off date of 3 April 2006. Given the available data it is not possible to appropriately adjust estimates of the treatment effect on PFS.

The use of lapatinib in women previously treated for advanced or metastatic breast cancer was subject to assessment by NICE, but this was suspended in October 2010 following a lengthy appraisal process. Instead of completing the appraisal, NICE planned to provide guidance on the use of lapatinib and trastuzumab in a later technology appraisal.¹⁷⁹ In the suspended NICE TA, GSK presented numerous analyses of the OS data. These are presented in Table 7.1.

Table 7.1: Existing analyses of EGF100151 OS data

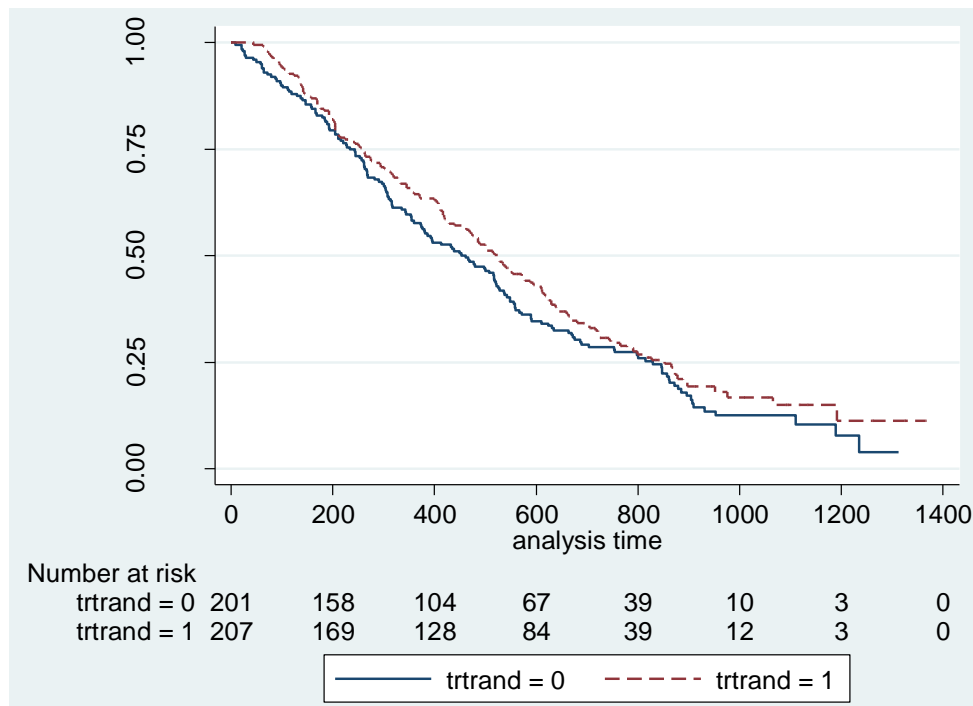
Dataset cut-off time	Patient population	Method description	OS Hazard Ratio (95% CI)
3 April 2006	399	ITT	0.78 (0.55, 1.12)
28 Sept 2007	408	ITT	0.90 (0.71, 1.12)
1 Oct 2008	408	ITT	0.87 (0.71, 1.08)
1 Oct 2008	408	ITT, adjusted for baseline prognostic factors*	0.81 (0.65, 1.00)
1 Oct 2008	372	Exclude crossover patients	0.78 (0.62, 0.97)
1 Oct 2008	408	Censor crossover patients	0.82 (0.66, 1.02)
1 Oct 2008	408	Cox model with crossover event modelled as time-dependent covariate, without baseline covariates*	0.80 (0.64, 0.99). Crossover HR = 0.63 (0.41, 0.98)
1 Oct 2008	408	Cox model with crossover event modelled as time-dependent covariate, with baseline covariates*	0.75 (0.60, 0.94). Crossover HR = 0.65 (0.41, 1.01)

*Note for these analyses the GSK submission does not state how many patients remained in the analysis Taken from GlaxoSmithKline UK submission to NICE¹⁷³

Hence it can be seen that GSK conducted numerous “naive” analyses in an attempt to adjust for treatment crossover. These methods present a consistent message – there is likely to be an OS advantage associated with lapatinib plus capecitabine combination therapy compared to capecitabine alone. However the exact size of the treatment effect is uncertain.

Figure 7.1 illustrates the OS Kaplan-Meier curves for the experimental (trtrand=1) and control (trtrand=0) groups.

Figure 7.1: Overall survival Kaplan-Meier curves from the EGF100151 study



7.4 Methods

In this section the applicability of alternative crossover adjustment methods to the EGF100151 dataset is discussed. Ideally, all of the methods tested in Chapter 6 would be applied to the dataset in order to compare a full range of OS survival estimates. Unfortunately, the data available limits the analyses that can be applied. Section 7.4.1 addresses naive methods, Section 7.4.2 addresses observational-based complex methods, and Section 7.4.3 addresses randomisation-based complex methods. Section 7.4.4 summarises the methods applied to the case study dataset.

7.4.1 Naive methods

A variety of naive methods were applied to the dataset in order to obtain a range of naive estimates of the treatment effect that could be compared to estimates resulting from the more complex methods. The naive methods applied matched those included in the simulation study presented in Chapter 6. These were:

- ITT analysis
- Censor crossover patients
- Exclude crossover patients

- Treatment as a time-dependent covariate
- Crossover indicator as a time-dependent covariate

7.4.2 Observational-based methods

The lack of comprehensive data on TTP in crossover patients made the application of observational-based methods such as SNM with g-estimation and IPCW problematic. Because disease progression is likely to be an important indicator of OS and some patients crossed over both before and after disease progression an indicator of disease progression is an important variable to include in any observational-based analyses. Failing to include this parameter would violate the “no unmeasured confounders” assumption made by these methods and bias would result. In addition, only limited data were available on potential time-dependent covariates – ECOG data were collected over time, but were not collected after 3 April 2006. Therefore these data could not be used to completely model patient characteristics up until crossover time because several patients crossed over substantially after this date (20 of the 36 patients crossed over at some point up to 7 months after April 2006). A further problem with the application of observational-based methods was that almost all patients who became at risk of crossover did in fact cross over (36 of 39 patients), making it very difficult to model the probability of crossover – particularly given that data on disease progression and ECOG were missing.

The treatment crossover mechanism observed in study EGF100151 meant that two-stage crossover adjustment methods had to be excluded. As crossover could occur before disease progression these methods were not applicable as there was not a suitable “secondary” baseline period upon which to base estimates of the treatment effect in crossover patients. This meant that the two-stage Weibull and SNM with g-estimation methods had to be excluded.

Given the lack of data on time to disease progression, and that patients crossed over both before and after disease progression, conducting an IPCW analysis was also problematic and probably inappropriate. However, as an exploratory analysis, the IPCW method was applied under certain explicit assumptions. I assumed that patients could only cross over if they had not experienced disease progression by 3rd April 2006, that all crossover patients crossed over before disease progression (thus TTP was greater than crossover day for all crossover patients) and that all patients who had the opportunity to cross over did so. Under these assumptions crossover is essentially random (because enrolment time is random). These assumptions removed the need to know and include all the potentially confounding covariates in an IPCW

analysis. In the EGF100151 dataset it appears reasonable to assume that crossover was essentially random in terms of patients offered the opportunity to crossover, because the risk-set for crossover was defined by calendar time. However, as discussed above, some patients went on to experience disease progression before crossing over, and not all patients offered the opportunity to cross over did so (3 out of 39 did not). Hence, my assumptions do not match the observed data and the crossover that actually occurred was not random – thus analyses based upon these assumptions are prone to bias. The assumption that crossover always occurred prior to disease progression may be partially justified because for patients that crossed over after disease progression, crossover time was likely to have been close to disease progression time.

The IPCW method was applied by first creating a panel dataset with an implied observation for every day for every patient, and crossover patients were indicated as censored at the time of crossover. The denominator of the stabilised IPCW was calculated for control group patients for each day up until the observed time to progression – the denominator was set to 1 for times after disease progression reflecting that these patients were not at risk of crossover. Previous analyses undertaken by GSK had highlighted that ECOG, presence of liver metastases and number of metastatic sites were statistically significant baseline covariates for survival, and these were included within the IPCW model.¹⁸⁰ Time (in days since randomisation) was also included in the model. The numerator of the stabilised IPCW was estimated using the same model (including baseline covariates) but was applied to all time intervals for all patients in the control group – this reflects a model for crossover based only upon baseline covariates – that is without the knowledge of time to progression that was used in the model used for the denominator of the weight.

The “survivor function” approach was used to estimate mean survival associated with the IPCW method. This approach was applied as described in Section 6.4.2.10 of Chapter 6, except that mean survival was not restricted to a certain time-point – it was based upon an extrapolation that continued until all patients were predicted to have died. A weighted Kaplan-Meier (WKM) was also estimated, using the IPCW weights. A suitable parametric model was then fitted to this in order to estimate mean survival.

7.4.3 Randomisation-based methods

The relative lack of data available highlights issues that are likely to be important in many real-world scenarios – often crossover is not “clear-cut” and data on time-dependent covariates may be sparse. This makes observational-based approaches problematic and open to bias, but

randomisation-based approaches such as RPSFTM and the IPE algorithm remain possible to apply. This highlights the comparatively low data requirements of these methods, which may make them practically more useful. These methods were applied to the EGF100151 data both with and without the inclusion of baseline covariates that had been shown in previous analysis to be statistically significant indicators of survival.¹⁸⁰ “Survivor function”, “extrapolation” and “shrinkage” approaches were all applied to estimate mean survival following application of the RPSFTM and IPE methods. These were applied as described in Section 6.4.2.10 of Chapter 6, except that mean survival was not restricted to a certain timepoint.

While it is “easy” to apply randomisation-based methods to RCT datasets confounded by crossover, due to their lack of reliance on covariate data, difficulties do arise when considering *how* these methods should be applied. This was discussed in Section 4.10.2.3 of Chapter 4. To briefly reiterate, there are four possible approaches that may be taken when applying RPSFTM and IPE methods. Strictly speaking, according to the RPSFTM counterfactual survival model given by equation [3] in Section 4.10.2 of Chapter 4, an “on treatment” approach should be taken. However this assumes that the treatment effect disappears immediately upon treatment discontinuation and provides a “causal” treatment effect that would only be relevant for economic analysis if it were applied only while patients were receiving treatment within the economic model. It also requires that treatment discontinuation times are known both for experimental group patients *and* crossover patients. In order to obtain a treatment effect more relevant for the economic evaluation, an “on treatment – observed” approach can be taken – where the counterfactual survival times estimated using the “on treatment” approach are compared to the observed experimental group survival times. An alternative is to apply the methods on a “treatment group” basis which provides an assessment of the treatment effect associated with being randomised to the experimental group that allows for a treatment effect that is maintained beyond treatment discontinuation. However, this approach may be prone to bias if relative post-treatment discontinuation times differ between the experimental group and crossover patients, or if post-study treatments received do not reflect realistic treatment pathways. An additional option is to apply the methods on an “ever treated” basis whereby if a patient was randomised to the intervention group, or if they crossed over onto the experimental treatment at any time-point, they are assumed to be in the “treated” group from time zero until death. However this seems artificial, as a treatment effect would be associated with crossover patients before the point of crossover, and implies the assumption that treatment duration is similar between experimental and control groups. Hence this approach seems unlikely to be appropriate.

In the case study presented in this chapter “on treatment”, “on treatment – observed” and “treatment group” approaches are applied, in order to assess the difference these can make to treatment effect, survival-time and cost-effectiveness estimates. However, as is often likely to be the case discontinuation times were not known for crossover patients and so they were assumed to take the experimental treatment until death or censoring. This could bias estimates of the “on treatment” and “on treatment – observed” treatment effects, because crossover patients are assumed to receive the treatment from the point of crossover until death, whereas experimental group patients are not. In addition, data on time of treatment discontinuation in experimental group patients were not included in the dataset provided by GSK, and thus I assumed that discontinuation occurred at the time of disease progression, as per the trial protocol. Cameron *et al* (2008) report that in fact 14% of the experimental group discontinued treatment prior to disease progression.¹⁷⁷ For these reasons, the “on treatment” and “on treatment – observed” analyses that I present can only be described as illustrative.

7.4.4 Crossover adjustment methods to be applied to the EGF100151 dataset

The crossover adjustment methods applied to the EGF100151 dataset were:

- ITT analysis
- Censor crossover patients
- Exclude crossover patients
- Treatment as a time-dependent covariate
- Crossover indicator as a time-dependent covariate
- IPCW (under explicit assumptions laid out above)
- RPSFTM (with and without covariates) “treatment group”, “on treatment” and “on treatment – observed” approaches
- IPE algorithm (with and without covariates) “treatment group”, “on treatment” and “on treatment – observed” approaches

Various “survivor function”, “extrapolation” and “shrinkage” approaches were applied to these in order to estimate mean survival. A “Methods key” is presented in Table 7.3, prior to the presentation of the results of the crossover analysis, in order to describe each analysis applied.

All analyses were run on the most recent dataset available, which had an OS cut-off of 1 October 2008. For each method applied adjusted treatment effects (in terms of a HR or AF) and confidence intervals were stored and are reported in Section 7.5. Estimates of mean survival associated with each method are also reported, using extrapolation approaches

described in Chapter 5. Confidence intervals of treatment effects and survival estimates are also reported, but it should be noted that it is not straightforward to produce confidence intervals around independently fitted survival curves – such as those fitted in “extrapolation” approaches that fit models to the counterfactual survival data, or to the WKM. Boot-strapping would be required to estimate the confidence intervals for mean survival associated with these approaches. Such analyses have not been conducted here since the main focus of this chapter (and the thesis as a whole) is upon bias in point-estimates of the mean. While confidence intervals and uncertainty are clearly of great importance in economic evaluation, it is first necessary to determine accurate estimates of the mean.

STATA code for the analyses are presented in Appendix 10.

7.5 Results

The results associated with the application of the various treatment crossover adjustment methods to the EGF100151 dataset are presented in this section. First, in Section 7.5.1 an analysis of the hazards observed in the trial is presented, and the fits of various parametric models are considered. This is necessary because I go on to combine crossover adjustment methods with extrapolation approaches in order that mean survival estimates and associated cost-effectiveness results can be obtained. Thus a suitable parametric model must be identified. This analysis also sheds light on the plausibility of the “common treatment effect” assumption, which helps with the interpretation of the results associated with randomisation-based crossover adjustment methods. In Section 7.5.2 I discuss the implications of the observed hazards and parametric analysis on the application of the crossover adjustment methods. Section 7.5.3 presents the results of the crossover adjustment methods on survival estimates, before Section 7.6 goes on to consider the potential impact of alternative methods on cost-effectiveness results.

7.5.1 Analysis of hazards and assessment of fit of parametric models

Before undertaking crossover-adjustment analyses, I assessed the fit of alternative parametric models to the trial data in order to identify suitable parametric models that could be combined with crossover adjustment methods for extrapolation purposes. This was required due to the censoring observed in study EGF100151, illustrated in Figure 7.1. I also attempted to assess the plausibility of the “common treatment effect” assumption.

7.5.1.1 The “common treatment effect” assumption

I attempted to assess the “common treatment effect” assumption based upon hazard plots and previous analyses undertaken by GSK. Log-cumulative hazard plots are illustrated in Figure 7.2. As explained by Collett (2003) these plots can be used to determine whether the proportional hazards assumption holds, and which parametric models are appropriate.¹² Importantly for the assessment of the “common treatment effect” assumption, if the lines in the “Plot for Weibull and Exponential” graph (which plots $\ln(-\ln(\text{survivor function}))$ against $\ln(\text{time})$) are approximately parallel, the proportional hazards assumption is observed to hold. In addition to this, if the plot illustrates scatters that are in straight lines, then the Weibull model is appropriate for modelling the data. If the lines are straight and their gradient is not significantly different from 1, the exponential model is potentially appropriate. Similar interpretations hold for the three other plots shown in Figure 7.2 – in each case if the plots represent straight lines then the respective parametric model is potentially appropriate for modelling the observed data.

In addition to the log-cumulative hazard plots presented in Figure 7.2 a quantile-quantile plot for the EGF100151 survival data was constructed (see Figure 7.3). This plots the survival times in the control group and the experimental group for each survival probability. If an accelerated failure time model is appropriate to model the data this plot should present a straight line with a gradient equal to the acceleration factor, and which passes through the (0,0) point.¹⁸¹

It is difficult to conclude whether proportional hazards or a constant acceleration factor was demonstrated in the trial. Firstly, the trial data plotted in the graphs in Figures 7.2 and 7.3 are confounded by treatment crossover, which may affect the hazards over time (which demonstrates that validating these assumptions in a trial confounded by crossover will always be problematic). In addition, previous analysis by GSK has demonstrated that there were baseline prognostic covariates that had statistically significant impacts on overall survival (Number of metastatic sites (>3 or ≤ 3) P-value 0.014; ECOG performance status (0 or ≥ 1) P-value <0.001 ; Liver metastases (No or Yes) P-value <0.001).¹⁸⁰ These may impact upon the estimated treatment effect and are not accounted for in the plots shown in Figures 7.2 and 7.3. The plots in Figure 7.2 suggest that the hazards may not be proportional over time as the lines are not parallel prior to approximately time $\ln(5)$ (approximately 150 days). However, these segments of the curves are based upon relatively little trial data (at 150 days 25 deaths had occurred in the experimental arm and 29 had occurred in the control group, out of total death numbers of 158 in the experimental group and 164 in the control group). After 150 days

the proportional hazards assumption does seem reasonable. In addition, the quantile-quantile plot appears to demonstrate that the acceleration factor is approximately constant over time.

Figure 7.2: Log plots to assess observed hazards

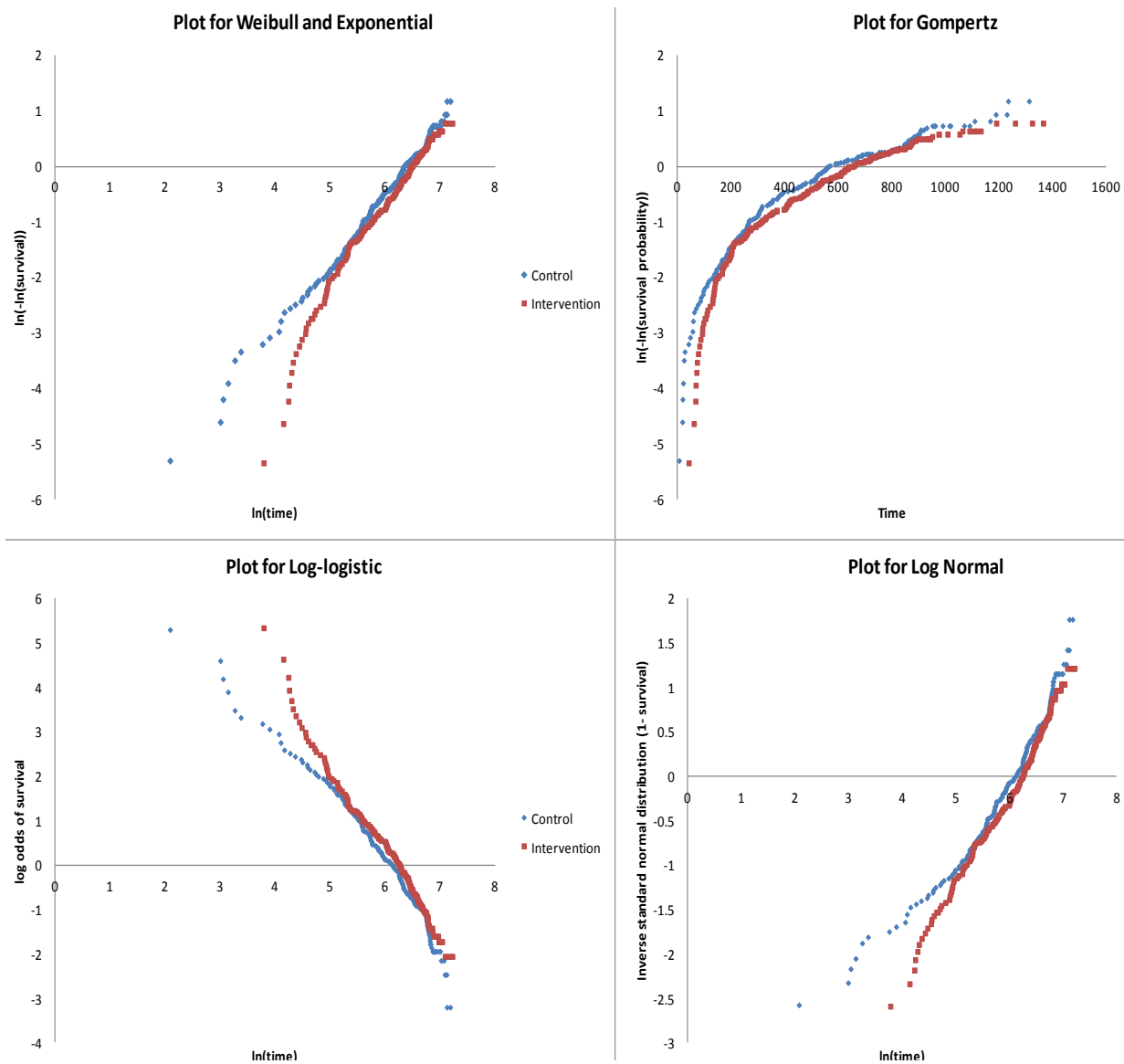
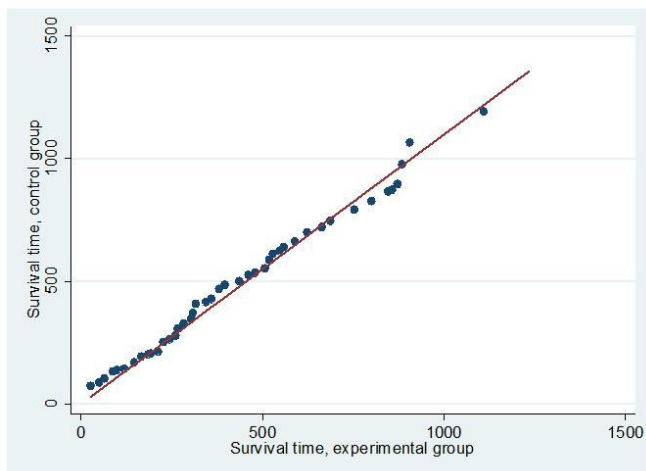


Figure 7.3: Quantile-quantile plot for study EGF100151



However, it is difficult to make any conclusions based upon these plots, as in the case of a clinical trial confounded by treatment crossover it may actually be expected that the observed treatment effect between randomised groups would be seen to reduce over time. If it does not this may signify that crossover patients did not benefit from crossing over. However this may not be the case if small proportions of control group patients crossed over (as in study EGF100151) or – as is usually likely to be the case – if the treatment crossover decision was not random. Hence, this analysis is of limited use when assessing the plausibility of the “common treatment effect” assumption.

It is noteworthy that in the previous analyses undertaken by GSK where the crossover variable was included as a time-dependent covariate there did not appear to be a decrement in treatment effect received by crossover patients (see Table 7.1). This analysis is imperfect and open to bias as time-dependent confounders may exist, but it appears at least that there is not strong evidence *against* the assumption of a common treatment effect. This is further justified by the fact that the majority of crossover patients had not experienced disease progression at the time that they received the experimental treatment, and therefore their capacity to benefit from it may not have been reduced. Hence in this case study the “common treatment effect” assumption may be reasonable, suggesting that RPSFTM and IPE crossover adjustment methods may be appropriate. However, it is difficult to make this conclusion confidently.

7.5.1.2 Assessment of fit of parametric models

In the context of modelling survival in the presence of treatment crossover it is of most importance to identify a model that accurately models survival in the experimental group, rather than the experimental group and the control group. This is because the trial data for the control group is confounded by treatment crossover, and thus the objective is to adjust for crossover and to extrapolate the counterfactual dataset appropriately, rather than to find a model that fits the observed control group data well. Hence, initially it is important to identify the most appropriate parametric model for extrapolating experimental group survival. Later, when a counterfactual dataset has been derived for the control group, extrapolation of this can be considered.

To determine the parametric distribution most suited to modelling the experimental group survival data, first the log-cumulative hazard plots and their derivatives shown in Figure 7.2 were considered. Considering only the plots for the experimental group in Figure 7.2 it can be seen that the Weibull and Gompertz plots do not appear to result in straight lines. The log normal and log-logistic plots appear to suggest that these models would be more appropriate

for modelling the experimental group survival times. Given that the log normal model is a special case of the generalised gamma model, this may also be appropriate for modelling experimental group survival.

Next, the alternative parametric models were fitted to the experimental group trial data. Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC) statistics are presented in Table 7.2. These provide information on the relative fit of alternative parametric models to the observed data.^{12;15}

Table 7.2: Alternative parametric model fits to the experimental group ITT data

Model	AIC	BIC
Exponential	534.18	537.52
Weibull	511.09	517.75
Gompertz	522.52	529.18
Log-logistic	509.86	516.52
Log normal	507.11	513.78
Generalised gamma	508.05	518.04

Table 7.2 suggests – as indicated by the log-cumulative hazard plots – that the log normal and generalised gamma models offer the best fits to the experimental group data. Importantly, the parameter values of the generalised gamma model indicated that exponential and Weibull models (which are special cases of the generalised gamma model) would not be appropriate for modelling experimental group survival, whereas a log normal model could not be ruled out. To investigate this further the models were inspected visually, compared to the experimental group Kaplan-Meier curve. These plots are shown in Figures 7.4 to 7.9.

Figure 7.4: Exponential model fitted to experimental group, compared to the Kaplan-Meier curve

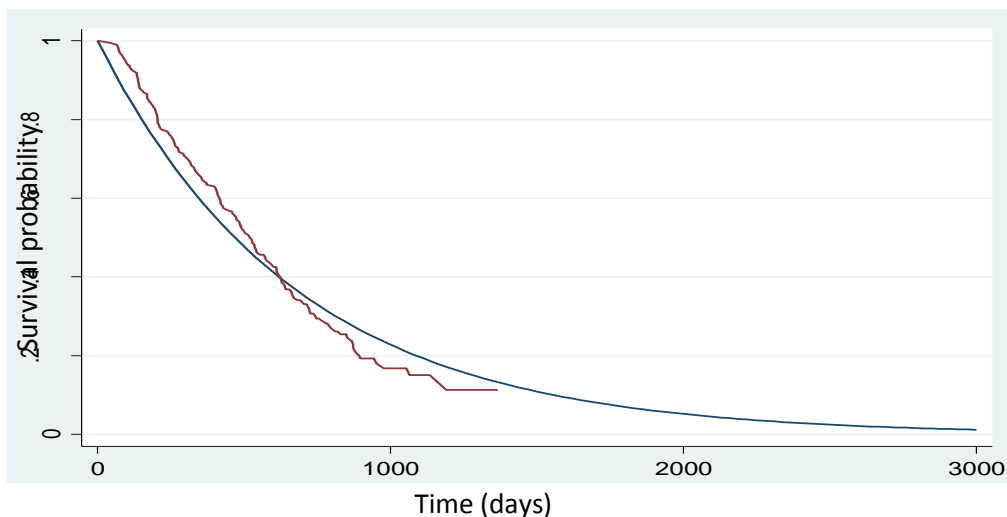


Figure 7.5: Weibull model fitted to experimental group, compared to the Kaplan-Meier curve

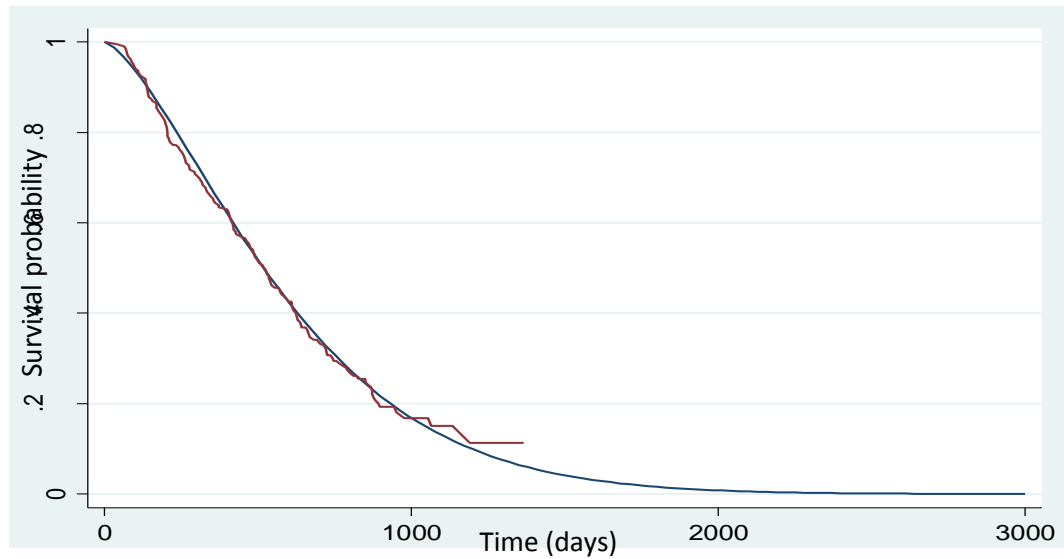


Figure 7.6: Gompertz model fitted to experimental group, compared to the Kaplan-Meier curve

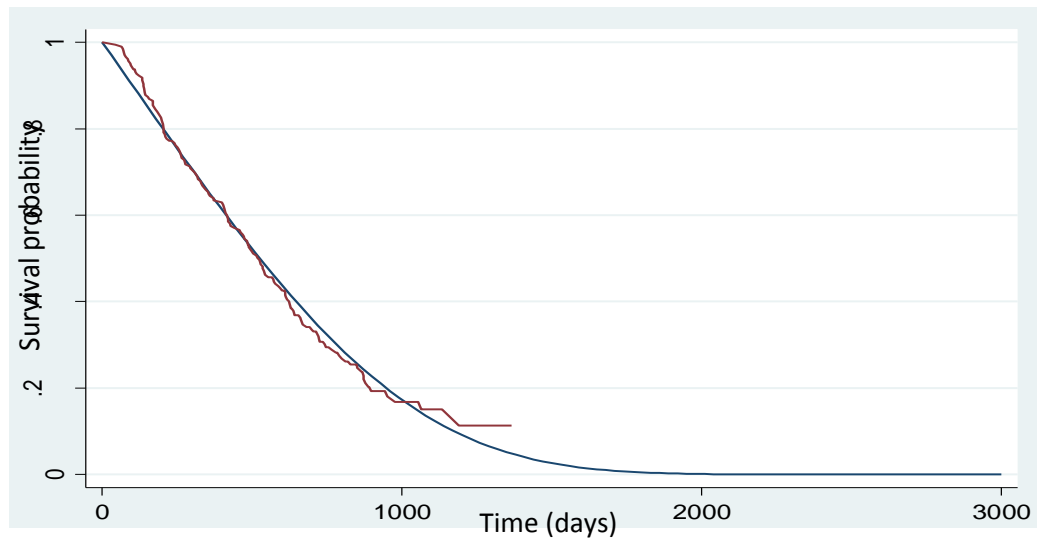


Figure 7.7: Log-logistic model fitted to experimental group, compared to the Kaplan-Meier curve

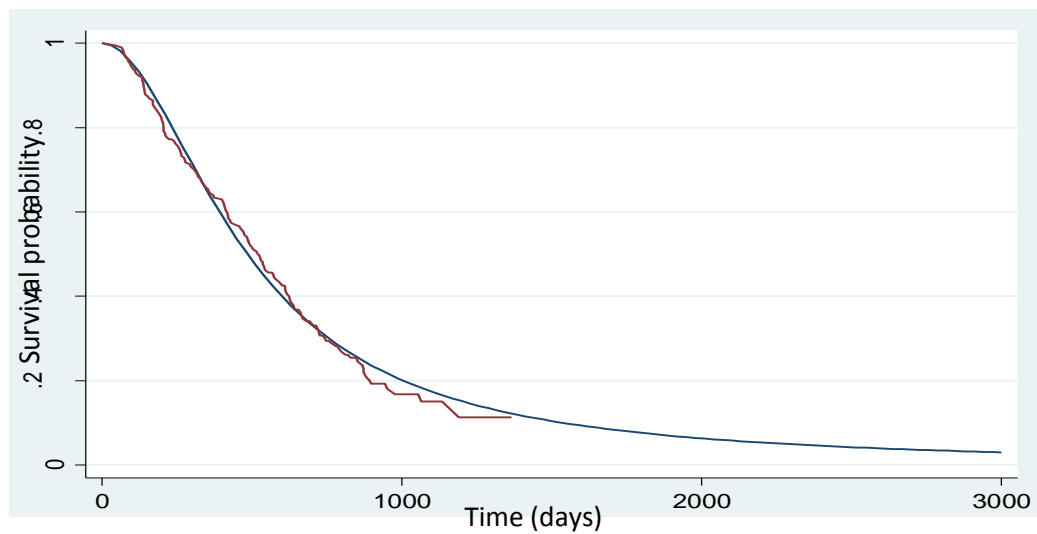


Figure 7.8: Log normal model fitted to experimental group, compared to the Kaplan-Meier curve

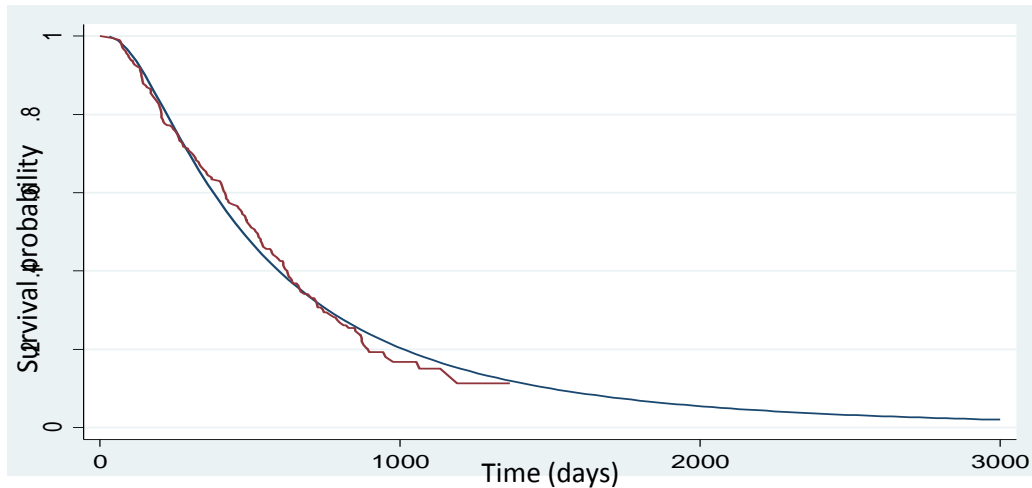
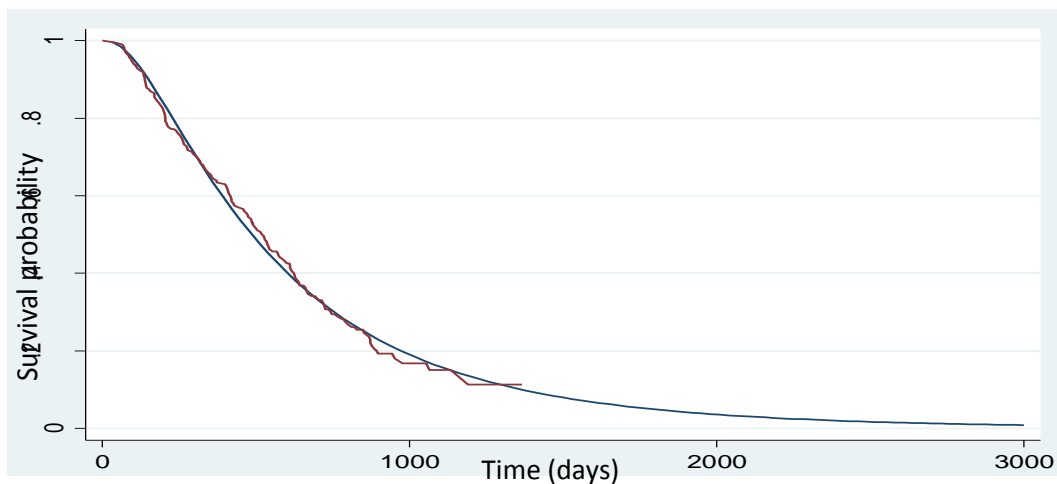


Figure 7.9: Generalised gamma model fitted to experimental group, compared to the Kaplan-Meier curve



Inspecting Figures 7.4 to 7.9 confirms that the log normal and generalised gamma models appear to be the most appropriate for modelling survival in the experimental group. The exponential model provides a poor fit, whereas the Weibull and Gompertz models appear to underestimate long-term survival compared to long-term trial data. The log-logistic and log normal models provide a very similar fit and appear to marginally overestimate long-term survival. The generalised gamma model provides a good fit and seems to predict long-term survival more appropriately. It is important to recognise that towards the tail of the Kaplan-Meier curve there remain very few patients at risk (as demonstrated by Figure 7.1). Therefore, it is debatable how much emphasis should be placed on selecting a parametric model that appropriately fits to this section of the Kaplan-Meier. However, the trial data represent the best evidence available on long-term survival for this particular patient cohort, and so it is reasonable to select a model that closely represents *all* segments of the Kaplan-Meier curve. Models such as the Weibull and Gompertz that seem to underestimate long-term survival are potentially inappropriate as they do not reflect the slight flattening of the OS curve

demonstrated by the Kaplan-Meier. This statement must be made with caution because uncertainty is high in the tails of the data, where relatively few patients remain at risk and few events are observed. However, such flattening of the OS curve is often observed in cancer trials, even in seriously ill patients, as small proportions of patients perform particularly well.

It is difficult to obtain data from external sources on expected long-term survival relevant for the patients that took part in study EGF100151, since the trial is in late-stage metastatic patients who have already received prior metastatic therapies. This makes it difficult to identify the length of time that the trial population had had cancer for, and also the stage of disease that patients had upon diagnosis is not known. Therefore, it is difficult to compare registry survival rates for metastatic cancer by stage at diagnosis to the trial population. However, UK statistics suggest that there is a flattening of survival curves between approximately 5 and 10 years after diagnosis in patients diagnosed with metastatic breast cancer.¹⁸² Overall, it seems reasonable to conclude that the generalised gamma model is likely to represent the best model for extrapolating experimental group survival.

7.5.2 Implications for crossover adjustment methods

Based upon the analysis described above, a number of conclusions can be drawn about the potential bias associated with the more complex crossover methods. Firstly, there is not strong evidence against the “common treatment effect” assumption, although this is difficult to conclude confidently because the data are confounded by treatment crossover. However if there is a time-dependent treatment effect it appears likely to be relatively weak. This suggests that the RPSFTM and IPE methods may result in low bias, and that the “survivor function” approach to estimating survival in the control group may be reasonable. The “survivor function” approach involves a constant acceleration factor being applied to the experimental group survivor function, whereas the “extrapolation” and “shrinkage” approaches allow independent parametric models to be fitted to counterfactual datasets after applying the RPSFTM or IPE methods (although the initial treatment effect used to obtain the counterfactual dataset is still limited by the “common treatment effect” assumption under these approaches). In this case study it appears that the extra flexibility associated with the “extrapolation” and “shrinkage” approaches with respect to the ability to fit independent parametric models to treatment arms may not be important. Estimating the control group survival curve using the “extrapolation” approach is affected by a loss of information associated with recensoring, and the “shrinkage” approach is prone to bias due to not fully recensoring. The “survivor function” approach avoids these problems, and so may represent the most appropriate version of the RPSFTM and IPE analyses in this case study.

The generalised gamma model is likely to represent the most appropriate model of survival in the experimental group. Hence for the “survivor function” approaches a generalised gamma model was used to estimate the experimental group survivor function over time, and the various crossover-adjusted treatment effects were applied to this to estimate survival in the control group. In addition, independent generalised gamma models were fitted when applying “extrapolation” approaches (that is, for the extrapolation of RPSFTM and IPE counterfactual datasets, and for the extrapolation of control group datasets obtained from “shrinkage” approaches and the IPCW weighted Kaplan-Meier).

It is noteworthy that the IPE method is only coded for use with a Weibull or exponential model – it is not available in the generalised gamma form. The Weibull form was used in this case study as it is likely to be more appropriate than the exponential, based upon the analysis described in Section 7.5.1. In the simulation study reported in Chapter 6 the choice of model used within the treatment effect estimation procedure was relatively unimportant in terms of resulting bias – more important was the model used to extrapolate over time in either the “extrapolation” or “survivor function” approaches. Hence, in this case study, when the IPE method was applied a Weibull model was used to estimate the adjusted treatment effect, but generalised gamma models were used for the estimation of survival over time.

In an extrapolation context, an important problem with the WKM provided by the IPCW method is that it provides a weighted version of the Kaplan-Meier but not an adjusted dataset (as discussed in Section 5.6.2.2 of Chapter 5) – hence there is no underlying dataset to which a parametric model can be fitted for purposes of extrapolation. This problem was addressed by fitting a generalised gamma model to the WKM by “creating” trial data that represented the WKM survivor function. To achieve this I used the WKM survivor function and calculated $(1 - \text{survivor function}) * 100000$ for each survival time and rounded up or down to the nearest whole number in order to estimate the number of patients (out of 100,000) that would have died at each survival time. For example, 8 days was the first observed death event in the control group of the trial, and the WKM survival probability at day 8 was 0.99496. Hence if there were 100000 patients in a trial $((1 - 0.99496) * 100000)$ 504 patients would be expected to have a survival time of 8 days. Taking this approach, a dataset was created that represented the WKM survivor function, and a generalised gamma model was fitted to this. It should be noted that there will inevitably be minor inaccuracies in this. A “survivor function” approach was also undertaken for the IPCW method, using the IPCW adjusted HR.

7.5.3 Results of crossover analysis

The results of the analyses are presented in Tables 7.4 and 7.5. Table 7.3 provides a “methods key” describing the approach taken in each of the analyses. Table 7.4 presents results when baseline covariates for ECOG, liver metastases and number of metastatic sites were not included in the analyses, and Table 7.5 presents results when these covariates were included. The results are split in this way because this allows comparable estimates between treatment groups to be compared more easily, under the assumption that if covariates are included in a model to estimate survival in the control group they should also be included in a model to estimate survival in the experimental group, and vice versa. It is noteworthy that when covariates were included in the analysis to estimate survival curves, their mean values were used. As demonstrated in Section 6.6.2 of Chapter 6, this may produce some error in the estimates of mean survival if covariate values are skewed. An alternative approach would have been to apply the crossover adjustment methods including covariates and then to estimate survival curves without incorporating covariates. However this would not have been possible for all methods and so for consistency the “mean covariate value” approach was taken. When estimating mean survival extrapolated models were truncated at 10 years, at which point the survival probability was 0.36% in the experimental group.

Table 7.3: Methods key

Method	Description
ITT	Intention to treat analysis
PP Exclude	Per protocol - exclude crossover patients
PP Cens	Per protocol - censor crossover patients at time of switch
TDCM	Incorporate treatment as a time-dependent covariate
TDCM XO	Incorporate crossover indicator as a time-dependent covariate
RPSFTM “treatment group”	RPSFTM applied on an “treatment group” basis
RPSFTM “on treatment / observed”	RPSFTM applied on an “on treatment” and “on treatment – observed” basis
IPE “treatment group”	IPE algorithm applied on an “treatment group” basis
IPE “on treatment / observed”	IPE algorithm applied on an “on treatment” and “on treatment – observed” basis
IPCW	Inverse probability of censoring weights
Cox	Cox regression model fitted to dataset
GG joint	Generalised gamma model fitted to dataset, with treatment as a covariate. Survival estimated directly from the resulting curves
GG ind	Generalised gamma model fitted independently to treatment arms. Survival estimated directly from the resulting curves
Shrinkage GG joint	Treatment effect applied to crossover patients to shrink survival times, then 'GG joint'
Shrinkage GG ind	Treatment effect applied to crossover patients to shrink survival times, then 'GG ind'
Survivor function	Treatment effect applied to experimental group survivor function
Extrapolation	Extrapolation of counterfactual survival dataset
WKM	Weighted Kaplan-Meier

For the RPSFTM and IPE “on treatment” and “on treatment – observed” analyses it is important to recognise what each analysis presents. The “shrinkage” analyses for these methods involve applying the estimated acceleration factor to shrink the survival times of crossover patients – the “full treatment Vs. no treatment” acceleration factors provided by the “on treatment” analyses are used because we assume that crossover patients receive the treatment until death (in the absence of more complete data). The “survivor function” analyses for these methods represent “on treatment – observed” analyses – the counterfactual survival times for control group patients estimated using the “on treatment” approach are compared to observed experimental group survival times using a Cox model, and the resulting hazard ratio is used within a “survivor function” approach in order to derive mean survival differences. The “extrapolation” analyses for these methods involve extrapolating the control group counterfactual survival times estimated using the “on treatment” approach. Therefore, in Tables 7.4 and 7.5 for each of the RPSFTM and IPE “on treatment / observed” analyses the acceleration factors presented represent “on treatment” acceleration factors (apart from for the “shrinkage” analyses, where AFs associated with the adjusted dataset are presented) and the hazard ratios represent “on treatment – observed” hazard ratios. The mean survival estimates all amount to “on treatment – observed” analyses. Therefore, these are comparable to the respective RPSFTM and IPE “treatment group” analyses. For each of the “on treatment / observed” “survivor function” analyses confidence intervals for the HR applied within the “survivor function” process are test-based – that is, they retain the p-value from the ITT analysis.

Table 7.4: Trial EGF100151 analysis – results when covariates are excluded

Approach	Method (estimation of treatment effect)	Method (estimation of survival)	HR			AF			Survival				Mean Survival Gain
			Mean	95% CI		Mean	95% CI		Control Group			Exp Group	
				LB	UB		LB	UB	Mean	LB	UB		
ITT	Cox	Survivor function	0.862	0.693	1.073	-	-	-	573.630	474.882	697.415	654.723	81.093
	GG joint	-	-	-	-	1.132	0.956	1.340	552.802	466.837	654.460	625.481	72.678
	GG ind	-	-	-	-	-	-	-	545.944	-	-	654.723	108.780
PP Exclude	Cox	Survivor function	0.772	0.616	0.968	-	-	-	521.092	430.302	636.195	654.723	133.631
	GG joint	-	-	-	-	1.240	1.034	1.488	509.269	424.547	610.786	631.321	122.052
	GG ind	-	-	-	-	-	-	-	506.040	-	-	654.723	148.683
PP Cens	Cox	Survivor function	0.819	0.653	1.028	-	-	-	548.566	451.861	671.040	654.723	106.158
	GG joint	-	-	-	-	1.171	0.984	1.393	533.045	448.061	634.055	624.054	91.010
	GG ind	-	-	-	-	-	-	-	528.371	-	-	654.723	126.352
TDCM	Cox	Survivor function	0.809	0.649	1.009	-	-	-	542.821	449.572	660.232	654.723	111.903
	GG joint	-	-	-	-	1.184	0.999	1.403	533.472	450.281	631.944	631.505	98.033
	GG ind	-	-	-	-	-	-	-	528.371	-	-	666.041	137.670
TDCM XO	Cox	Survivor function	0.822	0.655	1.031	-	-	-	549.993	453.000	672.816	654.723	104.731
	GG joint	-	-	-	-	1.171	0.984	1.394	542.463	455.873	645.384	635.336	92.873
	GG ind	-	-	-	-	-	-	-	528.608	-	-	654.723	126.115
RPSFTM "treatment group"	Shrinkage GG joint	-	0.834	0.807	0.870	1.157	1.123	1.186	540.692	528.393	557.305	625.273	84.581
	Shrinkage GG ind	-	-	-	-	-	-	-	534.245	522.910	550.449	654.723	120.479
	RPSFTM	Survivor function	0.785	0.550	1.120	1.178	0.949	1.458	557.162	450.997	689.383	654.723	97.561
	RPSFTM	Extrapolation	0.785	0.550	1.120	1.178	0.949	1.458	514.202	-	-	654.723	140.522
RPSFTM "on treatment / observed"	Shrinkage GG joint	-	0.829	0.797	0.880	1.163	1.114	1.198	538.031	523.601	562.568	625.387	87.356
	Shrinkage GG ind	-	-	-	-	-	-	-	531.749	518.629	555.829	654.723	122.974
	RPSFTM on treat	Survivor function	0.775	0.532	1.128	-	-	-	522.718	382.302	730.109	654.723	132.005
	RPSFTM on treat	Extrapolation	-	-	-	1.228	0.896	1.613	506.587	-	-	654.723	148.136

IPE "treatment group"	Shrinkage GG joint	-	0.832	0.813	0.857	1.159	1.135	1.180	539.796	530.669	550.996	625.305	85.508
	Shrinkage GG ind	-	-	-	-	-	-	-	533.402	524.970	544.161	654.723	121.322
	IPE	Survivor function	-	-	-	1.194	1.023	1.395	549.676	471.190	640.492	654.723	105.048
	IPE	Extrapolation	-	-	-	1.194	1.023	1.395	509.710	-	-	654.723	145.014
IPE "on treatment / observed"	Shrinkage GG joint	-	0.826	0.808	0.849	1.165	1.142	1.185	537.077	528.538	547.504	625.443	88.366
	Shrinkage GG ind	-	-	-	-	-	-	-	530.860	523.041	540.753	654.723	123.863
	IPE on treatment	Survivor function	0.768	0.521	1.133	-	-	-	518.763	375.725	732.867	654.723	135.960
	IPE on treatment	Extrapolation	-	-	-	1.247	1.070	1.453	503.806	-	-	654.723	150.917

Table 7.5: Trial EGF100151 analysis – results when covariates are included

Approach	Method (estimation of treatment effect)	Method (estimation of survival)	HR			AF			Survival			Mean Survival Gain	
			Mean	95% CI		Mean	95% CI		Control Group				Exp Group
				LB	UB		LB	UB	Mean	LB	UB		
ITT	Cox	Survivor function	0.814	0.652	1.016	-	-	-	532.452	445.785	643.029	634.222	101.770
	GG joint	-	-	-	-	1.159	0.990	1.358	526.633	449.682	616.716	610.450	83.818
	GG ind	-	-	-	-	-	-	-	518.776	-	-	634.222	115.446
PP Exclude	Cox	Survivor function	0.754	0.600	0.948	-	-	-	500.147	418.280	605.401	634.222	134.075
	GG joint	-	-	-	-	1.250	1.051	1.486	490.285	412.476	582.698	612.574	122.289
	GG ind	-	-	-	-	-	-	-	479.327	-	-	634.222	154.895
PP Cens	Cox	Survivor function	0.789	0.627	0.992	-	-	-	518.799	432.585	629.733	634.222	115.423
	GG joint	-	-	-	-	1.193	1.010	1.408	513.201	434.705	605.817	612.156	98.955
	GG ind	-	-	-	-	-	-	-	506.823	-	-	634.222	127.399
TDCM	Cox	Survivor function	0.780	0.624	0.974	-	-	-	514.025	431.203	619.777	634.222	120.197
	GG joint	-	-	-	-	1.202	1.025	1.409	564.272	481.248	661.545	677.879	113.607
	GG ind	-	-	-	-	-	-	-	558.026	-	-	711.136	153.110
TDCM XO	Cox	Survivor function	0.783	0.622	0.984	-	-	-	515.656	430.195	625.604	634.222	118.566

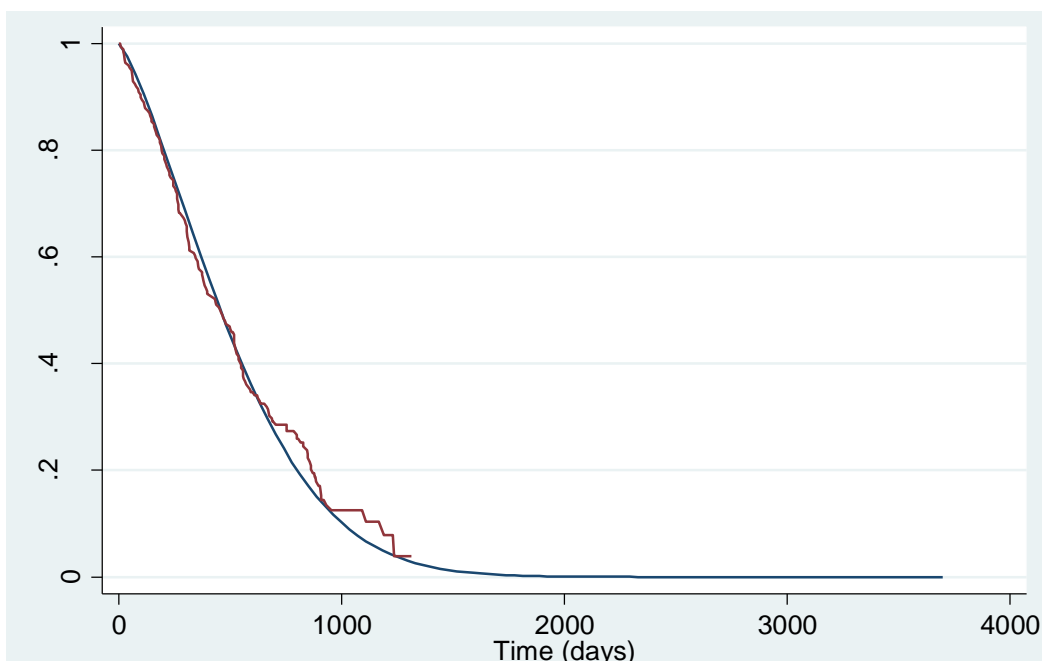
	GG joint	-	-	-	-	1.195	1.015	1.408	573.072	486.672	674.709	684.870	111.798
	GG ind	-	-	-	-	-	-	-	568.787	-	-	634.222	65.435
RPSFTM "treatment group"	Shrinkage GG joint	-	0.783	0.764	0.814	1.188	1.158	1.211	514.204	505.597	527.151	610.993	96.789
	Shrinkage GG ind	-	-	-	-	-	-	-	507.317	499.613	519.263	635.336	128.019
	RPSFTM	Survivor function	0.771	0.587	1.012	1.220	0.993	1.453	521.336	438.266	638.688	635.336	114.000
	RPSFTM	Extrapolation	0.771	0.587	1.012	1.220	0.993	1.453	491.673	-	-	635.336	143.663
RPSFTM "on treatment / observed"	Shrinkage GG joint	-	0.778	0.756	0.816	1.194	1.157	1.223	511.798	501.421	527.762	611.276	99.478
	Shrinkage GG ind	-	-	-	-	-	-	-	505.143	495.956	519.846	635.336	130.193
	RPSFTM on treat	Survivor function	0.726	0.515	1.025	-	-	-	485.431	373.798	647.945	635.336	149.905
	RPSFTM on treat	Extrapolation	-	-	-	1.276	0.984	1.606	487.512	-	-	635.336	147.824
IPE "treatment group"	Shrinkage GG joint	-	0.783	0.767	0.803	1.189	1.168	1.208	514.117	506.870	522.831	611.002	96.885
	Shrinkage GG ind	-	-	-	-	-	-	-	507.238	500.739	515.231	635.336	128.098
	IPE	Survivor function	-	-	-	1.222	1.057	1.412	520.517	450.789	619.964	635.336	114.819
	IPE	Extrapolation	-	-	-	1.222	1.057	1.412	488.461	-	-	635.336	146.875
IPE "on treatment / observed"	Shrinkage GG joint	-	0.778	0.763	0.797	1.194	1.174	1.213	511.828	504.100	520.003	611.272	99.444
	Shrinkage GG ind	-	-	-	-	-	-	-	505.170	499.086	512.617	635.336	130.166
	IPE on treatment	Survivor function	0.727	0.515	1.025	-	-	-	485.543	373.959	647.933	635.336	149.794
	IPE on treatment	Extrapolation	-	-	-	1.276	1.105	1.473	485.899	-	-	635.336	149.437
IPCW	IPCW	WKM	0.750	0.579	0.972	-	-	-	507.490	-	-	634.222	126.732
	IPCW	Survivor function	0.750	0.579	0.972	-	-	-	508.361	409.337	638.207	634.222	125.861

Tables 7.4 and 7.5 show that the treatment effect was estimated to be larger when covariates were included in the analysis. The patterns in differences in the estimated treatment effect generated by the different methods remained similar whether or not covariates were included. In this section the results are discussed with specific reference to the analysis that included covariates, but the discussion also holds for the analysis that excluded covariates.

In the ITT analysis, a Cox regression model produced a HR of 0.814, which when applied to the experimental group hazard function resulted in a mean survival time for the control group of 532 days, equating to a 102 day survival advantage for lapatinib. An ITT analysis using a generalised gamma model that included treatment as a covariate (GG joint) led to a reduced estimated survival gain of 84 days, largely due to a reduction in estimated survival in the experimental group. Independently fitting generalised gamma models to the control and experimental groups (GG ind) resulted in the largest ITT survival gain – 115 days.

It appears that a generalised gamma model fitted independently to the ITT control group may underestimate survival in this group – see Figure 7.10. The AIC and BIC statistics are higher for this model than for the corresponding model for the experimental group (for example, the AIC value increases from 462 for the experimental group model to 497 for the control group model, despite there being more patients in the experimental arm (all things being equal, a higher number of patients/observations leads to a higher AIC). Hence the “GG ind” analysis might not be appropriate.

Figure 7.10: Generalised gamma model fitted to ITT control group data, with covariates



When crossover patients were excluded from the analysis the estimated treatment effect (HR) according to a Cox model increased to 0.754, and the mean survival advantage associated with lapatinib increased accordingly. Due to the theoretical limitations associated with this technique and as shown by the simulations undertaken in Chapter 6, this result is very likely to be biased. However, it may be expected to lead to relatively small bias, considering that treatment crossover was potentially random, and the crossover proportion was small. Censoring crossover patients also led to an increase in the estimate of the treatment effect, with the HR falling to 0.789. In the simulation study reported in Chapter 6 censoring crossover patients exclusively led to underestimates of the treatment effect, no matter the treatment effect or the prognosis of switchers. However, that may have been a result of the simulated treatment crossover mechanism, as relatively large proportions of patients crossed over, and the poorest performers (those who died before 21 days) did not have the opportunity to cross over. Therefore, it cannot be confidently concluded that the “PP Cens” approach underestimated the treatment effect here. Like the exclusion approach though, fairly small levels of bias may be expected given the small proportion of crossover patients, and the apparently random nature of the crossover.

The “TDCM” and “TDCM XO” approaches led to estimates of the treatment effect that lay between the extremes of the ITT analysis and the “PP Exclude” analysis. These approaches would be expected to produce similar results as they both include treatment received as a time-dependent covariate. However while the “TDCM” approach uses information from the crossover patients to inform the treatment effect estimate the “TDCM XO” approach estimates a separate treatment effect relevant exclusively for crossover patients, and estimates the treatment effect in the experimental group without making use of information from the crossover patients. This is the reason that the “TDCM GG ind” approach led to a higher estimate of survival in the experimental group than the “TDCM XO GG ind” approach (signifying that crossover patients had relatively long survival times, particularly when baseline covariates are accounted for).

It is particularly important to note that the “TDCM XO” approach led to treatment effect estimates for crossover patients that were not statistically significantly different from the estimated treatment effect in the experimental group, and in fact the point estimates were very similar: the HR for the crossover indicator (in a model that included other coefficients) was 0.76 (95% confidence interval (CI) 0.48 – 1.20) compared to 0.78 (95% CI 0.62 – 0.98) for the experimental group. This analysis is open to bias as potential time-dependent confounders are not considered, but given that crossover was potentially random and occurred mainly

before disease progression, the associated bias might not be great. This analysis suggests that the “common treatment effect” assumption may be reasonable. In turn, this suggests that estimates of the treatment effect that result from RPSFTM and IPE methods may be prone to low levels of bias. The results of the “TDCM XO” analysis differ slightly from those previously presented by GSK, shown in Table 7.1. This is likely to be due to the fact that in the original dataset provided by GSK 8 patients had missing ECOG baseline data and so these patients would have been excluded from the analyses that included baseline covariates. However, I was able to obtain baseline ECOG data for these patients from a separate data source that provided ECOG data over time, hence in the analyses presented here these patients are not excluded.

Reassuringly, the RPSFTM and IPE “treatment group” methods gave results that were similar – but indicative of a slightly higher treatment effect – to those obtained from the “TDCM” and “TDCM XO” analyses. The acceleration factor obtained from the RPSFTM approach was 1.220, which was equivalent to a HR of 0.771 (based upon the counterfactual dataset compared to the experimental group using a Cox model). Using the “survivor function” method for estimating survival probabilities in the control group the mean survival gain was 114 days. 12.2 days longer than the corresponding ITT analyses (which used a Cox model to estimate the treatment effect, and then applied the “survivor function” approach). Results from the IPE method were very similar.

The RPSFTM and IPE “extrapolation” approaches estimated substantially shorter survival for the control group within the “treatment group” analyses. This follows the pattern observed in the simulation study presented in Chapter 6, where the loss of information due to recensoring led to extrapolated models that had a steeper gradient than is appropriate. For the case study analysis this is illustrated in Figures 7.11 and 7.12.

Figure 7.11 illustrates the Kaplan-Meier associated with the counterfactual dataset estimated by the RPSFTM approach (including covariates), compared to the intention to treat Kaplan-Meier. It appears that relatively little data is lost as the ITT Kaplan-Meier for the control group continues on only marginally from the counterfactual Kaplan-Meier. However, in the observed dataset the final data point was a censored patient at 1313 days. Under the RPSFTM recensoring approach this datapoint is recensored at 1076 days (1313 divided by the estimated acceleration factor) and the final datapoint becomes a death at 1235 days. This recensoring of long-term survivors has an important impact because it means that the counterfactual Kaplan-Meier drops to zero whereas the ITT Kaplan-Meier did not. Figure 7.12 demonstrates the

impact of this, showing the survival curve for the control group extrapolated directly from the counterfactual dataset, compared to the survival curve estimated for the control group using the “survivor function” approach. Clearly the former approach leads to a curve that falls more quickly. This illustrates that even when the treatment effect is relatively small recensoring can be very important for subsequent extrapolations if the data points that mark the end of the Kaplan-Meier change from being censored cases to observed deaths.

Figure 7.11: RPSFTM Counterfactual Kaplan-Meier

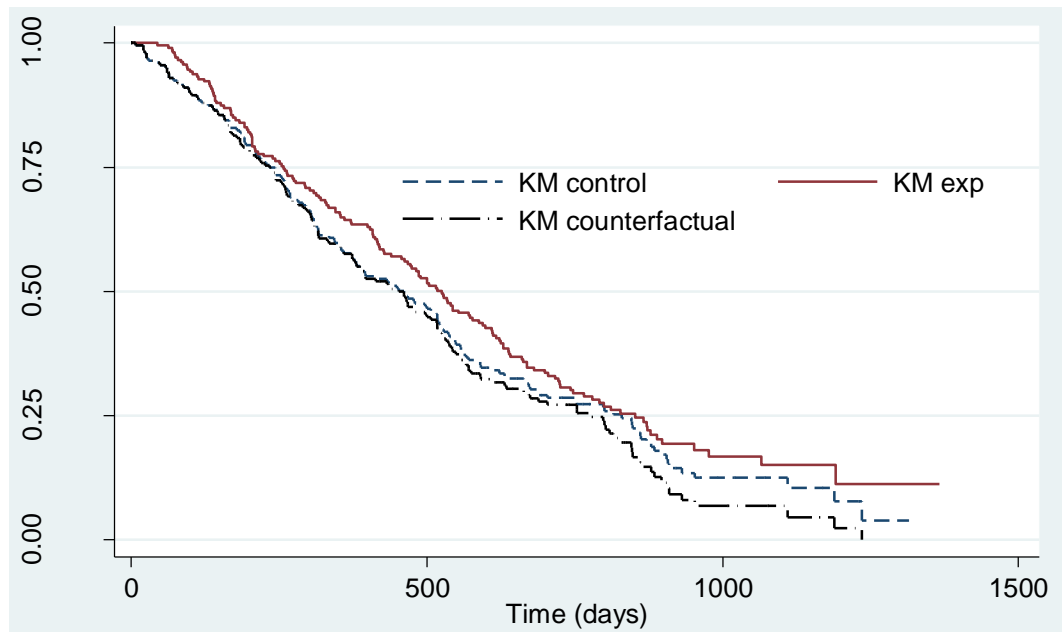
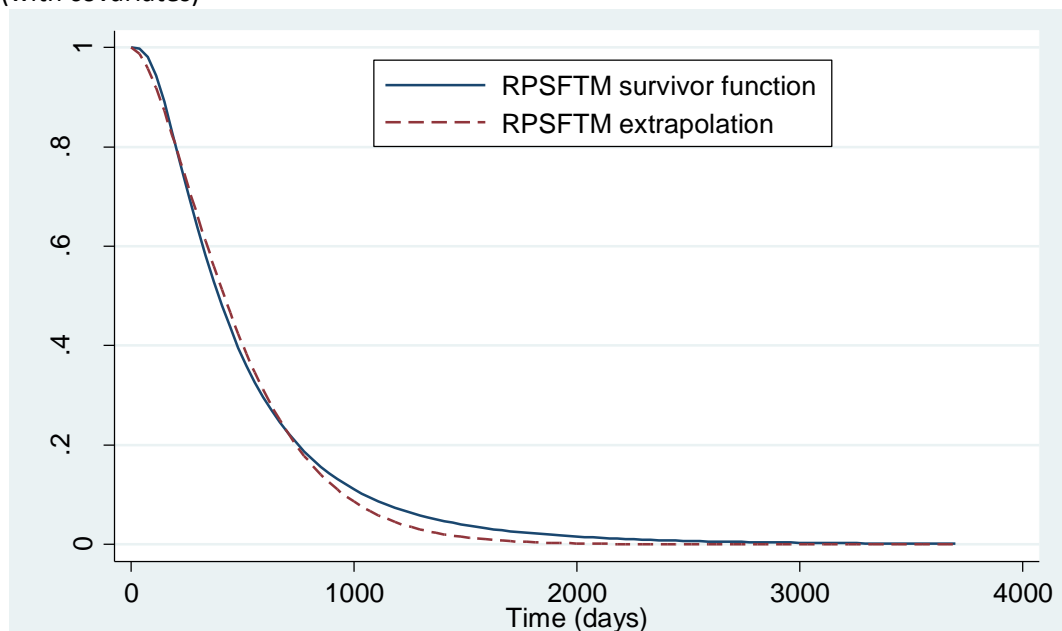


Figure 7.12: RPSFTM “survivor function” approach compared to “extrapolation” approach (with covariates)



The RPSFTM and IPE “on treatment / observed” approaches resulted in slightly lower estimates of control group mean survival, and correspondingly led to slightly higher estimates of the mean survival gain associated with the experimental treatment. This is despite the fact that the survival estimates associated with the “on treatment / observed” analyses were estimated using an “on treatment – observed” approach. This suggests that there is little evidence of a treatment effect that continues beyond treatment discontinuation, and that there is a possibility that survival is slightly worse in experimental group patients than in control group patients after treatment discontinuation (though this may also be attributable to any differences in post-study treatments received after treatment discontinuation).

Alongside the RPSFTM and IPE “on treatment / observed” approaches, the IPCW method produced the most favourable results for lapatinib plus capecitabine, with an estimated HR of 0.75 and a corresponding survival advantage of approximately 126 days. Given the seemingly reliable nature of the RPSFTM/IPE methods in this case it is difficult to recommend the IPCW method, owing to the assumptions that were made in order to apply it – in particular, all crossover patients were assumed to have crossed over prior to disease progression occurring. In the simulation study presented in Chapter 6 the IPCW method consistently overestimated the treatment effect in scenarios where good or average prognosis patients were most likely to switch, and this may be the case for study EGF100151. Given that the IPCW “survivor function” method was relied upon to estimate mean survival in the simulation study due to the difficulties associated with calculating the weighted Kaplan-Meier, it is reassuring that these two approaches gave very similar survival estimates in this case study.

It seems that the RPSFTM and IPE methods provide estimates of the treatment effect and survival gain for lapatinib plus capecitabine that are likely to be associated with low levels of bias. Although these methods are limited by the “common treatment effect” assumption the evidence from the trial seems to suggest that the treatment effect received by crossover patients was similar to that received by experimental group patients. Given this the treatment effect estimates associated with these methods are likely to be most robust. The simulation study presented in Chapter 6 demonstrated that the “survivor function” approach is likely to be optimal for extrapolating over time due to avoiding possible problems in the extrapolation associated with recensoring. Hence, it seems reasonable to conclude that RPSFTM or IPE “survivor function” approaches are most appropriate for estimating mean survival in the EGF100151 dataset. It is relevant to consider whether an “on treatment” or “treatment group” approach should be taken. This is a debatable question which should be addressed on

a case-by-case basis. For study EGF100151, in the absence of treatment discontinuation data the “treatment group” approach is likely to be more robust.

7.6 Impacts on cost-effectiveness

The various approaches for adjusting for crossover gave fairly similar estimates of the treatment effect and survival benefit associated with lapatinib plus capecitabine – for example the difference between the recommended RPSFTM/IPE analysis and the corresponding ITT analysis in terms of a HR was 0.04 (0.77 compared to 0.81), and 12 days (114 day advantage compared to 102). However, relatively small differences in survival benefits can have an important impact on cost-effectiveness results. This can be demonstrated through a simple indicative analysis of the cost-effectiveness results associated with the different crossover adjustment methods applied to the lapatinib case study. This analysis does not represent an attempt to accurately estimate the cost-effectiveness of lapatinib combination therapy compared to capecitabine alone. Instead the aim is simply to demonstrate that even when relatively few patients cross over and the estimated treatment effect is relatively low (as observed in trial EGF100151) the impact of crossover and methods to adjust for it can have important effects on cost-effectiveness estimates.

A utility score of 0.6 was assumed to be experienced by patients throughout the trial, representing an approximate average of the pre-progression and post-progression utility scores used in the suspended NICE appraisal of lapatinib for women with previously treated advanced or metastatic breast cancer.¹⁸³ The mean number of QALYs associated with each treatment group were calculated by multiplying the estimated mean overall survival time (in years) by 0.6 – thus no age effects were included. The incremental QALY gain associated with the experimental treatment was calculated compared to the control group. In a full economic evaluation the beneficial effect of lapatinib combination therapy on progression free survival and any quality of life impacts would also need to be taken into account, but for the purposes of this indicative analysis the QALY gain is based only on OS advantages. The mean incremental cost per patient associated with the experimental treatment was assumed to be £14,015 (taken from the NICE Final Appraisal Determination for the lapatinib technology appraisal¹⁸⁴). It is important to note that using this incremental cost for all the crossover adjustment scenarios is unrealistic – incremental cost estimates are likely to be affected by incremental survival estimates. Thus, it is likely that crossover adjustment methods that produce larger survival gains would also produce larger incremental costs – to some extent counteracting the impacts on the ICER. However the impacts of this are likely to be relatively

minor, as the survival increases are at a time of life at which care is likely to be palliative and relatively inexpensive. Using the incremental cost stated in the lapatinib technology appraisal and the estimated incremental QALY gain based upon the survival estimated by each crossover adjustment method, the mean ICER that would be obtained by each method was calculated. The results of this analysis are presented in Tables 7.6 and 7.7.

Table 7.6: Indicative cost-effectiveness results – methods excluding covariates

Approach	Method (estimation of treatment effect)	Method (estimation of survival)	QALYs		Inc. QALYs	Inc. Cost	ICER
			Control	Exp			
ITT	Cox	Survivor function	0.942	1.076	0.133	£14,015	£105,208
	GG joint	-	0.908	1.027	0.119	£14,015	£117,389
	GG ind	-	0.897	1.076	0.179	£14,015	£78,430
PP Exclude	Cox	Survivor function	0.856	1.076	0.220	£14,015	£63,845
	GG joint	-	0.837	1.037	0.200	£14,015	£69,902
	GG ind	-	0.831	1.076	0.244	£14,015	£57,381
PP Cens	Cox	Survivor function	0.901	1.076	0.174	£14,015	£80,367
	GG joint	-	0.876	1.025	0.150	£14,015	£93,744
	GG ind	-	0.868	1.076	0.208	£14,015	£67,523
TDCM	Cox	Survivor function	0.892	1.076	0.184	£14,015	£76,242
	GG joint	-	0.876	1.037	0.161	£14,015	£87,028
	GG ind	-	0.868	1.094	0.226	£14,015	£61,972
TDCM XO	Cox	Survivor function	0.903	1.076	0.172	£14,015	£81,462
	GG joint	-	0.891	1.044	0.153	£14,015	£91,864
	GG ind	-	0.868	1.076	0.207	£14,015	£67,650
RPSFTM "treatment group"	Shrinkage GG joint	-	0.888	1.027	0.139	£14,015	£100,869
	Shrinkage GG ind	-	0.878	1.076	0.198	£14,015	£70,815
	RPSFTM	Survivor function	0.915	1.076	0.160	£14,015	£87,449
	RPSFTM	Extrapolation	0.845	1.076	0.231	£14,015	£60,714
RPSFTM "on treatment"	Shrinkage GG joint	-	0.884	1.027	0.144	£14,015	£97,665
	Shrinkage GG ind	-	0.874	1.076	0.202	£14,015	£69,378
	RPSFTM on treat	Survivor function	0.859	1.076	0.217	£14,015	£64,631
	RPSFTM on treat	Extrapolation	0.832	1.076	0.243	£14,015	£57,593
IPE "treatment group"	Shrinkage GG joint	-	0.887	1.027	0.140	£14,015	£99,775
	Shrinkage GG ind	-	0.876	1.076	0.199	£14,015	£70,323
	IPE	Survivor function	0.903	1.076	0.173	£14,015	£81,217
	IPE	Extrapolation	0.837	1.076	0.238	£14,015	£58,833
IPE "on treatment"	Shrinkage GG joint	-	0.882	1.027	0.145	£14,015	£96,549
	Shrinkage GG ind	-	0.872	1.076	0.203	£14,015	£68,880
	IPE on treatment	Survivor function	0.852	1.076	0.223	£14,015	£62,751
	IPE on treatment	Extrapolation	0.828	1.076	0.248	£14,015	£56,532

Table 7.7: Indicative cost-effectiveness results – methods including covariates

Approach	Method (estimation of treatment effect)	Method (estimation of survival)	QALYs		Inc. QALYs	Inc. Cost	ICER
			Control	Exp			
ITT	Cox	Survivor function	0.875	1.042	0.167	£14,015	£83,833
	GG joint	-	0.865	1.003	0.138	£14,015	£101,788
	GG ind	-	0.852	1.042	0.190	£14,015	£73,902
PP Exclude	Cox	Survivor function	0.822	1.042	0.220	£14,015	£63,633
	GG joint	-	0.805	1.006	0.201	£14,015	£69,766
	GG ind	-	0.787	1.042	0.254	£14,015	£55,080
PP Cens	Cox	Survivor function	0.852	1.042	0.190	£14,015	£73,916
	GG joint	-	0.843	1.006	0.163	£14,015	£86,217
	GG ind	-	0.833	1.042	0.209	£14,015	£66,968
TDCM	Cox	Survivor function	0.844	1.042	0.197	£14,015	£70,981
	GG joint	-	0.927	1.114	0.187	£14,015	£75,098
	GG ind	-	0.917	1.168	0.252	£14,015	£55,722
TDCM XO	Cox	Survivor function	0.847	1.042	0.195	£14,015	£71,957
	GG joint	-	0.941	1.125	0.184	£14,015	£76,313
	GG ind	-	0.934	1.042	0.107	£14,015	£130,383
RPSFTM “treatment group”	Shrinkage GG joint	-	0.845	1.004	0.159	£14,015	£88,147
	Shrinkage GG ind	-	0.833	1.044	0.210	£14,015	£66,644
	RPSFTM	Survivor function	0.856	1.044	0.187	£14,015	£74,839
	RPSFTM	Extrapolation	0.808	1.044	0.236	£14,015	£59,386
RPSFTM “on treatment”	Shrinkage GG joint	-	0.841	1.004	0.163	£14,015	£85,764
	Shrinkage GG ind	-	0.830	1.044	0.214	£14,015	£65,531
	RPSFTM	Survivor function	0.797	1.044	0.246	£14,015	£56,913
	RPSFTM	Extrapolation	0.801	1.044	0.243	£14,015	£57,715
IPE “treatment group”	Shrinkage GG joint	-	0.845	1.004	0.159	£14,015	£88,059
	Shrinkage GG ind	-	0.833	1.044	0.210	£14,015	£66,603
	IPE	Survivor function	0.855	1.044	0.189	£14,015	£74,305
	IPE	Extrapolation	0.802	1.044	0.241	£14,015	£58,088
IPE “on treatment”	Shrinkage GG joint	-	0.841	1.004	0.163	£14,015	£85,793
	Shrinkage GG ind	-	0.830	1.044	0.214	£14,015	£65,544
	IPE	Survivor function	0.798	1.044	0.246	£14,015	£56,956
	IPE	Extrapolation	0.798	1.044	0.245	£14,015	£57,092
IPCW	IPCW	WKM	0.834	1.042	0.208	£14,015	£67,320
	IPCW	Survivor function	0.835	1.042	0.207	£14,015	£67,786

Again, with the patterns similar between Tables 7.6 and 7.7, the results illustrated for methods including covariates (Table 7.7) are focussed upon here. Most importantly, the methods that are likely to be most appropriate (RPSFTM or IPE “treatment group” “survivor function” approaches) gave ICERs of approximately £74,500 per QALY gained; almost £10,000 lower than the corresponding ITT analysis (ITT Cox “survivor function” approach, ICER = £84,000 per QALY

gained). This illustrates how very small differences in incremental QALY gains (in this case 0.187 compared to 0.167) can lead to relatively big differences in the ICER. The importance associated with the type of analysis undertaken is also highlighted by the large differences in the ICER calculated by the RPSFTM “treatment group” “extrapolation” and “survivor function” approaches – the “extrapolation” approach led to an ICER that was £15,000 lower than that obtained using the “survivor function” approach. These two approaches use the same method to estimate the treatment effect – the difference in survival estimate is purely to do with the method of transforming that treatment effect into a measure of mean survival gain. This highlights the importance of identifying appropriate extrapolation methods, and also demonstrates that great care must be taken when extrapolating from a censored dataset.

In the lapatinib case study it is difficult to predict exactly what the impact of applying crossover adjustment methods to the confounded dataset would be with respect to potential NICE guidance, particularly because I have not completed a detailed economic evaluation. However, it is reasonable to speculate that if lapatinib combination therapy was accepted to meet NICE’s “End-of-Life” criteria, it *may* be regarded as cost-effective based upon an analysis adjusted for treatment crossover (based upon an ICER of approximately £75,000 per QALY gained), whereas this would be very unlikely to be the case if no adjustment was made for crossover (leading to an ICER of approximately £84,000 per QALY gained). The case study also shows that if a naive analysis that excluded crossover patients was relied upon, the ICER would likely be underestimated, which – under “End-of-Life” criteria – could lead to an inappropriate treatment recommendation if NICE’s Appraisal Committee were willing to accept an ICER of approximately £64,000 per QALY gained, but not an ICER of approximately £75,000 per QALY gained.

7.7 Limitations

The case study presented in this chapter has been useful in identifying and illustrating key practical issues associated with the application of crossover adjustment methods in the context of an economic evaluation. However, there are several limitations associated with the case study.

First, the methods that could be applied to the case study dataset were limited due to the data available. However, rather than representing a weakness of the analysis undertaken in this chapter, this is informative because it demonstrates the practical limitations associated with

several of the crossover adjustment methods (in particular those that are observational-based).

Second, although a detailed analysis was undertaken to identify the parametric model most appropriate for extrapolating experimental group survival data, a similar approach was not taken when extrapolating counterfactual survival data for the control group. One reason for this is that guidance included in the NICE DSU Technical Support Document on survival analysis suggests that generally it is likely to be most appropriate to apply the same type of parametric model to different treatment groups within a trial, even if curves are fitted independently to them.¹⁵ This avoids fitting vastly different functional forms to patients who have the same underlying disease. Therefore, if a generalised gamma model is used to model survival in the experimental group a generalised gamma model should also be used to model survival in the control group. However, the analysis presented in Section 7.5.3 suggests that the generalised gamma model may provide a relatively less good fit to the control group counterfactual survival data compared to the experimental group data. This was not addressed further in this case study as such analysis lies beyond the scope of this thesis. However, in a full economic evaluation of the lapatinib trial this would warrant further investigation.

Third, the cost-effectiveness results presented in this chapter are only indicative because they are not based upon a full, detailed economic model. However, for the purpose of demonstrating the potential impact of crossover adjustment methods on cost-effectiveness results the analysis presented is sufficient.

Fourth, the way in which the IPCW WKM was constructed may be prone to small levels of bias. Approaches for reconstructing datasets to reflect Kaplan-Meier curves that are likely to be more robust than the approach taken in this chapter are now available.¹⁵⁵

Finally, a limitation of this chapter is that only one case study is included. It is therefore difficult to generalise about the practical issues associated with applying crossover adjustment methods in an economic evaluation context, or about the likely impact on cost-effectiveness results of using (or not) alternative adjustment methods. However, the chapter remains useful because a range of practical issues were identified, and these led to discussion surrounding the data that are required in order for different methods to be applied, and *how* different methods might be applied. In addition, the impact of adjusting for treatment crossover is highly likely to be case-specific. No amount of case studies would allow such conclusions as “adjusting for treatment crossover using an appropriate method will lead to substantial and important

reductions in the ICER”, because this will always depend upon the crossover proportion and the size of the treatment effect. Hence it could never be concluded that adjusting for treatment crossover will *always* be important. If the crossover proportion is low, and/or the treatment effect is low, and/or the unadjusted ICER is either extremely high or extremely low, adjusting for crossover is unlikely to alter treatment recommendations. However, my case study has shown that even when a relatively small proportion of control group patients cross over and receive what appeared to be a reasonably small treatment effect, adjusting for crossover results in reductions in the ICER that *could* be interpreted as potentially decision-changing. Given that often only small changes in QALY estimates are required to cause potentially important changes in the ICER, it seems likely that adjusting for treatment crossover is likely to often have an important impact on cost-effectiveness results.

7.8 Conclusions

The case study presented in this chapter has demonstrated the application of a range of crossover adjustment methods to a real-world dataset. It has shown that randomisation-based approaches like the RPSFTM and IPE methods are easier to apply and much less data intensive than observational-based approaches. It was not possible to apply the SNM method or a simple two-stage Weibull method to the case study dataset and the IPCW could only be applied under restrictive assumptions. This demonstrates the problems that may be encountered when attempting to apply these methods in a real-world context.

The case study demonstrated that it is not straightforward to apply complex crossover adjustment methods to real-world RCT datasets. Although randomisation-based methods can be applied relatively easily, these can be applied in many different ways and an analyst must be able to appropriately select suitable methods. This requires an understanding of the theoretical foundations of the alternative methods and so there is a risk that analysts could successfully apply an inappropriate method. This represents a danger that I will seek to negate through the recommendations made in Chapter 8, but the fact remains that a good understanding of the theoretical characteristics of the alternative methods is required if the methods are to be applied appropriately. In addition, some appreciation of the underlying disease processes would improve the likelihood that an appropriate modelling approach is chosen – thus there is a clear need for collaboration with clinical experts when these analyses are undertaken. The case study also led me to consider that in situations where the required data are available for observational-based methods to be applied, their application could be very time-consuming and difficult as the required data are likely to be stored in different

databases, would have to be pooled, cleaned and possibly rearranged. For analysts unfamiliar with the requirements of the crossover adjustment methods, this may provide scope for error.

The case study also demonstrated that it is important to analyse the trial data in detail in order to help identify which crossover adjustment methods are likely to be appropriate. Alternative methods can lead to diverging survival estimates that – even if the differences are small – can lead to important variations in cost-effectiveness results. In some situations alternative crossover adjustment methods could lead to changes in cost-effectiveness results that could lead to different treatment recommendation decisions. Although not the prime focus of this thesis the case study also shows that it is extremely important to identify appropriate extrapolation techniques, as different approaches can cause even more significant changes to the ICER. This has direct relevance for the crossover adjustment methods because the alternative methods are amenable to different extrapolation approaches. The mean survival differences associated with RPSFTM and IPE “survivor function” and “extrapolation” approaches led to substantial differences in ICERs, even though the estimated treatment effect was relatively small and the crossover proportion was low. This clearly highlights the importance for extrapolation of the information loss associated with recensoring. Hence the case study shows that issues previously discussed as theoretical issues throughout Chapters 4, 5 and 6 – such as data availability, approaches for applying adjustment methods, recensoring and approaches for performing extrapolation – are practically important.

Chapters 6 and 7 have assessed the performance of crossover adjustment methods in a high level of detail, using simulated scenarios and a real-world dataset. Together they complete Part 4 of this thesis. The findings of these chapters are wide-ranging and help to answer the research questions set out at the beginning of this thesis. Chapter 8 draws upon these findings in order to make recommendations on the use of crossover adjustment methods.

Chapter 8

Overview, recommendations, discussion, conclusions and future research priorities

8.1 Chapter overview

The purpose of this chapter is to draw together the findings of all previous chapters in order to make coherent recommendations regarding the use of crossover adjustment methods in the context of economic evaluation. The research questions are addressed and the contributions made by this thesis are outlined. Outstanding issues are discussed and future research priorities are considered.

In Section 8.2 I provide an overview of the thesis, summarising the findings of each chapter and part. In Section 8.3 I include some new material based upon discussion that arose when I presented my findings to a group of oncologists. This is valuable, as it sheds further light on practical issues associated with the application of crossover adjustment methods. This adds to a discussion on the findings of the thesis that is presented in Section 8.4. Recommendations on the use of crossover adjustment methods are given in the form of an analysis framework in Section 8.5 and recommendations for further research are presented in Section 8.6. Conclusions on the thesis as a whole are given in Section 8.7.

8.2 Thesis overview

This thesis is made up of five parts. Part 1 consisted of Chapters 1 and 2, and described the treatment crossover problem in the context of economic evaluation. Part 2 (Chapter 3) investigated how important the treatment crossover problem was in a practical, health technology assessment context. Part 3 included Chapters 4 and 5 and sought to identify methods that are potentially appropriate for adjusting for treatment crossover, given an economic evaluation context. Part 4 (Chapters 6 and 7) then investigated the performance of these methods in a simulation study, and in a real-world context. The fifth and final part of this thesis consists of the present chapter, in which I discuss my findings and consider what should be done in future to address the treatment crossover problem, in an economic evaluation context.

In Chapter 1 I discussed the theoretical and practical motivations for the thesis. I described why economic evaluation is required and the decision problem faced by resource allocation decision-makers. I then discussed the role that survival analysis plays in economic evaluation and introduced the treatment crossover problem. I introduced a range of treatment crossover adjustment methods and highlighted that research was required in order to better understand the suitability of these. Based upon this, I outlined the research question that formed the focus of my thesis. It asked which methods for dealing with treatment crossover are most appropriate in a range of different (and relevant) scenarios, given that a state of the world in which the novel intervention exists must be compared to one in which it does not exist. Linked to this was the hypothesis that crossover adjustment methods commonly used in health technology assessments are likely to lead to biased estimates of the treatment effect and therefore biased cost-effectiveness results. In addressing these questions a key aim of the thesis was to set out guidelines making recommendations regarding which crossover adjustment methods are likely to be appropriate in a range of different circumstances. The remainder of the thesis was set out to answer these questions and to achieve this aim.

In Chapter 2 I outlined the importance of addressing the treatment crossover problem with reference to the theoretical framework of health economic evaluation. The point was made that assessing the size of the treatment effect correctly, in the presence of crossover, was as important in a Welfarist context as in an Extra-Welfarist context. In fact, the implications of my thesis are broader than economic evaluation alone, since I address the problem of estimating the “true” clinical effectiveness of an intervention in the presence of treatment crossover. This is important, but evidence suggests that treatment crossover creates more problems for obtaining reimbursement for a therapy than for obtaining a licence. This is because regulatory agencies such as the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) recognise the difficulties associated with obtaining long-term overall survival data and accept that data on disease progression outcomes represent acceptable primary endpoints for drug approval.^{4;5} Hence, when treatment crossover is permitted after disease progression (as is often the case), the problems created are much less for regulatory agencies than for HTA authorities such as NICE, who require overall survival to be included within an economic evaluation. Hence, this thesis focused on the treatment crossover problem in the context of the economic evaluation decision problem.

The aim of Chapter 3 was to demonstrate the importance of the treatment crossover problem. This was achieved by conducting a review of recently completed NICE technology appraisals in the metastatic and/or advanced cancer disease area. This highlighted that the treatment

crossover problem occurred regularly in these appraisals (25 (55.6%) of 45 reviewed appraisals were affected by treatment crossover). In addition, the review highlighted that treatment crossover has been consistently inappropriately addressed in NICE technology appraisals, with naive censoring or exclusion approaches commonly used. Complex methods for adjusting for treatment crossover were seldom used. The review also demonstrated that inappropriately addressing the treatment crossover problem may have affected the treatment recommendations made in a number of appraisals. Chapter 3 is novel and important because it clearly demonstrates the extent to which treatment crossover is a problem in health technology assessment – such evidence on the problem has not previously been published.

After highlighting the extent of the treatment crossover problem, I attempted to identify potential solutions to the problem in Chapters 4 and 5. I presented a systematic review of crossover adjustment methods in Chapter 4. Relevant methods were described in detail with particular attention paid to their theoretical characteristics, assumptions and limitations. Two main branches of methods were found – those that have observational-analysis origins, and those that have been developed specifically for a randomised controlled trial context, which make use of the randomisation assumption. The review demonstrated that observational-based methods – such as Structural Nested Models with g-estimation, and inverse probability of censoring weights – make very different assumptions to randomisation-based approaches – such as the Rank Preserving Structural Failure Time Model and Iterative Parameter Estimation. Whereas observational-based methods rely on a high level of data availability and being able to model the treatment crossover probability, randomisation-based methods require a much lower amount of data but are instead reliant on an “common treatment effect” assumption. Due to these significant differences between the potentially relevant crossover adjustment methods they are likely to be appropriate (and associated with low levels of bias) in quite different circumstances. Chapter 4 is novel and important because a systematic review of methods potentially appropriate for adjusting for treatment crossover in an RCT and economic evaluation context has not previously been published. In addition, it introduces a novel “two-stage” method that has not before been used to adjust for treatment crossover.

Having identified crossover adjustment methods that were relevant for further consideration in Chapter 4, in Chapter 5 I examined these methods further, focussing upon how they might be combined with extrapolation methods. This is novel – previous research into crossover adjustment methods in the context of economic evaluation have not considered extrapolation.²¹ This is important because extrapolation is generally required for economic evaluation due to the limited follow-up typically observed in clinical trials. Therefore to

identify the most appropriate crossover adjustment methods for use in an economic evaluation context both their performance (and theoretical underpinnings) with respect to satisfactorily adjusting the treatment effect to take into account treatment crossover, *and* their amenability to being combined with suitable extrapolation approaches have to be considered. I showed that while the treatment crossover adjustment methods are generally amenable to combination with extrapolation techniques the ability to follow the survival model selection process algorithm advised by the NICE Decision Support Unit (and constructed by me as part of the research that led to this thesis) is restricted somewhat.

I showed that this was partly due to the difficulties associated with assessing the suitability of key assumptions such as proportional hazards in the presence of treatment crossover, which harms the ability of the analyst to justify fitting parametric models independently to each treatment group, or jointly. However Chapter 5 demonstrated that other problems exist due to the specific crossover adjustment methods themselves. For example, the recensoring associated with RPSFTM and IPE methods involves a loss of information which may harm extrapolation attempts. The IPCW does not provide an adjusted dataset, which makes fitting parametric models independently problematic. Even taking a simple proportional hazards modelling approach is complicated by the fact that the IPCW uses a Cox model rather than a parametric model, and there is no baseline control group survival curve to which one could apply the estimated treatment effect. In Chapter 5 I suggested solutions to these problems – including the application of the inverse of the IPCW HR to the experimental group hazard function in order to derive the control group hazard and survivor functions – although in some cases these were associated with theoretical limitations.

I introduced three main methods that could be combined with various of the crossover adjustment methods in order to undertake extrapolation – a “shrinkage” approach which could be followed by any extrapolation approach (but which had theoretical limitations) and could be applied to RPSFTM, IPE and SNM methods; a “survivor function” approach similar to proportional hazards/constant acceleration factor modelling that could be applied to RPSFTM, IPE and IPCW methods; and an “extrapolation” approach similar to an independently fitted parametric modelling approach which could be applied to RPSFTM and IPE counterfactual datasets (that are subject to recensoring) and the IPCW weighted Kaplan-Meier (provided a dataset is “created” to reflect the IPCW WKM).

In Chapter 6 I designed and implemented a simulation study specifically to address the research question posed in Chapter 1, and to assess the performance of relevant crossover

adjustment methods identified in Chapter 4 across a range of realistic scenarios. Naive methods were also included so that these could be compared to the more complex, potentially more appropriate methods. A simulation study was required because this allowed the bias associated with the methods to be compared given a known truth. The simulation study was necessarily complex, due to the inclusion of a time-dependent covariate and a time-dependent treatment effect, which allowed the data and the crossover mechanism to be simulated in a realistic way. This added complexity allowed the simulation study to build upon a previous simulation study in this area, upon which I was a co-author.²¹ In addition, my simulation study improved upon previous research by including a more comprehensive range of relevant methods.

The simulation study showed that all the crossover adjustment methods had important limitations that led to substantial bias in certain scenarios. As expected, naive methods (such as simple censoring and exclusion approaches) produced high levels of bias consistently across all scenarios. More complex randomisation-based methods produced very low levels of bias when the “common treatment effect” assumption held, but when this assumption was violated in the simulated datasets bias became progressively higher. The observational-based methods generally produced relatively moderate levels of bias no matter what happened to the treatment effect over time, except when treatment crossover proportions were extremely high (approximately 90% in a control group made up of 250 patients). This is unsurprising given the observational origins of these methods – it is noteworthy that in key methodological papers identified in Chapter 4 that applied SNM and MSM methods to datasets these were usually very large compared to what can be expected in an RCT. For example, Mark and Robins (1993) applied an SNM to a dataset that included 12,866 participants,¹³⁴ Wittelman *et al* (1998) analysed a dataset with 4,404 subjects;¹⁴⁷ and Fewell *et al* (2004) applied a MSM to a dataset with 1,240 patients and over 91,000 person-months of observations.¹⁷⁵ This has important implications for the use of these methods in practice, in an RCT context. These findings also confirmed those of other researchers – Howe *et al* found that IPCW was prone to bias in small samples, if selection bias was very strong, and if there were unmeasured confounders.¹²⁸ The bias I found was even more important, which is likely to be due to the complex data generation and crossover mechanism, particularly when there were very high levels of crossover.

In some scenarios, in which the treatment effect received by crossover patients was substantially different from that received by experimental group patients, and in which less than 90% of control group patients crossed over, the observational-based methods produced

least bias. However, often in scenarios such as these a standard intention to treat analysis was optimal, due to the low benefit received by crossover patients. Importantly, the simulation study also demonstrated that a simple two-stage Weibull method may consistently adjust for crossover with low levels of bias if the treatment crossover occurs exactly as defined in my data generating mechanism. Hence Chapter 6 provided important evidence on the use of crossover adjustment methods, and this is used to inform analytical recommendations made later in this chapter.

In Chapter 7 I demonstrated how treatment crossover adjustment methods could be applied to a real-world dataset, and what impact the different methods can have both on survival estimates and cost-effectiveness estimates. This helped identify practical problems associated with several of the methods. In particular two-stage SNM and Weibull methods were not applicable, and the IPCW could only be applied under restrictive assumptions which the data suggested were false (although, the bias associated with this may have been fairly small). Randomisation-based methods were comparatively easy to apply and analysis of the trial dataset suggested that these were likely to lead to low levels of bias. However, the case study demonstrated additional issues with the different methods that feed into the recommendations made later in this chapter. In particular, there are different approaches that can be taken when applying randomisation-based methods, and these can lead to importantly different cost-effectiveness results. Also, it was demonstrated that it may often not be practical to apply observational-based approaches and it may be problematic to justify the use of randomisation-based methods through standard analyses of the dataset. The case study also confirmed the findings of the simulation study regarding the importance of choosing an appropriate extrapolation approach in combination with the treatment crossover adjustment method.

8.3 Clinical opinion

Following completion of the simulation study reported in Chapter 6, I presented my research to oncologists at Weston Park Hospital in Sheffield, one of only three dedicated cancer hospitals in the UK. I used this opportunity to obtain the opinion of these clinical experts on several of the key assumptions that underpin the observational-based and randomisation-based crossover adjustment methods. While this session did not represent a formal interview process, it did allow me to obtain clinical opinions on key parts of my research.

First, I asked the clinicians about the plausibility of the “common treatment effect” assumption that is central to RPSFTM and IPE methods. I asked whether patients who receive treatment after disease progression are likely to receive the same treatment effect as patients who received that treatment prior to disease progression. The clinicians could not make generalisations and factors such as the specific drug, disease, patient-type and previous treatment were key. However, the consensus was that generally a lower treatment effect would be expected in patients with later stage disease. Although this represents weak “evidence”, incorporating such knowledge – combined with other scientific knowledge on the mechanism of action of the drug in question – in a systematic approach to identifying an appropriate crossover adjustment method is important. This also casts doubt on the general validity of the “common treatment effect” assumption.

I then asked the clinicians for their thoughts on the likelihood that a treatment crossover mechanism observed in a clinical trial could be accurately modelled, given the availability of baseline and time-dependent values of key prognostic covariates. The oncologists made several important comments that should be taken into account when considering the use of observational-based crossover adjustment methods. Firstly, it was suggested that where crossover is included in the trial protocol it may be helpful to specify which types of patients should be considered for crossover. These details would be very valuable in determining the crossover mechanism and it is likely that, given the availability of key data, treatment crossover could be modelled. However, often such specification of patients is not given in trial protocols (it was not for the lapatinib case study reported in Chapter 7) and in these circumstances there are several reasons why modelling the treatment crossover mechanism is likely to be problematic. Even when crossover-eligible patients are specified in the protocol there could be problems in applying observational-based methods, because if any prognostic patient characteristics mean that a patient definitely will or definitely will not cross over, these characteristics could not be incorporated within an observational-based crossover adjustment analysis – as discussed in Chapter 4.

In the absence of information on crossover-eligibility criteria in the trial protocol, the clinicians suggested a number of reasons why modelling the crossover mechanism may be difficult. For new cancer drugs there is likely to be significant between-doctor variability regarding which control group patients should be deemed appropriate for crossover. For such novel therapies there has not been time for clinical consensus to be reached and therefore some doctors may decide that patients with poor prognosis (who perhaps have the most to gain) should cross over, whereas others may decide that patients with good prognosis (who may find the new

treatment more tolerable) should cross over. This problem is likely to be exacerbated in multi-centre and multi-national trials where different groups of doctors may form different opinions based upon their initial experiences of using the novel therapy. Hence, while the treatment crossover mechanism is unlikely to be random, it may be very difficult to model, particularly in relatively small RCT datasets. In addition, the clinicians noted that, while data on prognostic covariates is likely to be important for the modelling of the treatment crossover mechanism, data on patient opinion are not collected in clinical trials and may be critical to the crossover decision. For example, a doctor may decide that a patient has good prognosis and should switch treatments, but the patient may refuse. Without the inclusion of a patient opinion indicator in a model of the crossover mechanism the resulting model of the crossover process is likely to be prone to error and will subsequently lead to biased estimates of the adjusted treatment effect.

8.3 Discussion

In this thesis it has been clearly demonstrated that no currently existing methods for adjusting for treatment crossover can be expected to consistently lead to low levels of bias across a range of realistic scenarios. Nevertheless, the research presented here contributes importantly to knowledge in this area, and also allows recommendations to be made that – if followed – will help increase the chances that an appropriate method will be chosen to address treatment crossover in individual economic evaluations. The thesis also addresses the issues associated with combining treatment crossover adjustment methods with extrapolation methods that are usually required within economic models – an area neglected by previous research.

A key message from my analyses relates to the critically important distinctions between observational-based crossover adjustment methods and randomisation-based methods. My experience of the use of these methods in economic evaluations suggests that analysts, reviewers and decision-makers alike do not understand them well. I presented my work at the NICE Guide to Methods update committee meeting that discussed treatment crossover and described the alternative methodologies, and it was quite clear that the RPSFTM and IPCW methods were both considered as complex methods that were appropriate for adjusting for treatment crossover. However there was very little knowledge on how these methods differ and in what circumstances they may or may not be appropriate. This highlights the importance of my research and of disseminating the results.

Observational-based methods such as the SNM with g-estimation and IPCW have potentially strong advantages. While randomisation-based methods retain the p-value associated with the ITT analysis and therefore have confidence intervals that may be regarded as inefficiently wide, observational methods may produce more precise estimates. An SNM applied in a two-stage approach, and the IPCW method also make no assumption about the relationship between the treatment effect received by crossover patients and patients initially randomised to the experimental group. This means that neither method is adversely affected if the “common treatment effect” assumption does not hold. This is a very important advantage of these methods, particularly in the context of crossover in clinical trials in the oncology setting, in which typically crossover is only permitted after disease progression. My discussions with clinical experts, described in Section 8.2, provides further support for the implausibility of the “common treatment effect” assumption.

However, the observational-based methods have important disadvantages that are clearly highlighted by my simulation study, and real-world case study. While they have the potential to produce precise estimates of the adjusted treatment effect, these may not come to fruition if the available dataset is relatively small. Observational datasets are typically much larger than RCT datasets, with many more events occurring. In my simulations of a 500-strong sample the IPCW method produced confidence intervals that were generally similar in size to those produced by randomisation-based methods. Similar was true in the lapatinib case study (in which the sample size was 408).

In addition, the flexibility of the observational-based methods with respect to the relationship between the treatment effect received in crossover and experimental groups comes at the cost of a requirement that it must be possible to accurately model the treatment crossover decision mechanism. This is essential for the methods to produce low bias, and is reliant both upon the “no unmeasured confounders” assumption, and the ability to correctly specify appropriate models. The lapatinib case study demonstrated that in practice it might not be possible to apply the observational-based methods – both due to incompatible treatment crossover mechanisms that may prevent the application of two-stage methods (for example, if the treatment crossover mechanism is not rigid in terms of only permitting crossover in patients that have experienced disease progression); and due to insufficient amounts of data availability (if due to censoring or loss to follow-up details on key covariates at the time of crossover are unknown). In addition, my discussions with clinical experts (described in Section 8.2) suggested that in practice it may be very difficult to model the crossover decision due to

the novelty of new interventions, and due to patient preferences not being captured in clinical trial datasets.

My simulation study showed that even when there were no unobserved confounders the observational-based methods typically led to moderately high levels of bias, and extremely high levels of bias when the crossover proportion was very high. Although the IPCW method performed much better than a naive censoring approach (which the IPCW would reduce to if all confounders were unknown), substantial bias remained. This is likely to be due to the relatively small number of patients in the control group who did not cross over, which makes it difficult to model the crossover mechanism accurately. Even in scenarios where only around 60% of control group patients crossed over, my simulation study suggested that the observational-based methods struggled to model the probabilistic nature of the crossover decision. In larger datasets it may be expected that this problem would reduce, but my simulations suggested that a sample size of 500 patients was not sufficient to ensure that observational methods could model the treatment crossover process accurately.

Theoretically, observational-based crossover adjustment methods have important advantages, but they are prone to important bias when applied to relatively small RCT datasets, and may be particularly difficult to apply appropriately in a real-world setting. Randomisation-based methods benefit from the fact that they were developed specifically to be applied to RCT datasets and provided randomisation has been performed adequately, they will produce low bias, if their other assumptions are satisfied. This was demonstrated in my simulation study, where the RPSFTM and IPE methods produced very low levels of bias in scenarios in which the treatment effect did not differ between the experimental group and crossover patients.

However, the potential strengths of the observational methods represent the key weaknesses of the randomisation-based methods. The RPSFTM and IPE methods rely upon the “common treatment effect” assumption and my simulation study demonstrated that when this assumption does not hold the methods will produce substantial bias. It is therefore critical to assess the justifiability of the “common treatment effect” assumption when considering the application of these methods, but in practice this is difficult due to the confounding caused by the crossover in the observed dataset. Time-dependent Cox models or similar can be fitted to estimate the treatment effect in the experimental group and the crossover group separately, and the resulting estimates can be compared, but such analyses are subject to bias in the presence of time-dependent confounders. Despite this, such analyses may remain of use when determining whether there are any obvious differences in the effects received by the

different groups. An alternative and possibly preferable approach would be to investigate the treatment effect in patients randomised to the trial in question at different stages of disease. However, in metastatic cancer trials this may often not be possible – for example in the lapatinib trial analysed in Chapter 7, 15 (8%) patients were randomised with stage IIIb or IIIc disease, with the remaining 92% simply classed as entering the trial with “metastatic” disease.¹⁷⁷ Hence the “common treatment effect” assumption may often be very difficult to test, and external data from other trials that may have included patients with different stages of disease, or clinical expert opinion and arguments based upon the biological process of the treatment action should be used to provide further evidence, where possible.

As stated in Chapter 4, various authors have attempted to apply multi-parameter versions of the RPSFTM, in order to allow a relaxation of the “common treatment effect” assumption.^{123;129;133;134;138} However, relying solely on the randomisation assumption to allow two different treatment effects to be estimated for different groups has proven unsuccessful, with meaningful point estimates difficult to determine. Hence this is an outstanding problem with randomisation-based methods.

If randomisation-based methods such as the RPSFTM and IPE algorithm are to be applied in practice, consideration must be given to *how* the method should be applied. As shown in Chapter 7, whether an “on treatment”, “on treatment – observed” or an “treatment group” approach is taken can have an important impact on the adjusted treatment effect and corresponding cost-effectiveness results. My conclusions are that an “treatment group” approach may offer a truer analysis of the effect associated with being randomised to the treatment group, which is what is required within an economic evaluation. However, if the treatment effect post discontinuation is minimal, and if the period between discontinuation and death is longer in the experimental group than in crossover patients, this approach may under-adjust for treatment crossover and therefore may lead to underestimates of the cost-effectiveness of the new treatment. On the other hand, “on treatment” and “on treatment – observed” approaches estimate the effect of treatment under the assumption that the effect disappears as soon as treatment is discontinued. Whether this is a reasonable assumption is essentially a clinical question, but this approach is also hindered by the fact that trial data may not be available on discontinuation times in crossover patients – and so while treatment discontinuation is taken into account for the experimental group it would not be for crossover patients. Therefore available data should also be taken into account when deciding which approach to take.

An extra “disadvantage” of randomisation-based methods might be perceived to be the fact that the ITT p-value is maintained. Analysts, drug manufacturers and even decision-makers may find this frustrating, as crossover adjustment may lead to an increased acceleration factor (or a reduced HR), but confidence intervals that are wide and which may not indicate a statistically significant effect. White *et al* (1999) discuss this and suggest that the fact that these methods respect the randomisation of the trial is actually an important advantage, because bias associated with observational inferences is avoided.¹²³ Crossover adjustment does not provide any new evidence, it is simply a way in which counterfactual results can be estimated – therefore it is theoretically appropriate that a non-significant ITT result should not become significant once crossover adjustments are made.

Recensoring represents an additional problem associated with randomisation-based methods, particularly in the context of extrapolation and economic evaluation. In circumstances where the treatment effect is time-dependent (even if the same average effect was received in the experimental group and crossover patients) recensoring is likely to result in biased estimates, particularly if the treatment effect is relatively large which causes recensoring to have a larger impact. Chapter 7 demonstrated that even small over- or under-estimates of the acceleration factor due to recensoring can be transformed into important under- or over-estimates of control group survival after extrapolation, which can cause cost-effectiveness estimates to change substantially. Even if the treatment effect is not time-dependent and the RPSFTM or IPE adjusted treatment effect is unbiased, the recensoring can create problems for extrapolation due to the loss of important longer-term data – particularly if parametric models are fitted independently to the counterfactual survival data, rather than undertaking extrapolation using a “survivor function” approach (applying the adjusted treatment effect to a parametric model fitted to the unconfounded experimental group data).

Extrapolation represents an additional difficulty for treatment crossover adjustment analyses. As well as recensoring issues, combining independently fitted parametric models and proportional hazards modelling techniques with treatment crossover adjustment method outputs is not always straightforward, as discussed in Chapter 5. Methods of analyses advocated by the NICE DSU Technical Support Document on survival analysis¹⁵ can be undertaken fairly simply in combination with two-stage SNM and Weibull methods, but the resulting extrapolation is prone to bias if the crossover adjustment is biased. Fitting a parametric model to an IPCW weighted Kaplan-Meier curve requires “re-creating” the WKM dataset, and “proportional hazards modelling” (as defined by the NICE DSU TSD) requires a slightly more complex approach than simply applying a hazard ratio to the hazard function

associated with a baseline survival curve. Instead, the inverse of an adjusted treatment effect (HR or acceleration factor) must be applied to an independently fitted experimental group hazard function or the survival times associated with the survivor function, and this may lead to some bias as the estimated treatment effect is not related to the independently fitted experimental group curve.

Importantly, my simulation study clearly demonstrated that naive methods for adjusting for treatment crossover – such as censoring or excluding crossover patients, or incorporating treatment received as a time-dependent covariate – were consistently associated with high levels of bias across all scenarios. Although there are important disadvantages associated with the more complex observational and randomisation-based methods and these methods were often associated with significant bias themselves, the naive methods consistently performed even more poorly. This supports the hypothesis stated in Chapter 1 that methods commonly used in NICE appraisals are likely to have led to biased survival estimates, cost-effectiveness results, and potentially inappropriate resource allocation decisions. As stated in Section 3.7 it is difficult to assess whether treatment recommendations would have changed if more appropriate crossover adjustment methods had been used in the NICE appraisals in which naive methods were used – this could only be accurately assessed if more appropriate analyses were run and the resulting ICERs were obtained. However, given that naive methods were used in several appraisals, and that in these appraisals there were a mixture of positive and negative treatment recommendations, it is clear that the use of highly biased methods may have caused inappropriate decisions to have been made. It is important to attempt to avoid the potential for such failures in decision making in the future by learning from the research presented in this thesis, and to that end Section 8.4 presents analytical recommendations regarding the use of crossover adjustment methods.

It is important that a simple two-stage Weibull (note that any other parametric accelerated failure time model could also be used) method that I included in my simulation study successfully produced least bias in a wide range of scenarios. Although this method does not represent a highly sophisticated approach for undertaking crossover adjustment, it may be particularly successful in circumstances in which crossover occurs as modelled in my simulations. The method partially controls for time-dependent confounders in a simple manner, by making use of a secondary “baseline” after which treatment crossover is permitted to occur. If crossover always occurs very soon after this secondary baseline very little time-dependent confounder bias would be expected. The method is not restricted by assumptions about the commonality of the treatment effect, but is reliant on suitable data on important

prognostic factors being available (that is, “no unmeasured confounders”) at the time of the secondary baseline. It is not possible to apply the method if there is no secondary baseline (for example, if crossover happens before disease progression for some patients), and it is likely to lead to substantial bias if crossover is permitted to occur large amounts of time after the secondary baseline. However, given that in the context of metastatic oncology RCTs crossover is often only permitted after progression, and that it seems more likely that crossover would occur sooner rather than later after progression if it is to occur at all, the two-stage Weibull method is worthy of consideration when selecting crossover adjustment methods.

Finally, the review conducted in Chapter 3 showed that it might not always be necessary to adjust confounded trial data to account for treatment crossover, rather external datasets might be available that could be used instead.⁶² This is a potentially valid approach, but is heavily reliant on data availability and the existence of suitable external datasets – because it is unlikely to represent a generalisable method, this type of approach was not considered further in this thesis.

8.4 Recommendations

Based upon the research contained within this thesis I have formulated the following recommendations regarding the use of crossover adjustment methods in the context of economic evaluation. These are in the form of an analysis framework – presented in Figure 8.1 – that may be used when the treatment crossover problem is being addressed. That I have been able to formulate such recommendations demonstrates that I have answered the research question set out in Chapter 1, and have met the aims specified in that chapter. Firstly though, important recommendations for clinical trialists are made.

8.4.1 Recommendations for trialists

The collection of appropriate data is extremely important if treatment crossover adjustment methods are to be applied. Going forward it would be extremely useful if, when it is planned that treatment crossover will be permitted, clinical trialists ensured that suitable data on baseline and time-dependent covariates were collected in order to allow all crossover adjustment methods to be applied at a later date. It would be particularly valuable if it was required that clinicians recorded the reason for whether or not an eligible patient was crossed over, including an insight into patient preferences. In combination with this, trialists should ensure that data collection is not stopped prematurely (i.e. before the point of treatment crossover). Combined, this may enhance the likelihood that the crossover mechanism could

be successfully modelled, which may allow the observational-based adjustment methods to produce lower levels of bias. This would lead to a higher probability that a suitable adjustment method could be identified on a case-by-case basis.

8.4.2 Recommendations for analysts

Step (1) involves assessing the treatment crossover mechanism using the trial protocol and the observed dataset. This should demonstrate whether and which adjustment methods are potentially applicable. For instance, it may become apparent whether data on relevant switching indicators were collected. The time at which patients became able to switch treatments is also important to determine, as this helps identify whether two-stage methods are likely to be applicable.

For Step (2), the crossover proportion amongst control group patients who became at-risk of crossover should be assessed. If this is greater than 90% the IPCW and SNM methods are highly prone to bias, given a sample size in the region of 500. This is likely to be the case for most cancer clinical trials, since sample sizes are rarely larger than the size of 500 (250 in each arm) tested in my simulation study. It is likely that the sample size would need to be substantially greater than 500 in order for the observational-based methods to produce unbiased results when the proportion of patients that cross over is as high as 90%. Randomisation-based methods are relatively less affected by high levels of switching and therefore should be given precedence (unless there is evidence of a strong time-dependent treatment effect or the comparator included in the RCT is active, rendering the counterfactual survival model inappropriate).

Step (3) involves drawing upon Steps (1) and (2) and further assessing the pivotal assumptions of each of the adjustment methods in order to further determine which may be potentially appropriate. For the RPSFTM and IPE algorithm the “common treatment effect” assumption should be assessed. Survival models with the randomised group included as a covariate and a switching indicator variable may be used, but the potential bias associated with these should be recognised. Depending upon the extent to which treatment switching occurred, log-cumulative hazard and quantile-quantile plots may remain useful for assessing the proportionality of hazards and the constancy of the acceleration factor over time. If patients with different stages of disease were randomised into the trial, the treatment effect in these subgroups should be investigated to offer further evidence on the “common treatment effect” assumption, although this may also be prone to bias due to switching.

Given the limitations associated with assessing the “common treatment effect” assumption using trial data, external data sources should be sought and expert opinion on the clinical and biological plausibility of the assumption must be routinely considered. It is important to harness what is known by a variety of scientists and clinicians about the impact of patient characteristics and disease progression on the effects of the drug being studied. If these analyses suggest that the “common treatment effect” assumption holds an RPSFTM or IPE approach should be used. An IPCW approach may also produce low bias, but this is less certain.

When using RPSFTM or IPE methods the duration of the treatment effect (i.e. whether it is likely to be maintained to any extent after treatment discontinuation) must be considered. If it is likely that the treatment effect may be maintained beyond treatment discontinuation a “treatment group” application might be considered. The decision of whether to take an “on treatment”, “on treatment – observed” or “treatment group” approach should be justified based upon the economic evaluation decision problem, clinical opinion, biological plausibility and data availability. It is likely to be appropriate to present each analyses, in order that the sensitivity to these can be shown. Clinical expert opinion on whether treatment advantages are likely to cease, continue, or be reversed after treatment discontinuation may be important in justifying the chosen approach. The comparator included in the RCT (i.e. whether it is active or not) must also be considered. If the comparator is active the RPSFTM and IPE methods may not be appropriate, although a “treatment group” approach may be justified based upon assumptions made about the treatment pathways observed in the trial.

For the IPCW the “no unmeasured confounders” assumption should be considered. The likelihood that data on important covariates were not collected should be informed by clinical expert opinion as well as an assessment of covariate data reported from other trials in similar disease areas. Combined with this, consideration should be given to whether the collection of covariate data stopped at any point during the trial (for example, at the point of disease progression) as this restricts the applicability of the IPCW method. These issues should be considered in combination with those specified in Steps (1) and (2).

When considering the use of two-stage methods the existence of an appropriate secondary baseline (such as disease progression) is pivotal. These will only exist if there is a time-point before which treatment switching could not occur. If such a time-point exists two-stage methods are possible to apply, but their potential bias will be related to how soon after this

point switching occurs – if there are long delays until switching the potential for bias associated with time-dependent confounding becomes important.

After applying the switching adjustment methods Step (4) involves a review of the output of the methods in order to help identify whether the methods are likely to have performed well. For RPSFTM and IPE methods this includes a consideration of the degree of recensoring, and a comparison of “on treatment” and “treatment group” RPSFTM and IPE results in order to identify whether the treatment effect may have continued beyond treatment discontinuation. It is also important to assess the g-estimation output in order to identify the success with which the RPSFTM method has identified a unique treatment effect, and whether RPSFTM and IPE methods produce treatment effects that result in equal counterfactual survival times between randomised groups (this can be assessed using a Cox regression model comparing counterfactual survival times). For the IPCW it is particularly important to assess the weights calculated for each patient over time – instances where certain patients are allocated particularly high weights are likely to lead to erroneous IPCW results. Outputs from two-stage methods may be used to help determine the appropriateness of other methods – for instance, if the two-stage methods produce estimates of the treatment effect in crossover patients that are (not) similar to the effect estimated for patients randomised to the experimental group the RPSFTM / IPE methods may (not) be appropriate.

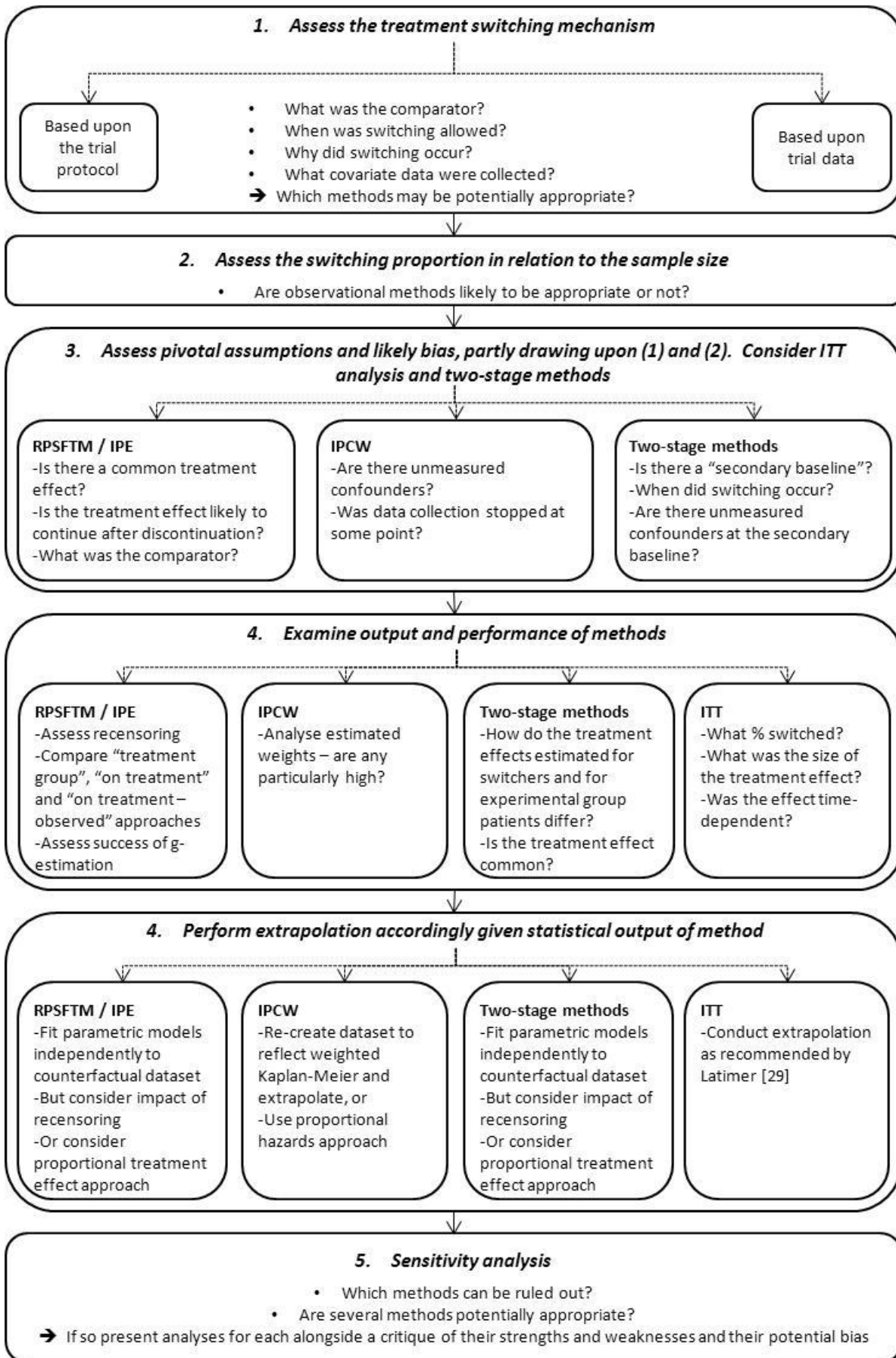
In tandem with a consideration of complex switching adjustment methods a standard ITT analysis should be considered, as if other methods are likely to have performed poorly the ITT analysis may provide least bias. Given the limitations associated with switching adjustment methods the ITT analysis should always be presented. If there is evidence of a time-dependent treatment effect the strength of that effect should be assessed if possible. If it is feasible to apply a two-stage method the suitability of this approach should be considered based upon the treatment crossover mechanism, however if this is not feasible the trial characteristics and crossover mechanism must be considered carefully in order to determine which methods are prone to least bias. Based upon my simulations, if the decrement in the treatment effect received by crossover patients is likely to be approximately 15%, there is good data availability for prognostic covariates and the crossover proportion of at-risk patients is less than 90% (assuming a trial sample size of approximately 500), the IPCW method is likely to generate similar levels of bias to RPSFTM/IPE methods. If it is applicable, a SNM can be expected to deliver marginally higher levels of bias. However, if the treatment effect is small (with an acceleration factor of approximately 1-1.3 in the experimental group) the ITT analysis is likely to be preferable (although this will still contain bias).

If the decrement in the treatment effect received by crossover patients is likely to be around 25%, RPSFTM and IPE methods are certainly unsuitable. Under the conditions specified above the IPCW method may represent the preferable analysis, but only if the treatment effect is high. If the treatment effect in the experimental group is equivalent to an AF of around 1.3 an ITT analysis is likely to be preferable, whereas with an AF of 1.8 the IPCW method is likely to produce less bias. For AF's between 1.3 and 1.7 it is currently unknown which method is likely to be preferable based upon the research presented in this thesis. Again, an SNM may provide a reasonable alternative to the IPCW method, but is less reliable, particularly if disease severity is high.

After adjustment methods have been assessed based upon their theoretical and practical suitability as well as their performance in Steps 1-4, Step (5) addresses combining the adjustment methods with an extrapolation approach (if required). This is based upon the statistical output of the applied adjustment method. For RPSFTM, IPE and two-stage methods an analysis investigating the impact of recensoring on the tail of the counterfactual Kaplan-Meier curve should be undertaken to identify whether recensoring is likely to lead to inappropriate extrapolations. A "survivor function" approach whereby the treatment effect is applied to an extrapolation of uncensored experimental group survival times may be preferable. However the choice of extrapolation method should follow the advice offered by the NICE DSU Technical Support Document where possible, given the crossover adjustment method utilised.¹⁵ For IPCW appropriate methods should be used to recreate a dataset to reflect the weighted Kaplan-Meier if a proportional hazards approach to extrapolation is not to be taken.

Finally, when preliminary analysis of trial data suggests that the choice of preferable adjustment method is unclear, sensitivity analysis should be undertaken to demonstrate the uncertainty associated with the methodology used.

Figure 8.1: Treatment crossover analysis framework



8.5 Further research

Important limitations of my simulation study and the real-world case study are discussed in Chapters 6 and 7 respectively. Of particular importance is the fact that a simulation study can never cover every conceivable scenario that may be of interest, and that the results might be dictated in some way by the techniques used to simulate the data. There is therefore obvious potential for further research to help confirm the results of the study presented in this thesis.

It would be of value to run further scenarios to supplement the simulation study presented here, particularly in order to learn more about the performance of alternative methods when the crossover proportion is lower (perhaps 5% to 50%), to investigate lower strengths of time-dependent treatment effects, and to investigate different treatment effects. Potentially this could help to more accurately determine the point at which randomisation-based methods become inappropriate, and would demonstrate if (as expected) observational methods are as prone to bias with very low levels of treatment crossover as they are with very high levels.

It would also be helpful to test scenarios that include different sample sizes, as it is likely that observational-based methods will work better when there are more data available, and may be less sensitive to high proportions of treatment crossover when the sample size is larger. This may help define what size of dataset is required for observational methods to be able to produce low levels of bias and relatively high levels of precision.

Although a complex data generating technique was employed in order to generate data in as realistic a manner as possible, it would be useful to complete a similar simulation study utilising an alternative data generating technique. For instance, a model other than a Weibull could be used to generate the underlying survival times, and the time-dependent covariate could be incorporated differently. However, the use of the Weibull should not have caused bias in my results, since the incorporation of the time-dependent covariate dictated that the final survival times did not follow a Weibull distribution. Linked to this, it may be of interest to re-run the simulation study incorporating alternative methods for calculating the treatment effect applied to crossover patients, since the method I used may be perceived as arbitrary. However, as pointed out in Chapter 6, the method for estimating the treatment effect applied to crossover patients would not be expected to impact importantly upon the performance of the crossover adjustment methods. The IPCW method censors crossover patients, while the RPSFTM and IPE methods do not seek to model the treatment effect specific to crossover

patients – what matters for the latter methods is the extent to which the average effect in crossover patients differs from that in the experimental group, not how this effect is estimated. The application of the same constant treatment effect in all crossover patients may benefit two-stage SNM and Weibull methods, but even if a time-dependent effect was applied in these patients these two-stage methods would be expected to estimate the average effect with similar success.

The real-world case study illustrated that the IPE algorithm is currently only coded in STATA for use with Weibull and exponential accelerated failure time models. Given the importance associated with accurately extrapolating data, extending the coding of this method for use with alternative models would be beneficial. In order to understand more about the practicalities of applying the crossover adjustment methods in real-world scenarios it would be useful to undertake further case studies applying the methods to additional real-world datasets.

Most importantly, it is clear that further research into novel methods for adjusting for treatment crossover would be very valuable. Available methods perform relatively poorly when there is a time-dependent treatment effect. In particular, in circumstances where a suitable two-stage technique is not possible, data availability is poor, crossover proportions are very high, and where the treatment effect is strongly related to time (and therefore the average effect is different in crossover patients compared to the experimental group) there do not exist theoretically sound methods that can accurately adjust for treatment crossover. In these circumstances the ITT analysis may produce least bias. Even where data availability is good and crossover proportions are moderate, if there is an important time-dependent treatment effect, existing methods are likely to lead to sub-optimally high levels of bias. Randomisation-based methods that can successfully estimate different treatment effects for the experimental group and crossover patients have not yet been developed, but such methods could be extremely valuable for use in economic evaluations.

It would also be interesting to investigate ways in which information and advice from clinical and scientific experts could be incorporated into the treatment crossover adjustment process more systematically, such that involvement could be made reproducible across studies. My discussions with clinicians from Weston Park hospital were extremely helpful with regard to understanding whether or not it is reasonable to expect it to be possible to model the treatment crossover mechanism. However, my discussions were not systematic and my conclusions from these may be open to bias. Investigating the use of more appropriate

methods to extract robust information from experts to inform treatment crossover analyses would be valuable.

Finally, it is important that the findings of my research are disseminated in order to promote more appropriate and consistent use of crossover adjustment methods in future health technology assessments. To this end, I have already presented my work at international conferences and national workshops. I have begun to submit my findings to high quality peer reviewed journals. Importantly, my work has already informed the draft update to the NICE methods guide,¹⁸⁵ and I have been asked to write a new NICE DSU Technical Support Document on the use of crossover adjustment methods. Therefore my research is highly likely to influence the analyses undertaken in future health technology assessments.

8.6 Conclusions

The aim of this thesis was to explore the use of crossover adjustment methods within economic evaluations to inform resource allocation decisions. The main hypothesis was that methods commonly used in health technology assessments were prone to substantial amounts of bias. The thesis therefore considered a) why it is important to address the treatment crossover problem, b) what methods are commonly used in economic evaluations to address treatment crossover, c) what methods are available from the literature to address treatment crossover, d) which methods work best across a wide range of scenarios, and e) which methods are practical for use in a real-world setting, specifically in the context of extrapolation and economic evaluation.

It was well known prior to the completion of this thesis that naive methods for adjusting for treatment crossover – such as censoring or excluding crossover patients from the analysis – are highly prone to bias due to their theoretical limitations. This had been demonstrated in a previous simulation study upon which I was a co-author.²¹ The research reported in this thesis provides further confirmation of this finding, and based upon a review of previous NICE technology appraisals it is clear that this may have affected previous resource allocation decisions. What this thesis adds is a much more complex and complete analysis and evaluation of crossover adjustment methods – not just regarding their performance in a simulation study, but also their application to real-world datasets and their amenability to be combined with extrapolation modelling, as is usually required within economic evaluations. Previous research had not conducted a systematic review of potentially appropriate crossover adjustment methods, had not compared all relevant crossover adjustment methods, had not

tested methods in scenarios that violate their key assumptions, and had not considered how the results provided by the methods could be incorporated within an economic evaluation. These gaps in the literature are addressed in detail in this thesis.

While several complex treatment crossover adjustment methods exist, this thesis demonstrates that these have important limitations. A wide range of scenarios designed to reflect real-world RCT datasets were considered, and I have shown that the performance of the methods was very sensitive to dataset characteristics – particularly when certain characteristics caused one or more methodological assumptions to be violated. In some circumstances – for example, where the treatment effect received by crossover patients is very low relative to the effect received by patients randomised to the experimental group – a simple ITT analysis may provide lower bias than any of the alternatives. However, in other circumstances more complex methods provide much lower bias – particularly the randomisation-based methods, when the “common treatment effect” assumption holds. Given the currently available set of limited methods, datasets and crossover mechanisms should be analysed on a case-by-case basis in order to identify which crossover adjustment method is likely to result in least bias. Key assumptions such as the “common treatment effect” assumption and the “no unmeasured confounders” assumption should be investigated as far as possible. The ability to model the treatment crossover mechanism is also pivotal, and the likelihood that this will be possible should be considered for each trial confounded by crossover that is analysed.

I have also demonstrated that it is important to consider how the outputs that are provided by the crossover adjustment methods may be used within an extrapolation exercise for use within an economic model. This topic has not been discussed in any previously published literature. Given the substantial impact that alternative approaches to extrapolation can have on cost-effectiveness results,^{11;15-17;25} it is important that crossover adjustment methods are used in combination with appropriate extrapolation techniques. The use of crossover adjustment methods should not prevent the ability to achieve this, although both jointly and independently fitting parametric models to RCT treatment groups is made slightly more complex when data have been adjusted for crossover.

Overall, I have met the objectives that I set out at the beginning of this thesis. My research has filled several important gaps within the literature, and I have been able to formulate guidance on the use of crossover adjustment methods in an economic evaluation context. Although there is scope for further research, my findings are valuable and – if disseminated effectively –

will allow treatment crossover to be addressed more consistently and robustly in future health technology assessments. In turn, this should lead to a higher probability that consistent and appropriate resource allocation decisions are made, allowing a more cost-effective use of health care resources.

Reference List

- (1) Department of Health. The NHS Cancer Plan. 2000. London, The Stationary Office.
- (2) Sorenson C, Drummond M, Kanavos P. Ensuring value for money in health care: The role of health technology assessment in the European Union. 2008. World Health Organisation, on behalf of the European Observatory on Health Systems and Policies. Observatory Studies Series No.11. 18-9-2012.
- (3) National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. 2008.
- (4) U.S.Department of Health and Human Services Food and Drug Administration. Guidance for Industry: Clinical trial endpoints for the approval of cancer drugs and biologics. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research, editors. 2007.
- (5) Committee for Medicinal Products for Human Use (CHMP). Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man (CHMP/EWP/205/95 REV.3). Methodological considerations for using progression-free survival (PFS) as primary endpoint in confirmatory trials for registration. 201. European Medicines Agency.
- (6) Briggs A, Claxton K, Scuplher M. Decision modelling for health economic evaluation. New York: Oxford University Press Inc.; 2006.
- (7) Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in health and medicine. New York: Oxford University Press, Inc.; 1996.
- (8) Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 3rd Edition. 2006.
- (9) Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? *British Medical Journal* 2000; 320:1197-1200.
- (10) Cook J, Drummond M, Heyse JF. Economic endpoints in clinical trials. *Statistical Methods in Medical Research* 2004; 13(2):157-176.
- (11) Davies A, Briggs A, Schneider J, Levy A, Ebeid O, Wagner S et al. The Ends Justify the Mean: Outcome Measures for Estimating the Value of New Cancer Therapies. *Health Outcomes Research in Medicine* 2012; 3(1):e25-e36.
- (12) Collett D. Modelling Survival Data in Medical Research, 2nd ed. Boca Raton: Chapman & Hall/CRC CRC Press LLC; 2003.
- (13) Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* 2009; 45:228-247.
- (14) Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* 2011; 30:2409-2421.

- (15) Latimer NR. NICE DSU Technical support document 14: Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. Report by the Decision Support Unit, editor. 2011.
- (16) Guyot P, Welton NJ, Ouwens M, Ades AE. Survival time outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. *Value in Health* 2011; 14:640-646.
- (17) Connock M, Hyde C, Moore D. Cautions regarding the fitting and interpretation of survival curves. *Pharmacoeconomics* 2011; 29(10):827-837.
- (18) Demeris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Statistics in Medicine* 2006; 25:1960-1975.
- (19) Jackson CH, Sharples LD, Thompson SG. Survival models in health economic evaluations: Balancing fit and parsimony to improve prediction. *International Journal of Biostatistics* 2010; 6(1):34.
- (20) White IR. Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research* 2005; 14(4):327-347.
- (21) Morden JP, Lambert PC, Latimer NR, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Medical Research Methodology* 2011; 11.
- (22) Robins JM, Tsiatis AA. Correcting for Noncompliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. *Communications in Statistics-Theory and Methods* 1991; 20(8):2609-2631.
- (23) Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Statistics in Medicine* 2002; 21(17):2449-2463.
- (24) Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; 56(3):779-788.
- (25) Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S. Methodological issues in the economic analysis of cancer treatments. *European Journal of Cancer* 2006; 42(17):2867-2875.
- (26) Fuchs V. Who shall live? Health, economics and social choice. New Jersey: World Scientific Publishing; 1982.
- (27) Weinstein MC. Theoretical issues in cost-effectiveness analysis. *Journal of Health Economics* 1997; 16:121-128.
- (28) McGuire A. Theoretical concepts in the economic evaluation of health care. In: Drummond M, McGuire A., editors. Economic evaluation in health care. Oxford: Oxford University Press; 2001.
- (29) Hicks J. Value and capital. 1939. Oxford, Clarendon Press.

- (30) Tsuchiya A., Williams A. Welfare economics and economic evaluation. In: Drummond M, McGuire A., editors. *Economic evaluation in health care*. Oxford: Oxford University Press; 2001.
- (31) Brouwer WBF, Culyer AJ, van Exel NJA, Rutten FFH. Welfarism vs. extra-welfarism. *Journal of Health Economics* 2008; 27:325-338.
- (32) Culyer AJ. The normative economics of health care finance and provision. *Oxford Review of Economic Policy* 1989; 5(1):34-58.
- (33) Sen A. The possibility of social choice. *American Economic Review* 1999; 89(3):349-378.
- (34) Broome J. Qalys. *Journal of Public Economics* 1993; 50(2):149-167.
- (35) Williams A. Economics of Coronary Artery Bypass Grafting. *British Medical Journal* 1985; 291:326-329.
- (36) Williams A. The Cost-Benefit Approach to the Evaluation of Intensive Care Units, in Miranda, D.R. and Langrehr, D., eds. *The ICU - a cost benefit analysis*. 1986. Amsterdam, Elsevier.
- (37) Littlejohns P. The Establishment of the National Institute for Clinical Excellence: Its Importance and Implications for the Pharmaceutical Industry. *Drug Information Journal* 2001; 35:181-188.
- (38) Department of Health. *A first class service: Quality in the new NHS*. 1998.
- (39) National Institute for Health and Clinical Excellence. *Guide to the single technology appraisal process*. 2009.
- (40) National Institute for Health and Clinical Excellence. *Guide to the multiple technology appraisal process*. 2009.
- (41) National Institute for Health and Clinical Excellence. *Appraising life-extending, end of life treatments*. 2009.
- (42) Sculpher M. Identifying subgroups and exploring heterogeneity: Workshop briefing paper. 2007. London, National Institute for Health and Clinical Excellence.
- (43) Donaldson C, Currie G, Mitton C. Cost effectiveness analysis in health care: contraindications. *British Medical Journal* 2002; 325(7369):891-894.
- (44) Birch S, Gafni A. Cost-effectiveness/cost-utility analyses: Do current decision rules lead us where we want to be? *Journal of Health Economics* 1992; 11:279-296.
- (45) Birch S, Gafni A. Information created to evade reality (ICER). Things we should not look to for answers. *Pharmacoeconomics* 2006; 24(11):1121-1131.
- (46) Birch S, Gafni A. On being NICE in the UK: guidelines for technology appraisal for the NHS in England and Wales. *Health Economics* 2002; 11:185-191.
- (47) Birch S, Gafni A. The biggest bang for the buck or bigger bucks for the bang. *Journal of Health Services Research and Policy* 2006; 11(1):46-51.

- (48) Williams A. The Cost Benefit Approach, Chapter 2 in "Being Reasonable About the Economics of Health: Essays by Alan Williams" edited by Culyer A.J. and Maynard A. 1997. Lyme, N.H., Edward Elgar Publishing.
- (49) Forbes C, Shirran L, Bagnall AM, Duffy S, Ter Riet G. A rapid and systematic review of the clinical effectiveness and cost effectiveness of topotecan for ovarian cancer. 2001. NHS Centre for Reviews and Dissemination, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (50) Lloyd Jones M, Hummel S, Bansback N. A Review of the Evidence for the Clinical and Cost-effectiveness of Irinotecan, Oxaliplatin and Raltitrexed for the Treatment of Advanced Colorectal Cancer. 2001. University of Sheffield School of Health and Related Research, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (51) Lewis R, Bagnall AM, Forbes C, Shirran E, Duffy S, Kleijnen J et al. A rapid and systematic review of the clinical effectiveness and cost effectiveness of trastuzumab for breast cancer. 2001. NHS Centre for Reviews and Dissemination, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (52) National Institute for Health and Clinical Excellence. Sunitinib for the treatment of gastrointestinal stromal tumours, TA 179. 2009.
- (53) National Institute for Health and Clinical Excellence. Pazopanib for the first-line treatment of advanced renal cell carcinoma. Final Appraisal Determination. 24-12-2010.
- (54) National Institute for Health and Clinical Excellence. Final appraisal determination: Everolimus for the second-line treatment of advanced renal cell carcinoma. 2010.
- (55) National Institute for Health and Clinical Excellence. Guidance on the use of imatinib for chronic myeloid leukaemia, TA70. 2003.
- (56) Wilson L, Connock M, Song F, Yao G, Fry-Smith A, Raftery J et al. Imatinib for the treatment of patients with unresectable and/or metastatic gastro-intestinal stromal tumours - a systematic review and economic evaluation. 2003. West Midlands Health Technology Assessment Collaboration, University of Birmingham, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (57) National Institute for Health and Clinical Excellence. Final Appraisal Determination: Imatinib for the treatment of unresectable and/or metastatic gastro-intestinal stromal tumours, TA86. 2004.
- (58) Janssen-Cilag Ltd. STA submission to NICE: Velcade (Bortezomib) for the treatment of multiple myeloma patients at first relapse. 2006.
- (59) Abrams K, Palmer S, Wailoo A. Bevacizumab, sorafenib, sunitinib, and temsirolimus for renal cell carcinoma. 2008. National Institute for Health and Clinical Excellence Decision Support Unit.

- (60) National Institute for Health and Clinical Excellence. Final Appraisal Determination: Sunitinib for the first-line treatment of advanced and/or metastatic renal cell carcinoma, TA169. 2009.
- (61) Hoyle M. Sunitinib for GIST: additional notes for the ACD meeting from PenTAG. 2009.
- (62) Hoyle M, Rogers G, Garside R, Moxham T, Stein K. The clinical and cost effectiveness of lenalidomide for multiple myeloma in people who have received at least one prior therapy: An evidence review of the submission from Celgene. 2008. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (63) Bond M, Hoyle M, Moxham T, Napier M, Anderson R. The clinical and cost-effectiveness of sunitinib for the treatment of gastrointestinal stromal tumours: a critique of the submission from Pfizer. 2009. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (64) Hind D, Tappenden P, Tumor I, Eggington S, Sutcliffe P, Ryan A. The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation (review of Guidance No. 33), Addendum: Economic evaluation of irinotecan and oxaliplatin for the treatment of advanced colorectal cancer. 2005. School of Health and Related Research, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (65) Merck Serono LTD. Single Technology Appraisal Submission: Erbitux (cetuximab) for the first-line treatment of metastatic colorectal cancer. 2008.
- (66) National Institute for Health and Clinical Excellence. Final Appraisal Determination: Imatinib for the treatment of unresectable and/or metastatic gastro-intestinal stromal tumours, TA86. 2004.
- (67) National Institute for Health and Clinical Excellence. Guidance on the use of trastuzumab for the treatment of advanced breast cancer, NICE Technology Appraisal Guidance No.34. 2002.
- (68) Coon J, Hoyle M, Green C, Liu Z, Welch K, Moxham T et al. Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: A systematic review and economic evaluation: Addendum to the report submitted on 2nd May 2008. 2008. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (69) National Institute for Health and Clinical Excellence. Final Appraisal Determination: Sunitinib for the treatment of gastrointestinal stromal tumours, TA179. 2009.
- (70) Collins R, Fenwick E, Trowman R, Perard R, Norman G, Light K et al. A systematic review and economic model of the effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer. 2005. Centre for Reviews and Dissemination,

University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.

- (71) Jones J, Takeda A, Tan SC, Cooper K, Loveman E, Clegg A et al. Gemcitabine for metastatic breast cancer. 2006. Southampton Health Technology Assessments Centre, University of Southampton, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (72) Garside R, Pitt M, Anderson R, Rogers G, Dyer M, Mealing S et al. The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high grade glioma: A systematic review and economic evaluation. 2005. PenTAG, University of Southampton, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (73) Tappenden P, Jones R, Paisley S, Carroll C. The use of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer. 2006. School of Health and Related Research, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (74) Lister-Sharp D, McDonagh M, Khan KS, Kleijnen J. A Systematic Review of the Effectiveness and Cost Effectiveness of the Taxanes used in the Treatment of Advanced Breast and Ovarian Cancer. 2000. NHS Centre for Reviews and Dissemination, University of York, January 2000, Report commissioned by the NHS HTA Programme on behalf of the National Institute for Clinical Excellence.
- (75) National Institute for Health and Clinical Excellence. Appeal Panel Decision Document (Aventis). 2000.
- (76) Walker S, Palmer S, Erhorn S, Brent S, Dyker A, Ferrie L et al. Fludarabine phosphate for the first-line treatment of chronic lymphocytic leukaemia. 2006. Centre for Health Economics, University of York, and NHS Northern and Yorkshire Regional Drug and Therapeutics Centre, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (77) Hind D, Tappenden P, Tumor I, Eggington S, Sutcliffe P, Ryan A. The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation (review of Guidance No. 33). 2005. School of Health and Related Research, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (78) Robins JM, Tsiatis AA. Correcting for Noncompliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. *Communications in Statistics-Theory and Methods* 1991; 20(8):2609-2631.
- (79) National Institute for Health and Clinical Excellence. Final Appraisal Determination: Lenalidomide for the treatment of multiple myeloma in people who have received at least one prior therapy, TA171. 2009.
- (80) Hoyle M, Rogers G, Garside R, Moxham T, Stein K. The clinical and cost effectiveness of lenalidomide for multiple myeloma in people who have received at least one prior therapy: An evidence review of the submission from Celgene: Addendum to the report submitted on 1st September 2008. 2009. Peninsula Technology Assessment

Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.

- (81) Ishak KJ, Caro JJ, Drayson MT, Dimopoulos M, Weber D, Augustson B et al. Adjusting for patient crossover in clinical trials using external data: a case study of lenalidomide for advanced multiple myeloma. *Value in Health* 2011; 14(5):672-678.
- (82) National Institute for Health and Clinical Excellence. NICE issues Guidance on Taxanes for Ovarian Cancer. 2000.
- (83) National Institute for Health and Clinical Excellence. NICE Issues Guidance on Chemotherapy Agents for Breast Cancer and Leukaemia. 2001.
- (84) National Institute for Health and Clinical Excellence. Guidance on the use of paclitaxel in the treatment of ovarian cancer (TA55). 2003.
- (85) National Institute for Health and Clinical Excellence. Paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan for second-line or subsequent treatment of advanced ovarian cancer: Review of Technology Appraisal Guidance 28, 45 and 55, TA91. 2005.
- (86) National Institute for Health and Clinical Excellence. Pemetrexed for the treatment of non-small-cell lung cancer, TA124. 2007.
- (87) National Institute for Health and Clinical Excellence. Cetuximab for the treatment of recurrent and/or metastatic squamous cell cancer of the head and neck, TA172. 2009.
- (88) National Institute for Health and Clinical Excellence. Erlotinib for the treatment of non-small-cell lung cancer, TA162. 2008.
- (89) Bagust A, Boland A, Dundar Y, Davis H, Dickson R, Green J et al. Pemetrexed for the treatment of relapsed non-small-cell lung cancer: ERG Report. 2006. Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (90) Vermorken JB, Mesia R, Rivera F, et al. Platinum-Based Chemotherapy plus Cetuximab in Head and Neck Cancer. *New England Journal of Medicine* 2008; 359:1116-1127.
- (91) Roche Products Ltd. Achieving clinical excellence in the treatment of relapsed non-small cell lung cancer, Tarceva (erlotinib) NICE STA Submission. 2006.
- (92) Bagust A, Boland A, Dundar Y, Davis H, Dickson R, Green J et al. Erlotinib for the treatment of relapsed non-small-cell lung cancer: ERG Report. 2006. Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence.
- (93) National Institute for Health and Clinical Excellence. NICE issues Guidance on Taxanes for breast cancer. 2000.
- (94) National Institute for Health and Clinical Excellence. NICE recommends selective use of drugs for advanced colorectal cancer. 2002.

- (95) National Institute for Health and Clinical Excellence. Irinotecan, oxaliplatin and raltitrexed for advanced colorectal cancer: Review of Technology Appraisal 33. 2005.
- (96) National Institute for Health and Clinical Excellence. Gemcitabine for the treatment of metastatic breast cancer, TA 116. 2007.
- (97) National Institute for Health and Clinical Excellence. Sunitinib for the first-line treatment of advanced and/or metastatic renal cell carcinoma, TA169. 2009.
- (98) National Institute for Health and Clinical Excellence. NICE Issues Guidance on Topotecan for Ovarian Cancer. 2001.
- (99) National Institute for Health and Clinical Excellence. Docetaxel for the treatment of hormone-refractory metastatic prostate cancer, TA101. 2006.
- (100) National Institute for Health and Clinical Excellence. Carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma, TA121. 2007.
- (101) National Institute for Health and Clinical Excellence. Final Appraisal Determination, Bortezomib monotherapy for relapsed multiple myeloma, TA129. 2007.
- (102) National Institute for Health and Clinical Excellence. Lenalidomide for the treatment of multiple myeloma in people who have received at least one prior therapy. NICE technology appraisal guidance 171. 2009.
- (103) National Institute for Health and Clinical Excellence. Cetuximab for the first-line treatment of metastatic colorectal cancer, TA176. 2009.
- (104) National Institute for Health and Clinical Excellence. Bevacizumab and cetuximab for the treatment of metastatic colorectal cancer, TA 118. 2007.
- (105) National Institute for Health and Clinical Excellence. Fludarabine monotherapy for the first-line treatment of chronic lymphocytic leukaemia, TA119. 2007.
- (106) National Institute for Health and Clinical Excellence. Final Appraisal Determination, Renal cell carcinoma – bevacizumab, sprafenib, sunitinib and temsirolimus, TA178. 2009.
- (107) GlaxoSmithKline UK. Pazopanib (Votrient(R)) for the first-line treatment of patients with advanced renal cell carcinoma (RCC): Addendum to GSK's submission to NICE. 2010.
- (108) Kilonzo M, Hislop J, Elders A, Fraser C, Bissett D, McClinton S et al. Pazopanib for the first line treatment of patients with advanced and/or metastatic renal cell carcinoma: A Single Technology Appraisal. 2010. Aberdeen HTA Group, Institute of Applied Health Sciences, University of Aberdeen.
- (109) National Institute for Health and Clinical Excellence, National. Health Technology Appraisal Appeal Hearing: Advice on Everolimus for the second-line treatment of advanced renal cell carcinoma. 2011.
- (110) Novartis Pharmaceuticals UK Ltd. Single technology appraisal (STA) for everolimus (Afinitor(R)) in advanced renal cell carcinoma. 2010.

- (111) Pitt M, Crathorne L, Moxham T, Bond M, Hyde C. Everolimus for the second-line treatment of advanced and/or metastatic renal cell carcinoma. Peninsula Technology Assessment Group PMSUoE, editor. 2009. A report commissioned by the NIHR HTA Programme on behalf of the National Institute for Health and Clinical Excellence.
- (112) Pitt M. PENTAG response to the Novartis updated submission. 2010.
- (113) National Institute for Health and Clinical Excellence. Pazopanib for the first line treatment of metastatic renal cell carcinoma, TA 215. 2011.
- (114) National Institute for Health and Clinical Excellence. Everolimus for the second-line treatment of advanced renal cell carcinoma, TA219. 2011.
- (115) Hutton JL, Ashcroft R. What does 'systematic' mean for reviews of methods? Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. 1998. London, BMJ Books.
- (116) Edwards SJL, Lilford RJ, Kiauka S. Different types of systematic review in health services research. In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: BMJ Books; 1998.
- (117) Gough D, Oliver S, Thomas J. An introduction to systematic reviews. London: Sage; 2012.
- (118) Ramer SL. Site-ation pearl growing: methods and librarianship history and theory. *Journal of the Medical Library Association* 2005; 93(3):397-400.
- (119) Schlosser RW, Wendt O, Bhavnani S, Nail-Chiwetalu B. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. *International Journal of Language & Communication Disorders* 2006; 41(5):567-582.
- (120) Hartley RJ, Michael Keen E, Large JA, Tedd LA. Online Searching: Principles and Practice. Epping, UK: Bowker-saur; 1990.
- (121) Dolan P, Shaw R, Tsuchiya A., Williams A. QALY maximisation and people's preferences: a methodological review of the literature. *Health Economics* 2005; 14:197-208.
- (122) Tsuchiya A, Dolan P. The QALY model and Individual Preferences for Health States and Health Profiles over Time: A Systematic Review of the Literature. *Medical Decision Making* 2005; 25:460-467.
- (123) White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. *Statistics in Medicine* 1999; 18(19):2617-2634.
- (124) Mittlbock M, Whitehead J. The interpretation of clinical trials of immediate versus delayed therapy. *Lifetime Data Analysis* 1998; 4(3):253-263.
- (125) White IR, Walker S, Babiker AG. strbee: Randomization-based efficacy estimator. *The Stata Journal* 2002; 2(2):140-150.

- (126) Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal* 2[2], 164-182. 2002.
- (127) Young JG, Hernan MA, Picciotto S, Robins JM. Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis* 2010; 16:71-84.
- (128) Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias. *American Journal of Epidemiology* 2011; 173(5):569-577.
- (129) Robins JM, Greenland S. Adjusting for Differential Rates of Prophylaxis Therapy for Pcp in High-Dose Versus Low-Dose Azt Treatment Arms in An Aids Randomized Trial. *Journal of the American Statistical Association* 1994; 89(427):737-749.
- (130) Law MG, Kaldor JM. Survival analyses of randomized clinical trials adjusted for patients who switch treatments. *Statistics in Medicine* 1996; 15(19):2069-2076.
- (131) Walker AS, White IR, Babiker AG. Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. *Statistics in Medicine* 2004; 23(4):571-590.
- (132) Tanaka Y, Matsuyama Y, Ohashi Y. Estimation of treatment effect adjusting for treatment changes using the intensity score method: Application to a large primary prevention study for coronary events (MEGA study). *Statistics in Medicine* 2008; 27(10):1718-1733.
- (133) Mark SD, Robins JM. A Method for the Analysis of Randomized Trials with Compliance Information - An Application to the Multiple Risk Factor Intervention Trial. *Controlled Clinical Trials* 1993; 14(2):79-97.
- (134) Mark SD, Robins JM. Estimating the Causal Effect of Smoking Cessation in the Presence of Confounding Factors Using A Rank Preserving Structural Failure Time Model. *Statistics in Medicine* 1993; 12(17):1605-1628.
- (135) Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* 2001; 96(454):440-448.
- (136) Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* 2000; 19(14):1849-1864.
- (137) Loeys T, Vansteelandt S, Goetghebeur E. Accounting for correlation and compliance in cluster randomized trials. *Statistics in Medicine* 2001; 20(24):3753-3767.
- (138) Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part I: Structural nested models and marginal structural models to test and estimate treatment arm effects. *Statistics in Medicine* 2004; 23(13):1991-2003.
- (139) Huang XL, Cormier JN, Pisters PWT. Estimation of the causal effects on survival of two-stage nonrandomized treatment sequences for recurrent diseases. *Biometrics* 2006; 62(3):901-909.

- (140) Shao J, Chang M, Chow SC. Statistical inference for cancer trials with treatment switching. *Statistics in Medicine* 2005; 24(12):1783-1790.
- (141) Hsu CH, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine* 2006; 25(20):3503-3517.
- (142) Lee MLT, Chang M, Whitmore GA. A Threshold Regression Mixture Model for Assessing Treatment Efficacy in a Multiple Myeloma Clinical Trial. *Journal of Biopharmaceutical Statistics* 2008; 18(6):1136-1149.
- (143) Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; 52(1):137-151.
- (144) Robins JM. Structural Nested Failure Time Models. Andersen PK, Keiding N, editors. *Survival Analysis*. 4372-4389. 1998. Chichester, UK, John Wiley and Sons. The Encyclopedia of Biostatistics. Armitage, P. and Colton, T.
- (145) Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, editors. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. New York: Springer-Verlag; 1999. 95-134.
- (146) Hsu CH, Taylor JMG. A robust weighted Kaplan-Meier approach for data with dependent censoring using linear combinations of prognostic covariates. *Statistics in Medicine* 2010; 29(21):2215-2223.
- (147) Witteman JCM, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MAJ et al. G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *American Journal of Epidemiology* 1998; 148(4):390-401.
- (148) White IR. Survival analysis of randomized clinical trials adjusted for patients who switch treatments. *Statistics in Medicine* 1997; 16(22):2619-2620.
- (149) White IR. Estimating treatment effects in randomized trials with treatment switching. *Statistics in Medicine* 2006; 25(9):1619-1622.
- (150) Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. *Statistics in Medicine* 2004; 23(13):2005-2022.
- (151) Cox DR, Oakes D. *Analysis of Survival Data*. Boca Raton: Chapman & Hall/CRC; 1984.
- (152) Ouwens MJNM, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Research Synthesis Methods* 2010; 1:258-271.
- (153) Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Medical Research Methodology* 2011; 11(61).
- (154) Demeris N, Lunn D, Sharples LD. Survival extrapolation using the poly-Weibull model. *Statistical Methods in Medical Research* 2011; 0(0):1-15.
- (155) Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology* 2012; 12(9).

- (156) Thompson Coon J, Hoyle M, Green C, Liu Z, Welch K, Moxham T et al. Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: A systematic review and economic evaluation. 2008. Produced by Peninsula Technology Assessment Group, Universities of Exeter and Plymouth on behalf of NICE.
- (157) Dalziel K, Round A, Stein K, Garside R, Price A. The effectiveness and cost effectiveness of imatinib for first line treatment of chronic myeloid leukaemia in chronic phase. 2003. Produced by Peninsula Technology Assessment Group, Universities of Exeter and Wessex Institute for Health Research and Development, University of Southampton on behalf of NICE.
- (158) Main C, Ginnelly L, Griffin S, Norman G, Barbieri M, Mather L et al. Topotecan, pegylated liposomal doxorubicin hydrochloride and paclitaxel for secondline or subsequent treatment of advanced ovarian cancer. 2004. Produced by Centre for Reviews and Dissemination, University of York on behalf of NICE.
- (159) Roche Products Ltd. Rituximab for the treatment of relapsed follicular lymphoma, Roche submission to the National Institute for Health and Clinical Excellence. 2007.
- (160) Roche Products Ltd. Rituximab for the 1st line treatment of chronic lymphocytic leukaemia, Roche submission to the National Institute for Health and Clinical Excellence. 2008.
- (161) Pfizer Limited. Single Technology Appraisal of Sunitinib for the treatment of gastrointestinal stromal tumours. 2008.
- (162) Dundar Y, McLeod C, Boland A, Walley T, Hounscome J, Bagust A et al. Rituximab for the first line treatment of stage III-IV follicular non-Hodgkin's lymphoma. 2006. Produced by Liverpool Reviews and Implementation Group, University of Liverpool on behalf of NICE.
- (163) Ortho Biotech Limited. Velcade (R) (Bortezomib) for the treatment of multiple myeloma patients at first relapse, manufacturer submission. 2006.
- (164) Green C, Bryant J, Takeda A, Cooper K, Clegg A, Smith A et al. Bortezomib for the treatment of multiple myeloma patients. 2006. Produced by Southampton Health Technology Assessments Centre, University of Southampton on behalf of NICE.
- (165) Gelber RD, Goldhirsch A, Cole BF. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Controlled Clinical Trials* 1993; 14:485-499.
- (166) Liverpool Reviews and Implementation Group. ERG Addendum: Pemetrexed for the first-line treatment of locally advanced or metastatic non-small cell lung cancer (NSCLC). 2009. University of Liverpool.
- (167) Royston P, Parmar MKB. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002; 21:2175-2197.
- (168) Loeyes T, Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics* 2003; 59(1):100-105.

- (169) Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; 25:4279-4292.
- (170) Crowther MJ, Lambert PC. Simulating complex survival data. *The Stata Journal* 2012; (In Press).
- (171) Stata statistical software intercooled, Version 11.0 [Texas, USA: 2009.
- (172) Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; 24:1713-1723.
- (173) GlaxoSmithKline UK. Manufacturer Submission to the National Institute for Health and Clinical Excellence. Submission to address the question of whether and how lapatinib falls within the Supplementary Advice to Appraisal Committees on appraising treatments that extend life at the end of life. 25-8-2009.
- (174) Merck Serono LTD. Single technology appraisal submission: Erbitux (cetuximab) for the first-line treatment of recurrent and/or metastatic squamous cell carcinoma of the head and neck. 2009.
- (175) Fewell Z, Hernan MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal* 2004; 4(4):402-420.
- (176) Geyer CE, Forster J, Lindquist D, et al. Lapatinib plus Capecitabine for HER2-Positive Advanced Breast Cancer. *New England Journal of Medicine* 2006; 355:2733-2743.
- (177) Cameron D, Casey M, Press M, et al. A phase III randomized comparison of lapatinib plus capecitabine versus capecitabine alone in women with advanced breast cancer that has progressed on trastuzumab: updated efficacy and biomarker analyses. *Breast Cancer Research and Treatment* 2008; 112:533-543.
- (178) GlaxoSmithKline UK. Manufacturer submission to the National Institute for Health and Clinical Excellence. Single technology appraisal of lapatinib for the treatment of women with previously treated advanced or metastatic ERbB2- (HER2) over-expressing breast cancer. 17-4-2007.
- (179) National Institute for Health and Clinical Excellence. Breast cancer (advanced or metastatic) - lapatinib [ID20]. 2010. 26-9-2012.
- (180) European Medicines Agency. Assessment Report for Tyverb. EMEA/302222/2008. 2008. European Medicines Agency.
- (181) Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis - choosing a model and assessing its adequacy and fit. *British Journal of Cancer* 2003; 89(4):605-611.
- (182) Cancer Research UK. Breast cancer - survival statistics. Figure 3.7, taken from West Midlands Cancer Intelligence Unit data. 2012.
- (183) Jones J, Takeda A, Picot J, von Keyserlingk C, Clegg A. Lapatinib for HER2 over-expressing breast cancer. 2007. National Institute for Health and Clinical Excellence. Evidence Review Group Report commissioned by the NHS R&D HTA Programme.

- (184) National Institute for Health and Clinical Excellence. Final appraisal determination: Lapatinib for the treatment of women with previously treated advanced or metastatic breast cancer. 10-6-2010.
- (185) National Institute for Health and Clinical Excellence. Guide to the Methods of Technology Appraisal: Draft Methods Guide. 2012.

Appendix 1: NICE metastatic and/or advanced cancer appraisals – evidence tables

1. TA3: Ovarian cancer - taxanes (replaced by TA55), May 2000.

Guidance: Paclitaxel (Taxol) should be used to treat women with ovarian cancer as a standard initial therapy and should be used to treat women who have not previously received paclitaxel, whose cancer has recurred or been resistant to other forms of treatment.

Source: Only source available was the HTA report: http://www.nice.org.uk/nicemedia/pdf/hta_ov_taxanes.pdf

Lister-Sharp D, McDonagh M, Khan KS, Kleijnen J. A Systematic Review of the Effectiveness and Cost-effectiveness of the Taxanes used in the Treatment of Advanced Breast and Ovarian Cancer, NHS Centre for Reviews and Dissemination, University of York, January 2000, Report commissioned by the NHS HTA Programme on behalf of the National Institute for Clinical Excellence

Note that the manufacturer (BMS) made a submission, but this is not available from the NICE website, and details from it that had been included in the AC's version of the HTA report have been removed from the public version.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>An economic model was not developed, instead the literature was reviewed. The manufacturer's submission is not available, and all reference to it has been removed from the HTA report.</p> <p>First Line: 4 RCTs found: GOG 111, GOG 132, OV10 and ICON 3. A total of 3770 patients were included. All evaluated paclitaxel combined with either cisplatin or carboplatin.</p> <p>9 CEAs described. Of these; 2 based survival gain on median from one study; 2 based survival gain on mean from one study (not explained how the mean was calculated); 1 used pooled mean from more than one study (not explained how mean was calculated); 1 used median survival gain from one study, but then used Declining Exponential Approximation of Life Expectancy (DEARE) to calculate specific life expectancy based on actuarial methods; 1 used 'simple linear modelling' assuming that average survival time was 50% longer for paclitaxel; 2 used survival curve fitting, normalised to a population of 100 patients to extrapolate study data to lifetime.</p> <p>3 CUAs. The survival analysis methods described in the HTA for the CUAs are not detailed. 1 study used survival curve fitting, 1 appears to be based only upon incremental progression free months, while it is unclear how survival was measured in the other study.</p>
Evidence synthesis (pool survival estimates?)	<p>The two UK based CEAs assumed that the effectiveness of carboplatin alone was the same as that of cisplatin plus cyclophosphamide, in the GOG-111 trial in their primary analysis. A secondary analysis used efficacy rates for carboplatin found in the literature, in studies comparing carboplatin to a non-taxane containing regimen. In an update of the original UK CEA paper response rates for paclitaxel plus cisplatin from the OV 10 (ECOCIT) trial were substituted for those of the GOG-111 trial. The OV10 trial included patients diagnosed with stage II ovarian cancer, whereas the cost-effectiveness exercise is based on only grade III-IV patients.</p>
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>10 CEAs found, and 3 CUAs found. Details on 9 CEAs and 3 CUAs given. Mainly based on the results of the GOG 111 trial (at the time most of these studies were completed GOG 111 was the only complete trial comparing paclitaxel plus cisplatin to any standard comparator. Modelling was used to extrapolate effectiveness from the trial length (48 months) to life years gained, or to estimate resource use in a 'real-world' scenario. All investigated paclitaxel combined with cisplatin. Most compared against cisplatin plus cyclophosphamide because this was the comparator in GOG 111, although in England carboplatin is more often used. No data exists directly comparing paclitaxel combined with carboplatin and carboplatin.</p> <p>For studies comparing paclitaxel plus cisplatin to cyclophosphamide plus cisplatin, the range of cost-effectiveness ratios for life years gained was £3,960 to £13,360. The low estimate was for Spain and the high was for Japan. Two cost-effectiveness studies done in England compared carboplatin alone to paclitaxel plus cisplatin. The range of costeffectiveness ratios for life years gained was £7173 to £12,417.</p> <p>One CUA completed in England compared paclitaxel plus cisplatin, carboplatin alone and also with cyclophosphamide plus doxorubicin plus cisplatin (CAP) to no treatment. Although superficially similar to the ICON 3 trial, data on response rates was obtained from a variety of disparate trials. Very few details on how QALYs gained were derived were given. Cost per quality adjusted life years were calculated for each regimen compared to no treatment, but an incremental analysis comparing treatments to each other was not done. Using the costs and quality of life estimates given in this analysis the incremental cost per QALY gained can be calculated. The incremental cost per</p>

	<p>QALY gained comparing paclitaxel/platinum to CAP is £5433, and versus carboplatin alone is £5273. The two non-British CUAs also address quality adjusted life years. The cost per QALY gained in the Messori study using the Q-TWIST method was £11,269. In the Ortega study, incorporating patient preferences, the cost per quality adjusted progression free life year gained ranged from a low of £6860 to a high of £10,377. In sensitivity analysis, the maximum cost per quality adjusted progression free life year gained was £18,000.</p> <p>The HTA report states that generalisability could be a problem because of a lack of specific information, source of efficacy, resource use and cost data and the assumptions that were made.</p>
Other issues noted (eg crossover)	All of the 4 RCTs identified allowed treatment crossover to alternate treatment. However this is not considered in the economic section. It seems likely that the reviewed papers did not account for crossover.

2. TA6: Breast cancer - taxanes (replaced by TA30), June 2000.

Guidance: As patients reach the appropriate stage in their treatment for advanced breast cancer, they should be offered either docetaxel (Taxotere) or paclitaxel (Taxol). The decision as to which product should be used should be taken by the responsible clinician in discussion with the patient taking into account the clinical trial data set out in full in the guidance. The use of Taxanes for adjuvant treatment of early breast cancer, or for the first-line treatment of advanced breast cancer, should be limited to clinical trials.

Source: HTA report: http://www.nice.org.uk/nicemedia/pdf/hta_ov_taxanes.pdf

Lister-Sharp D, McDonagh M, Khan KS, Kleijnen J. A Systematic Review of the Effectiveness and Cost-effectiveness of the Taxanes used in the Treatment of Advanced Breast and Ovarian Cancer, NHS Centre for Reviews and Dissemination, University of York, January 2000, Report commissioned by the NHS HTA Programme on behalf of the National Institute for Clinical Excellence

Appeal Panel Decision Document (Aventis), June 2000, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32042>

Appeal Panel Decision Document (BMS and CancerBACUP), May 2000, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32039>

Note that the manufacturer (BMS) made a submission, but this is not available from the NICE website, and details from it that had been included in the AC's version of the HTA report have been removed from the public version.

The Guidance above was made after the Appeals were heard. The Appeal points relevant for the survival analysis in the appraisal are described below:

Appeal (Aventis): Aventis argued that response rates and OS had been excluded from the Guidance, and thus that NICE had failed to consider all relevant data, and had failed to provide adequate reasoning for this. They also argued that the difference in cost-effectiveness conclusions in the ACD and FAD went unexplained. Aventis agreed that the Appraisal Committee properly included in the clinical criteria set out in the Guidance: response rate; progression free survival; overall length of survival; and quality of life. They complained, however, that the current draft of the Final Appraisal Determination ("FAD"), and the Institute's Guidance, ignored the considerable body of evidence relating to response rates and overall length of survival.

The Appeal Panel were informed that the Appraisal Committee had considered data relating to response rates and overall length of survival in considerable detail. The Committee considered, however, that response rates were a surrogate measure for clinical effectiveness, and that progression free survival was a more reliable and relevant measure. The Committee also considered that, in this situation, overall length of survival was a less reliable measure of the clinical effectiveness of an anti-cancer drug owing to numerous other factors affecting overall survival. Regarding the cost-effectiveness analysis the AC representative noted that upon their review of the cost-effectiveness of docetaxel and paclitaxel the Appraisal Committee considered the evidence tendentious. They considered that, in the light of their review, such comparisons were inappropriate; and that it was only possible to provide a broad range of cost per additional life year for taxanes generally.

Aventis also appealed that due to their concerns, the Guidance made was perverse.

The Appeal Panel were satisfied that the Appraisal Committee had not ignored the data on response rates and overall survival, and that they had fully considered the cost-effectiveness of the taxols, and so dismissed the appeal.

Appeal (BMS and Cancer BACUP): This appeal was specifically to do with the use of paclitaxel at 2nd line. Grounds 1a, 1c and 2a of the BMS appeal was that the Guidance has over-emphasised a single end-point (overall survival) in assessing clinical effectiveness in advanced breast cancer, applied unequal assessment of study data, and failed to provide adequate reasoning. The Appeal Panel noted that the Appraisal Committee had limited its appraisal of the clinical effectiveness of paclitaxel to those studies enrolling patients who had previously been treated with anthracyclines since this is the patient population for which the product is authorised in the UK. The Panel was unable to ascertain, however, whether the Committee had fully considered the additional studies shown in Table 2 (page 40) of BMS's Response to the Provisional Appraisal Determination. The Panel was also concerned that the Committee's possible failure to give adequate consideration to this data might have denied it the opportunity to appraise, fully, the clinical effectiveness of paclitaxel in the authorised indication. Therefore the appeal was upheld on this point. Ground 1b of BMS's appeal was that the Guidance incorrectly assessed the RCT evidence, and had not taken account of the crossover nature of the CA139-278 trial. The Panel decided that the AC had outlined the difficulties associated with assessing the evidence appropriately, and this appeal point was not upheld. BMS (ground 2b) and BACUP both appealed on the ground that the Guidance fails to recognise that paclitaxel and docetaxel are not the same. They have different structures and toxicities and patients would lose out if only docetaxel was

recommended. The Panel accepted that the OS evidence and toxicity evidence may not have been interpreted correctly (because the conclusion that there was no OS gain was based on a study which was not designed to find an OS gain, and due to mis-interpretation of a toxicity paper), and hence this appeal point was upheld.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>First Line Paclitaxel: Four RCTs were identified: EORTC, TITGANZ, E1193 and CA139-278. A total of 1425 patients were included. Of these, EORTC, E1193 and TITGANZ evaluated single agent paclitaxel and E1193 and CA139-278 evaluated combination paclitaxel/anthracycline. There were no economic evaluations identified.</p> <p>First Line Docetaxel: One Phase III trial of docetaxel as a first-line treatment for advanced breast cancer was identified. This was available only as a conference abstract and randomisation was not specifically mentioned. Consequently, the results should be treated with caution. Although a combination of docetaxel and doxorubicin produced greater overall response than doxorubicin and cyclophosphamide combined, there were no long-term results such as progression free or over-all survival.</p> <p>Second Line Paclitaxel: One phase II RCT was identified: CA139-047. A total of 81 patients were included. Seven economic evaluations were identified, 6 of which compared paclitaxel to docetaxel based on indirect comparisons. The only economic evaluation comparing paclitaxel to control (mitomycin) was submitted in confidence and thus removed from the HTA report.</p> <p>Second Line Docetaxel: Three completed RCTs were identified: 303 Study, 304 Study, and Scand. A total of 1092 patients were included. There were six economic evaluations.</p> <p>Overall, little data is given with regard to the precise survival data used – ie little comment is made on whether OS was used. It appears that the majority of models relied upon response rates for efficacy, rather than survival measures.</p> <p>1 study estimated quality adjusted PFS using a retrospective chart review and a decision analytic model. 1 study estimated QALYs using pooled estimates from RCTs for response rates and expert opinion using a decision analytic model with a 3-year timeframe. 2 studies used a Markov Model (one with a 3-year timeframe. The other with an unspecified timeframe) to estimate QALYs, using efficacy data from 2 RCTs. 1 study used a Markov Model with pooled estimates of response rates from RCTs, and clinical opinion, to estimate QALYs. 1 study used a Markov Model with pooled estimates of response rates from RCTs, and clinical opinion, to estimate quality adjusted progression free days.</p>
Evidence synthesis (pool survival estimates?)	No details.
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>7 CUAs, one of which included a CEA. 1 was excluded from the report because it was submitted in confidence.</p> <p>For studies comparing docetaxel to paclitaxel, the range of cost-utility ratios for QALYs was £1990 to £5233. Two studies did not present an incremental analysis. One found docetaxel to be the dominant strategy over paclitaxel, while the other found vinorelbine to be dominant over either taxane.</p> <p>In the three studies comparing docetaxel to vinorelbine, the one UK study found the cost per QALY gained of docetaxel was £14,050. A Canadian study found vinorelbine to be dominant. However, the third study comparing docetaxel and vinorelbine (from France) found the opposite; that docetaxel was dominant to vinorelbine. Although these two studies used similar rates of response for vinorelbine (16%) and paclitaxel (21-29%) the rates used for docetaxel were quite different (57% in the French study and 30% in the Canadian study).</p>
Other issues noted (eg crossover)	<p>Paclitaxel 1st Line: TITGANZ was analysed on an intention-to-treat basis and gave details of length of follow-up: 26 months. EORTC and E1193 allowed cross-over to alternate treatment and TITGANZ recommended treatment with epirubicin on progression. Patients crossing over in this way are violating the randomisation; however, no details were given as to whether such patients were censored.</p> <p>Paclitaxel 2nd Line: It was not clear whether trial CA139-047 was analysed on an intention to treat basis and no details were given of length of follow-up. However, the authors stated that most of the patients were alive at the time of analysis. Cross-over to alternate treatment was allowed - more than half the patients in the control arm crossed over to the paclitaxel arm; none crossed the other way. No details were given as to whether such patients were censored. In none of the economic evaluations was the estimation of benefits based on a direct clinical comparison.</p> <p>Docetaxel 2nd Line: The 303 and 304 studies were analysed on an intention to treat basis; Scand excluded a single patient. Length of follow-up ranged from 11 months</p>

	<p>(Scand) to 23 months (303 Study). At least two thirds of the participants of these studies had died. The Scand study recommended crossover to alternate treatment on objective signs of disease progression. Patients crossing over in this way are violating the randomisation; however, no details were given as to whether such patients were censored. In the economic analyses, there were no direct comparisons for the estimation of benefits.</p> <p>The HTA report states that generalisability could be a problem because of a lack of specific information, source of efficacy, resource use and cost data and the assumptions that were made in the economic evaluations. However no specific mention about survival analyses included in the evaluations is made in the economic section of the report.</p>
--	---

3. TA23: Brain cancer - temozolomide, April 2001.

Guidance: Patients with recurrent malignant glioma (brain cancer) who have failed first-line chemotherapy treatment with other agents (either because of lack of efficacy or because of side effects) may be considered for treatment with temozolomide. Such patients must have a histologically proven malignant glioma (WHO grades III and IV, or transformed grade II) at first relapse, recurrence or progression (as assessed by imaging), Karnofsky performance status greater than or equal to 70 and a projected life expectancy of 12 weeks or more, at initiation of temozolomide treatment.

Temozolomide is not recommended for first-line chemotherapy treatment for patients with malignant glioma who have failed primary therapy (surgery and/or radiotherapy), except in the context of a randomised controlled trial against a standard-treatment comparator.

As temozolomide is not currently licensed for adjuvant chemotherapy treatment of malignant glioma, its use in this indication has not been considered in this appraisal.

Source: Guidance on the use of temozolomide for the treatment of recurrent malignant glioma (brain cancer), NICE Technology Appraisal Guidance No. 23, April 2001
<http://guidance.nice.org.uk/TA23/Guidance>, accessed 23/12/09

Dimnes J, Cave C, Huang S, Major K, & Milne R, The effectiveness and cost-effectiveness of temozolomide for the treatment of recurrent malignant glioma. Draft report to NICE, Wessex Institute for Health Research and Development, 2000. <http://www.nice.org.uk/nicemedia/pdf/temozolomidehta.pdf>, accessed 23/12/09

Note: The manufacturer was Schering Plough. However no mention of the manufacturer submission is made in the guidance document or the HTA report.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>GBM. Data from the only RCT identified was used (Yung <i>et al</i> 2000, n=225). This presented data on the median PFS for each group, and the OS difference between treatment groups, so a separate study which demonstrated the median OS for patients on a range of different treatments was used to estimate survival when temozolomide is not given (Wong <i>et al</i>, n=458).</p> <p>AA. There was no RCT data. Median data from an uncontrolled study (Yung <i>et al</i> 1999, n=162) were used to estimate PFS and OS for temozolomide whereas the Wong <i>et al</i> study was again used to estimate survival when temozolomide is not given.</p>
Evidence synthesis (pool survival estimates?)	No details.
Survival model(s) fitted (Weibull, exponential etc)	None.
Independent survival models, or hazard ratio (proportional hazards) modelling	None.
Justification for survival model used?	None.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>A very simple economic model was used. This simply estimated PFS, OS and direct treatment costs and estimated an incremental cost-effectiveness ratio. CEA using progression free weeks and life years gained were undertaken, as well as CUA using QALYs.</p> <p>GBM. Cost per progression free week gained: £691-£1011 Cost per life year gained: £22,159 - £47,943 Cost per QALY gained: £24,454 - £175,256</p> <p>AA. Cost per progression free week gained: £554-£737 Cost per life year gained: £16,441 - £52,856</p>

	Cost per QALY gained: £24,089 - £127,743
Other issues noted (eg crossover)	<p>The authors recognised that using only median data was not ideal. They also highlighted the existence of only one RCT as important, particularly as the methods used in it were not well reported. The economic evaluation included data from uncontrolled studies and so the results are very uncertain.</p> <p>The authors note general problems with using survival analysis from trials. They state that “measures of progression and survival depend importantly on the timing of the baseline and follow-up evaluations. The point at which recurrence is detected and further treatment is initiated will affect the estimates of PFS and survival. Furthermore, when imaging is being performed more regularly than in normal practice, initial recurrence may be detected earlier producing longer estimates of survival. Likewise, however, additional progression after recurrence may also be detected earlier than in routine practice, thereby underestimating progression-free survival. Therefore, the results for both PFS and survival may not be directly generalisable to clinical practice.”</p> <p>The authors also note that using results from an RCT for one arm in an economic model and data from a separate uncontrolled trial in the other arm of the model is unlikely to be accurate, and can only give an idea of the potential cost-effectiveness of the treatment.</p> <p>Crossover is not mentioned.</p>

4. TA25: Pancreatic cancer - gemcitabine, May 2001.

Guidance: Gemcitabine may be considered as a treatment option for patients with advanced or metastatic adenocarcinoma of the pancreas and a Karnofsky performance score of 50 or more, where first line chemotherapy is to be used.

Gemcitabine is not recommended for patients who are suitable for potentially curative surgery, or patients with a Karnofsky score of less than 50.

There is insufficient evidence to support the use of gemcitabine as a second line treatment in patients with pancreatic adenocarcinoma.

Source: Guidance on the use of gemcitabine for the treatment of pancreatic cancer, NICE Technology Appraisal Guidance No. 25, May 2001, <http://www.nice.org.uk/nicemedia/pdf/gemcitabineguidance.pdf>, accessed 05/01/10

Ward S, Bansback N, Morris E, Calvert N. A review of the Clinical and Cost-effectiveness of Gemcitabine for the Treatment of Pancreatic Cancer. Health Technology Assessment Report, 4th December 2000. <http://www.nice.org.uk/nicemedia/pdf/GemcitabineHTARreport.pdf>, accessed 05/01/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>In the first line setting, five published randomised controlled trials (RCT) were identified. One trial compares gemcitabine with a bolus infusion of 5-FU and another with intra-arterial 5-FU (abstract only). In addition, three trials compare gemcitabine to metalloproteinase inhibitors (two trials involve marimastat and one BAY12 9566).</p> <p>There is thus only one fully reported single-blind RCT, reported by Burris <i>et al</i>, which compares gemcitabine to treatment with 5-FU as a first line treatment used on patients with a Karnofsky score of 50 or more (n=126). In this trial, patients randomised to gemcitabine had significantly better one-year survival (18% vs 2%, p=0.0002), significantly better median survival (5.65 vs 4.41 months, p=0.0025), improved median progression free survival (2.33 vs 0.92 months, p=0.0002) and a longer time to treatment failure (2.04 vs 0.92 months, p=0.0004.). However, in this trial the comparator, 5-FU, was given by bolus injection, not the usual current means of administration; and the results of this trial may be prone to bias due to lack of blinding of the investigators. Furthermore, the 12-month survival rate of 2% in 5-FU group is unusually low when compared with other published 5-FU studies.</p> <p>No relevant RCTs were identified which examine the effect of gemcitabine as a second line treatment in patients with relapsed disease.</p> <p>The review group state that the area under a survival curve indicates the overall survival time</p>	<p>Eli Lilly's economic evaluation for first line treatment compares gemcitabine with 5-FU using outcome and resource use data made available to them from the Burris trial. This data was not available to SchARR.</p> <p>Three endpoints were considered including progression-free survival, clinical benefit response, and the primary outcome of total survival time. Outcome results from the Burris trial were used to derive estimates for these 3 variables. Kaplan-Meier survival curves were used to estimate area under the curve and 95% CI for mean survival. Discounting of benefits at 6% appears to have been calculated though this is not clear from the table of results. They estimate an overall mean survival for gemcitabine patients of 6.79 months, compared with 4.52 months for 5-FU patients, an incremental benefit of 2.27 months (0.19 life years). This figure is considerably higher than the median difference of 1.24 months reported by Burris <i>et al</i>.¹² The mean survival gain for Gemcitabine patients, estimated by Eli Lilly, is 1.14 months longer than the median figure reported by Burris (an 83% increased difference). The incremental differences for progression free survival and CBR are reported at 1.39 months and 19% respectively. Only OS is used in the survival analysis,</p>

	<p>experienced by the cohort. Therefore, the area between the gemcitabine and 5-FU curves indicates the mean difference in survival experienced by the two groups. The survival gain, taken from the Eli Lilly industry submission, estimated in this way gave 2.27 months (6.79 months vs 4.52 months). SchHARR did not have access to the raw data on which this figure was based. However their estimate, based on the same methodology, but using the published curves from the Burris paper produced a figure close to the Eli Lilly value (6.734 (KM), 6.744 (T) vs 4.345 (KM), 4.351 (T)). Thus the Eli Lilly value was used in the cost-effectiveness calculations. For the review group's analysis only OS was used, not time until progression, for the analysis of survival. However time until progression was used to estimate treatment duration. And A Q-TWIST analysis was used in an illustrative cost per QALY analysis – although for this the estimate of time spent in each state is not described – proportions spent in each state taken from the Burris trial may have been used.</p> <p>The group note that due to the shape of the survival curves in the Burris trial, the median survival, although a useful measure in assessing clinical efficacy, under estimates the area under the curve. In the economic analysis, the mean survival gain used in conjunction with the mean cost gives a better indication of the cost-effectiveness ratio.</p>	not time until progression.
Evidence synthesis (pool survival estimates?)	Not applicable.	Not applicable.
Survival model(s) fitted (Weibull, exponential etc)	<p>Of the two published economic evaluations found, for one the survival analysis method is not reported in the HTA report. For the other (Trippoli and Messori, 1999) it is reported that a Gompertz model was fitted to survival data from Burris <i>et al</i> to extrapolate survival estimates. This resulted in a survival gain estimate of 2.9 months compared to the medial gain of 1.24 from the clinical trial.</p> <p>Kaplan Meir curves were used to estimate mean survival. Appendix 3 shows that KM curves were complete, so this method may be reasonable. The Kaplan-Meier product-limit estimator was used to estimate the mean and, then, this was verified by using the Trapezoidal rule so to define errors. In the first method, the mean survival time is calculated by multiplying the difference in each time step by the proportion of patients still alive. Datapoints were extrapolated at each step.</p> <p>In the second method data points were extrapolated from the graph at very small time steps. The area was then calculated using the Trapezoidal rule, a simple numerical integration technique.</p>	Area under Kaplan Meir curves were used to estimate mean survival.
Independent survival models, or hazard ratio (proportional hazards) modelling	No models.	No models.
Justification for survival model used?	None given.	None given.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>Two published economic evaluations of gemcitabine as first line therapy for pancreatic cancer were identified, one of which was only available in abstract form. An economic evaluation of gemcitabine, in both first and second line treatment, was included in the Eli Lilly submission. SchHARR, on behalf of NICE, undertook its own analysis.</p> <p>All of the economic analyses submitted drew on the effectiveness data from the single RCT by Burris <i>et al</i>. For first line treatment the estimates for cost per life year gained ranged from approximately £7,200 to £18,700 dependent on the 5-FU regimen used as comparator. These figures are very sensitive to reduced estimates of survival benefit over comparators.</p> <p>An area under the curve based model was used by the review group – mean survival estimates were combined with cost estimates. The base case measure cost per LYG, however as an illustration a Q-TWIST analysis is also run, using three health states identified in the Burris trial: time in clinical benefit, time before progressive disease when not in clinical benefit, and time from disease progression to death. No utility scores were available for these so scores of 1, 0.5 and 0.5 were used and sensitivity analysis was run around these.</p>	Area under the curve based model. Results are based on cost per LYG. Costs are applied to the estimated survival benefit.

Other issues noted (eg crossover)	Sensitivity analysis was run reducing survival gains proportionately, based on lower median survival seen for gemcitabine in other studies. Time on treatment (time until progression) was also reduced proportionately in these analyses. No mention is made of possible treatment crossover within the trial used to inform the economic models.	The review group notes that the Eli Lilly survival assumptions are higher than the median figures reported in the trial, and there is some doubt about the survival gains demonstrated in the Burris trial (discussed earlier in the current report). These factors may mean that Eli Lilly's central cost per LYG estimate of £12,200 is a significant under-estimate of the actual ratio. This is a justifiable worry based on concerns about the trial, but the review group do state that mean is better.
-----------------------------------	---	---

5. TA26: Lung cancer - docetaxel, paclitaxel, gemcitabine and vinorelbine (updated by and incorporated into CG24 Lung cancer), June 2001.

Guidance: The use of these chemotherapy drugs for non-small cell lung cancer (NSCLC) should be considered as an option for people who are not suitable for, or who are unlikely to respond to, potentially curative treatment.

Gemcitabine, paclitaxel, and vinorelbine should each be considered as part of initial (first-line) chemotherapy treatment for people with advanced non-small cell lung cancer. The use of any one of these drugs in combination with a chemotherapy that is platinum-based, is likely to be the most-effective form of treatment.

Docetaxel used on its own should be considered as a treatment for people who's cancer is 'locally advanced' or has spread to other parts of the body (metastatic cancer), but only if they have suffered a relapse after receiving initial chemotherapy with other agents

Source: NICE Issues Guidance on Chemotherapy Drugs for Lung Cancer, Press Release, 12 June 2001, NICE 2001/020, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32199>, accessed 06/01/10

Scott DA, Clegg A, Sidhu M, Hewitson P, Waugh N. Clinical and Cost-effectiveness of Paclitaxel, Docetaxel, Gemcitabine and Vinorelbine in Lung Cancer. Wessex Institute for Health Research and Development, Report commissioned by the NHS HTA Programme on behalf of the National Institute for Clinical Excellence, 9th January 2001. <http://www.nice.org.uk/nicemedia/pdf/lungcancerhtareport.pdf>, accessed 06/01/10

Note: There is no reference to any manufacturer submissions in the HTA report. The guidance document is not available because the appraisal has been updated, hence the guidance noted above is taken from the accompanying press release from the original guidance.

Assessment Group Model / Evidence Review Group Alterations	
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>Median rather than mean survival data has been used in the analysis. Efficacy data (median survival and number of drug cycles) was calculated by averaging across a number of studies rather than be reliant on one particular study. Summary statistics from trials were used.</p> <p>Efficacy was presented in terms of median survival rather than response, which some studies have used, since response is not necessarily indicative of increased length of life.</p> <p>A reliance on trial summary statistics was the reason behind using median data. The authors note that it is debatable whether mean or median values should be used in cost-effectiveness calculations. They note that means are certainly preferable when considering budgetary impact but that they were forced to use median survival since that is what is reported in the literature. Only one paper identified by the authors had recalculated mean survival from the raw data reported in trials which they state is beyond the scope of this review. In all cases, mean LYS was higher than medians, though the least difference was for BSC (0.49 mean, 0.43 median). The HTA group used median based survival rather than the available mean data in their baseline model for a number of reasons: the inability to check the authors data; mean survival is only calculated for some of the regimens that needed to be considered; their mean calculations comprise data from only one trial (a phase II in the case of GEM) whilst the HTA group estimates were based on the average across a number of trials; and in their recalculations of the raw data they have only considered stage IV patients, an approach inconsistent with the rest of the HTA group model. Given the mean calculations all being greater than the medians the effect of introducing means would be to increase the favourability of the regimens in the model – this was confirmed by sensitivity analysis using mean data.</p>
Evidence synthesis (pool survival estimates?)	Data on median cycles, efficacy (in terms of median survival), dosages (in mg/m ²), and number of patients per study arm were extracted from each phase III trial and certain robust phase IIs. For the survival analysis, data was gathered from 23 studies. These form the backbone for the construction of the UK model. Where a range of findings are given, the baseline has followed the majority, in terms of median cycles and dosage or the study with the largest sample size. In the case of median survival, the decision was taken to obtain an average across studies reporting the data rather than anchor the model to one particular set of data. Baseline was therefore determined by

	survival weighted by number of patients, best and worst estimates were defined by the upper and lower bounds of the data. Numbers of patients were used to weight the data rather than Jadad given the lack of variation of the Jadad scores, and the problems with such scores. In the case of PAX, where dosages (and thus drug cost) varied markedly between studies, this approach was not taken. Instead, several strategies were examined using the different PAX dosages. Although dosages varied, survival remained fairly consistent.
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	None.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	The HTA group found 16 economic evaluations of interest (none of which were UK based). These were reviewed briefly, but survival analysis techniques used within them are not well described. The HTA group went on to develop their own economic model, designed for the UK setting. The HTA group developed a simple model which estimated overall survival gains for each regimen based on median survival taken from a range of studies. They applied drug, administration, adverse event and other costs to estimate incremental costs per life years gained. Thus there is implicitly only 2 health states – alive and dead. Due to a lack of good utility data, they reviewed any available quality of life data for each regimen and though this did not allow utility scores to be estimated, the authors concluded that each regimen would at least not worsen quality of life.
Other issues noted (eg crossover)	Treatment crossover is not mentioned in the economic section of the report. However, in the efficacy section it is noted that two studies for one of the drug regimens used a crossover design. In one other study no crossover was allowed.

6. TA28: Ovarian cancer - topotecan (replaced by TA91), July 2001

Guidance: Topotecan should be considered as one of the treatment options for women with advanced ovarian cancer if first-line chemotherapy has not been successful.

Topotecan is not recommended for women who have an ECOG (Eastern Cooperative Oncology Group) score of 3 or below (see notes for editors for further details)

Topotecan is also not recommended for women who have an obstruction in their bowel which is caused by the cancer or who have already been treated with topotecan or another drug of the same type.

A woman's response to treatment should be monitored carefully. If there is evidence that the cancer has progressed, then treatment with topotecan should be stopped. Also a reduction in the ECOG (described above) may also be a reason to stop treatment.

Source: NICE Issues Guidance on Topotecan for Ovarian Cancer, Press Release, 6 August 2001, NICE 2001/028, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32219>, accessed 06/01/10

Forbes C, Shirran L, Bagnall AM, Duffy S, Ter Riet G. A rapid and systematic review of the clinical effectiveness and cost-effectiveness of topotecan for ovarian cancer. NHS Centre for Reviews and Dissemination, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 15th February 2001. <http://www.nice.org.uk/nicemedia/pdf/lungcancerhta-report.pdf>, accessed 06/01/10

Note: SmithKline Beecham and Schering Plough were the manufacturers, and both made submissions. However the data contained in these submissions was regarded as commercial in confidence and therefore very few details of the economic evaluations conducted by the manufacturers' are given in the HTA report. The HTA group did not conduct their own economic evaluation and thus few details can be obtained regarding how the survival analysis was conducted in the economic section of this NICE appraisal.

The guidance document is not available because the appraisal has been updated, hence the guidance noted above is taken from the accompanying press release from the original guidance.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	The published cost consequence analysis used effectiveness data from a number of published studies identified through a literature search. The data from these studies was presented for a number of relevant outcomes including response rate, the presence of progressive disease, median time to progression and median survival. The data clearly showed differences in efficacy between the drugs considered, but the authors' assumed equivalence and conducted a cost minimisation analysis anyway. The SmithKline Beecham submission used data from the findings of trial 039 comparing topotecan with paclitaxel. From the HTA report we can gather that a number of

	<p>different outcome measures were used in the cost-effectiveness analyses. Results in terms of 'cost per...' are reported for per week of survival, response rates, and per TWIST (time without toxicity or symptoms). It is of note that in this evaluation cost ratios are only presented for patients who respond to treatment.</p> <p>The Schering Plough submission assumed clinical equivalence between topotecan and caelyx (or rather, that caelyx was at least as good as topotecan, based on the 30-49 trial), and so no survival analysis was included in the economic model. The analysis was based on 474 patients in the 30-49 trial.</p> <p>Kaplan-Meier overall survival data, as well as median values and hazard ratios are presented in the clinical section of the report, with permission from SmithKline Beecham and Schering Plough. Response data and time to progression data is also presented. Given this data, it would have been possible for the manufacturers to produce economic evaluations with 3 or more health states using survival data. However no information regarding what the two companies used in their economic evaluations is given due to commercial in confidence data. Therefore we do not know how the survival analysis in the economic models was conducted. However because Schering Plough conducted a cost minimisation analysis we can assume that no survival analysis was included within their economic model.</p>
Evidence synthesis (pool survival estimates?)	No details.
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>Seven published manuscripts reporting details of two studies of clinical effectiveness and one economic evaluation were selected for inclusion. Further details of the two clinical effectiveness studies and two new economic evaluations were identified from confidential drug company submissions.</p> <p>The three economic evaluations included in the review comprised one cost minimisation analysis comparing topotecan with caelyx, one cost-consequence analysis comparing topotecan with three other drugs (paclitaxel, etoposide and altretamine), and one cost-effectiveness analysis comparing topotecan with paclitaxel.</p> <p>The HTA group reported that the cost-consequence study (comparing topotecan with three comparators) was of poor quality and was of little relevance to the UK NHS. The two remaining evaluations (one cost minimisation and one cost-effectiveness analysis) were of reasonable quality overall. The HTA group found further limitations to the cost-effectiveness evidence, but these are marked as commercial in confidence.</p> <p>The cost consequence analysis was from a published paper. The cost-effectiveness analysis was the manufacturer submission from SmithKline Beecham comparing topotecan and paclitaxel and modelled a hypothetical group of 1000 patients. The cost minimisation was the manufacturer submission from Schering Plough, comparing topotecan and caelyx.</p>
Other issues noted (eg crossover)	<p>It is of note that in the SmithKline Beecham evaluation cost ratios are only presented for patients who respond to treatment.</p> <p>Crossover is not mentioned in the economic section of the report. However the appendix data on trial 039 shows that it had a crossover design, as after the randomised phase of the trial patients' crossover over treatments. It is stated that: "this part of the trial is referred to as the crossover trial but is not considered here". This appears to mean that only data up until the end of the randomised phase were included, based on the ITT analysis. It is also stated that in this trial patients who progressed during treatment were removed from the study. Those whose best response was stable disease after 6 courses were removed or switched to the other treatment – it is these patients who were then included in the 'crossover' section of the study (n=110). Because data presented in the HTA report is only for the ITT phase, these data will not be included in the OS analysis presented. It is not clear how this was dealt with in the economic analysis.</p>

7. TA29: Leukaemia (lymphocytic) - fludarabine (replaced by TA119), Sept 2001

Guidance: Oral fludarabine is recommended as second line therapy for B-cell chronic lymphocytic leukaemia (CLL) for patients who have either failed, or are intolerant of, first line chemotherapy, and who would otherwise have received combination chemotherapy of either:

- o cyclophosphamide, doxorubicin, vincristine and prednisolone (CHOP)
- o cyclophosphamide, doxorubicin and prednisolone (CAP) or
- o cyclophosphamide, vincristine and prednisolone (CVP).

The oral formulation of fludarabine is preferred to the intravenous formulation on the basis of more favourable cost-effectiveness. Intravenous fludarabine should only be used when oral fludarabine is contra-indicated.

Source: NICE Issues Guidance on Chemotherapy Agents for Breast Cancer and Leukaemia, Press Release, 25 September 2001, NICE 2001/031, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32242>, accessed 07/01/10

Wake B, Bryan S, Barton P, Fry-Smith A, Davenport C, Song F, Hyde C. Fludarabine as Second Line Therapy for B-Cell Chronic Lymphocytic Leukaemia. Birmingham: University of Birmingham, Department of Public Health and Epidemiology, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 9th January 2001. <http://www.nice.org.uk/nicemedia/pdf/fludarabine1.pdf>, accessed 07/01/10

Fischer AJ. Fludarabine Annex: Cost-effectiveness. Prepared by AJ Fischer of the NICE Appraisal Team, May 2001. <http://www.nice.org.uk/nicemedia/pdf/fludarabine3.pdf>, accessed 07/01/10

Note: Schering were the manufacturers, and they made a submission. Although only a small amount of data in the HTA report is marked as commercial in confidence, there are relatively few details of the manufacturer's model presented, and the manufacturer's submission is not available on the NICE website. The guidance document is not available because the appraisal has been updated, hence the guidance noted above is taken from the accompanying press release from the original guidance.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>Only one economic evaluation was found by the HTA group – that was the manufacturer's submission. The effectiveness data used in the evaluation was from a Phase III trial (French Cooperative Group on CLL, 1996) for iv fludarabine vs CAP, and "expert opinion" for CHOP. The effectiveness data used in the economic model was based on response rates, response durations and expected disease free days. Thus it is not clear if overall survival estimates were used.</p> <p>There were 96 relevant patients included in the RCT. The clinical section of the HTA report shows that median survival was 728 days in the fludarabine group and 731 days in the CAP group, whereas the time to progression was a median of 324 days versus 179 days. Therefore it is to be expected that the cost-effectiveness analysis did not focus on overall survival.</p> <p>An accompanying NICE report on the cost-effectiveness analysis notes that the manufacturer's submission concentrates on expected time in remission in the cost-effectiveness analysis. However, it is noted that it is not clear whether this time is a median or a mean.</p>
Evidence synthesis (pool survival estimates?)	None.
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	None.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	The manufacturer conducted a cost-effectiveness analysis. Results were reported as cost per year of remission gained. It is not clear what type of model was used. Cost data from a separate observational resource use study were combined with effectiveness evidence from the RCT to estimate cost-effectiveness. The HTA group did not conduct an independent economic analysis.
Other issues noted (eg crossover)	Treatment crossover is not mentioned in the HTA report. However given that the survival of included patients was relatively high, it seems likely that patients will have received post progression treatments.

TA30: Breast cancer - taxanes (review)(replaced by CG81), Sept 2001

Guidance: This appraisal reviewed TA6, completed in June 2000. New evidence was considered but the original guidance was not altered.

The use of docetaxel in combination with an anthracycline in first-line treatment of advanced breast cancer is not currently recommended. As paclitaxel is not licensed for first-line use with anthracycline its use has not been considered in this indication.

Docetaxel and paclitaxel are recommended as an option for the treatment of advanced breast cancer where initial cytotoxic chemotherapy (including an anthracycline) has failed or is inappropriate.

The taxanes are not currently licensed in the UK for adjuvant treatment of early breast cancer therefore their use in this indication should be limited to randomised clinical trials.

Source: NICE Issues Guidance on Chemotherapy Agents for Breast Cancer and Leukaemia, Press Release, 25 September 2001, NICE 2001/031, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32242>, accessed 07/01/10

Bagnall AM, Forbes C, Lewis R, Golder S, Kleijnen J. An update of a rapid and systematic review of the effectiveness and cost-effectiveness of the taxanes used in the treatment of advanced breast cancer. NHS Centre for Reviews and Dissemination, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 27th April 2001. <http://www.nice.org.uk/nicemedia/pdf/taxanesreviewhtareport.pdf>, accessed 07/01/10

Note: Bristol-Myers Squibb and Aventis were the manufacturers, and they both made submissions. Aventis included an economic evaluation alongside a clinical trial for docetaxel, but no details of this are given in the publicly available HTA report due to data being commercial in confidence. The manufacturer submissions are not available on the NICE website.

The guidance document is not available because the appraisal has been updated, hence the guidance noted above is taken from the accompanying press release from the original guidance.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	No new economic data for paclitaxel alone or the combination of paclitaxel / anthracycline at first line were found. There had been no economic evaluations of this combination in the original report. There was a small amount of new clinical data (2 interim reports of RCTs presented as abstracts). The manufacturer made a submission for docetaxel at first line, including an economic evaluation. This was not in the original appraisal. All data referring to this are removed from the HTA report. No new economic evaluations of clinical data were found for either taxane at second line. The original report included a review of 7 economic evaluations.
Evidence synthesis (pool survival estimates?)	No details.
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	See TA6.
Other issues noted (eg crossover)	As stated in the review of TA6, above, treatment crossover was an issue in a number of the trials used within the published economic evaluations reviewed by the HTA group. This update review did not offer any description of this and thus any treatment crossover and how it was dealt with in the published economic evaluations was not considered.

9. TA34: Breast cancer - trastuzumab, March 2002

Guidance: Trastuzumab in combination with paclitaxel (combination trastuzumab is currently only licensed for use with paclitaxel) is recommended as an option for people with tumours expressing human epidermal growth factor receptor 2 (HER2) scored at levels of 3+ who have not received chemotherapy for metastatic breast cancer and in whom anthracycline treatment is inappropriate.

Trastuzumab monotherapy is recommended as an option for people with tumours expressing HER2 scored at levels of 3+ who have received at least two chemotherapy regimens for metastatic breast cancer. Prior chemotherapy must have included at least an anthracycline and a taxane where these treatments are appropriate. It should also have included hormonal therapy in suitable oestrogen receptor positive patients.

HER2 levels should be scored using validated immunohistochemical techniques and in accordance with published guidelines. Laboratories offering tissue sample immunocytochemical or other predictive tests for therapy response should use validated standardised assay methods and participate in and demonstrate satisfactory performance in a recognised external quality assurance scheme.

Source: Guidance on the use of trastuzumab for the treatment of advanced breast cancer, NICE Technology Appraisal Guidance No. 34, March 2002 <http://www.nice.org.uk/nicemedia/pdf/advancedbreastcancerno34PDF.pdf>, accessed 07/01/10

Lewis R, Bagnall AM, Forbes C, Shirran E, Duffy S, Kleijnen J, ter Riet G, Riemsma R. A rapid and systematic review of the clinical effectiveness and cost-effectiveness of trastuzumab for breast cancer. NHS Centre for Reviews and Dissemination, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 4th October 2001. <http://www.nice.org.uk/nicemedia/pdf/trastuzumabassessmentreport.pdf>, accessed 07/01/10

Note: Roche were the manufacturers, and made a submission. However comprehensive details of the economic evaluations submitted are not given in the assessment report and the manufacturer's submission is not available on the NICE website.

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The industry submission data included two economic evaluations. One evaluated trastuzumab as combination therapy with paclitaxel versus paclitaxel alone and one evaluated trastuzumab as monotherapy versus vinorelbine. Several important elements relating to the methods of both economic evaluations were classified as confidential. The industry submissions were the only relevant economic evaluations that were found.</p> <p>The HTA group reported that the economic evaluation of trastuzumab as combination therapy (for first line therapy for MBC) was relatively well conducted. The incremental cost per LYG for trastuzumab in combination with paclitaxel was £14,069 and the cost per QALY was £29,448. However, it is important to note that the data on survival was extrapolated from survival curves that only included participants who did not switch to trastuzumab on disease progression, all of whom had very poor prognosis and died during the trial. The effectiveness data was derived from a single RCT which included 469 participants. However, only a subset of these participants were included in the economic evaluation, i.e. only those who had received paclitaxel chemotherapy (n=188/469), and had MBC overexpressing HER2 at level 3+ (n=349/469). It was not stated exactly how many were included in the economic evaluation. For ethical reasons, all participants with confirmed disease progression in the effectiveness trial were entitled to enroll on the follow-on protocol (study H0659g) which meant that they could receive trastuzumab. Only participants that did not switch to receive trastuzumab upon progression were included in the analysis. Clinical effectiveness was assessed using time to disease progression as the primary endpoint. Secondary endpoints included response rate, duration of response, 1-year survival and quality of life. It is not stated how and which of these endpoints were used in the economic model.</p> <p>However, it is stated that median overall survival data was used: The overall median survival data used in the economic evaluation of trastuzumab combination therapy was based on a sub-population of participants included in the RCT (those relevant to the licensed indication). This subset of participants included those who had received paclitaxel, with or without trastuzumab for the treatment of MBC overexpressing HER3 at level 3+ and did not cross over to the follow-on study (H0659g). When considering this sub-population of patients, who did not switch to trastuzumab on disease progression, the addition of trastuzumab to paclitaxel resulted in an increase in the median survival of 17.9 months (6.2 months for paclitaxel alone and 24.1 months for trastuzumab plus paclitaxel). However, when considering all HER2 3+ participants who received paclitaxel, the survival advantage resulting from the addition of trastuzumab was 7 months (median survival was 18 months for paclitaxel alone and 25 months for trastuzumab plus paclitaxel). This large difference in the overall median survival advantage when only including participants who did not cross over to the follow-on study (H0659g) was not explored in the sensitivity analysis.</p> <p>We are also told that survival data used in the economic evaluation was extrapolated from survival curves that only included participants who did not switch to trastuzumab on disease progression, all of whom had very poor prognosis and died during the trial – therefore some kind of extrapolation occurred. This would not seem to tally with using median survival estimates, as for these extrapolation would not be required (unless the median survival had not been reached at the time at which the study reported – which does not appear to have been the case).</p> <p>In the FAD, we are told that after extrapolating the trial results for the relevant selection of patients, approximately 10 months mean survival advantage was imputed into the economic evaluation. In the FAD an ICER of £37,500 is quoted, different from that included within the assessment report. We therefore have to assume that further updates to the economic analysis were made after completion of the assessment report, details of which are not published on the NICE website. It is possible that one of the updates included using mean survival estimates rather than medians – although it is not entirely clear which estimates were originally used in the economic model.</p> <p>The HTA group reported that the economic evaluation of trastuzumab as monotherapy was not considered to be as good and the cost analysis was of limited validity as was the effectiveness evidence it was based on. The effectiveness data relating to trastuzumab was derived from a non-randomised study that included 222 women who had received prior chemotherapy for MBC (Roche study H0649g). Some supportive data were also derived from preliminary analysis of a study using trastuzumab as first line therapy for MBC (Roche study H0650g). The health outcome used in the economic evaluation was median survival.</p>
Evidence synthesis (pool survival estimates?)	<p>For monotherapy two separate studies were used – no head-to-head data were available.</p> <p>For combination therapy one RCT was used, but only a subset of the data was included in the economic model.</p>
Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	No details.

Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>For monotherapy, cost-effectiveness analysis using life years gained was conducted.</p> <p>For combination therapy both a cost-effectiveness analysis (using life years gained) and a cost-utility analysis were presented. It is stated that a state transition model was used, although the health states are not explicitly described. It seems likely that the health states were stable disease, progressive disease and death.</p>
Other issues noted (eg crossover)	<p>Only a subset of the patients included in the RCT which was used for the combination therapy economic evaluation were included in the economic model. It was not stated how many patients were left in the analysis included in the economic model once only patients who received paclitaxel, had HER2 3+ and didn't switch onto trastuzumab following progression were included. The HTA report states that the RCT included 469 participants, of whom only 188 received paclitaxel. Of these, only participants who had MBC overexpressing HER2 at level 3+ were included in the economic evaluation. Furthermore, 75% of participants who were randomised to receive paclitaxel alone switched to trastuzumab plus paclitaxel on disease progression, and only the remaining 25% were included in the economic evaluation. Heavy assumptions will therefore have been made about the survival data.</p> <p>In addition, the HTA group report that although limiting the analyses to the licensed group of patients seems justified, this large difference in the overall median survival advantage when only including participants who did not cross over to the follow-on study (H0659g) should have been explored in the sensitivity analysis.</p> <p>The HTA report does not make clear for the combination therapy analysis whether median survival data or extrapolated survival curves were used.</p> <p>The use of a non-randomised trial for trastuzumab and an RCT for the comparator drug (vinorelbine) in the monotherapy evaluation was a problem and a potential source of selection bias – the underlying populations may not have been similar.</p> <p>The HTA group bring up another possible cause of bias: For the RCT of trastuzumab combination therapy (study H0648g) and one case series of trastuzumab monotherapy (study H0649g), the primary outcome measure and the incidence of congestive heart failure was assessed by an independent committee that was blinded to treatment group assignment. However, other outcomes were assessed by the investigators who were not reported to have been blinded to treatment group assignment. None of the included studies reported blinding of the administrators or participants (to having received trastuzumab). Whilst blinding in cancer trials is acknowledged to be difficult to undertake due to the nature of the disease and of the drugs being given, blinding is important in that it avoids observer bias and is therefore essential for any subjective clinician evaluating outcome measures such as alleviation of symptoms and QOL. Previous research has shown that non-blinded studies can overestimate the treatment effect. Non-blindness of administrators can result in biased administration of co-interventions.</p> <p>In addition the HTA group note that it is important in any trial that baseline characteristics are comparable between intervention groups. The most important baseline characteristics, as determined by the expert panel for the review, were not all reported on for the trastuzumab combination trial or studies of monotherapy. It cannot therefore, be assumed that the participants in each treatment group did not differ with respect to these factors.</p> <p>They also state that when reporting a RCT with survival-type data the recommended appropriate summary statistics that should be used are the log hazard ratio and its variance. For the trastuzumab combination therapy trial no hazard ratio or measure of its variance were reported. However, the analysis relating to median survival and duration of response, for the trial and one case series of trastuzumab monotherapy, were reported to have been based on Kaplan-Meier methodology, which means that the time to event was explicitly considered for each individual in the study. For the RCT only the P value of the log rank test was reported along with the median time, and for the case series only the median time was given.</p> <p>The HTA group also state that response to treatment is a surrogate outcome measure for assessing the effects of treatment on survival or quality of life. Because women with MBC have such poor prognosis, tumour shrinkage may alleviate symptoms (especially pain) and improve quality of life, which means that information relating to complete or partial response would be important but not independent from quality of life. However, alleviation of symptoms was not addressed by most included studies, which is surprising as these outcomes are probably the most important for this patient group. Therefore, as partial response is a surrogate measure for complete response, conclusions about effectiveness should be drawn from the complete response findings. Conclusions should not be drawn on the findings of partial response when used as a surrogate measure, unless outcomes relating to symptom relief are also reported or the results of both partial and complete response are in the same direction.</p> <p>It is worthy of note that in the NICE FAD the cost-effectiveness of trastuzumab in combination therapy is estimated as £37,500, which is different from the £29,448 quoted in the HTA report. £37,500 does not appear anywhere in the assessment report, and no other documents on the NICE website show how this figure was reached. The FAD tells us that after extrapolating the trial results for the relevant selection of patients, approximately 10 months mean survival advantage was imputed into the economic evaluation. We therefore have to assume that further updates to the economic analysis were made after completion of the assessment report, details of which are not published on the NICE website. It is possible that one of the updates included using mean survival estimates rather than medians – although it is not entirely clear which estimates were originally used in the economic model.</p>

10. TA33: Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (replaced by TA93), March 2002

Guidance:

NICE has recommended oxaliplatin as a first line combination treatment with 5-fluorouracil and folinic acid (5FU/FA) for patients where the cancer has only spread to the liver and may be operable after treatment.

Routine first-line treatment using Oxaliplatin and 5FU/FA combination treatments are not recommended.

Irinotecan is recommended as a second line monotherapy for patients where 5FU containing treatment has failed or where 5FU is inappropriate. Irinotecan combination treatments with 5FU/FA are not recommended as routine first-line treatment.

Raltitrexed is not recommended for use outside appropriately designed clinical studies.

Source: NICE recommends selective use of drugs for advanced colorectal cancer, Press Release, 2002, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32310>, accessed 07/01/10

Lloyd Jones M, Hummel S, Bansback N. A Review of the Evidence for the Clinical and Cost-effectiveness of Irinotecan, Oxaliplatin and Raltitrexed for the Treatment of Advanced Colorectal Cancer, University of Sheffield School of Health and Related Research, Report commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 15th January 2001. <http://www.nice.org.uk/nicemedia/pdf/AdvancedcolorectalcancerAssessmentReport.pdf>, accessed 07/01/10

Note: Aventis, Sanofi-Synthelabo and Astra Zeneca were the manufacturers, and all made a submission. However comprehensive details of the economic evaluations submitted are not given in the assessment report and the manufacturer's submission is not available on the NICE website. Note that Astra Zeneca did not provide a de novo economic analysis. Note that the assessment report states that Aventis had responded to the group's comments regarding their model and made alterations, but no further report documents these and what is contained in the assessment report relates to version 1 of the Aventis model. The guidance document is not available because the appraisal has been updated, hence the guidance noted above is taken from the accompanying press release from the original guidance.

A number of appeals were made against the Institute's Final Appraisal Determination on the use of irinotecan, oxaliplatin and raltitrexed for colorectal cancer. Elements of the appeals were upheld. The appeal did not directly concern how the survival analysis was conducted in the economic evaluations. In December 2001 a further appeal was lodged against the Institute's second Final Appraisal Determination. This appeal was not upheld. This appeal did contain elements relevant for the cost-effectiveness analysis and also the estimates of survival used in the cost-effectiveness analysis, linking to the attributability of OS gains to particular treatments in the face of treatment crossover. This shows the controversy on this matter.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>For first line therapy the assessment report notes that the estimates of cost-effectiveness are based on progression-free survival (PFS), rather than survival, as chemotherapy subsequent to the allocated first-line regimens means that survival cannot be uniquely related to the allocated therapy. The use of progression-free survival in place of survival has considerable implications on the results of the economic analysis. Oxaliplatin shows greater improvement in PFS compared to 5FU than irinotecan, based on the assessment group analysis of the PFS curves, although no survival benefit has been shown in clinical trials with oxaliplatin, whereas it has with irinotecan. Estimates for second-line treatment (where lower proportions of patients had further chemotherapy) on both PFS and survival show different results.</p> <p>The marginal cost per progression-free year (PFY) for oxaliplatin compared to the de Gramont 5FU regimen is £23,000. The same figure for irinotecan is £58,400. Second-line treatment with irinotecan (single-agent therapy) is cheaper than inpatient de Gramont. If it assumed that all treatments are given on an outpatient basis, the marginal cost per PFY is unchanged for oxaliplatin, £49,000 for irinotecan, and £26,400 for irinotecan second-line.</p> <p>For second-line treatment, the marginal cost per LYG (i.e. based on survival benefit) is zero when irinotecan is compared to inpatient treatment with de Gramont, £11,180 when compared to outpatient de Gramont, and between £17,700 and £28,200 when compared to BSC.</p> <p>As there is no benefit in either PFS or survival when treatment with raltitrexed is compared to 5FU, a cost-effectiveness analysis was not deemed appropriate.</p>	<p>Aventis: It appears that medians have been used in the manufacturer's economic model.</p> <p>Sanofi-Synthelabo Ltd: Two economic analyses are presented. The first, comparing first-line treatment with 5FU with and without oxaliplatin, is based on data from the de Gramont trial. The second compares the cost-effectiveness of oxaliplatin and irinotecan (both in combination with 5FU) with 5FU alone, using the Douillard publication as the source of data for irinotecan treatment. Both analyses use median differences in progression-free years as the benefit measure.</p>

	<p>The assessment group rely on summary statistics and plots in order to fit curves and extrapolate, allowing them to estimate mean survival differences. All other reviewed evaluations are reported to have used median data. The area between survival curves was estimated both for survival and for PFS curves using the method of trapezoids (limited to the extent of the published curves) and by fitting theoretical curves, to allow projection. The group note that the availability of survival curves limits this analysis to a subset of the trials, although this includes all the large multi-centre trials.</p> <p>The assessment group applied survival curves to estimate mean time spent pre progression and time spent with progressive disease (thus overall survival), but based their economic evaluation on the PFS estimates.</p>	
Evidence synthesis (pool survival estimates?)	The assessment group applied survival curves to data presented in 6 studies separately. Each study uses a slightly different regimen, and so the trials are not synthesised.	<p>Aventis: The model treatment and adverse event probabilities are described as being based on a meta-analysis of trial data for irinotecan and oxaliplatin. However, it is not clear what the source is of all the parameters used in the model.</p> <p>Sanofi-Synthelabo Ltd: One model uses data all from one trial. The other uses data from two separate trials.</p>
Survival model(s) fitted (Weibull, exponential etc)	<p>In order to undertake their analysis, the assessment group scanned the published graphs into digitising software which allows data points to be easily read off the curves. From these data, the areas under the curves at 3, 6 and 12 months were estimated using the trapezoidal rule. Three commonly used curves in survival analysis, the exponential, Weibull and Gompertz, were fitted to the data using a least squares minimisation procedure in Excel.</p> <p>It was found that the Weibull curve gave the most reasonable fit to all OS and PFS data (except for one small study) and so was used. Sensitivity analysis was not undertaken using the other curves.</p>	None.
Independent survival models, or hazard ratio (proportional hazards) modelling	Not clearly specified, but appears to be independent survival models for each treatment arm.	None.
Justification for survival model used?	The sum of square deviations, maximum deviation and comparison of actual and predicted areas at 3, 6 and 12 months were used to assess the appropriateness of the fitted curves.	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>The assessment group use a cost-effectiveness analysis, with cost per progression free year gained used for the first line analysis, and cost per life year gained used for the second line analysis. They include a cost-utility analysis only as sensitivity analysis due to a lack of confidence about available utility data.</p> <p>All published and reviewed papers which presented cost-effectiveness analysis also used median overall survival as the effectiveness measure, but little other details about the type of economic model are given – it is likely that they simply apply costs to the survival gain to estimate the cost-effectiveness ratio.</p>	<p>Aventis: Aventis submitted a model of treatment for advanced metastatic disease which follows patients through first- and second-line treatments to death. Alternative treatments included are first-line treatment with irinotecan or oxaliplatin in combination with 5FU, and 5FU alone by the de Gramont regimen. It is assumed all patients have second-line therapy with either 5FU or single-agent irinotecan. It is assumed that all patients receive chemotherapy for 6 weeks, after which they are assessed as to whether they are responding to treatment or not, or are stable. It is assumed that those not responding have no further first-line therapy, whilst the responders and stabilised patients continue on therapy for differing times. Subsequent to therapy, patients have a period of time in remission (not on therapy, but stable). At progression a proportion of patients (60%) go on to second-line therapy with either 5FU or single-agent irinotecan. On this, the assessment group state: It has previously been noted that it is difficult to assess the cost-effectiveness of the treatments on the basis of survival, as patients in all first-line trials have gone on to have further (uncontrolled) therapy. It is therefore difficult both to cost the treatment and to know to what extent survival differences shown (or not) are dependent on the first-line therapy. Furthermore in attempting to</p>

		<p>compare the cost-effectiveness of irinotecan with oxaliplatin there are no fully published trials that include both treatments.</p> <p>Sanofi-Synthelabo Ltd: Cost-effectiveness analysis using median PFS as the effectiveness measure – only first-line treatment is considered so the model is likely to have been quite simple.</p>
<p>Other issues noted (eg crossover)</p>	<p>The assessment group note two particular issues regarding the survival analysis in the economic evaluations they reviewed, as well as their own. The following text is taken from their report: Survival is a clearly unambiguous and highly relevant clinical measure. Median survival based on Kaplan-Meier curves is consistently reported across all clinical trials. However there are two difficulties in the use of median survival.</p> <p>The first is with regard to the median as the measure of survival. The median certainly has the benefit of simplicity, and avoids having to make any explicit assumptions about the survival distributions. However there is an implicit assumption about the relative shapes of the two curves, and this does not necessarily reflect the actual survival difference between treatments. The true difference is the area between the survival curves. However, this is not as easily measured. It can be simply estimated from survival curves using the trapezoidal rule, but as the curves are usually incomplete (censored) they need to be projected in order to be able to estimate the total area between the survival curves. Both methods (trapezoids and curve fitting to allow projection) were used to estimate survival benefits for studies where survival curves have been published. Note all the other studies that calculate a cost-effectiveness ratio use the median measure.</p> <p>The second issue with respect to many of the trials included in this review is the problem of crossover between treatments. Once patients had progressed on their allocated therapy, some received further therapy with a different agent.</p> <p>It is clear that, with only one exception, for all first line trials where data is provided, over 50% of patients went on to have further chemotherapy after progression on their initial allocated therapy, and in one study as many as 79% of patients in one treatment arm received further chemotherapy. It is equally clear that the survival benefit of the first line allocated therapy can not be estimated from the survival differences shown, as the effect of the second line therapy is unknown. Where differences in survival have been shown between treatment with a new agent and the control, one interpretation is that it reflects the benefit of earlier treatment with the more active agent.</p> <p>Survival is therefore a measure of sequential chemotherapy regimes, and the influence of the initial allocated therapy on overall survival difficult to ascertain. As the survival of patients in the different control arms can not be uniquely related to their allocated therapy, progression free survival will therefore be used as the primary measure of benefit, despite the recognised problems with the measure, discussed below.</p> <p>Thus the assessment group provide an answer to their issues – they fit curves and estimate mean PFS and OS, and also base their analysis on PFS rather than OS.</p> <p>The problem with PFS that is noted by the assessment group is that: Note, however, that the determination of patient progression is not a completely objective measure, and the estimated length of progression-free time may be affected by the frequency of check-ups.</p> <p>The authors also note an issue with using response rates as an outcome measure in economic evaluations: Perhaps because of the difficulties discussed above in the measurement of survival, there is debate as to what extent response is an indicator of survival benefit. Buyse <i>et al</i> carried out an analysis with the aim of identifying the relationship. Using patient-level data from several trials of different 5FU regimens, they found response highly and significantly predictive of survival, when comparing hazard ratios at several different times from one to twelve months. However, they also found that for individual trials only 38% of the variation in survival rates could be explained by the variation in response rates.</p> <p>Regarding the ideal economic evaluation in the circumstances, the assessment group note: Ideally the cost-effectiveness of giving sequential treatments (i.e. based on lifetime costs and benefits) to patients would be included in the analysis. However there is such scant data available on the second-line treatment given to patients after failure of first-line treatment in the clinical trials that an estimate could not be made.</p> <p>The authors compare mean and median survival estimates and show that for some trials the median is greater than the mean, for some it is similar, and for some it is less.</p> <p>Although the median is not an ideal estimate of the PFS benefit, there is also uncertainty in the estimation of the mean survival benefit based on the projection of the PFS curves.</p> <p>The authors note that extrapolation will mean uncertainty – they assess this to some extent with one-way sensitivity analysis.</p>	

11. TA37: Lymphoma (follicular non-Hodgkin's) - rituximab (replaced by TA137), March 2002

Guidance: Rituximab is recommended for the treatment of stage 3 or 4 follicular NHL when a patient is considered to be unsuitable for conventional chemotherapy treatment ('last-line treatment'). A record of each patient treated and how they have responded should be kept to help obtain more information on how well this drug works.

Rituximab is not recommended for patients with stage 3 or 4 follicular NHL for whom otherconventional chemotherapy treatments are still available.

Source: NICE recommends the selective use of rituximab for follicular non-Hodgkin's lymphoma, Press Release 2002/018, 28th March 2002, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32334>, accessed 29/03/10

Wake B, Bryan S, Barton P, Fry-Smith A, Davenport C, Song F and Hyde C. Rituximab as Third Line Treatment for Refractory or Recurrent Stage III or IV Follicular Non-Hodgkins Lymphoma, Department of Public Health and Epidemiology, University of Birmingham on behalf of the National Institute for Clinical Excellence, 9th January 2001. <http://www.nice.org.uk/nicemedia/pdf/Rituximab-AssessmentReport.pdf>, accessed 29/03/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The Review Group state that “We emphasise that the nature of the evidence precludes accurate quantitative estimates of net effect, although qualitatively we acknowledge that net benefit is likely to accrue, despite the considerable uncertainties. We believe the uncertainties make it impossible to assess reliably whether the net benefit associated with rituximab is the same, less or more than alternatives. The implications of this are greatest in attempting to decide whether rituximab should be used at the earliest stage allowed by the current licence (i.e. as a third line treatment option, and least when it is being used as a treatment of last resort.”</p> <p>The systematic review undertaken by the Review Group identified no RCTs or comparative studies. 4 prospective case-series were included, incorporating information on 387 patients. All were open to substantial bias and considerable caution was applied in interpreting the results. No information was available on overall survival, nor were there direct measurements of impact on quality of life. Rituximab did achieve clinical responses in some patients, but most of these were partial (generally defined as ≥50% decrease in size of lesions and no new lesions). The duration of responses appeared to be of a length that would be clinically useful.</p> <p>Two of the included case report studies presented data on median time to progression, but only for subsets of included patients (those who had partial or complete responses to treatment). One case report study gave data on median time to progression for the whole cohort included in the study, but the study did not include a comparator. An indirect comparison was not conducted as the reviewers believed that this would result in unreliable estimates due to the variability in the clinical data with respect to patient population, line of treatment, lack of blinding/randomisation/control in trials.</p>	
Evidence synthesis (pool survival estimates?)	None.	No details.
Survival model(s) fitted (Weibull, exponential etc)	None.	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	None.	No details.
Justification for survival model used?	None.	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	No economic model was built, sensitivity analysis based on the cost model developed by the manufacturer was undertaken. The Review Group stated “Reliable estimates of the relative cost-effectiveness and cost-utility of rituximab cannot currently be provided given the uncertainties surrounding the level of net benefits.”	<p>The economic analysis reported in the Roche submission considers the use of rituximab in patients with stage III-IV follicular lymphoma who are chemoresistant or in their second or subsequent relapse after chemotherapy. The comparators used in the incremental analysis are two alternative forms of chemotherapy, which represent “standard clinical practice in the NHS” (Roche, 2000, p40): fludarabine and CHOP. The central assumption is that there are equivalent clinical outcomes for the three interventions of interest (rituximab, CHOP and fludarabine). This is held to be the case both for the response rate to therapy and, for those patients who do respond, the duration of the response. On the basis of this assumption, a cost-minimisation analysis is undertaken where the focus is solely upon the costs associated with the alternative treatments. The perspective is that of the NHS and the main result is that, overall, rituximab is associated with a lower cost, because of its favourable side effects profile, and is therefore defined as the 'dominant' alternative.</p> <p>The manufacturer also presents an "illustrative analysis" where quality of life issues are explicitly</p>

		<p>considered. This represents an attempt to extend the earlier analysis using a cost-utility framework. The argument is made that all treatments considered in the analysis are associated with some level of toxicity and so the quality of life experienced during the treatment period is poorer than that experienced during remission. This is clearly an advantage for rituximab since the duration of the treatment period is shorter. Whilst the logic of the argument is sound, there are some weaknesses in the analysis reported.</p> <p>The utility data used in the analysis are taken from patients with early breast cancer. The relevance of such data to a patient group with NHL has to be questioned.</p> <p>The estimates of time in treatment and remission health states are given without any indication of the uncertainty in these point estimates. The results of the utility analysis are clearly sensitive to variation in these time intervals and we know from other sources that not all patients receiving CHOP or fludarabine undergo a full course of 6 cycles. –Note that the assessment report does not state what survival periods were assumed, and the manufacturer’s submission is not available on the NICE website.</p>
Other issues noted (eg crossover)	The ERG note that the evidence supporting the assumption of equivalent clinical outcomes is very weak. Very poor efficacy data existed and so an economic analysis including survival outcomes was not performed.	

12. TA45: Ovarian cancer (advanced) - pegylated liposomal doxorubicin hydrochloride (replaced by TA91), July 2002

Guidance: The National Institute for Clinical Excellence has recommended that Pegylated liposomal doxorubicin hydrochloride (PLDH also known as Caelyx) as one option for the treatment of women with advanced ovarian cancer.

In summary the guidance recommends that:

- PLDH should be considered as one of the treatment options for women with advanced ovarian cancer if first-line (initial) chemotherapy has not worked or if the cancer has stopped responding to the platinum-based chemotherapy.
- PLDH is **not** recommended for women whose cancer has resulted in very poor health. For example women who are only able to carry out limited self care and are mainly confined to bed or a chair
- PLDH is also not recommended for women whose bowel is blocked because of the cancer, or for women who have already been treated with PLDH and have not responded to the treatment.
- Only oncologists who specialise in the use of chemotherapy for the treatment of ovarian cancer should supervise of treatment using PLDH and the indications for treatment, clinical outcomes and adverse effects should be carefully recorded
- A woman's response to the treatment should be monitored carefully. If there is evidence that the cancer has started to grow or spread then treatment with PLDH should be stopped. A reduction in a woman's overall health and ability may also be a reason to stop treatment.

Source: NICE issues guidance on the use of PLDH (Caelyx) for ovarian cancer, Press Release 2002/040, 19th July 2002, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32411>, accessed 29/03/10

Forbes C, Wilby J, Richardson G, Mather L Sculpher M and Riemsma R. A rapid and systematic review of the clinical effectiveness and cost-effectiveness of pegylated liposomal doxorubicin hydrochloride (Caelyx[®] UK, Doxil[®] USA) for ovarian cancer, NHS Centre for Reviews and Dissemination, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 21st November 2001. <http://www.nice.org.uk/nicemedia/pdf/HTAovariancancerreport.pdf>, accessed 29/03/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	The review group states that “Although the assumption of equivalent overall survival duration was justifiable on the basis of the trial design and its results, the use of cost-minimisation analysis effectively assumes away uncertainty in survival duration. Hence the uncertainty in cost-effectiveness is not fully represented in a cost-minimisation analysis. Although this is usually not a problem if the economically preferred intervention is based on mean costs and mean effects, the cost-minimisation approach fails to describe all uncertainty to decision makers and limits the opportunity to estimate the potential cost-effectiveness of additional research. It was,	The review group found two economic papers (one was the company submission) that used a similar approach and data from the same clinical trial (Smith 2001). This trial was an RCT designed to show equivalence in overall survival between the two drugs. As the majority of the clinical outcomes showed no significant difference

	<p>therefore, decided to reintroduce a measure of survival duration into the analysis and undertake a full cost-effectiveness analysis, relating differential mean costs to differential mean survival duration. A fully stochastic analysis was developed. Uncertainty in mean costs was characterised as a log normal distribution based directly on the results reported in the Smith <i>et al</i> paper. Uncertainty in mean survival duration was also characterised as a log normal; other distributions, such as the Weibull distribution, could not be used as patient-level data would be required to generate the scale and shape parameters. Monte Carlo simulation was used to propagate uncertainty in these inputs, to generate a graphical representation of uncertainty in differential costs and life-years on a cost-effectiveness plane. Cost-effectiveness acceptability curves were then used to present the probability that pegylated liposomal doxorubicin hydrochloride was more cost-effective than topotecan for a range of maximum values the NHS might be willing to pay for an additional life-year in these patients.”</p> <p>“The survival data presented in the clinical details of [study] 30-49 did not provide a estimate of mean overall survival, instead median survival duration was reported. If a full cost-effectiveness analysis was to be undertaken, it was necessary to take the data presented on median overall survival duration, together with some explicit assumptions, to estimate mean overall survival durations, together with their variances, in the two arms of the trial. This involved the assumption that overall survival followed an exponential distribution which implied a fixed hazard rate.”</p> <p>Details of the method used for estimating a mean from a median using an exponential distribution were given in the appendix to the assessment report.</p> <p>Using an exponential distribution, it was estimated that mean survival was 1.91 in the PLDH arm, and 1.79 in the topotecan arm. Unlike the two economics papers identified, reported survival from trial 30-49 was based on the population of evaluable patients rather than on intention to treat.</p> <p>Only overall survival was modelled, not progression free survival and OS.</p>	<p>between pegylated liposomal doxorubicin hydrochloride and topotecan, a cost minimisation analysis was adopted. Both studies considered only the costs occurring within the treatment period and assume that similar levels of resources are utilised outside the time horizon of their analysis.</p>
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	Exponential.	None.
Independent survival models, or hazard ratio (proportional hazards) modelling	Independent.	None.
Justification for survival model used?	Lack of patient-level data, thus constant hazards had to be assumed.	None.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	EEACT.	Cost minimisation analysis.
Other issues noted (eg crossover)	<p>The ERG noted “The major limitation of the economic analyses in the review is the fact that effectiveness has not been quantified in terms of QALYs. Although 30-49 has demonstrated similar survival durations between the two arms, equivalence in health-related quality of life has not been established.”</p> <p>Another issue noted was that 481 participants were originally randomised in trial 30-49 but the analysis performed by the trialists which reports to be an ITT analysis is only based on 474 participants and is therefore not a true ITT analysis. These 7 patients were lost prior to the beginning of treatment, but reasons for this and their treatment allocation were not given.</p> <p>It is not clear why the evaluable population (n=416) was used for the review group’s mean survival estimate – the clinical section of the report and the appendices seem to suggest that median overall survival estimates were available for several patient groups – including the ITT group (n=474). The evaluable population were patients who met the study enrolment criteria and received at least two doses of the randomised treatment. The report seems to suggest that perhaps initially only survival data for the evaluable population was available, hence why it was used, it states “The data provided in the company submission presented overall median survival for the evaluable population”. The evaluable population lived longer and took longer to progress than the full ITT analysis patients.</p>	

13. TA50: Leukaemia (chronic myeloid) - imatinib (replaced by TA70), October 2002

Guidance: Imatinib should be a treatment option for adults with the Philadelphia chromosome type of CML who are in the chronic phase if either:

- interferon- α treatment is potentially causing harmful effects, or
- treatment with interferon- α is not controlling the leukaemia.

Imatinib should also be an option for adults with the Philadelphia-chromosome type of CML in the accelerated or blast-crisis phases provided that they haven't previously had treatment with imatinib.

Source: Imatinib (Glivec) is NICE's 50th Technology Appraisal Guidance, Press Release 2002/052, 2002, http://www.nice.org.uk/nicemedia/pdf/50_press_release.pdf, accessed 30/03/10

Garside R, Round A, Dalziel K, Stein K, Royle P. The Effectiveness and Cost-Effectiveness of Imatinib (STI 571) in Chronic Myeloid Leukaemia, Peninsula Technology Assessment Group, University of Exeter, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 21st February 2002. <http://www.nice.org.uk/nicemedia/pdf/imatinibassessmentreport.pdf>, accessed 30/03/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The Assessment Group believed that the rate of progression and survival for imatinib assumed by the manufacturer after year 1 may have been overestimated because it was based on IFN data from a trial in which the difference between IFN and HU was greater than in any other trial. Also patients were younger and at an earlier disease stage in the IFN trial than in the imatinib trial. Therefore the Assessment Group re-ran the model basing post 1-year survival and progression on an RCT with 7-year follow-up which was in an older and more severe population, and which showed no significant survival difference between IFN and HU. This increased the ICER from around £33k to around £268k.</p> <p>The Assessment Group therefore concluded that the model results were highly reliant on a model parameter (long-term survival and progression) for which there was no empirical data. They compared survival estimates using the manufacturer's model and their adjusted estimates to a meta analysis of IFN survival and found that neither fitted the meta analysed data well.</p> <p>The median length of response was used as a measure of disease free progression in the model. The Assessment Group tested the use of PH (Haematological Response) median response time compared to CR (Cytogenetic Response) median response time in sensitivity analysis</p>	<p>For imatinib, survival data beyond 12 months was not available, and pre 12 months is based on three case series. It was assumed that imatinib would assume the same survival curve as IFN beyond 12 months. However the assessment group state that the RCT with 10 year follow-up from which IFN survival data was taken was the trial which showed the biggest difference between IFN and HU. Patients in the IFN trial were younger and an earlier disease-stage than those in the imatinib trial.</p> <p>Also, it was assumed that the rate of disease progression after year 10 was the same as in year 10. The assessment group believed this to bias in favour of imatinib because the IFN progression and death rates were particularly low in year 10 of the relevant RCT.</p> <p>Compared to survival curves taken from the relevant trials, the rates assumed in the model overestimate survival.</p> <p>As a sensitivity analysis the manufacturer ran the model where progression after 1 year was based on fitting a Weibull curve to the first 12 months of imatinib data. This led to a £11,000 increase in the manufacturer's ICER (£33-35k to £44-46k).</p> <p>Note that subsequent to the ACD 24 month survival data for imatinib was submitted by the manufacturer.</p>
Evidence synthesis (pool survival estimates?)	Some use of external data.	No details.
Survival model(s) fitted (Weibull, exponential etc)	No details.	Weibull for sensitivity analysis. Other analyses based upon empirical rates.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.	No details.
Justification for survival model used?	Alternative scenarios based on different data sources.	Assumed imatinib survival equivalent to IFN in the long-term.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Reviewed the manufacturer's Markov model.	The manufacturer used a Markov model to compare the cost-utility of imatinib against HU for all phases of CML second line use. The Markov Model had 7 states – 'chronic', 'CR', 'CHR', 'PHR', 'Accel', 'Blast' and 'Death'. A monthly cycle was used and the model run until all the modelled cohort were dead. Effectiveness of imatinib is based on three unpublished case series studies, with short follow-up. Response rates and survival after one year are based on assumptions. Effectiveness of HU is based on data from an RCT with 10 year follow-up.
Other issues noted (eg crossover)	This Appraisal shows an example of attempting to use external data to base long-term survival estimates. However, in this case this was not done successfully. However, this is a difficult situation in this appraisal due to the absence of long-term data, and the use of survival for a different treatment as a proxy for survival of the treatment of	

	interest. The manufacturer believed that their cost-effectiveness estimates would be underestimated because imatinib would have favourable survival compared to IFN in reality. However the Assessment Group believed that the survival had been overestimated for imatinib (or rather, that the IFN survival data had been overestimated, and applied to imatinib). It should be noted though, that the Appraisal Committee reported in the FAD that they accepted the use of long-term survival from IFN studies as a proxy for imatinib, although precisely which IFN study survival was taken from was an important issue. They reported that the pooled estimate of survival from similar studies gave a survival estimate closer to the manufacturer's estimate than the assessment group's.
--	--

14. TA54: Breast cancer - vinorelbine (replaced by CG81), December 2002

Guidance: Vinorelbine by itself should not be used as the first treatment for advanced breast cancer.

Vinorelbine on its own should be one option for treating advanced breast cancer when initial treatment with chemotherapy using drugs called anthracyclines has not worked or is unsuitable for the patient. The patient and the doctor who is responsible for the cancer treatment should choose the follow-up treatment together, after they have discussed the benefits and possible side effects of the drugs available.

NICE cannot recommend routine use of vinorelbine together with other chemotherapy drugs (combination therapy).

Source: NICE issues guidance on vinorelbine for advanced breast cancer, Press Release 2002/066, 16 December 2002, <http://www.nice.org.uk/guidance/index.jsp?action=article&o=32531>, accessed 30/03/10

Lewis R, Bagnall A, King S, Woolacott N, Forbes C, Shirran L, Duffy S, Kleihnen J, ter Riet G, Riemsma R. A rapid and systematic review of the clinical effectiveness and cost-effectiveness of vinorelbine for breast cancer, NHS Centre for Reviews and Dissemination, University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, February 2002. http://www.nice.org.uk/nicemedia/pdf/Assessmentreport_Vinorelbine.pdf, accessed 30/03/10

Note: The manufacturer did not conduct an economic analysis – they just submitted a review of economic evaluations, all of which were also reviewed by the assessment group.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	Four economic evaluations of vinorelbine used as monotherapy were found to meet the inclusion criteria. No economic evaluation that investigated vinorelbine as combination therapy was found. One economic evaluation was only available in a conference abstract and three studies were available as published papers. The majority of the studies included a cost-utility analysis, and one study also included a cost-effectiveness analysis. Two economic evaluations investigated the use of vinorelbine in the treatment of anthracycline resistant MBC, one of which also included patients with MBC resistant to paclitaxel. Other chemotherapy agents evaluated by these two studies included the following drugs used as monotherapy: docetaxel, paclitaxel, 5-fluorouracil, and gemcitabine. A third study also looked at the cost-effectiveness of docetaxel, paclitaxel and vinorelbine as second line treatment in participants with MBC, but no details were given about previous therapy. One study reported evaluating the use of docetaxel in comparison with vinorelbine and paclitaxel as salvage therapy in patients with anthracycline resistant advanced breast cancer.	
Evidence synthesis (pool survival estimates?)	<p>The source of the effectiveness data was clearly stated for three studies that looked at the use of vinorelbine, docetaxel and paclitaxel. Leung <i>et al.</i> reported using three separate phase III RCTs for each drug. For the second economic evaluation, reported by Launois <i>et al.</i>, the data for docetaxel was based on the results of the drug registration file which included pooled results from three non-comparative phase II studies. For paclitaxel, interim results from one trial were used and vinorelbine data taken from a single published non-controlled study.</p> <p>There were no head to head trials of vinorelbine, paclitaxel, and docetaxel. Therefore, both economic evaluations have taken effectiveness data for individual drugs from separate studies and brought them together in a comparison. This is not ideal, as the study populations may not be comparable and therefore differ in terms of prognosis and responsiveness to treatment. The results should therefore be treated with caution. Despite the effectiveness data for both economic evaluations having been taken from cohorts of patients from separate trials/studies, the effectiveness data used by Leung <i>et al.</i> (derived from RCTs) will represent better and more conservative estimates than those taken from non comparative studies by Launois <i>et al.</i></p> <p>Similarly, the third economic evaluation based on the UK NHS, conducted by Brown <i>et al.</i> (looking at: vinorelbine, paclitaxel, and docetaxel) also reported using published phase III and phase II trials as the source of effectiveness data and used data for individual drugs from separate studies and brought them together in a comparison. Again the effectiveness data was not based on a head to head comparison, rather these data were derived from weighted average efficacy and adverse event rates for each drug. For the final economic evaluation, information relating to capecitabine was reported to have been taken from the registration trial and information relating to vinorelbine, 5-fluorouracil, and gemcitabine were derived from the literature and discussed by a panel of North American oncologists (a modified delphi approach). No information was reported on the type of literature used to derive this information and reference details were not provided.</p>	

Survival model(s) fitted (Weibull, exponential etc)	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.
Justification for survival model used?	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>Brown et al., 2000: Modelling based on response rate, time to progression, median survival, rate of grade IV febrile neutropenia with hospitalisation. Assessment Group state: The study also uses median survival rather than mean survival which would be more appropriate. Given that the clinical trial had been completed, mean survival data may have been available. If not median survival can be adjusted using statistical techniques to more accurately reflect likely mean survivals.</p> <p>Launois et al., 1996 Modelling based on Type of response (complete response, partial response, no change, disease progression) and the nature of toxicity/adverse reactions. Also median time to progression and median duration of response.</p> <p>Not mentioned by the Assessment Group, but again this analysis is based on median survival times.</p> <p>Leung et al, 1999 Modelling was based on disease progression. The measures of effectiveness required for the decision model were toxic death rates, treatment-limiting toxicity rates, and tumour response rates.</p> <p>It is not stated whether time to progression values were based on means or medians.</p> <p>Silberman et al., 1999 Modelling was based on response rates (method of valuation not stated) and adverse effects (method of valuation not stated). Only a short time frame was used and little information on this as only published as an abstract.</p>
Other issues noted (eg crossover)	For several of the comparisons in the economic model single arms of RCTs were used, or phase II data. Therefore comparisons likely to be unreliable.

15. TA55: Ovarian cancer - paclitaxel (review), January 2003

Guidance: It is recommended that paclitaxel in combination with a platinum-based compound or platinum-based therapy alone (cisplatin or carboplatin) are offered as alternatives for first-line chemotherapy (usually following surgery) in the treatment of ovarian cancer.

The choice of treatment for first-line chemotherapy for ovarian cancer should be made after discussion between the responsible clinician and the patient about the risks and benefits of the options available. In choosing between treatment with a platinum-based compound alone or paclitaxel in combination with a platinum-based compound, this discussion should cover the side-effect profiles of the alternative therapies, the stage of the woman's disease, the extent of surgical treatment of the tumour, and disease-related performance status.

When relapse occurs after an initial (or subsequent) course of first-line chemotherapy, additional courses of treatment with the chosen chemotherapy regimen (re-challenge therapy) should be considered if the initial (or previous) response has been adequate in extent and duration. Once the tumour fails to respond adequately to the chosen first-line regimen, different treatment options should be considered as part of second-line therapy (see 1.4).

Paclitaxel is not recommended as second-line (or subsequent) therapy in women with ovarian cancer who have received the drug as part of their first-line treatment. For women who have not received paclitaxel as part of first-line treatment, it should be considered as one option alongside other drugs licensed for second-line treatment of ovarian cancer.

Only oncologists specialising in ovarian cancer should supervise the provision of chemotherapy in ovarian cancer.

Source: Guidance on the use of paclitaxel in the treatment of ovarian cancer (TA55), 2003, <http://guidance.nice.org.uk/TA55/Guidance/Recommendation>, accessed 31/03/10

Bagnall A, Forbes C, Lewis R, Golder S, Riemsma R, Kleijnen J. An update of a rapid and systematic review of the effectiveness and cost-effectiveness of the taxanes used in the treatment of advanced ovarian cancer, NHS Centre for Reviews and Dissemination, University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, March 2002. <http://www.nice.org.uk/nicemedia/live/11486/32543/32543.pdf>, accessed 31/03/10

Note: The manufacturer submitted an economic model, but this was the same as in the original appraisal, just with updated costs. Few details are given on this evaluation, other than it is based on the GOG111 clinical trial. The manufacturer's submission is not available on the NICE website.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>In the original review, there were nine cost-effectiveness analyses and three cost-utility analyses of paclitaxel as first-line therapy for advanced ovarian cancer. Two UK evaluations used carboplatin rather than cisplatin.</p> <p>In the update two new economic evaluations were found of paclitaxel combined with cisplatin versus cyclophosphamide combined with cisplatin as a first-line treatment for women with advanced ovarian cancer. Both were abstracts from conference proceedings (ASCO 2000). Both were cost-effectiveness analyses. Both evaluations derived effectiveness data from the same RCT (OV10). One evaluation used a subset of participants included in the RCT (those from Canadian centres only) for effectiveness outcomes for the economic evaluation and the other used all trial participants for effectiveness outcomes. Clinical effectiveness of paclitaxel combined with cisplatin was estimated using progression free survival and overall survival derived directly from the RCT in one evaluation, and using overall survival estimated using both a restricted means analysis and a parametric Weibull model in the other.</p>	
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	Restricted means and Weibull model in one analysis, and direct RCT data in the other analysis.	
Independent survival models, or hazard ratio (proportional hazards) modelling	<p>In the Canadian evaluation survival estimates included in the model appeared to be based on medians: Median progression free survival 17 months versus 10.1 months; Median overall survival 36.8 months versus 25.6 months.</p> <p>In the Belgian evaluation survival estimates included in the model appeared to be based on hazard ratios: Overall survival significantly higher in paclitaxel arm (HR 0.73, 95% CI: 0.60, 0.89).</p> <p>The Belgian evaluation tested both a restricted means approach and an approach using a Weibull model . The cost-effectiveness results differed importantly for these results: ICER per LYG (restricted means analysis): Paclitaxel/ cisplatin vs cisplatin/ cyclophosphamide 0.31 years, 855,000 BF ICER per LYG (Weibull model): p/ c vs c/ c 0.79 years, 335,000 BF</p> <p>Both of these represent ICERs below £30,000 in UK terms, so there was not discussion about which method may have been best.</p>	
Justification for survival model used?	No details	
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Few details.	
Other issues noted (eg crossover)	<p>Both evaluations were reported only as abstracts from conference proceedings and so a lot of details were missing.</p> <p>There were updates of three of the clinical trials originally included in the review (including ICON3, the largest trial). Patients in two of the trials had significantly greater progression free survival and overall survival than controls, however the largest trial by far (ICON3) found no significant differences between groups. Therefore the assessment group states that "Economic evaluations based on the OV10 treatments (paclitaxel/ cisplatin versus cisplatin/ cyclophosphamide) found the paclitaxel combination to be cost-effective (matrix score 'A'). However if there is no survival benefit, as indicated by ICON3, these evaluations would not be based on valid data and in fact the confidence intervals for cost per QALY would include infinity, making paclitaxel less cost-effective than the control treatments."</p> <p>As stated in the review of the original appraisal (TA3): All of the 4 RCTs identified allowed treatment crossover to alternate treatment. However this is not considered in the economic section. It seems likely that the reviewed papers did not account for crossover.</p>	

16. TA62: Breast cancer - capecitabine (replaced by CG81), May 2003

Guidance: If a person has locally advanced or metastatic breast cancer capecitabine in combination with docetaxel should be used in preference to docetaxel on its own if:

- the person has already tried an anthracycline and this has failed, or
- treatment with an anthracycline is unsuitable.

Capecitabine monotherapy is recommended as an option for locally advanced or metastatic breast cancer if:

- a person has not already taken capecitabine in combination with docetaxel but has tried an anthracycline and a course of medicines that has included a taxane without success, or
- further treatment with an anthracycline is unsuitable.

Source: Capecitabine for locally advanced and metastatic breast cancer, Media Briefing, 28th May 2003, <http://www.nice.org.uk/nicemedia/live/11500/32642/32642.pdf>, accessed 31/03/10

Jones L, Westwood M, Wright K, Riemsma R, Hawkins N and Richardson G. A rapid and systematic review of the clinical effectiveness and cost-effectiveness of capecitabine (Xeloda) for locally advanced and/or metastatic breast cancer, NHS Centre for Reviews and Dissemination, University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 23rd September 2002 <http://www.nice.org.uk/nicemedia/live/11500/32640/32640.pdf>, accessed 31/03/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>An independent model was not built and other full economic evaluations were not found for either monotherapy or combination therapy. Therefore the assessment group critiqued the manufacturer's models:</p> <p>Monotherapy</p> <p>The estimates of effectiveness in the submission for vinorelbine and capecitabine treatment were based on estimates of the median time to progression and the median survival time from a number of clinical trials. All of these trials were non-comparative single arm trials. It is not possible to exclude the possibility of bias in the comparison of treatment arms from different studies. Differences, both observed and unobserved, in the patient characteristics and treatment between studies may lead to biased estimates of treatment effect. The assessment group stated that mean survival would have been preferable: "The use of median times in the calculation of total QALYs is not consistent with the 'QALY' paradigm, e.g. 1 year of life with a utility index of 0.2 is regarded as equivalent to 0.2 years of life with a utility index of 1. Calculations using mean times would be consistent in this respect."</p> <p>The Assessment Group ran sensitivity analysis replacing medians with means: "The total QALYs for each study were recalculated using means and standard errors for time to progression and survival time calculated assuming an exponential survival curve. This derivation of mean survival is theoretically valid assuming an exponential survival curve is appropriate." The estimated means were always higher than the medians. The impact this had on the ICERs was not reported because capecitabine appears to be dominant.</p> <p>Combination Therapy</p> <p>Comparative effectiveness was determined from a single 511 patient randomised controlled trial, SO149999, comparing docetaxel in combination with capecitabine with docetaxel monotherapy. The trial was open label and, therefore, we cannot discount bias arising from an awareness of the treatment that the patient received. There was an increased mean time to progression and mean survival time associated with the combination treatment. Thus mean survival was used as the survival benefit measure. However it is not stated how mean survival was estimated.</p>	
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	Exponential.	No details.
Independent survival models, or hazard ratio (proportional hazards) modelling	No details.	No details.
Justification for survival model used?	Estimated from medians without access to data.	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Appears to have been state transition models.	
Other issues noted (eg crossover)	The Assessment Group note that "Although survival may be considered an objective endpoint and should not be subject to observation bias, there may be systematic differences between trials and centres in the case-mix of enrolled subjects and their ancillary treatment, which could lead to differences in the observed survival times of subjects."	

In the trial used to inform the combination economic model (SO149999) patients were withdrawn from the trial upon documented disease progression. Thus it is likely that other treatments were received at this point. Although it is not mentioned, treatment crossover could therefore be an issue.

17. TA61: Colorectal cancer - capecitabine and tegafur uracil, May 2003

Guidance: Oral therapy with either capecitabine or tegafur with uracil (in combination with folinic acid) is recommended as an option for the first-line treatment of metastatic colorectal cancer.

The choice of regimen (intravenous fluorouracil/folinic acid [5-FU/FA] or one of the oral therapies) should be made jointly by the individual and the clinician(s) responsible for treatment. The decision should be made after an informed discussion between the clinician(s) and the patient; this discussion should take into account contraindications and the side-effect profile of the agents as well as the clinical condition and preferences of the individual.

The use of capecitabine or tegafur with uracil to treat metastatic colorectal cancer should be supervised by oncologists who specialise in colorectal cancer.

Source: Guidance on the use of capecitabine and tegafur with uracil for metastatic colorectal cancer, TA61, May 2003, <http://www.nice.org.uk/nicemedia/live/11498/32624/32624.pdf>, accessed 31/03/10

Ward S, Kaltenthaler E, Cowan J, Brewer N. A Review of the Evidence for the Clinical and Cost-effectiveness of Capecitabine and Tegafur with Uracil for the Treatment of Metastatic Colorectal Cancer, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 23rd September 2002
<http://www.nice.org.uk/nicemedia/live/11497/32622/32622.pdf>, accessed 31/03/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>Studies were identified through a systematic search of medical databases. Two economic evaluations by Murad <i>et al.</i>, were found, based on the same South American study comparing UFT with 5-FU. Two resource use studies were also identified: one relating to medical resource use in the two main capecitabine trials and one to resource use in the Carmichael UFT/LV trial. No published cost-effectiveness evaluations were found for capecitabine. In addition to the published studies an economic evaluation was included as part of the sponsor submissions from Roche and Bristol Myers Squibb.</p> <p>UFT Murad <i>et al.</i>, 1997 presented an economic evaluation of the treatment of patients with colorectal cancer in Brazil and Argentina. This study estimated the total cost of a course of treatment over 18 months with UFT/LV compared to a course of treatment of 5-FU. The treatment regimen of 5-FU was not given. Therapeutic equivalence was assumed and thus this was a cost minimisation analysis.</p> <p>Ollendorf, 1999 present a cost study of inpatient and outpatient services in an international phase III trial comparing UFT with LV to 5-FU with LV. No survival measures were included.</p> <p>Capecitabine Twelves <i>et al.</i>, 2001 analysed the resource use of 602 patients with advanced or metastatic colorectal cancer in an international trial comparing capecitabine treatment with Mayo regimen 5-FU/LV treatment. No survival measures were used.</p> <p>Assessment Group Model</p> <p>UFT An economic evaluation was undertaken to compare the cost-effectiveness of UFT/LV and capecitabine to intravenous 5-FU/LV. The Group state that “Mean survival was calculated from the survival curve published in the sponsor submission using area under</p>	<p>BMS BMS presented an economic evaluation comparing UFT/LV with intravenous 5-FU/LV treatment. This used a Markov model over a five-year time horizon to estimate costs of treatment with 5-FU/LV and UFT/LV. The model included first and second-line chemotherapy costs, costs of palliative care, treatment of adverse events, hospitalisations not due to adverse events and monitoring. The submission also included two economic evaluations of UFT/LV as a first-line treatment for advanced colorectal cancer, one based on each of the studies funded by BMS: Douillard <i>et al</i> and Carmichael <i>et al</i>. The effectiveness measure was toxicity endpoints. This is because the only trials comparing UFT/LV to a recognised regimen (Mayo) were designed to show non-inferiority, and so the only situation in which UFT/LV has a proven superiority to 5-FU/LV is in selected adverse events. Therefore none of the economic models included survival endpoints – implicitly assuming equivalent survival. The Assessment Group states that this ignores “significantly reduced time to progression (although only amounting to 0.3 months), and a statistically non-significant but possibly clinically important reduced overall survival (1.0 months in the Douillard study)”.</p> <p>Roche The submission included an economic evaluation of first line treatment with capecitabine for patients with advanced colorectal cancer. Roche used outcome and resource use data from the Van Cutsem and Hoff trials. The Assessment Group state that “It seems that the intention was to use survival, progression-free survival and quality adjusted survival as outcomes, however since the survival difference was negligible, a cost-minimisation analysis was performed instead. Outcome results were used from the trials mentioned above. Although capecitabine patients experienced a higher response rate, there was no statistical difference in time to progression or overall survival, so therapeutic equivalence was assumed.” The Assessment Group state that “The decision to perform a cost-minimisation analysis was reasonable, since there was no difference in survival outcomes.”</p>

	<p>the curve analysis. The area under a survival curve gives the mean overall survival experienced in the trial. Therefore, the area between the UFT/LV survival curve and the 5-FU/LV survival curve gives the mean survival benefit of UFT/LV over 5-FU/LV. Calculated in this way, the mean survival of UFT/LV was 15.3 months and the mean survival of 5-FU/LV was 15.7 months.”</p> <p>“In the Carmichael study (study 012)...Mean survival was calculated from the survival curve published in the sponsor submission using area under the curve analysis. Mean time to progression was 4.3 months for UFT/LV and 4.6 months for 5-FU/LV, and mean survival was 14.0 months for UFT/LV and 12.7 months for 5-FU/LV.” Due to concerns over the 5-FU regimen in the Carmichael study the Douillard data was used in the economic model.</p> <p>Capecitabine Two international RCTs with identical protocols compared capecitabine to the Mayo Clinic regimen. The data from these two trials were pooled in a report by Twelves (2002) and the Assessment Group used this data in their model. The mean survival for capecitabine, estimated using AUC analysis, was 15.7 months and the mean survival of Mayo was 15.1 months.</p> <p>Cost-minimisation analyses were performed for comparisons of capecitabine and UFT/LV with 5-FU/LV because the survival benefits associated with the different treatments were shown to be statistically equivalent.</p>	
Evidence synthesis (pool survival estimates?)	Some pooling within evidence used.	No details.
Survival model(s) fitted (Weibull, exponential etc)	None – restricted means.	None.
Independent survival models, or hazard ratio (proportional hazards) modelling	None.	None.
Justification for survival model used?	No details.	Assumed therapeutic equivalence.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Evaluations alongside trials. One cost minimisation analysis.	Markov models, but one manufacturer (Roche) conducted a cost minimisation analysis and the other (BMS) only differentiated effectiveness based upon AEs.
Other issues noted (eg crossover)	<p>It is not stated how the assessment group produced the survival curves with which they estimated mean survival – whether this was from complete KM curves, extrapolated curves, or some other method. Considering there were some differences in mean survival estimates it is surprising that a cost minimisation analysis was used. In sensitivity analysis the Assessment Group completed a cost-effectiveness analysis based on median progression free and overall survival gains associated with an alternative 5-FU regimen (ie assuming survival benefits for this alternative regimen compared to UFT and capecitabine, based on the survival benefits shown in a comparison of the two 5-FU regimens. Therefore this doesn't take into account any differences in survival shown between UFT and 5-FU and capecitabine and 5-FU in the clinical trials (eg equivalence had been assumed, but an alternative 5-FU regimen had been shown to have a 5.2 week OS advantage over the standard 5-FU regimen, so a cost-effectiveness analysis was run assuming a survival disbenefit of 5.2 weeks for UFT and capecitabine).</p> <p>The Assessment Group state that no information was given for either trial or the pooled data for capecitabine regarding cross over to other treatments nor information concerning the addition of other chemotherapeutic agents, as regards secondary chemotherapy.</p> <p>However, in the Douillard study, secondary chemotherapy was administered to 52% of patients in the UFT/LV group and 50% in the 5-FU/LV group, although data on type of drugs was not collected. In the Carmichael study 41% of patients in the UFT/LV group and 39% in the 5-FU/LV group received secondary chemotherapy including fluoropyrimidines, irinotecan and oxaliplatin.</p> <p>Hence treatment crossover may have influenced overall survival estimates.</p>	

18. TA65: Non-Hodgkin's lymphoma - rituximab, September 2003

Guidance: Rituximab is recommended for use in combination with a regimen of cyclophosphamide, doxorubicin, vincristine and prednisolone (CHOP) for the first-line treatment of people with CD20-positive diffuse large-B-cell lymphoma at clinical stage II, III or IV (see Section 2.3). Rituximab is not recommended for use when CHOP is contraindicated.

The clinical and cost-effectiveness of rituximab in patients with localised disease (Stage I, see Section 2.3) has not been established. It is recommended that rituximab be used in these circumstances only as part of ongoing or new clinical studies.

A specialist in the treatment of lymphomas should supervise the use of rituximab in combination with CHOP for the treatment of diffuse large-B-cell lymphoma.

Source: Rituximab for aggressive non-Hodgkin's lymphoma, TA65, September 2003, <http://www.nice.org.uk/nicemedia/live/11506/32679/32679.pdf>, accessed 01/04/10

Knight C, Hind D, Brewer N and Abbott V. Rituximab (MabThera) for aggressive non Hodgkin's lymphoma: systematic review, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 23rd May 2003 <http://www.nice.org.uk/nicemedia/live/11505/32676/32676.pdf>, accessed 01/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The proportion of patients that achieved a complete response upon receiving CHOP for DLBCL and the duration of overall survival of patients who have received a CHOP regimen has been derived from the Scottish and Newcastle lymphoma group (SNLG) database acquired by Roche and kindly provided to SchARR. The observed survival data from the SNGL database has been used to reflect the transitions between the health states over time. The relative effectiveness of R-CHOP compared to a CHOP only treatment regimen for patients with diffuse-large B-cell lymphoma (DLBCL) has been derived from the published literature based on the Group d'Etude des Lymphomes de l'Adulte (GELA) studies.</p> <p>“survival curves for the CR and NR populations who received the CHOP were derived for the model based on data from the SNGL database. The survival curves for CR and NR populations who received RCHOP were then created by applying the relative improvement in the proportion of complete responders and disease-free survival that R-CHOP provides compared to CHOP alone based on evidence reported on the GELA trial.</p> <p>The Kaplan-Meier survival curves derived from the SNLG were the Overall Survival of all patients receiving CHOP and Disease-Free survival of patients who were CR to CHOP therapy. Overall survival is normally calculated from the date of randomisation to the date of death, regardless of the cause of death and any patients that are censored are alive at the time of the analysis. Therefore, summing the area under an overall survival curve gives the total life years gained (LYG) for that particular disease, in our case DLBCL patients receiving CHOP.</p> <p>Disease-free survival is normally calculated for patients who are CRs from the date of randomisation to the date of the first event, where events are classed as relapses and death from the disease. However, unrelated deaths are not considered to be events and are usually censored at date of death. Therefore, summing the area under the disease free survival curve does not give the true LYG for patients who are CR. Further review of the disease-free survival curve from the SNLG suggests that this method of creating the disease-free curve was likely as there were only two types of cases occurring, relapses and non-relapses.</p> <p>The SchARR model divides the population that received CHOP chemotherapy into two populations or disease states, the complete responders (CR) and non-complete responders (NR), which includes those not responding to the initial CHOP therapy and those relapsing after being a CR.</p>	<p>The model and data used by the manufacturer is the same as that used by the assessment group. However the methods for estimating survival differ: “The relative risk difference from adding rituximab to a CHOP regimen has been applied to the increase in complete responders, reduction in disease-free survival and reduction in overall survival. In the reviewer's opinion including all three improvements to the inclusion of rituximab in the CHOP regimen is over-estimating the effect. Coiffier <i>et al</i> stated in reporting on the GELA study that the longer survival in the R CHOP group was due to a higher response rate during therapy and fewer relapses among patients who had a complete response. Applying an improvement in complete response rate and an increase in disease-free survival (reduction in relapse) by implication will bring about an improvement in overall survival. Therefore, by adding a further improvement to overall survival over-states the effect and due to the model methodology employed introduces an assumption that patients who fail to respond to R-CHOP therapy achieve an improvement in survival over patients who fail to respond to CHOP only therapy. However, any improvement in survival of patients who fail to respond to treatment is not reported by Coiffier.</p> <p>The survival curves derived from the SNGL data relate to disease free and overall survival. However, the method by which the disease-free survival curve is derived seems unsuitable to be used to measure life years gained. There is evidence to suggest that the disease-free survival curve has excluded deaths from any cause other than lymphoma and hence using this survival curve as a source of measuring life years gained for the total population seems inappropriate. The relative risk improvements of R-CHOP over CHOP have only been applied for the 2.8 years that patients on the GELA trial were followed for. However, some of the relative risks are reported after only 2 years and not from the end of the trial.”</p>

	<p>A survival curve for the CR and NR was derived using the following assumptions with regards to the SNLG data:</p> <ul style="list-style-type: none"> • The initial proportion of CR and hence NR was taken from the SNLG data. • A probability distribution was created to determine for every death at time t along the overall survival curve whether it came from the CR or NR populations. No published evidence could be found that compared relative risk of death between a CR and NR. It was assumed that there was a 90% chance that each death at time t came from the NR population, as the prognosis for patients who do not respond to initial CHOP chemotherapy is poor (sensitivity around this assumption is addressed later). When all the NR population had died all further deaths at time t from the overall survival curve then came from the CR population. • Every relapse from the CR population at time t on the disease-free survival curve from the SNGL data was added to the NR population. <p>It should be noted that these “survival curves” created for the CR and NR health states are not true Kaplan-Meier survival curves as the proportion of patients left alive in the NR health state can increase at a given time t if the number of relapses from the CR health state is greater than the deaths from the NR health state.</p> <p>Monte-Carlo simulation was employed to determine the sensitivity on the pseudo survival curves of assuming 9 in 10 deaths occur in the NR health state. Although each simulation run produced different survival curves for both the CR and NR populations, this method ensured that the total LYG from summing the areas under each of the CR and NR survival curves always equalled the total LYG from the original overall SNGL survival curve.”</p> <p>“The <i>GELA</i> study research report has shown that the addition of rituximab to the CHOP regimen has increased the complete response rate and prolongs disease-free and overall survival. The relative improvement in complete response rate between R-CHOP and CHOP was calculated from Coiffier <i>et al</i> (Table 3), where complete response was defined as complete responders and unconfirmed complete responders, and showed that there was a relative increase of 19.5% for the R-CHOP group compared to CHOP alone (p=0.009). The relative improvement in disease free survival for patients treated with R-CHOP has been derived from the <i>GELA</i> study research report, which states that R-CHOP reduced the risk of progression by 53% (risk ratio 0.47). The relative improvement in complete response rates and disease free survival have been applied to the CR survival curve for patients receiving the CHOP regimen to create complete responder survival curves for patients receiving R-CHOP.</p> <p>The improved complete response rate has been used to alter the proportion of the total population who after completion of R-CHOP treatment are in the CR disease state and those that are in the NR disease state. The NR survival curve is applicable to patients in the NR disease state following either CHOP or R-CHOP treatment as we assume that patients who fail to respond or relapse from the R-CHOP regimen have the same probability of survival as those who fail to respond or relapse from the CHOP regimen. The model calculates the mean duration of survival by adding together the disease-free survival among patients who achieved a complete response to the mean survival among patients who failed to be a complete responder or relapsed after being a complete responder. The mean survival for each of the disease states is calculated by summing the area under each curve.</p> <p>Total Mean Survival = Mean Survival Complete Responder * Percentage Complete Responder + Mean Survival Non-responder * Percentage Non-Responder</p> <p>In the model the relative survival benefits of R-CHOP are assumed to last for the firsts 3</p>	
--	--	--

	years only as the trial on which these assumptions are made had a follow up period of 3 years. For years 3 through to 15, the survival rate of patients in the CR health state following R-CHOP is assumed to be the same as the survival rate of patients in the CR health state following CHOP.	
Evidence synthesis (pool survival estimates?)	Use of external data.	Use of external data.
Survival model(s) fitted (Weibull, exponential etc)	Rates based upon empirical data rather than parametric models.	Rates based upon empirical data rather than parametric models.
Independent survival models, or hazard ratio (proportional hazards) modelling	Relative treatment effect applied to base survival curve.	Relative treatment effect applied to base survival curve (but slightly differently to the approach taken by the assessment group.
Justification for survival model used?	Attempt to make use of external information.	Attempt to make use of external information. Relative treatment effects only applied for period of the trial, as uncertain what would happen after that.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	The model evaluates the cost-effectiveness of introducing rituximab to the treatment regimen of cyclophosphamide, vincristine, prednisolone and doxorubicin (R-CHOP) compared to a cyclophosphamide, vincristine, prednisolone and doxorubicin (CHOP) only treatment regimen. The model is a Markov transition model with 3 health states that split into two age cohorts, those aged 60 and over and those aged less than 60 years old. The 3 states are complete responder (CR) to treatment, non-responder and relapse from complete responders (NR) to treatment, and death.	Same model as the assessment group.
Other issues noted (eg crossover)	Note model timeframe is 15 years, and treatment benefit assumed to last 3 years. Therefore survival curves aren't fully extrapolated until death and hence mean survival may be an underestimate. Regarding second line therapy it is assumed that 20% of patients who fail to respond to CHOP or R-CHOP treatment and are aged less than 60 years receive HDC/ABMT with a 25% success rate. This assumption has been based on a personal communication with Professor Barry Hancock at the Weston Park NHS Hospital Trust, Sheffield. Given observational data combined with trial data is used to estimate overall survival, it is likely that some confounding due to differential treatment pathways may be an issue, although this is not mentioned by the assessment group. This is an interesting use of external data combined with relative risks / hazard ratios for survival.	

19. TA70: Leukaemia (chronic myeloid) - imatinib, October 2003

Guidance: Imatinib is recommended as first-line treatment for people with Philadelphia-chromosome-positive chronic myeloid leukaemia (CML) in the chronic phase.

Imatinib is recommended as an option for the treatment of people with Philadelphia-chromosome-positive CML who initially present in the accelerated phase or with blast crisis. Additionally, imatinib is recommended as an option for people who present in the chronic phase and then progress to the accelerated phase or blast crisis if they have not received imatinib previously.

There is currently no evidence on clinical and cost-effectiveness on which to base guidance on the continued use of imatinib that has been initiated in the chronic phase of CML but has failed to stop disease progression to either the accelerated phase or blast crisis. Therefore, under these circumstances the use of imatinib is recommended only in the context of further clinical study. The data for this study should be collected systematically to allow aggregation and analysis at a national level in order to inform the appraisal review.

For people in chronic-phase CML who are currently receiving interferon alpha (IFN- α) as first-line treatment, the decision about whether to change to imatinib should be informed by the response of the disease to current treatment and by the tolerance of the person to IFN- α . This decision should be made after informed discussion between the person with CML and the clinician responsible for treatment, taking full account of the evidence on the risks and benefits of imatinib and the wishes of the person.

Source: Guidance on the use of imatinib for chronic myeloid leukaemia, TA70, October 2003, <http://www.nice.org.uk/nicemedia/live/11516/32754/32754.pdf>, accessed 01/04/10

Dalziel K, Round A, Stein K, Garside R, Price A. The Effectiveness and Cost-Effectiveness of Imatinib for First Line Treatment of Chronic Myeloid Leukaemia in Chronic Phase, Peninsula Technology Assessment Group, University of Exeter, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 28th March 2003 <http://www.nice.org.uk/nicemedia/live/11515/32751/32751.pdf>, accessed 01/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	One RCT comparing imatinib with IFN- α +Ara-C was identified. Four RCTs comparing IFN- α to HU were included along with five studies comparing BMT and IFN- α . The study comparing IFN- α +Ara-C to imatinib was of reasonable quality with the main potential biases being the lack of blinding (patient, physician, outcome measurement and data analysis), the potential for bias in the assessment of quality of life, and the high crossover and attrition rates.	<p>Novartis</p> <p>The purpose of this economic evaluation was to compare the cost-effectiveness of imatinib with IFN-α+Ara-C for the treatment of newly diagnosed patients with CML in whom BMT was not considered a therapeutic option. The model used a Markov structure containing the following health states: chronic phase, complete HR, partial HR and complete CR, accelerated phase and blast crisis death. In this model it is only possible to die from CML from the blast state. From other states, deaths from non-CML causes are permitted.</p> <p>The model crosses patients from imatinib to IFN α+Ara-C and vice versa when they progress or lose a response. Third line treatment for all patients is HU. The model runs for 30 years.</p> <p>The cost-effectiveness analysis is based on data from Novartis study 0106. Two different methods were used to estimate survival: CR method: Modelling the relationship between CR and survival based on the IFN-α literature and then applying this to the response rates seen in study 106. After 2 years survival is based on whether patients had a CR (during the first 2 years), independently of treatment (data taken from study by Bonifazi). PFS method: The imatinib group uses progression free survival (PFS) data from the Novartis study 010660 for the first 12 months, and then assumes the progression free survival of IFN-α+Ara-C for subsequent years (The Italian Cooperative Study Group, 1998).</p>
Evidence synthesis (pool survival estimates?)	None, but some use of external data.	None, but some use of external data.
Survival model(s) fitted (Weibull, exponential etc)	<p>Survival data were obtained from published studies of the effectiveness of various drug treatments for CML. We based the economic model on survival curves and progression curves. The following transition probabilities were modelled as being cycle dependent (i.e. the transition probability changes as the time spent by the cohort in the model increases).</p> <ul style="list-style-type: none"> • Chronic to accelerated/ blast • Chronic/ accelerated/ cytogenetic response to death • Chronic to cytogenetic response <p>In order to obtain the transition probabilities we electronically scanned the survival curves and used the program TechDig to obtain coordinates for a number of points along the curve. These coordinates were used to estimate a Weibull distribution of the following formula: $=EXP-\lambda*(time/year^\gamma)$</p> <p>$\lambda$ and γ were estimated using a least squared method to achieve best fit with data taken from survival and progression curves. Transition probabilities were calculated from the cumulative survival function given a cycle length of three months.</p> <p>The following transition probabilities were constant each cycle and were derived from the Literature:</p> <ul style="list-style-type: none"> • Accelerated to blast • Blast to death • Chronic to cytogenetic response <p>When calculating the transition probabilities for imatinib as second line treatment, we used data from the published chronic phase 2 trial for the first 5 cycles (1.25 years) after which</p>	Appears to be based upon empirical data rather than survival models.

	<p>we used the IFN-α data derived from the Italian trial as a conservative estimate.</p> <p>In order to estimate transition probabilities for HU as first line treatment we calculated a hazard ratio compared to IFN-α. The scanned survival or progression curves were compared in Stata, assuming an appropriate distribution (Weibull, gamma, exponential, or log normal) to estimate the hazard ratio and standard error. This was used as an estimate of the relative risk. Separate hazard ratios were calculated for mortality, progression and cytogenetic response. For imatinib, insufficient long-term data were available. A survival function was estimated from point data provided at 6, 9, 12, and 18 months from the Novartis study 0106, and then a similar procedure as with HU was undertaken. It was not possible to estimate the standard error using the point data, so for survival a large standard error was assumed in order that the confidence interval crossed 1, to reflect the lack of statistical significance demonstrated so far. Sensitivity analyses were used to explore the effect of uncertainty around all parameters.</p> <p>In sensitivity analysis using data from different studies for progression and survival with IFN-α was tested.</p>	
Independent survival models, or hazard ratio (proportional hazards) modelling	<p>Mixture of Weibull and exponential models and PH modelling.</p> <p>Because the model is quite complex with lots of health states (accelerated, blast, chronic, complete response, first line, second line, third line, death) a range of methods are used. For two survival estimates (transition from accelerated to blast; chronic to CR) a median survival estimate is transformed into a cycle transition rate.</p> <p>It is stated that “appropriate distributions” used to acquire hazard ratios were Weibull, gamma, exponential, or log normal.</p>	<p>The Assessment Group compare the survival results from their method to the methods used by Novartis. They show that the Novartis CR approach and the independent model are reasonably similar with the independent model giving a higher death rate in the first couple of years and after approximately 12 years. The Novartis PFS approach gives a lower death rate than the other two techniques.</p> <p>The Group state that “At the end of 20 years, a higher proportion of the Novartis cohort (7%) remain in CCR, compared to the independent model (2%). Somewhat perversely, it appears that prolonged survival is associated with a higher ICER, presumably because costs continue to accrue at a greater rate than benefits.</p> <p>The independent model results in an ICER for imatinib and IFN-α that is similar to the Novartis PFS approach (despite survival being more similar to the CR approach). One of the main differences between the three models is the modelling of survival and progression. The Novartis models assume IFN-α+Ara-C survival rates for imatinib after 12 months. The independent economic model applies a continuing relative risk of benefit (for the length of the chronic phase) for imatinib. There is no long-term empirical data to support or refute either technique.”</p>
Justification for survival model used?	<p>In order to validate the survival and progression data used in the model, model-derived data were plotted on the same graph as original data. Comparisons showed similar curves for progression and survival (Italian Cooperative Study Group, 1998), as well as time to loss of cytogenetic response (Bonifazi, 2001). A more detailed description and figures are shown in Appendix 10.9 (page 128). The models seemed to fit the real data well.</p>	<p>No details.</p>
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>A Markov model was developed to determine the incremental cost-effectiveness ratio of Imatinib compared to HU and IFN-α, and of HU compared to IFN-α in terms of cost per QALY, using a 20 year time period. The three treatment pathways that are compared in the model each consist of first line treatment, treatment when disease progresses to accelerated phase, treatment when disease progresses to blast phase and treatment for those who lose their cytogenetic response. It is assumed that patients who progress after starting IFN treatment switch onto imatinib, and vice versa. Patients who start on HU do not switch to IFN or imatinib at any stage.</p>	<p>Markov structure containing the following health states: chronic phase, complete HR, partial HR and complete CR, accelerated phase and blast crisis death.</p>
Other issues noted (eg crossover)	<p>The assessment group note that high rates of crossover is an issue with the RCT comparing imatinib with IFN+Ara-C - crossover due to intolerance was 0.7% for imatinib compared to 22.8% for IFN-α+Ara-C. Thus treatment crossover was an important issue. The Committee asked the assessment group to re-run economic analyses using</p>	

per protocol estimates rather than ITT estimates. This reduced the relevant ICER from £87k to £60k.
 In the relevant trials overall survival was often high at the end of follow-up (eg around 90% at 12 months), hence required extrapolation was substantial.

20. TA86: Gastro-intestinal stromal tumours (GIST) - imatinib, October 2004

Guidance: Imatinib treatment at 400 mg/day is recommended as first-line management of people with KIT (CD117)-positive unresectable and/or KIT (CD117)-positive metastatic gastro-intestinal stromal tumours (GISTs).

Continuation with imatinib therapy is recommended only if a response to initial treatment (as defined in Section 1.5) is achieved within 12 weeks.

Responders should be assessed at intervals of approximately 12 weeks thereafter. Continuation of treatment is recommended at 400 mg/day until the tumour ceases to respond, as defined in Section 1.5.

An increase in the dose of imatinib is not recommended for people receiving imatinib who develop progressive disease after initially responding (see Section 1.5).

For the purpose of this guidance, response to imatinib treatment should be assessed on the basis of the results of diagnostic imaging to assess size and density of the tumour(s), patients' symptoms and other factors, in accordance with the Southwest Oncology Group (SWOG) criteria detailed in Appendix D. For the purpose of this guidance, response to therapy is defined as the SWOG classifications of complete response, partial response or stable disease.

The use of imatinib should be supervised by cancer specialists with experience in the management of people with unresectable and/or metastatic GISTs.

Source: Imatinib for the treatment of unresectable and/or metastatic gastrointestinal stromal tumours, TA86, October 2004, <http://www.nice.org.uk/nicemedia/live/11548/32969/32969.pdf>, accessed 01/04/10

Wilson J, Connock M, Song F, Yao G, Fry-Smith A, Raftery J and Peake D. Imatinib for the treatment of patients with unresectable and/or metastatic gastro-intestinal stromal tumours – a systematic review and economic evaluation (Commercial-in-confidence [CIC] data removed), West Midlands Health Technology Assessment Collaboration, University of Birmingham, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 20th October 2003 <http://www.nice.org.uk/nicemedia/live/11548/32972/32972.pdf>, accessed 01/04/10

NICE, Final Appraisal Determination Imatinib for the treatment of unresectable and/or metastatic gastro-intestinal stromal tumours, TA86, August 2004, <http://www.nice.org.uk/nicemedia/live/11547/32967/32967.pdf>, accessed 01/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The Assessment Group concluded: “The original Novartis model has overestimated the cost-effectiveness of imatinib for patients with unresectable and/or metastatic GIST because of (1) disproportion of survival and TTF in the imatinib arm, and (2) the survival curve for patients in the control arm may have been biased against long-term survivors.”</p> <p>To correct for this the Group made two adaptations to the Novartis model:</p> <p>First, the number of patients in the state of imatinib treatment is estimated according to the same TTF curves, as in the original Novartis model. It is assumed that all patients in the state of imatinib treatment are alive. Patients who fail to respond to imatinib are moved to the state of progressive disease, and start to follow the same survival process as the new control patients. The number of surviving patients over time is calculated as the sum of patients in the state of imatinib treatment and surviving patients in the state of progressive disease. That is, in the modified model, the survival outcome in the imatinib arm is determined by both the TTF curves and the survival curve for progressive patients. An important</p>	<p>There were no controlled trials that directly compared imatinib with current treatment for unresectable and/or metastatic GIST. Thus, results from cohort trials or case series studies had to be used. The Novartis model used data from a single trial (CSTI571-B2222) to estimate survival curves for patients treated with imatinib. This open-label, multicentre trial compared two imatinib doses (400 mg or 600 mg/day) in 147 patients with malignant unresectable and/or metastatic GISTs. The advantage of using this trial is that it provides the most complete available survival data for imatinib treated patients, with a follow-up of up to 25 months. The survival rate was 88% after one year and 78% after two years.</p> <p>The median follow up of patients in the trial (CSTI571-B2222) was 25 months. The Novartis model used exponentially fitted curves to project the survival and the time to treatment failure for patients treated with imatinib (Figure 6, page 46) beyond the observed data. The exponential curves were fitted using data for the first 90 weeks for survival and data for the first 60 weeks for TTF because heavily censored data from longer follow-up was considered unreliable. According to the Novartis submission sensitivity analyses suggested no difference if all data available was used.</p> <p>“It is more problematic to obtain good survival data for control patients because of the following difficulties. Firstly, the molecular marker KIT was introduced in the diagnosis of GIST from 2000, but was not used in the previous studies. Other than by retrospective immunotesting this makes it generally impossible to separate KIT positive GIST from other gastrointestinal sarcomas in older studies. A second problem is that there is a lack of objective definition of</p>

	<p>advantage with the Modified-A model is that both the imatinib arm and the control arm will use the same survival curve for patients in the state of progressive disease.” This substantially reduces the survival advantage associated with imatinib (ICER increases from £14k to £22k)</p> <p>Second, there is some CiC discussion about the Goss study removed from the report. However, it seems that the Group thought that the use of survival curves for patients who never received imatinib in the Goss study underestimated the survival of control patients, possibly due to patients with better life expectancy at some point crossing over onto imatinib (this is confirmed by the FAD –see final box in this table). Therefore, they further modified the Novartis model (additional to the change in Modified-A) by using the survival curve for all patients with metastatic or recurrent GIST in the Goss study. This increased the ICER from £22k to £30k.</p> <p>The Group also built their own model. They modelled survival in quite a similar way to that in the amended Novartis model. The Birmingham model (as in the modified Novartis models) assumed that patients leaving imatinib treatment had the same state of progressive disease as patients in the control arm.. The model used data for the first 40 months to project long-term survival for this group of patients (exponential fit). For sensitivity analysis we used the first 80 months of data instead of 40 months.</p> <p>For TTF the Group used the same TTF data as in the Novartis model. The TTF curve was an exponential function fitted to the Kaplan-Meier data of Study CST571-B2222. Sensitivity analyses was undertaken to explore the effects of different TTF curves to trial data. They used Weibull and exponential fitted curves and also lower and upper values around the fitted exponential curve.</p>	<p>unresectability for the recurrent or metastatic GIST. The authors of the Novartis submission identified five published studies that reported survival outcomes of patients with advanced GIST. It was reported that the median survival for patients with advanced GIST is about 12 months, ranging from 2 to 20 months. An unpublished study by Goss <i>et al</i> employed histological confirmation of CD117 in the diagnosis of GIST, and may be considered the most relevant. In the Novartis model, the survival curve based on the unpublished Goss study was used in the baseline scenario, and survival curves from Clary <i>et al</i> was used for sensitivity analysis. The fitted exponential curves were well matched with the observed survival curves for the control patients.”</p> <p>The number of patients in each state is calculated every 4 weeks. The reported outcomes are up to 10 years, though the results after 2 years are of great uncertainty. In the control arm, the number of surviving patients (i.e., the number of patients in the state of progressive disease) over time is determined by the survival curve of historical patients who have not received imatinib treatment. For imatinib the model estimates the number of surviving patients according to the survival curve from a clinical trial.</p> <p>The Assessment Group notes that “An important weakness of the Novartis model is that the (time to failure) TTF and survival curves are independently calculated, and no efforts have been made to calibrate the outcomes of the two curves”. “the small proportion of patients in the state of imatinib treatment is disproportionate to the great proportion of surviving patients during the period of modelling. For example, the proportion of patients in the state of imatinib treatment and the overall survival are 44% and 79% respectively after 2 years; 13% and 55% respectively after 5 years; and 2% and 30% respectively after 10 years (baseline scenario). This is possible only if the progressive patients in the imatinib arm had a good survival prognosis, which is contrary to the assumption that the majority of patients in the state of progressive disease will die in two years”</p>
Evidence synthesis (pool survival estimates?)	None.	Some use of external data for control survival.
Survival model(s) fitted (Weibull, exponential etc)	Exponential models as used by the manufacturer, but also tested Weibull models.	Exponential, not including last data points.
Independent survival models, or hazard ratio (proportional hazards) modelling	Appear to be independent, but in the DSU analysis it is noted that constant HRs were not assumed, suggesting that this may have been the case in the analyses undertaken by the manufacturer / assessment group. There is little information on this.	Appear to be independent, but in the DSU analysis it is noted that constant HRs were not assumed, suggesting that this may have been the case in the analyses undertaken by the manufacturer / assessment group. There is little information on this.
Justification for survival model used?	Last data points not included because heavy censoring made them unreliable. Limited details on any other justifications.	Last data points not included because heavy censoring made them unreliable.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	State-transition model similar to that produced by the manufacturer.	The model is a state-transition model, and has two arms: the control and the imatinib treatment arm. The patients in the control arm have only two states in the model (progressive disease or death) based on the assumption that patients who do not receive imatinib have a gloomy prognosis. The patients in the state of progressive disease may remain in this state, or move to the state of death. In the imatinib arm, a state of imatinib treatment is added into the model. Patients in the state of imatinib treatment include those who have a stable disease or who achieve a partial response, because evidence suggested that the cost and survival consequences were the same with the stable disease or partial response.
Other issues noted (eg crossover)	Interesting example of where the survival modelling methods of the manufacturer were shown not to be justifiable or valid, as they came up with unreasonable survival estimates.	

	<p>Crossover was an issue in this appraisal, given the CiC details removed about the way control survival was estimated. The FAD shows that the manufacturer modelled only using patients who did not switch onto imatinib once it became available (67 of 132 patients did switch once imatinib became available). The Assessment Group then adjusted this by using all patients, and showed that overall survival estimates based on this were reasonable. Also, because once treatment had failed patients in the imatinib arm of the model follow the same survival curve as the control group this alteration increased survival in both arms (though this was likely to be more than proportional in the control group).</p> <p>The assessment group states: "Because of the immaturity of the data and trial design, evidence for survival has considerable uncertainties associated with it, which makes it difficult to answer the crucial question of how and if these clinical responses translate into patient benefit in terms of prolonged survival and quality of life. It is clear from comparing the survival curve for patients in an imatinib trial (Demetri 2002, n=147) with curves from a variety of sources describing survival of similar groups of patients not treated with imatinib that imatinib does indeed confer survival benefit. However, estimating the extent of this benefit is fraught with difficulties particularly with regard to considerable extrapolation beyond available data for imatinib-treated patients and to the selection of the most appropriate "control" survival curve for comparison." And "The questions that remain are (1) what is the most accurate estimation of survival in control groups; (2) what is an accurate long-term projection of survival and time-to-treatment failure beyond observed trial data; and (3) what potential biases can arise in the indirect comparison of survival of patients with and without imatinib."</p> <p>The FAD states: "At the instruction of the Appraisal Committee, the Assessment Group, in conjunction with the NICE Decision Support Unit (DSU), was commissioned to develop its own economic model. Additional data from the cohort study (survival estimates, censored at the time imatinib became available) were sourced to improve the estimates of survival with progressive disease. One of the key differences between the DSU model and the other economic models was that the DSU model was structured so that all patients started in the same health state of progressive disease. Also, all the relevant censored data from the cohort study were used to estimate survival following progressive disease – that is, not just patients who died before they could be prescribed imatinib, but also survivors up to the point at which they were transferred to imatinib treatment. Another key difference was that the extrapolations of both the trial data and the censored cohort study data were based on all the data available and did not assume a constant hazard ratio. The estimates of utility were the same as those included in the manufacturer's economic model. The model was also structured to estimate the cost-effectiveness of different policies regarding dose escalation." The results of the DSU model suggest that the incremental cost per additional quality-adjusted life year (QALY) is approximately £32,000 for patients on 400 mg/day estimated over 10 years – slightly higher than the amended manufacturer estimates. The Committee concluded that data on the survival of the control group, censored when imatinib became available, was the least prone to bias and the best estimate of prognosis of untreated GIST.</p> <p>Interesting attempt to control for crossover.</p>
--	---

21. TA91: Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review), May 2005

Guidance: Paclitaxel in combination with a platinum-based compound (carboplatin or cisplatin) is recommended as an option for the second-line (or subsequent) treatment of women with platinum-sensitive or partially platinum-sensitive advanced ovarian cancer, except in women who are allergic to platinum based compounds.

Single-agent paclitaxel is recommended as an option for the second-line (or subsequent) treatment of women with platinum-refractory or platinum-resistant advanced ovarian cancer, and for women who are allergic to platinum-based compounds.

PLDH is recommended as an option for the second-line (or subsequent) treatment of women with partially platinum-sensitive, platinum-resistant or platinum-refractory advanced ovarian cancer, and for women who are allergic to platinum-based compounds.

Topotecan is recommended as an option for second-line (or subsequent) treatment only for those women with platinum-refractory or platinum-resistant advanced ovarian cancer, or those who are allergic to platinum-based compounds, for whom PLDH and single-agent paclitaxel are considered inappropriate.

Within these recommendations, the choice of treatment for second-line (or subsequent) chemotherapy should be made after discussion between the responsible clinician and the patient about the risks and benefits of the options.

Source: Paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan for second-line or subsequent treatment of advanced ovarian cancer: Review of Technology Appraisal Guidance 28, 45 and 55, TA91, May 2005, <http://www.nice.org.uk/nicemedia/live/11554/33024/33024.pdf>, accessed 08/04/10

Main C, Ginnelly L, Griffin S, Norman G, Barbieri M, Mather L, Stark D, Palmer S, Riemsma R. Topotecan, pegylated liposomal doxorubicin hydrochloride and paclitaxel for second-line or subsequent treatment of advanced ovarian cancer (report contains no commercial in confidence data), Centre for Reviews and Dessimation, University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 20th September 2004 <http://www.nice.org.uk/nicemedia/live/11554/33028/33028.pdf>, accessed 08/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
<p>Survival data used (patient-level data, or summary statistics – mean, median etc)</p>	<p>Four economic evaluations as well as 3 manufacturer submissions were reviewed. The assessment group also conducted their own economic evaluation.</p> <p>Of the 4 published evaluations, 3 were cost minimisation analyses (which the Group did not deem appropriate based on the clinical data not showing equivalence) and 1 was a cost consequence analysis which did not combined costs and effects. This study used clinical data from a non-randomised study, and used median survival as the clinical outcome measure.</p> <p>AG model The AG used a 3-state Markov Model. Mean estimates for PFS and OS were calculated, which then led to estimates of the mean time spent in the progressed health state being calculated.</p> <p>The Group states: “As previously noted, PFS and overall survival were reported in several ways, including median number of weeks, Kaplan-Meier survival curves and hazard ratios. For the purposes of the model it was necessary to identify the most appropriate summary measure that would enable comparisons to be made between the alternative treatments. The hazard ratio represents the most accurate of these measures for comparing survival between treatments, as it is specifically designed to allow for censoring and time to an event. Furthermore, the use of the (log) hazard ratio and its variance allows studies to be pooled using conventional meta-analytic approaches. However, since some of the studies did not report the hazard ratio, alternative methods were explored which would allow the ratio (and variance) to be estimated from other data reported in these studies. For those studies that did not report the hazard ratio, it is possible to estimate the statistic from either the median number of weeks (PFS and overall survival) or the published Kaplan-Meier curves, however both methods are subject to limitations. In order to estimate the hazard ratio from the median number of weeks’ survival, it is necessary to make an assumption about the distribution of survival times, for example that they follow an exponential distribution with a constant hazard. If the actual survival curves do not follow an exponential distribution, then it is possible that the use of the approach will not reflect the actual survival difference between treatments.”</p> <p>“In order to estimate the hazard ratio from survival curves, one must measure the area under the survival curves, which must be extrapolated beyond the published curves to eliminate right- censoring. This method depends on the quality of the published curves, i.e. the accuracy with which they can be measured, and the validity of the assumed distribution used to extrapolate those curves. Due to the limitations of these approaches and potential bias that the use of alternative approaches might introduce in estimating the relative treatment effects, the base case model (analysis 1) only included those studies which reported hazard ratios for survival data.”</p> <p>Thus a PH modelling technique was used, due to using clinical data from a number of trials. 5 RCTs were eligible for inclusion in a mixed treatment comparison. However a baseline survival needed to be estimated to apply the HRs to, and this was done by estimating mean survival from a median, assuming an exponential distribution: “In order to build the model, it was necessary to select one of the three treatments to</p>	<p>GSK submission This submission used updated clinical data for topotecan and PLDH from the 30-49 trial. A cost minimisation analysis was performed, assuming equivalence in efficacy. The Assessment Group stated that “the use of a cost minimisation framework is not ideal. Although significant differences between PLDH and topotecan were not apparent there were differences for platinum sensitive patients and drug toxicity. The submission highlighted the impact that different adverse events could have on patients on quality of life. By using a cost-minimisation framework the impact on QoL has not been considered. Furthermore, in a 3-year follow-up of patients included in trial 30-49, it was reported that survival was higher for PLDH compared with topotecan (hazard ratio: 1.216; 95% CI: 1.000-1.478; p=0.05) and significantly prolonged in platinum-sensitive patients (hazard ratio: 1.432; 95% CI: 1.066-1.923; p=0.0017)”.</p> <p>Schering Plough This submission used short-term clinical data for topotecan and PLDH from the 30-49 trial. In addition, results of long-term survival analysis were presented (which were not available during the previous NICE appraisal), showing a trend favouring PLDH both in terms of overall survival and for a sub-group of platinum-sensitive patients. On the basis of these results it was assumed that PLDH was at least as effective as topotecan, and a cost-minimisation analysis was performed. Although not used in the economic model, in discussion survival estimates quoted were medians. Again, the Assessment Group highlighted the disadvantages associated with cost minimisation analysis.</p> <p>BMS In this submission four alternative 2nd line chemotherapies for patients with advanced ovarian cancer were compared: paclitaxel monotherapy, topotecan, PLDH and paclitaxel in combination with a platinum agent. The rationale for this analysis was that while paclitaxel as monotherapy has shown similar results in terms of response rate, PFS and overall survival, compared to topotecan and PLDH, paclitaxel in combination with platinum had demonstrated survival and tolerability benefits over platinum monotherapy. Given the lack of head to head comparisons for the 4 therapies under study, a model was constructed to estimate costs and effects associated with these alternatives. Life years gained over 3 years were calculated for the four chemotherapies based on clinical trial survival curves obtained from the literature.</p> <p>Effectiveness evidence for the comparison between PLDH and paclitaxel as monotherapy were taken from trial 30-49 and data for the comparison between paclitaxel monotherapy and topotecan were obtained from trial 039. In addition efficacy and tolerability of paclitaxel in combination with platinum has been recently shown in 2 parallel, international trials (ICON4). Kaplan-Meier curves were used to estimate PFS and overall survival and hazard ratios (HR) were presented.</p> <p>No direct comparisons of paclitaxel monotherapy, paclitaxel in combination with platinum therapy, topotecan and PLDH, are available from the trial data. In order to compare effectiveness evidence from the 3 trials (30-49, 039 and ICON4), a model was developed. Life-years gained with the 4 chemotherapeutic agents were estimated on the basis of the survival curves found in the trials. The ICON4 trial was used to provide survival data for paclitaxel combination, 039 was used for topotecan</p>

	<p>provide a baseline PFS and overall survival against which the hazard ratios of the other two treatments could be compared. Trial 30-57 was ruled out as a provider of baseline information, as it was terminated early, and the length of follow-up was not available and was likely to have been 1 year or less. The data concerning median weeks' survival from trial 039 was available for a longer period of follow-up, of around 4 years. However it was also limited, that is it would not allow specification of subgroup-specific baselines. Trial 30-49 provided 4 years follow-up, by which time 87% of patients had died, and allowed specification of subgroup-specific baselines, and baselines stratified by other covariates. Consequently trial 30-49 was chosen as the source of baseline data, and topotecan acted as the common comparator between the two completed trials (30-49 and 039).</p> <p>Since none of the trials provided estimates of the absolute hazard of progression or death, it was necessary to estimate the baseline hazard using median weeks and an exponential approximation. This approach has been used in a previous technology assessment report looking at treatment for advanced breast cancer. The baseline hazard (λ) and its variance are calculated according to the following formulae:</p> $\lambda = -\text{LN}(0.5)/t$ $\text{Var}(\lambda) = \lambda^2/r$ <p>Where t = median weeks survival; r = number of events. Using this approach, the baseline hazard (λ) can then be converted into a mean survival time, for progression-free survival and overall survival, by simply taking the inverse of the hazard ($1/\lambda$). This represents the mean survival times for topotecan. The mean PFS and overall survival for the 2 treatment comparators are then estimated by applying the hazard ratio (relative to topotecan) to the baseline hazard, in order to estimate the absolute hazard for each of the comparators. Mean survival times for the comparators are then estimated using a similar approach to that described for topotecan (i.e by taking the inverse of the absolute hazards for each of the comparators)."</p> <p>This technique seems at odds with the Group's earlier rationale for not estimating HRs in studies where they are not reported (an exponential distribution would have had to have been assumed). However, possibly this is less important in this case as this calculation is for baseline survival, rather than the treatment effect.</p> <p>The hazard ratios for PFS and overall survival of paclitaxel and PLDH compared to topotecan are then multiplied by the respective absolute hazard for the baseline to calculate the absolute hazard of PFS and overall survival for paclitaxel and PLDH. The absolute hazards (λ) for each treatment are then converted into mean PFS and overall survival by taking the inverse of the hazard ($1/\lambda$).</p>	<p>and paclitaxel monotherapy and trial 30-49 was used for PLDH. Survival curves were available from the ICON 4 and 039 trial reports. For PLDH, survival curves observed in trial 30-49 were obtained from a previous Schering-Plough submission to NICE.</p> <p>Life-years gained for the 4 agents were calculated as follows: (Proportion of patients alive yearn + Proportion of patients alive yearn+1)/(2* 100) That is, the proportion alive between two periods was averaged and then divided by 100 to obtain per patient values. Given the different follow-up period for the four agents (4.5 year for paclitaxel/platinum and for topotecan, 4 years for paclitaxel monotherapy, 3 years for PLDH), survival curves were truncated at 3 years to allow direct comparability. This analysis showed that paclitaxel/platinum was associated with the highest number of YG (1.83) followed by PLDH (1.45), topotecan (1.37) and paclitaxel monotherapy (1.295) in a 3-year time-horizon. Therefore only four timepoints were used (0, year 1, year 2, year 3), and it was assumed that patients died in a linear fashion between each time point.</p> <p>The Assessment Group stated that there were numerous drawbacks with the BMS analysis: "The main limitation of the study appears to be associated with the different characteristics of the patient population in the trials considered for the effectiveness analysis. The ICON4 study, from which survival rates were obtained for paclitaxel/platinum, included only platinum-sensitive patients who relapsed more than 6 months after the completion of a previous platinum chemotherapy. Instead, the patient population of the clinical trials from which survival rates were obtained for topotecan (trial 039), paclitaxel monotherapy (trial 039) and PLDH (mainly from trial 30-49) included both platinum-sensitive and platinum-refractory patients. It is well known that survival rates associated with platinum-sensitive patients are (on average) much higher than those for platinum-resistant and platinum-refractory patients (as illustrated, for example, in Table 54). It therefore seems inappropriate to compare the survival rates found in the ICON4 study with those of the 2 other clinical trials. It would have been more appropriate to compare ICON4 with the sub-group of platinum-sensitive patients in the 30-49 and 039 trials. Using the survival rates for PLDH found in trial 30-49 (table 54) for platinum sensitive patients only, we estimated a total of 1.89 YG over 3 years for PLDH. This value is higher than the figure found for paclitaxel/platinum in this submission (1.83). Also, considering only platinum-sensitive patients in trial 039 trial, YG associated with paclitaxel alone would rise to 1.74 and to 1.56 for topotecan. It should however be noted that, while in trial 30-49 censored patients were excluded from the analysis and survival rates were calculated considering only patients alive and dead for every year, survival rates used in this submission for paclitaxel/platinum were obtained considering all patients at risk for each year (thus, including censored patient). This approach tends to underestimate the actual survival rate for paclitaxel/platinum because censored patients are treated as dead. In addition, YG were calculated by taking the average number of patients alive between two periods (and then divided by 100 to obtain per patient values). This approach assumes that all deaths occurred exactly at the middle of the two periods considered (at month 6 of each year). This is clearly not the case in practice and the results obtained should be considered as an approximation of the real per patient YG. It is not clear whether the approach used in this submission was due to lack of patient-level data or it was a methodological choice." This type of analysis seems to be an attempt to estimate mean survival (it is a type of area under the curve estimate), but is likely to be inaccurate.</p>
Evidence synthesis (pool survival estimates?)	Mixed treatment comparison.	Some use of data across studies.

Survival model(s) fitted (Weibull, exponential etc)	The Group note that "in order to estimate mean survival times from the estimated hazards it was necessary to assume that the survival data was approximately exponential (i.e. a constant hazard) in form. In the absence of patient-level data it is difficult to establish the validity of this assumption or to determine whether alternative distributional forms would be more appropriate (e.g. Weibull). However, since overall survival times in this patient population were relatively short, it is unlikely that alternative assumptions would significantly impact on the results presented here."	All manufacturers except BMS undertook cost minimisation analysis. BMS estimated survival seemingly under exponential assumptions, but this is not explicitly stated.
Independent survival models, or hazard ratio (proportional hazards) modelling	PH modelling.	It is noted that HRs were presented, but it appears that BMS modelled agents independently.
Justification for survival model used?	Acknowledge exponential may not be ideal, but suggest that because survival times are short an exponential is likely to be sufficient.	Few details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	3-state Markov Model (PFS, Progressed, Death)	Cost minimisation analysis, except BMS who undertook an analysis based upon life years gained.
Other issues noted (eg crossover)	Regarding crossover, the Group notes that "Finally, the analysis presented here does not directly consider the impact of treatments provided as part of third-line and subsequent therapies. It is possible that the differences observed in the various trials may be partly confounded by the different therapies received after 2nd line drugs. For example, patients receiving PLDH as 2nd line therapy might have received topotecan as 3rd line therapy, while the same pathway may not be possible for patients receiving topotecan as 2nd line drug. In other words, differences in the long-term results could also depend on treatments received after the 2nd line therapies." Treatment crossover following treatment failure was reported to have occurred in an RCT comparing the effectiveness of single agent paclitaxel to a combination of cyclophosphamide, doxorubicin and cisplatin, and also in trial 039, comparing paclitaxel and topotecan. Survival data from both these trials were used in the economic model. Thus although the Group acknowledged treatment crossover as a potential problem, they did not attempt to adjust for this.	

22. TA93: Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (review), August 2005

Guidance: Irinotecan and oxaliplatin, within their licensed indications, are recommended as treatment options for people with advanced colorectal cancer as follows:

- irinotecan in combination with 5-fluorouracil and folinic acid as first-line therapy, or irinotecan alone in subsequent therapy
- oxaliplatin in combination with 5-fluorouracil and folinic acid as first-line or subsequent therapy.

Raltitrexed is not recommended for the treatment of patients with advanced colorectal cancer. Its use for this patient group should be confined to appropriately designed clinical studies.

Source: Irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: Review of Technology Appraisal 33, August 2005, <http://www.nice.org.uk/nicemedia/live/11562/33132/33132.pdf>, accessed 09/04/10

Hind D, Tappenden P, Tumur I, Eggington S, Sutcliffe P and Ryan A. The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation (review of Guidance No. 33), School of Health and Related Research, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 10th January 2005 <http://www.nice.org.uk/nicemedia/live/11561/33129/33129.pdf>, accessed 09/04/10

Hind D, Tappenden P, Tumur I, Eggington S, Sutcliffe P and Ryan A. The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation (review of Guidance No. 33), Addendum: Economic evaluation of irinotecan and oxaliplatin for the treatment of advanced colorectal cancer, School of Health and Related Research, University of Sheffield, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, 25th January 2005 <http://www.nice.org.uk/nicemedia/live/11561/33130/33130.pdf>, accessed 09/04/10

NICE, Final Appraisal Determination Imatinib Irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer (review of Technology Appraisal 33), TA93, July 2005, <http://www.nice.org.uk/nicemedia/live/11561/33131/33131.pdf>, accessed 09/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary)	The AG state that their economic evaluation is an improvement compared to others because of the data they use: "The analysis improves upon previous economic evaluations of irinotecan- and oxaliplatin containing chemotherapy regimens as it synthesises published and unpublished evidence on overall survival and resource use	Sanofi First line model and a sequencing model. A range of trials were used: de Gramont <i>et al</i> , Giacchetti, Douillard, Saltz, Köhne, Goldberg, Tournigand, FOCUS.

<p>statistics – mean, median etc)</p>	<p>relating to sequences of chemotherapies from the current FOCUS trial and the GERCOR trial reported by Tournigand and colleagues.”</p> <p>“The analysis includes economic comparisons of irinotecan and oxaliplatin using three clinical benefit measures: overall survival, first-line progression-free survival, and second-line progression-free survival. At the time of writing, progression-free survival curves for secondline chemotherapy regimens within the FOCUS trial were not available, therefore only the cost-effectiveness of second-line FOLFOX6 versus FOLFIRI is evaluated”</p> <p>“Kaplan-Meier curves giving empirical estimates of overall survival and progression-free survival in each treatment arm were obtained from the trial reported by Tournigand and colleagues and from unpublished data made available to the TAR group (Personal communication, G.Griffiths, MRC Clinical Trials Unit, London).</p> <p>The sequence of chemotherapies recommended within the 2002 NICE Guidance1 (FOCUS Treatment Plan A: first-line 5-FU/FA followed on progression by second-line irinotecan) was taken as the baseline for the Weibull regression analysis of overall survival and first-line progression-free survival. Due to the absence of evidence on the effectiveness of second-line therapies from the FOCUS trial⁵ the FOLFOX6/FOLFIRI sequence evaluated within the trial reported by Tournigand and colleagues was taken as the baseline for the regression analysis of second-line progression-free survival.”</p> <p>The AG noted that “It should also be noted that the median survival and progression-free survival benefit is lower than mean benefit estimated as the area under the curve; this highlights the importance of using the mean benefit rather than the median as a measure of overall survival.”</p>	<p>“Estimates of mean and median time on treatment (measured in months) for each study were estimated at 3 and 6 months, using Kaplan-Meier survival curves (extrapolated using Weibull curve to account for censoring). The analysis of sequences draws on two studies which evaluated planned sequences of chemotherapies. However, there are important problems with the use of these data; firstly, the preliminary results from FOCUS reported only grouped data for first- and second-line irinotecan/oxaliplatin. Thus, the analysis presented within the submission makes the explicit assumption that oxaliplatin in combination with 5-FU/FA and irinotecan in combination with 5-FU/FA are equivalent in terms of survival and progression free survival, and that the two drugs have identical adverse event profiles. The key difficulty in using the Tournigand trial is that the time on first- and second-line therapies is unknown and must therefore be estimated. These difficulties lead to important weaknesses in the analysis of sequences of chemotherapies.”</p> <p>Aventis</p> <p>A range of trials were used: de Gramont <i>et al</i>, Goldberg <i>et al</i>, Douillard <i>et al</i>, Saltz <i>et al</i>, Cunningham <i>et al</i>, Rougier <i>et al</i>, Andre <i>et al</i>, Rothenberg <i>et al</i>.</p> <p>A first-line evaluation was submitted, as well as an evaluation of three alternative sequences. “The authors used a state transition approach to simulate chemotherapy sequences using data from both first and second-line clinical trials in an attempt to remove the confounding in trial data that resulted from treatment crossovers and mixed salvage treatments”</p> <p>“Estimates of the effectiveness of first- and second-line chemotherapy regimens were derived from progression-free survival curves and overall survival curves reported within clinical trials. Survival and progression-free survival curves were extrapolated using survival analysis, whereby Weibull curves were fitted to empirical survival and progression-free survival data using a least squares approach to estimate the final portion of each curve.”</p> <p>“The 3-month probability of dying whilst on first-line therapy was calculated using extrapolated overall survival curves reported within first-line clinical trials of 5-FU/FA and oxaliplatin. The probability of dying was calculated as 1 minus the proportion of patients surviving at time t+1 divided by the proportion of patients surviving at time t. However, the use of survival curves for individual chemotherapies to estimate the probability of death during each Markov cycle results means that the results remain confounded; it is unknown how much of the observed survival benefit was actually attributable to the allocated treatment.”</p> <p>Due to this the Assessment Group stated:</p> <p>“In summary, whilst the approach adopted within the submission to NICE from Sanofi-Synthelabo119 attempted to prevent the confounding arising from patients crossing over to other chemotherapy regimens following disease progression, the use of a Markov approach does not overcome this problem. The cost-effectiveness results relating to sequences of chemotherapies remain confounded and should be considered unreliable.”</p> <p>AZ (did not submit an economic evaluation)</p>
<p>Evidence synthesis (pool survival estimates?)</p>	<p>None.</p>	<p>A range of trials were used in the economic model.</p>
<p>Survival model(s) fitted (Weibull, exponential etc)</p>	<p>All survival curves and progression-free survival curves were digitally scanned using TECHDIG™ software which is designed to replicate published survival curves. All scanned Kaplan-Meier curves were subsequently imported into Microsoft EXCEL™. As some patients were still alive at the end of the trial duration (i.e. right-censored), the final portion of each survival curve was extrapolated using regression analysis to estimate the parameters of a Weibull survival curve.</p> <p>Means were used, using area under the curve. Also a type of ‘restricted mean’ analysis was tested:</p> <p>“As discussed within the Technology Assessment Report, the best measure of survival is</p>	<p>Both manufacturers used Weibull models.</p>

	<p>the mean rather than the median. Mean overall survival and progression-free survival benefits were calculated for each of the seven treatment arms using the formula: Mean survival = $(1/\alpha \cdot \lambda)(1/\gamma) \times \Gamma[1+(1/\gamma)]$ where Γ is the mathematical gamma function. Additional analyses were undertaken using only the empirical Kaplan-Meier curves (thus ignoring the missing final portion of the curve), and mean overall survival and progression free survival were estimated by calculating the area under each curve (AUC) using the trapezoidal rule. At the time of the analysis, second line progression-free survival curves were available only from the Tournigand trial.”</p>	
Independent survival models, or hazard ratio (proportional hazards) modelling	<p>Proportional hazards modelling was used: “In order to take account of correlations between the effectiveness of regimens and sequences of chemotherapy regimens, survival curves and first-line progression-free survival curves for the remaining six sequences (i.e. FOCUS Treatment Plans B-E, and the two Tournigand treatment arms) were estimated using the Weibull survivor function for the baseline FOCUS Treatment Plan A together with a log-rank hazard ratio describing the survival difference between the experimental curve and the baseline curve. The log-rank hazard ratios were treated as relative hazards between the experimental arms as compared to the baseline. The same approach was used in the analysis of second-line therapies albeit using second-line FOLFIRI as the baseline survivor function.”</p>	Unclear.
Justification for survival model used?	<p>The AG state: “The Weibull survivor function $S(t)$ is given by the formula: $S(t) = \exp[-\lambda t^\gamma]$ where λ = scale parameter, t = time, and γ = shape parameter. Transforming the survivor function $S(t)$ gives the linear relationship: $\Rightarrow \ln[-\ln S(t)] = \ln \lambda + \gamma \ln t$ where $\ln(t)$ is the independent variable and $\ln[-\ln(S(t))]$ is the dependent variable. If the Weibull assumption is valid, this transformation applied to the Kaplan-Meier survival estimates results in a straight line whereby $\ln[-\ln S(t)] = y$, $\ln \lambda$ = intercept, γ = gradient and $\ln t = x$. All of the regression analyses resulted in an approximately straight line relationship between $\ln(t)$ and $\ln[-\ln(S(t))]$ which justifies the Weibull assumption.”</p> <p>Also, regarding the PH modelling:</p> <p>“The analysis of overall survival and progression-free survival within the model makes an explicit assumption of proportional hazards between the patients evaluated within the FOCUS trial³ and patients evaluated within the Tournigand trial. Put simply, the analysis is based upon the assumption that the hazard of death at any given time for an individual within the Tournigand trial is proportional to the hazard of death at that time for a similar individual in the FOCUS trial.³ Log-rank hazard ratios for FOCUS treatment plans B,C,D, and E versus FOCUS Plan A (MdG+Ir) in terms of overall survival and first-line progression-free survival were made available to the Assessment Group by the MRC (Personal communication, G. Griffiths, MRC Clinical Trials Unit, London). Log-rank hazard ratios comparing the FOLFOX6/FOLFIRI and FOLFIRI/FOLFOX6 sequences evaluated within the Tournigand study to the baseline FOCUS Plan A (MdG+Ir) were not available, thus an implied relative risk for each of the Tournigand treatment arms was estimated using a least-squares approach, and was tested by undertaking a separate regression analysis for the Tournigand treatment arms.”</p>	Limited detail.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	State transition model.	<p>Sanofi Sequencing model had the following time periods:</p> <ol style="list-style-type: none"> 1. Time on first-line treatment; 2. Time following treatment cessation until disease progression; 3. Time from first-line disease progression until start of second-line treatment; 4. Time on second-line treatment; 5. Time following cessation of second-line treatment until disease progression;

		<p>6. Time from disease progression until death.</p> <p>Aventis Markov Model</p> <ol style="list-style-type: none"> 1. Progression-free on first-line therapy 2. Progression on first line therapy 3. Progression-free on second-line therapy 4. Progression on second-line therapy 5. Dead
<p>Other issues noted (eg crossover)</p>	<p>The Assessment Group state: “The addition of oxaliplatin to first line 5-FU is associated with: no significant difference in overall survival (but see caveat below); significantly improved progression-free survival (p<0.00001); significantly higher response rates (p<0.0001); more serious gastrointestinal and haematological toxicities; no significant overall improvement of quality of life. Schedules which offer treatment breaks do not appear to reduce clinical effectiveness but may reduce toxicity. Caveat: confounding by crossover from 5-FU monotherapy to oxaliplatin combination in all trials may mask a real survival advantage for the latter.”</p> <p>Also: “The current NICE recommendation, 5-FU monotherapy followed by irinotecan monotherapy, appears to be inferior to any other planned sequence. Combination irinotecan and 5-FU as first line therapy, significantly improved overall survival and time to first progression. However, although this plan did not have an official secondline therapy, some patients received ‘salvage’ oxaliplatin and capecitabine (oral 5- FU), which will have affected the treatment effect size for overall survival to an unknown extent. Staged combination therapy (oxaliplatin and 5-FU followed by irinotecan and 5-FU or vice versa) appears to provide the best overall and progression-free survival, although there has been no head-to-head comparison against other treatment plans. In the only trial (GERCOR), to use all three active chemotherapies (5-FU, Irinotecan and Oxaliplatin), overall survival was over 20 months in any staged combination. In the FOCUS trial (the other study which planned sequences of treatment), the longest recorded median overall survival from a treatment plan using only two active agents was 16.2 months.”</p> <p>And: “Over half of first-line trial participants across all studies except two, were treated with unplanned second-line therapies: it is unknown to what extent estimates of overall survival are confounded as a result.”</p> <p>12 studies relevant for the economics review were found, but the Group state: “the principal limitation of existing economic evidence relates directly to flaws in the design and reporting of the clinical trials from which evidence of effectiveness is drawn. Within the majority of clinical trials, patients received further chemotherapy following disease progression, thus the survival benefits in patients within these trials cannot be uniquely attributed to the allocated therapy. Consequently, economic analyses that draw on evidence from these trials were limited either to the use of progression-free survival which may be considered at best a surrogate outcome, or were subject to confounding due to patients crossing over to alternative chemotherapy agents following disease progression.”</p> <p>In addition, some of these studies used median measures, rather than mean, which the AG states is incorrect for economic evaluations.</p> <p>This is noted to also have been a problem in TA33, which this appraisal reviews. However, in this appraisal a sequencing model has been used in order for this confounding to be less of a problem – this approach was not taken in the earlier appraisal.</p> <p>The authors present a brief review of economic endpoints such as overall survival, quality adjusted survival, progression-free survival, quality adjusted progression-free survival, tumour response and adverse events avoided. The Group note that overall survival can be confounded by treatment crossover, and also mean rather than median survival is required in economic evaluations – usually requiring extrapolation using parametric curves, which can lead to some error.</p> <p>Crossover is noted as a particular problem by the authors. The see sequencing models as an answer, based on sequencing trials (although such trials may often not be available): “Thus, overall survival can be evaluated only as a measure of sequences of chemotherapy regimens. Only the trial reported by Tournigand and the FOCUS trial have evaluated the overall survival benefits of planned chemotherapy sequences in advanced colorectal cancer. It should be noted however that whilst the FOCUS trial incorporated a protocol change which resulted in sequences whereby all three drugs were used; this was not planned from the outset, hence only the trial reported by Tournigand <i>et al</i> planned from the outset to compare sequences containing all three active agents. To illustrate the magnitude of this problem, Table 54 shows the percentage of patients who received further chemotherapy with a different agent following disease progression. Of the eight trials that reported the number of patients who received further chemotherapies following progression, in all but one trial this proportion was greater than 50%.”</p> <p>The authors state that the sequencing trials allow inclusion of OS in the modelling, but that confounding still remains in one of the trials: “To date, only two trials have used planned sequences of chemotherapies; these studies allow for the analysis of overall survival data (although due to the late amendment to the study protocol, there remains some confounding within the results of the FOCUS trial”</p> <p>The Group also consider PFS as an endpoint: “The primary advantage of this outcome measure is that it is not confounded by patients receiving other chemotherapy agents following disease</p>	

<p>progression. However, there exist problems in interpreting progression-free survival results from existing clinical trials...Time to progression is dependent on the frequency of checkups. ... Median progression-free survival may not represent true progression-free survival benefits”</p> <p>The AG also state: “For economic studies in which evidence of effectiveness is drawn from clinical trials with unplanned crossovers following disease progression, the only means by which to avoid confounding is through the use of progression-free survival as the measure of benefit. However, as discussed in Section 4.1.2., this benefit measure may at best be considered a surrogate outcome, hence the interpretation and generalisability of the cost-effectiveness results are limited.”</p> <p>And: “The most significant improvement to existing economic evaluations of these therapies would be the analysis of overall survival, whereby evidence of effectiveness would be drawn from clinical trials which have evaluated planned sequences of chemotherapies. The use of planned sequences of chemotherapy would also enable the analysis of progression-free survival for first- and second-line therapies. Mean overall survival should be estimated as the area under the survival curve.”</p> <p>The FAD showed that treatment crossover remained a problem even for the AG sequencing model: “However, the Committee was aware of a number of uncertainties around the modelling, such as the data on resource use and quality of life taken from the FOCUS study that were not yet validated, and – most critically – that the cost of salvage therapies was not included. The Committee heard testimony from clinical experts that a significant proportion of patients in the FOCUS trial received further treatment after progression and that the extent to which this happened was unevenly distributed between the treatment arms.”</p>
--

23. TA101: Prostate cancer (hormone-refractory) - docetaxel, June 2006

Guidance: Docetaxel is recommended, within its licensed indications, as a treatment option for men with hormone-refractory metastatic prostate cancer only if their Karnofsky performance-status score is 60% or more.

It is recommended that treatment with docetaxel should be stopped:

- at the completion of planned treatment of up to 10 cycles, or
- if severe adverse events occur, or
- in the presence of progression of disease as evidenced by clinical or laboratory criteria, or by imaging studies.

Repeat cycles of treatment with docetaxel are not recommended if the disease recurs after completion of the planned course of chemotherapy.

Source: Docetaxel for the treatment of hormone-refractory metastatic prostate cancer, TA101, June 2006, <http://www.nice.org.uk/nicemedia/live/11578/33348/33348.pdf>, accessed 09/04/10

Collins R, Fenwick E, Trowman R, Perard R, Norman G, Light K, Birtle A, Palmer S, Riemsma R. A systematic review and economic model of the effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer, Centre for Reviews and Dissemination, University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, September 2005 <http://www.nice.org.uk/nicemedia/live/11577/33344/33344.pdf>, accessed 09/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>One paper was found that assessed the cost-effectiveness of mitoxantrone plus prednisone compared to prednisone alone. The evaluation was based on data from the CCI-NOV22 clinical trial. Due to the extent of the crossover within the trial, cumulative costs over time were presented for the two treatment groups (as initially randomised) and for those in the prednisone group who did not crossover (intention to treat). Mean survival was estimated based on patient-level data – it is not clear if any extrapolation was required.</p> <p>The Assessment Group developed their own model, which included far more comparators than were included in the manufacturer’s model. An indirect treatment comparison was used. A two-state Markov model was used, and so overall survival was the clinical endpoint. The cycle length was 1 month, and transition probabilities to death were estimated. For three of the comparators (those included in the Sanofi trial (TAX 327)) ((D+P (3-weekly), M+P and D+P (weekly)) the results of the Sanofi survival analysis using a Weibull extrapolation were used.</p>	<p>Sanofi</p> <p>The economic analysis evaluated the cost-effectiveness of docetaxel plus prednisone (3-weekly regimen) compared to mitoxantrone plus prednisone. The evaluation was based on an analysis of patient-level data derived from prospective collection of resource use and patient outcome data from the TAX 327 clinical trial. The primary outcome for the cost-effectiveness analysis was lifeyears gained based on a comparison of overall survival in the different intervention groups. Separate life-years gained estimates were provided based on a within-trial comparison (using median survival data) and a lifetime comparison (using mean survival data). The lifetime comparison was based on an extrapolation approach using parametric survival-analysis (using a Weibull).</p> <p>Survival estimates using the different approaches were: Median survival from Kaplan-Meier: Docetaxel plus prednisone =18.9 months Mitoxantrone plus prednisone =16.5 months <i>Difference = 2.4 months</i></p>

	<p>The AG state “As previously stated in the economic review section, the submission by Sanofi-Aventis presented the mean estimate for the coefficients for the intercept and scale parameters for two interventions: D+P (3-weekly) and M+P. Since this analysis was based on a patient-level analysis of survival data from the TAX 327 study, it was decided that this approach would provide the most reliable approach to quantifying mean survival for these interventions. Additional data were therefore requested in order to extend the approach used by Sanofi-Aventis to facilitate the inclusion of other relevant comparators and to ensure that uncertainty surrounding the coefficients was incorporated in the final decision model. Furthermore, the use of the Markov model to estimate mean survival enabled discounting to be incorporated. Details of the intercept and scale parameters for the D+P (weekly) arm of TAX 327 were requested in addition to the standard errors for these coefficients for each of the three comparators in this trial.”</p> <p>“For the purposes of the probabilistic analysis it is also important to reflect the covariance between the intercept and scale parameters from the Weibull regression. The covariance matrix for each intervention was supplied on request by Sanofi-Aventis. This matrix was used to derive the Cholesky decomposition matrix which was then used to allow for correlation when generating the random normal draws for the intercept and scale parameters in the probabilistic simulation”</p> <p>“Since hazards are instantaneous these need to be converted to a transition probability for a given time period (e.g. cycle) and require use of the integrated hazard function. For the Weibull distribution the integrated hazard function is: $H(t) = \int_0^t h(u) du = \lambda u^\gamma$.</p> <p>Using this formula, the hazard rate was estimated for each of the monthly cycles of the model. Following this procedure, the hazard rates were then converted into transition probabilities using standard techniques”. A number of the hazards and associated transition probabilities were then presented by the author.</p> <p>An indirect comparison approach was used to estimate hazard ratios for the other comparators so that they could be included in the model: “Since patient-level data were not available for any of the other comparators it was necessary to derive an estimate of the relative treatment effect for these to be applied in the model. Using the Bucher approach outlined in the clinical effectiveness review, indirect hazard ratios were estimated in order to include other comparators in the economic model. In order to reflect the potential correlation between the different interventions, docetaxel-based regimens were assessed via an estimate of the indirect hazard ratio versus D+P (3-weekly) and mitoxantrone/prednisone strategies were assessed via the indirect hazard ratio in relation to M+P. The indirect hazard ratios for these additional comparators are shown in Tables 31 and 32. The uncertainty associated with each hazard ratio was characterised by assigning a normal distribution to the (log) hazard. The hazard ratio was then applied to the absolute hazard for either D+P (3-weekly) or M+P and then converted in order to obtain the required transition probability.”</p>	<p>Mean survival based on extrapolation using parametric survival model using Weibull distribution (95% CI): Docetaxel = 22.38 (20.38-24.62) months Mitoxantrone = 18.65 (17.30 – 20.12) months Difference=3.73 months</p>
Evidence synthesis (pool survival estimates?)	Mixed treatment comparison.	None.
Survival model(s) fitted (Weibull, exponential etc)	For three comparators the manufacturer’s Weibull extrapolation was used.	Weibull applied to both treatment arms (although an analysis based on medians was also presented).
Independent survival models, or hazard ratio (proportional hazards) modelling	Appears that some Weibull models were fitted independently, but that HRs were used so that other comparators could be included in the model.	Independent.
Justification for survival model used?	Limited to HRs for other comparators because patient-level data not available. Noted that the manufacturer partially justified their choice of a Weibull model.	The AG state that “A Weibull model was applied to the survival data based on a visual check of a plot of log-cumulative hazard against log time. A Weibull model is used in

		situations in which the assumption of a constant hazard with respect to time is not appropriate (i.e. the risk of mortality is increasing/decreasing).” Thus the manufacturer did justify their choice to some extent.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Two-state Markov Model (alive and dead health states)	Model based upon life years.
Other issues noted (eg crossover)	<p>The AG note that one of the trials of one of the comparators allowed treatment crossover: “In addition to the differences in population, one trial allowed patients to cross over during the trial, this resulted in 50 out of 81 patients randomised to prednisone receiving additional mitoxantrone; the other two trials did not allow crossovers. Including crossovers in intention to treat analyses can result in ‘dilution’ of the true effects of a treatment, as patients are analysed as randomised. However, in this case the study that allowed crossovers had a stronger treatment effect in favour of mitoxantrone plus prednisone than the two studies that did not allow crossovers.”</p> <p>Hence no adjustment was made for crossover, but in this case it appears that the treatment effect may not have been impacted by the crossover (or that may be due to other differences in the trial).</p> <p>It is also stated in the clinical section of the report that in an RCT used in the analysis comparing docetaxel + prednisone with mitoxantrone + prednisone there was substantial crossover: “There was a high level of crossover between groups in this trial, 27% of patients randomised to the 3-weekly docetaxel group received mitoxantrone, 24% of patients randomised to the weekly docetaxel group received mitoxantrone and 20% of patients randomised to the mitoxantrone group received docetaxel”</p> <p>Similar was true in a trial comparing docetaxel plus prednisone plus estramustine with mitoxantrone plus prednisone: “There was a high level of crossover between groups in this trial, 16% of patients randomised to the one dose docetaxel group crossed over, 10% of patients randomised to the two dose docetaxel group crossed over and 48% of patients randomised to the mitoxantrone group crossed over. The difference in crossover between the treatment groups was statistically significant (P=0.00001)”</p> <p>Treatment crossover, and alternative post progression treatments were also allowed in other relevant trials (including TAX327). Some consideration was given to the impact of crossover on costs, but this was not seen to make a significant difference and was not a large issue. No adjustments to survival estimates were made.</p>	

24. TA105: Colorectal cancer - laparoscopic surgery (review), August 2006

Guidance: Laparoscopic (including laparoscopically assisted) resection is recommended as an alternative to open resection for individuals with colorectal cancer in whom both laparoscopic and open surgery are considered suitable.

Laparoscopic colorectal surgery should be performed only by surgeons who have completed appropriate training in the technique and who perform this procedure often enough to maintain competence. The exact criteria to be used should be determined by the relevant national professional bodies. Cancer networks and constituent Trusts should ensure that any local laparoscopic colorectal surgical practice meets these criteria as part of their clinical governance arrangements.

The decision about which of the procedures (open or laparoscopic) is undertaken should be made after informed discussion between the patient and the surgeon. In particular, they should consider:

- the suitability of the lesion for laparoscopic resection
- the risks and benefits of the two procedures
- the experience of the surgeon in both procedures.

Source: Laparoscopic surgery for colorectal cancer, Review of NICE technology appraisal 17, TA105, August 2006, <http://www.nice.org.uk/nicemedia/live/11588/33495/33495.pdf>, accessed 09/04/10

Murray A, Lourenco T, de Verteuil R, Hernandez R, Fraser C, McKinley A, Krukowski Z, Vale L, Grant A. Systematic review of the clinical effectiveness and cost-effectiveness of laparoscopic surgery for colorectal cancer, University of Aberdeen, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, November 2005 <http://www.nice.org.uk/nicemedia/live/11587/33469/33469.pdf>, accessed 09/04/10

	Assessment Group Model / Evidence Review Group Alterations
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>4 cost consequence analyses were found. Thus costs and effects are not compared together, and these studies appear to have largely been cost minimisation analyses. ‘Complications’ was the most commonly used outcome measure in the analyses, with some studies including other outcomes such as mortality and survival (eg, after 3 years).</p> <p>The Assessment Group conducted two analyses: “The first compares laparoscopic with open resection using a balance sheet approach and the second more formally synthesises the available data in an economic model. With the balance sheet the differences between interventions, in terms of costs and natural and clinical measures of effectiveness are</p>

	<p>presented. Such an approach served to highlight the choices and trade-offs between the two forms of resection.”</p> <p>In the cost consequence balance sheet it was estimated that the disease-free survival relative risk was 1.01 (0.95 to 1.07), and the OS relative risk was 1.03 (95% CI 0.98 to 1.09). Hence this was not assumed to be a benefit for either treatment as they were not different.</p> <p>For the Markov Model: “Estimation of the risk of death was based on the survival curve for open resection provided by Bonjer and colleagues... These data provided estimates of survival up to three years post surgery. Overall survival for open resection for each six month time period up to 36 months was estimated from these curves. From these data a mortality rate for each six-month cycle length was calculated. As interpreting rates from these curves is an imprecise method, and the mortality rates for each six-month period were similar, a constant mortality rate was assumed... The risk of recurrence of local or of metastatic disease was based on data on disease-free survival also provided by Bonjer and colleagues ...These data were estimated using the same methods as described for the risk of death described above. As with the risk of death a constant risk of recurrence was assumed. The risk of death following the recurrence of non-operative cancer was based on data derived from Benoist and colleagues. This study is a case-matched study set in France, which had the aim of determining the best treatment strategy for patients with asymptomatic colorectal cancer and irresectable synchronous liver metastases. Patients were recruited between 1997 and 2002 with 27 patients being treated with chemotherapy, without an initial primary resection, compared with 32 patients who were initially treated by resection of the primary tumour. The 27 chemotherapy patients (intervention group) were matched by age, sex, performance status, primary tumour location, number of liver metastases, nature of disease and the type of chemotherapy to the 32 patients who underwent resection of the primary tumour (control group). The mean age of the chemotherapy and resection groups was 61 and 60 respectively. Whilst this study is currently the best available data for this particular subset of patients, it should be noted that the very small sample size may result in imprecise estimates. The study setting might also impact upon the generalisability of results for the UK as this study, set in France, may have treatment regimes that differ from standard treatment in the UK.</p> <p>For the purposes of the model, the risk of death for patients with inoperable cancer was based on the interpretation of the survival curve for the “chemotherapy group” from the aforementioned study. This population was deemed to have similar characteristics to the patients undergoing non-operative management of recurrent disease within the model. The actuarial survival for the total time period of 24 months, divided into six month time periods, was estimated from this curve. A mortality rate for each six month cycle length was calculated and from this, a constant mortality rate obtained. Based on these data, a mortality rate for inoperable cancer with the value of 0.2 was calculated ... In order to reflect the statistical imprecision surrounding the occurrence of an event a Beta distribution was used. This distribution was used as it has been argued that it provides realistic representations of proportions”</p> <p>Estimating rates from curves like this may be imprecise, but these do represent rates that will lead to estimates of mean survival. Once the baseline rates of progression and death had been estimated, relative risks based on 3-year trial data were applied to estimate survival with laparoscopic surgery. The relative risks were close to 1 and the confidence intervals crossed 1.</p>
Evidence synthesis (pool survival estimates?)	None.
Survival model(s) fitted (Weibull, exponential etc)	Appears the Markov model was constructed assuming exponential survival curves.
Independent survival models, or hazard ratio (proportional hazards) modelling	Relative risks applied to baseline rates estimated from curves.
Justification for survival model used?	Mortality rates appeared to be similar over time.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>Markov model with the following states:</p> <ul style="list-style-type: none"> • Disease-free; • Recurrence of the disease where it may be possible to have a second operation or some form of non-operative management; • Disease-free (after a recurrence); where a patient following a successful second operation remains until they have a second recurrence/metastasis or die; • Non-operable recurrence resulting in non-curative management of the disease; and • Death <p>The cycle length was 6 months and the model ran for 60 cycles.</p>
Other issues noted (eg crossover)	None noted

25. TA110: Follicular lymphoma - rituximab, September 2006

Guidance: 1.1 Rituximab within its licensed indication (that is, in combination with cyclophosphamide, vincristine and prednisolone) is recommended as an option for the treatment of symptomatic stage III and IV follicular lymphoma in previously untreated patients.

Source: Rituximab for the treatment of follicular lymphoma, TA110, September 2006, <http://www.nice.org.uk/nicemedia/live/11592/33547/33547.pdf>, accessed 12/04/10

Dundar Y, Hounscome J, McLeod C, Bagust A, Boland A, Davis H, Walley T, Dickson R. Rituximab for the first line treatment of stage III-IV follicular non-Hodgkin's lymphoma, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, April 2006, <http://www.nice.org.uk/nicemedia/live/11591/46698/46698.pdf>, accessed 12/04/10

Note: STA – manufacturer submission not on website

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	No use of any other data.	Transition probabilities used in the economic model were based on one RCT (trial M39021) which compared RCVP to CVP alone. “The probability of remaining in the PFS health state was time dependent based upon log logistic extrapolation of PFS trial curves.” “Survival in the progressed state was estimated by fitting an exponential curve to the Scotland and Newcastle Lymphoma Group (SNLG) registry data on survival from second line chemotherapy. The SNLG database since 1994 has captured comprehensive treatment and outcomes data on more than 95% of lymphomas presenting in a population of 8.5 million across northern England and Scotland”
Evidence synthesis (pool survival estimates?)	No novel analyses.	Some use of external data.
Survival model(s) fitted (Weibull, exponential etc)	Tested the use of a Weibull model.	Log Logistic for PFS, exponential for post-progression. In sensitivity analysis a Weibull was fitted to PFS, which increased the ICER by approx 1k (£8k to £9k).
Independent survival models, or hazard ratio (proportional hazards) modelling	Unclear.	Unclear.
Justification for survival model used?	The ERG also note that the exponential model fit to the SNLG data is a very poor fit, and show that a Weibull fits much better. In addition, they correctly identify that when using external data it is important to consider the comparability of the trial and external populations – the manufacturer had not done this: “The fit of this model to the data is particularly poor, which is surprising considering the great effort put into modelling alternative functional forms for PFS. A suitable alternative model would be the Weibull function, which is commonly compatible with long-term survival in incurable cancers. The chart below shows how a Weibull model provides an excellent representation of the data.” “The exponential projection in submitted model yields a lifetime expected mean survival of 59.3 months, whereas the Weibull model projects a value of 71.6 months. This alteration has the effect of reducing the apparent survival benefit between RCVP and CVP, and also reduces the incremental cost of R-CVP versus CVP. Consequently, the ICER is increased, but is still below the £30,000 threshold. No information describing the characteristics of the SNLG patients used in the company submission is provided. This is of critical importance because the modelled population is so unrepresentative (by age group) of the normal incident population, implying that the mortality parameters in different parts of the model are almost certainly incompatible. It is not possible to correct the mortality rates applied to progressed patients for such age differences.”	No details.

	With changes made to the model, including those discussed above, the base ICER increased to £9k, when 64% of PFS benefit was estimated to extrapolate to OS benefit. When there was 0% extrapolated to OS benefit the ICER was £20k.	
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	No novel model.	The model is a three-state Markov model, with the health states being defined as: <ul style="list-style-type: none"> • “Progression free survival” • “Progressed” in which patients have relapsed. • “Death” which is an absorbing state. Cycle length is 1 month. Movement between health states is governed by transition probabilities. The probabilities applied to the PFS health state vary over time, but are generally similar between the two arms. The probabilities applied to the progressed health state are constant, and do not differ between the two arms.
Other issues noted (eg crossover)	<p>The ERG “calls into question the applicability of the long-term projective models of PFS (e.g. log-logistic) if they are unable to generate results which are compatible with mortality risks in the general population. A more appropriate analysis would be to model progression and death within a competing risks framework, ideally distinguishing between non-Hodgkin’s lymphoma deaths and all other causes of death. This would ensure that survival cannot be accidentally overestimated when projected to the end of life.”</p> <p>This relates to the type of economic model used.</p> <p>In addition, the ERG questions how OS has been estimated. Very little is said about the SNLG data, which is an interesting use of external data, but the ERG do discuss the implicit PFS to OS extrapolation.</p> <p>“Progression free survival and overall survival: A strong implicit assumption within the submitted model is that estimated progression free survival gains can be equated with equivalent overall survival benefits. From our own reading of the clinical literature, there is little evidence available in follicular lymphoma to support this hypothesis. The company’s own submission states that there was no statistically significant overall survival benefit (demonstrated by the M39201 RCT), but a statistically significant benefit in terms of PFS. It may be that future longer term studies will not show an overall survival benefit for patients receiving R-CVP compared to CVP, or may show only a reduced benefit. In other cancers, there is conflicting evidence on the correlation between PFS and OS. Colorectal cancer studies indicate that PFS is correlated with overall survival, but that each incremental month in PFS leads to only 0.68 month of additional OS. Evidence from ovarian cancer suggests that although progression free survival is improved, overall survival may not be affected. A recent study on chronic lymphocytic leukaemia (CLL) patients treated with rituximab and fludarabine versus fludarabine alone, showed that both progression free survival and overall survival did improve. However, for follow-up periods of 2-4 years only 30-40% of the estimated PFS gain was translated into OS gain. This is a more comparable haematological cancer with a similar treatment regimen, and therefore supports the suspicion that it would not be appropriate to infer that PFS benefit automatically confers OS benefit. Hence, the question of how much (if any) of the progression free survival (seen in the M39021 trial of NHL patients treated with rituximab and CVP versus CVP monotherapy) will translate into a survival gain remains unanswered. The submitted model does not address this issue, but assumes that the improvement in PFS automatically leads to a survival gain; this is because the mortality rates for progressed patients are identical for both arms. In the base case results this implies that 79% of PFS gain is translated into OS gain. Thus the gain in PFS is ameliorated only by differential mortality attrition as patients’ progress at different times in the arms.”</p> <p>It is not mentioned by the ERG why the manufacturer chose not to estimate post progression survival on data from their own trial, other than that the data was quite incomplete (median had not been reached). Crossover is not mentioned, but it could be a reason for not using trial OS data.</p>	

26. TA116: Breast cancer - gemcitabine, January 2007

Guidance: Gemcitabine in combination with paclitaxel, within its licensed indication, is recommended as an option for the treatment of metastatic breast cancer only when docetaxel monotherapy or docetaxel plus capecitabine are also considered appropriate.

Source: Gemcitabine for the treatment of metastatic breast cancer, TA116, January 2007, <http://www.nice.org.uk/nicemedia/live/11610/33875/33875.pdf>, accessed 12/04/10

Jones J, Takeda A, Tan SC, Cooper K, Loveman E, Clegg A, Murray N. Gemcitabine for metastatic breast cancer, Southampton Health Technology Assessments Centre, University of Southampton, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, July 2006, <http://www.nice.org.uk/nicemedia/live/11609/33865/33865.pdf>, accessed 12/04/10

Eli Lilly & Co., Gemcitabine for the treatment of metastatic breast cancer, Single Technology Appraisal Submission to the National Institute for Health and Clinical Excellence (Confidential information removed), May 2006

Note:

STA – manufacturer submission on website

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
<p>Survival data used (patient-level data, or summary statistics – mean, median etc)</p>	<p>“The model assumed a constant risk for disease progression and for mortality (p. 146). The ERG estimated the survival probabilities and risk of disease progression for patients in the paclitaxel arm of the trial from survival plots reported in the conference presentation by Albain and colleagues, and fitted a parametric survival function to these data using the outputs from an ordinary least squares regression on a log-cumulative hazard. These suggest that the survival functions have non-constant risk over time.”</p> <p>“The ERG estimated the external validity of the model by comparing it with survival estimates from the JHQG trial. The model was run with median survival times for gemcitabine/paclitaxel and paclitaxel as shown in the JHQG. ... The model shows a reasonable fit with the Kaplan-Meier curves, with a slightly increased survival for the model compared to the trial data for gemcitabine/paclitaxel and paclitaxel. The model and trial estimated an average survival benefit of 7.6 and 11.2 weeks respectively for gemcitabine/paclitaxel versus paclitaxel. Thus the model underestimates the treatment effect of gemcitabine/paclitaxel versus paclitaxel by about 30% compared to the JHQG trial.”</p> <p>“The pooled estimate for overall survival with docetaxel monotherapy is lower than the pooled estimate for overall survival with paclitaxel (see column 4, Table 11) – these are the values used to derive the transition probabilities in the base case analysis. However the only clinical trial included in the MS that reports a head-to-head comparison of paclitaxel monotherapy against docetaxel monotherapy (Jones and colleagues) shows a longer survival duration for docetaxel (see column 2, Table 11). The survival benefit for gemcitabine/paclitaxel over paclitaxel in the JHQG trial is similar to that reported for docetaxel over paclitaxel by Jones and colleagues. The ERG undertook an illustrative analysis to examine the possible impact of using the relative (rather than absolute) effects observed in the two clinical trials, adopting a method similar to the classical method for indirect comparisons. This analysis is presented for illustrative purposes and is not a recommendation for adopting this method for conducting indirect comparisons of this type. Using relative effects produces an estimate for overall survival with docetaxel monotherapy that is closer to that observed for gemcitabine/paclitaxel in the JHQG trial (column 5, Table 11). Using this value in the model generates an ICER for gemcitabine/paclitaxel compared to docetaxel monotherapy of £45,811 per QALY gained.” This is compared to the manufacturer’s based case of £17k.</p>	<p>Treatment effects used in the economic model are derived from the overall survival duration, time to disease progression, overall response rate and toxicity data reported in 15 clinical trials. Absolute values for each of these parameters for each intervention were extracted from relevant trial reports and weighted averages were calculated – each value was weighted by the number of cases in the relevant trial arm.</p> <p>“Definitions for time to disease progression are reported in 10 of the 14 Phase III RCTs”. There was not a common comparator, so the manufacturer acknowledges that unadjusted comparisons of survival in the trials is subject to bias.</p> <p>Comparators were:</p> <ol style="list-style-type: none"> 1. Gemcitabine 1250mg/m2 plus paclitaxel 175mg/m2 2. Docetaxel monotherapy 100mg/m2; 3. Paclitaxel monotherapy 175mg/m2; 4. Docetaxel 75mg/m2 plus capecitabine 1250mg/m2 bid. 5. Gemcitabine 1000mg/m2 plus docetaxel 75mg/m2. <p>“1. It was assumed that the time to disease progression would be different for responding patients vs. those who did not achieve a response. 2. The time from response to disease progression was also determined for responding patients. This was achieved using the data from the S273 trial on: a) Overall time to disease progression for each treatment b) Time to disease progression for responding patients c) Response rates for each treatment a) and b) were used to determine the time to disease progression for non-responders using a weighted combination of mean times from assumed exponential time to disease progression curves.”</p> <p>As stated by the ERG “Cycle probabilities for use in the Markov model were derived using standard transformations”. Times to events included in the model were estimated by transforming medians in to means, assuming an exponential distribution.</p> <p>ERG: “The pooled median overall survival durations for each intervention were converted to cycle probabilities for use in the Markov model using standard transformations (deriving rates from the medians, as described by Beck and colleagues, and transition probabilities from the rates as described by Miller and Homan). Time to progression was estimated separately for the responders and non-responders using the pooled median time to progression, pooled overall response rates and additional trial data.”</p> <p>This approach assumes that “Survival functions (overall survival and time to progression) can be fully inferred from the median, assuming an exponential function;”</p>
<p>Evidence synthesis (pool survival estimates?)</p>	<p>Presented a simple indirect comparison for sensitivity analysis.</p>	<p>Naïve, pooling and weighting of absolute values.</p>
<p>Survival model(s) fitted (Weibull, exponential etc)</p>	<p>As an illustrative analysis ran the model using median survival times, and using relative treatment effects – did not fit any other survival models.</p>	<p>Exponential.</p>
<p>Independent survival models, or hazard ratio (proportional hazards)</p>	<p>Presented an analysis based upon relative treatment effects as a sensitivity analysis of the manufacturer’s analysis.</p>	<p>Independent.</p>

modelling		
Justification for survival model used?	Determined that an exponential was unlikely to be reasonable, but did not state which type of model did fit well. Also used external data to show that the model lacked external validity.	No details.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Did not present a novel economic model.	A Markov Model, based upon a previous model (Cooper 2003), was constructed. 4 health states: Stable; Response; Progressive; and Death. 3 week model cycle.
Other issues noted (eg crossover)	<p>The manufacturer states: "1. It is assumed that 50% of patients would receive at least one more line of therapy post-study chemotherapy and 25% will go on to receive 2 lines of therapy. Although post-study chemotherapy will vary, it is assumed based on clinical opinion that patients will receive either vinorelbine or capecitabine. As such, the average cost for both of these products will be used to cost these treatments. It was assumed that post-study treatments would be given for 6 cycles for capecitabine and 9 for vinorelbine. This assumption was based on clinical trial data on poststudy treatments clinical trial data from Albain <i>et al.</i>, 2004. In addition market research in UK patients showed vinorelbine and capecitabine most commonly used in patients 2nd line."</p> <p>Thus post-progression treatment is considered. It is unclear how this matches up with the OS data used in the economic model, or whether any post-progression treatment occurred in the clinical trial.</p> <p>The manufacturer's submission states that one trial included in the economic analysis was subject to treatment crossover (docetaxel patients switched onto docetaxel plus trasuzumab). This is noted by the ERG, who note that OS in this trial was above the 95% pooled CI for docetaxel. Regarding the manufacturer's JHQG trial: "Subsequent therapies were at investigator's discretion so patient may have crossed-over but this was not a prospective cross-over study."</p> <p>In addition, there was a problem of some missing OS data, leading to some use of external data:</p> <p>"For the clinical trial where Gemcitabine 1000mg/m2 plus docetaxel 75mg/m2 vs. capecitabine 1250mg/m2 (x 2 daily) plus docetaxel 75mg/m2 was studied (Chan <i>et al.</i>, 2005), data on patients' median overall survival is not as yet determined. In the absence of any interim data, an estimate of the most likely survival estimate for this treatment combination had to be determined using data from two trials (O'Shaughnessy <i>et al.</i>, 2004; Albain <i>et al.</i>, 2004). The rationale for choosing these two studies was as follows: The O'Shaughnessy <i>et al.</i>, (2004) trial reported median overall survival and time to disease progression for patients randomised to receive capecitabine / docetaxel in the same doses as those administered in the Chan <i>et al.</i>, (2005) trial. The Albain <i>et al.</i>, (2004) trial also reported median overall survival and time to disease progression for patients randomised to receive Gemcitabine / paclitaxel. This was the only trial where Gemcitabine had been given in combination with a taxane-based therapy. Had the median time to disease progression been very similar for the capecitabine / docetaxel arms in both the Chan <i>et al.</i>, (2005) trial and the O'Shaughnessy <i>et al.</i>, (2004) trial, it might have sufficiently justifiable to have assumed similar overall survival rates for the Chan <i>et al.</i>, (2005) trial. There was however a difference in the times to disease progression between the two trials. For this reason alone, the relationship between the time to disease progression and overall survival (expressed as a ratio) reported in the O'Shaughnessy <i>et al.</i>, (2002) trial for the combination arm (capecitabine / docetaxel) and the Albain <i>et al.</i>, (2004) trial for the combination arm (Gemcitabine / paclitaxel) was studied. The mid-point between these 2 ratios was used to estimate median overall survival for the Gemcitabine / docetaxel arm of the Chan <i>et al.</i>, (2005) trial."</p> <p>Evidence synthesis caused problems – no common comparators, so estimates of survival based on weighted pooled (absolute – not relative) estimates are subject to bias. The ERG stated: "Data from clinical trials – fully published [except JHQG]. Only the JHQG trial was quality assessed. Overall survival and TTP advantage for gemcitabine/paclitaxel combination over paclitaxel monotherapy established by direct comparison in JHQG trial. All other comparisons indirect – with questionable validity of pooling method. No formal assessment of heterogeneity. Data from trials with anthracycline naïve patients included in base case. Inconsistent use of independent versus investigator assessed response."</p>	

27. TA118: Colorectal cancer (metastatic) - bevacizumab & cetuximab, January 2007

Guidance: Bevacizumab in combination with 5-fluorouracil plus folinic acid, with or without irinotecan, is not recommended for the first-line treatment of metastatic colorectal cancer.

Cetuximab in combination with irinotecan is not recommended for the second-line or subsequent treatment of metastatic colorectal cancer after the failure of an irinotecan containing chemotherapy regimen.

People currently receiving bevacizumab or cetuximab should have the option to continue therapy until they and their consultants consider it appropriate to stop.

Source: Bevacizumab and cetuximab for the treatment of metastatic colorectal cancer, TA118, January 2007, <http://www.nice.org.uk/nicemedia/live/11612/33930/33930.pdf>, accessed 12/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
<p>Survival data used (patient-level data, or summary statistics – mean, median etc)</p>	<p>Roche</p> <p>The assessment group based survival estimates in their economic model on PFS data from the relevant clinical trials, 2nd line PFS data from a relevant sequencing trial, and OS data from the bevacizumab trials. The group stated that post-hoc analyses from trial AVF2107 shows that the survival of patients beyond disease progression was unaffected by the use of bevacizumab in subsequent lines of therapy, which rationalises their modelling approach. Data showed that of the patients initially randomised to IFL plus bevacizumab, those who received 2nd line bevacizumab had a survival duration of 9.99 months, compared to 9.40 months in those who did not receive 2nd line bevacizumab. Patients in the IFL group who did not receive 2nd line bevacizumab had a survival duration of 10.09 months, and patients who were randomised to 5FU/FA plus bevacizumab who received 2nd line bevacizumab had a survival duration of 10.97 months.</p> <p>The manufacturer’s model resulted in more favourable results when compared to one treatment because the difference in PFS in the clinical trial was greater than the difference in mean overall survival. Conversely the assessment group’s model gave more favourable results when comparing against another comparator because the difference in OS in the clinical trial was greater than the difference in PFS.</p> <p>The assessment group fitted Weibull curves to the PFS Kaplan-Meier. Mean PFS durations were estimated from these curves. Second-line PFS was estimated using previous sequencing trials in metastatic CRC (Tournigand <i>et al</i>), and this was assumed to be the same in each treatment group. Total time without progression was then estimated by adding 1st and 2nd line PFS durations. The duration for which patients experience progressive disease was estimated as the OS duration minus the total PFS as calculated above.</p> <p>The differing modelling approaches resulted in different cost-effectiveness estimates. The Roche model resulted in more favourable cost-effectiveness estimates when the comparator was FU/LV, because the difference in mean PFS was greater than the difference in mean OS in trial AVF2192. The opposite was true for trial AVF2107, for which the assessment group model came up with more favourable cost-effectiveness estimates.</p> <p>“Kaplan-Meier curves giving empirical estimates of overall survival in each of the four treatment groups were obtained from the trial publications. These empirical survival curves were digitally scanned using TECHDIG software, and subsequently imported into Microsoft EXCEL. As some patients were still alive at the end of the AVF2107g and AVF2192g trials (i.e. the curves were right-censored), the final portion of each survival curve was extrapolated using regression analysis to estimate the parameters of a Weibull survival curve. Independent regression models were constructed to describe the probability of overall survival over time within each of the four treatment groups.”</p> <p>“The duration for which patients are free from disease progression whilst receiving first-line treatment was estimated using progression-free survival curves reported</p>	<p>Roche</p> <p>“The Roche submission to NICE65 included details of two mathematical models used to estimate the cost-effectiveness of bevacizumab in combination with irinotecan plus 5-FU/FA versus irinotecan plus 5-FU/FA alone, and bevacizumab in combination with 5-FU/FA versus 5-FU/FA alone”</p> <p>“The two cost-effectiveness models are based upon effectiveness evidence and resource use data collected within studies AVF2107g and AVF2192g respectively”</p> <p>“Whilst bevacizumab is currently indicated only for the first-line treatment of patients with metastatic CRC, the analysis includes additional long-term costs and health outcomes associated with unspecified subsequent-line therapies and other palliative treatments received beyond disease progression.”</p> <p>“Evidence relating to the additional survival benefits resulting from the use of bevacizumab in combination with first-line IFL and 5-FU/FA compared to chemotherapy alone was derived from trials AVF2107g and AVF2192g... Within studies AVF2107g and AVF2192g, patients who were randomised to receive bevacizumab as a first-line treatment were also subsequently allowed to receive bevacizumab as a subsequent-line therapy following disease progression. This is currently outside of the current licensed indications for bevacizumab. In an attempt to avoid this potential confounding, which could result in additional survival benefits for the bevacizumab-including treatment groups of studies AVF2107g and AVF2192g, the Roche models assigned the same risk of death following disease progression on first-line treatment to all patients irrespective of treatment group. This was modelled as the risk of death following disease progression over the entire clinical trial population. Second-line therapies were controlled for as a covariate in estimating survival beyond disease progression. The assumption implied by this approach is that all of the benefit attributable to bevacizumab is derived whilst the patient is on treatment, and that post-progression chemotherapy does not include bevacizumab. As the same post-progression survival curve is applied to all treatment groups, the models assume that the additional benefit of bevacizumab on overall survival is exactly equivalent to the additional benefit of bevacizumab on progression-free survival. Regression analysis was used to estimate Weibull coefficients describing progression-free survival time and post-progression survival time, using evidence from the trial datasets. Pre-progression mortality was assumed to be zero (i.e. patients must progress before they die), although within the clinical trials 4-9.5% patients died prior to documented disease progression; this represents a bias in all modelled treatment groups. The submission states that this assumption was tested within the sensitivity analysis, however, no results for this particular analysis were presented. The parametric progression-free survival curves were used to estimate the probability of transiting to the post-progression health state during any given cycle for each treatment arm. The proportion of patients who make this transition are then weighted by time in order to estimate the contribution of patients in the progression-free health state to overall survival within that treatment arm. This provides an estimate of the area under the curve. Within the Roche cost-effectiveness models, the contribution to overall survival of patients in the post-progression health state is estimated by multiplying the proportion of patients who progress during each month by post-progression survival probabilities.”</p> <p>“The Roche submission notes that bevacizumab has been shown to confer a survival advantage when administered alongside second-line chemotherapy in bevacizumab-naïve patients. Whilst adjusting the survival benefits observed within the intervention trial arms due to patients receiving</p>

<p>within studies AVF2107g and AVF2192g. Parametric Weibull curves were fitted to empirical Kaplan-Meier progression-free survival curves using the same method described in Section 6.2.1.4.1. Mean progression-free survival durations for each treatment group were estimated by calculating the area under the progression-free survival curves. Second- and subsequent-line progression-free survival durations were not measured for patients following disease progression on first-line therapy within studies AVF2107g or AVF2192g. Second-line progression-free survival durations were assumed to reflect the experience of patients allocated to the FOLFIRI-FOLFOX treatment group within the GERCOR trial reported by Tournigand <i>et al.</i> Given the absence of evidence to the contrary, the models assume that progression-free survival whilst receiving second-line treatment is the same for each treatment group, based upon the progression-free survival duration observed within the Tournigand trial. Total time without disease progression was estimated by adding first-line and second-line progression-free survival durations. The duration for which patients experience progressive disease was calculated as the overall survival duration observed within treatment groups within studies AVF2107g58 and AVF2192g60 minus the estimated total progression-free survival periods calculated above.”</p> <p>Merck “The mathematical model developed by the Assessment Group is centred around the methodology and data used within Merck’s submission to NICE, but incorporates more plausible assumptions concerning the expected survival of patients beyond the duration of the trial. The model also explores the impact of alternative assumptions concerning the survival of patients receiving active/best supportive care. It is crucial to note from the outset that the development of the Assessment Group model should be interpreted in the light of the absence of available evidence on the comparative efficacy of cetuximab plus irinotecan versus active/best supportive care. The review of the clinical effectiveness of cetuximab plus irinotecan (See Section 5.3) highlighted the complete absence of empirical evidence to demonstrate whether cetuximab plus irinotecan improves either health-related symptoms or overall survival in patients with EGFR-expressing metastatic CRC who have previously failed on irinotecan-containing therapy”</p> <p>“Owing to the dearth of direct clinical evidence, the primary health economic analysis is presented as a threshold analysis, which attempts to elucidate the degree of additional overall survival benefit required in order for cetuximab plus irinotecan to achieve an acceptable level of cost-effectiveness and cost-utility in comparison to active/best supportive care.”</p> <p>“The effectiveness of treatment with cetuximab plus irinotecan was estimated using patient-level data collected within the BOND trial. Owing to the questionable validity of the extrapolation of overall survival outcomes for patients estimated within the Merck model (See Section 6.1.4.2), an alternative method of extrapolation using Weibull regression analysis was used to adjust for censoring of patients outcomes within the cetuximab plus irinotecan arm of the BOND trial. Kaplan-Meier curves were constructed for patients allocated to the cetuximab plus irinotecan group of the BOND trial using the empirical patient-level survival outcomes reported within the Merck cost-effectiveness model. The parameters of a Weibull survivor function were then estimated using linear regression analysis.”</p>	<p>bevacizumab following disease progression appears intuitively appropriate, the Roche submission presents <i>post-hoc</i> analyses from AVF2107g which suggests that the survival of patients beyond disease progression was unaffected by the use of bevacizumab in subsequent lines of therapy.”</p> <p>“The similar mean survival durations and overlapping confidence intervals between treatment groups suggests that treatment with bevacizumab alongside second-line chemotherapy in patients who have previously received bevacizumab alongside first-line chemotherapy does not confer additional survival benefits over and above other available chemotherapies. In addition, the Roche submission notes that adjusting for second-line therapy within the regression analysis made little difference to the post-progression Weibull model coefficients. In the light of this evidence, the justification for adjusting the observed overall survival estimates for the bevacizumab-including treatment groups within the AVF2107g and AVF2192g trials for use in the model is unclear, and may have been unnecessary.”</p> <p>The AG found that the approach of assuming the same probability of death upon progression for both treatment arms was reasonable in one case – where OS gains were similar to PFS gains in the trial, but not in another, where PFS gains in the trial seemed larger than OS gains: “consideration of differences in mean progression free survival and mean overall survival (See Sections 5.2.2.1 and 5.2.2.2) suggests that this may be reasonable for the economic analysis of study AVF2107g. However, the impact of censoring on progression-free survival outcomes for study AVF2192g resulted in a notably larger difference in mean progression-free survival than mean overall survival between the treatment groups (See Sections 5.2.2.1 and 5.2.2.2); for this study, the use of progression-free survival is likely to result in cost-effectiveness estimates that are biased in favour of the bevacizumab-including treatment group.”</p> <p>Merck “The Merck submission to NICE reported details of a mathematical model used to estimate the cost-effectiveness of second- and subsequent-line treatment using cetuximab plus irinotecan versus active/best supportive care”</p> <p>“The expected overall survival duration of patients receiving cetuximab plus irinotecan was estimated using patient-level data collected within the BOND trial. Survival modelling techniques were used to extrapolate overall survival curves beyond the duration of the BOND study to account for censoring of patients outcomes in both arms of the trial. Parametric curves were estimated using empirical Kaplan-Meier overall survival curves at the point at which the intervention and comparator curves diverged, based upon methods detailed by Gelber <i>et al.</i> The expected overall survival time for each patient was estimated as the total survival duration up to the point at which the patient was censored <i>plus</i> the additional survival duration beyond the censored survival duration predicted by the parametric curve. The Merck submission states that this process was not undertaken for progression-free survival as almost all patients progressed during the follow-up period”</p> <p>The cetuximab submission considered a treatment continuation rule, and thus survival estimates from a trial which did not include this rule were adjusted: “For the cetuximab plus irinotecan arm of the BOND trial, empirical and projected survival estimates were adjusted in order to account for those patients who continued to receive cetuximab plus irinotecan within the BOND trial, who it is anticipated would be withdrawn from treatment in usual clinical practice according to Merck’s proposed continuation rule. Under the continuation rule, patients would continue to receive treatment with cetuximab only if they have either a complete or partial tumour response at the 6-week CT scan, or if there is no change at the 6-week scan and there is evidence of the presence of a grade 2 or higher acne-like rash. The expected survival duration for those patients who continue to receive treatment with</p>
--	---

		<p>cetuximab according to the continuation rule is calculated as the mean observed survival probability, with additional survival benefits attributed to those patients whose outcomes were censored. For those patients who have stable disease but do not have an acne-like rash, expected survival is calculated as their mean survival duration multiplied by an adjustment factor of 0.906. This adjustment factor represents the relative survival of patients with no change in CT scan at 6-weeks and without grade 2 or above acne-like rash as compared to the survival of patients who did not go on to achieve a complete or partial tumour response beyond 6-weeks in the BOND trial. Following this adjustment, mean overall survival for patients receiving cetuximab plus irinotecan is estimated to be 10.76 months (undiscounted). Without this adjustment, the model estimates that the mean survival duration of these patients is 11.01 months (undiscounted).”</p> <p>Based on plots of estimated OS compared to empirical OS, it seemed that the extrapolation of OS undertaken by Merck was not reasonable. The adjustment for treatment continuation made minimal difference.</p> <p>“There is no comparative evidence to demonstrate either an improvement in HRQoL or overall survival duration in patients receiving cetuximab plus irinotecan compared to active/best supportive care or indeed any alternative chemotherapy except for cetuximab monotherapy (See Section 5.3). The expected survival duration of patients receiving active/best supportive care was modelled using data collected within the cetuximab monotherapy arm of the BOND trial. The duration of overall survival for those patients receiving cetuximab monotherapy whose outcomes were censored was estimated using the approach reported by Gelber and colleagues. Survival durations for patients receiving active/best supportive care were modelled using an assumption based upon an RCT of second-line irinotecan versus best supportive care reported by Cunningham and colleagues. Within this study, the hazard ratio describing the relative survival of patients receiving best supportive care as compared to those receiving irinotecan was reported to be 1.71. This hazard ratio was applied to the observed survival duration of patients receiving cetuximab monotherapy within the BOND trial. The model therefore assumes that the relative hazard of overall survival between cetuximab monotherapy and active/best supportive care as second- and subsequent-line treatment is exactly equivalent to the relative survival hazard between irinotecan and BSC as second-line treatment. This is a crucial assumption and a key determinant of the cost-effectiveness of cetuximab plus irinotecan which cannot be justified using existing empirical evidence.”</p> <p>No trials comparing cetuximab to current standard comparators were identified; only single arm trials and one trial comparing cetuximab monotherapy with cetuximab combination therapy were found, and therefore cetuximab did not form a main focus of the appraisal.</p>
Evidence synthesis (pool survival estimates?)	Some use of data from a variety of trials.	Mostly trial based, although some learnings taken from external data.
Survival model(s) fitted (Weibull, exponential etc)	Weibull.	<p>Weibull, but said to be a poor fit, so the manufacturer conducted sensitivity analysis using an exponential. However the AG states:</p> <p>“Further analysis was undertaken using an exponential distribution to estimate progression-free survival durations instead of the Weibull curve used in the base case analyses. However, as the exponential distribution is a special form of the Weibull distribution which is restricted to a single parameter and constant hazard rate, the justification for this particular sensitivity analysis is unclear”</p> <p>The exponential lowered the ICER as estimated by the manufacturer from £60k to £44k, showing what a difference this can make.</p>
Independent survival models, or hazard ratio (proportional hazards) modelling	Independent for the bevacizumab analysis, less clear for the cetuximab analysis.	<p>Unclear, but at least some use of indirect proportional hazards modelling in the cetuximab model. The AG note that:</p> <p>“The suitability of the adjustment in overall survival in the cetuximab monotherapy arm is highly dubious, as this assumes that the benefits conferred by cetuximab monotherapy and irinotecan are</p>

		exactly equivalent.” Little detail.
Justification for survival model used?	For bevacizumab the key difference between the assessment group model and the manufacturer’s model was that the assessment group based their model around overall survival data from studies AVF2107 and AVF2192 rather than PFS. The assessment group state that post-hoc analyses from AVF2107 shows that the survival of patients beyond disease progression was unaffected by the use of bevacizumab in subsequent lines of therapy, which rationalises their modelling approach. The assessment group noted that their Weibull models seemed to fit well visually.	
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Bevacizumab Survival model Cetuximab Survival model using threshold analysis – eg how much more survival is required to be cost-effective. OS was estimated for cetuximab, and then different values were tested for BSC to see what was required for cetuximab to appear CE.	Roche The Roche models use a simple state transition approach based on three health states using a monthly cycle length: (1) Pre-progression (alive and without disease progression) (2) Post-progression (alive following disease progression) (3) Dead Merck Survival based model
Other issues noted (eg crossover)	<p>Treatment crossover was an issue that Roche attempted to adjust for by assuming the same probability of death for both model arms once a patient has reached the disease progression health state. However the AG note that evidence from the trial actually showed that treatment crossover did not have a significant effect on survival in this instance, and so use OS data from the trial, unadjusted, for their analysis.</p> <p>In one of the main bevacizumab clinical trials (AVF2107) approximately 55% of patients received some form of 2nd line therapy. 25% received oxaliplatin, 10% irinotecan, 23% capecitabine, while less than 2% went on to undergo metastasectomy. Approximately 33% of patients in the bevacizumab arms continued to receive bevacizumab as 2nd line therapy.</p> <p>In the other main bevacizumab clinical trial (AVF2192) 53% of patients received 2nd line treatment. 42% were treated with oxaliplatin, irinotecan or both agents.</p> <p>In both bevacizumab trials patients randomised to receive bevacizumab as a 1st line treatment were allowed to receive bevacizumab as a subsequent-line therapy upon disease progression. This is outside of the current license.</p> <p>The Merck submission considered a treatment discontinuation rule. Because this rule was not followed in the clinical trial, survival estimates had to be adjusted to incorporate it. Based on plots of estimated survival in both the cetuximab and BSC arm, the AG suspected that the OS extrapolation was not reasonable, as OS was estimated to be much higher after 9 months than had been observed in the trial. However, incremental estimates may not have been too biased, as the overestimate impacted upon both cetuximab and BSC. The adjustment for treatment continuation made minimal difference.</p> <p>The Roche model did not make explicit assumptions concerning cytotoxic therapies received following disease progression on first-line bevacizumab-containing therapy, and instead assumed a mean cost of £2,000 per month following disease progression on bevacizumab-containing therapy. The AG assume that “For the purpose of transparency, the Assessment Group models assume that patients would receive oxaliplatin in combination with 5-FU/FA following progression on first-line therapy; this is consistent with UK marketing authorisation, and current guidance issued by NICE on the use of chemotherapy for advanced CRC. The Assessment Group models also assume that a small proportion of patients subsequently receive third-line treatment with Mitomycin C and protracted 5-FU” Implicitly, the AG use a sequencing model to cope with any sequencing/crossover issues.</p>	

28. TA119: Leukaemia (lymphocytic) - fludarabine, February 2007

Guidance: No recommendations have been made with respect to fludarabine plus cyclophosphamide combination therapy because the current marketing authorisation does not specifically provide a recommendation that fludarabine should be used concurrently with other drugs for the treatment of chronic lymphocytic leukaemia.

Fludarabine monotherapy, within its licensed indication, is not recommended for the first-line treatment of chronic lymphocytic leukaemia.

Source: Fludarabine monotherapy for the first-line treatment of chronic lymphocytic leukaemia, TA119, February 2007, <http://www.nice.org.uk/nicemedia/live/11614/33943/33943.pdf>, accessed 12/04/10

Walker S, Palmer S, Erhorn S, Brent S, Dyker A, Ferrie L, Horsley W, Macfarlane K, White S, Thomas S. Fludarabine phosphate for the first-line treatment of chronic lymphocytic leukaemia, Centre for Health Economics, University of York, and NHS Northern and Yorkshire Regional Drug and Therapeutics Centre, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, October 2006, <http://www.nice.org.uk/nicemedia/live/11613/33937/33937.pdf>, accessed 12/04/10

Note: The manufacturer sought a recommendation for fludarabine combined with cyclophosphamide, however this combination was not licensed and therefore could not be recommended. STA – manufacturer’s submission available on website

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
<p>Survival data used (patient-level data, or summary statistics – mean, median etc)</p>	<p>“A central criticism of the manufacturer’s submission by the ERG relates to the approaches used to estimate relevant transition probabilities based on the individual patient data from the CLL4 trial. A number of specific issues were identified by the ERG; however, due to the current structure of the electronic model, it was not considered possible to explore the robustness of the manufacturer’s results to alternative approaches to estimating the transition probabilities. However, a key assumption within the manufacturer’s submission is that the transition probabilities remain constant over the time horizon of the model. This implies a constant hazard (i.e. following an exponential distribution), which is a strong assumption and could be influencing the cost-effectiveness results. The ERG noted the divergence between the cost-effectiveness results presented for alternative time horizons and outlined that the extrapolation approach is likely to be a key determinant of the final cost-effectiveness estimates. In order to assess the validity of the assumption of constant hazards, the ERG has undertaken additional analysis on the individual patient data contained in the manufacturer’s submission using survival analytic techniques. This analysis was undertaken to formally test the underlying assumption of a constant hazard and to deal with censored data.”</p> <p>“The ERG has separated the data into two groups: responders and non-responders. Two separate transitions were considered representing the transitions to progression and death following first-line treatment. A separate Weibull distribution has been fitted to both groups, and to each separate transition, to test the assumption of a constant hazard assumed in the submission”</p> <p>“When $\gamma=1$ the Weibull distribution reduces to the exponential distribution and hence would indicate that the assumption of constant transition probabilities would be appropriate. Therefore, the ERG has tested the value of the parameter γ for the two groups using survival analysis undertaken using the statistical package STATA.”</p> <p>“The t test on $\ln(\gamma)$ is the test for the exponential distribution (i.e. a null hypothesis of $\gamma=1$). Clearly here we have to reject the null hypothesis exponential model at 0.0000 significance (as $\ln(\gamma)$ does not equal zero and hence γ does not equal 1) and, therefore, we can reject the assumption of a constant hazard. As γ is greater than one, it implies that the hazard of progression increases with time.”</p> <p>The ERG did this for both PFS and time from progression to death, for responders and non-responders. They found that the exponential could be rejected for all these at less than 5% levels, apart from for the time from progression until death for non-responders, which could be rejected at the 10% level.</p> <p>The ERG conclude:</p> <p>“The results of the survival analysis undertaken by the ERG suggest that the assumption of constant transition probabilities do not appear to be appropriate based on the individual patient-level data from the CLL4 trial. Since the assumption of constant transition probabilities underpins the manufacturers approach to</p>	<p>“The model uses patient-level clinical data from CLL4 to model first-line treatment including the response rates applied in the model... In summary, this study included 783 patients with an age range from 35 to 86 years. A total of 194 patients (24.8%) were treated with F, 387 (49.4%) with Chl, and 196 (25%) with FC, leaving 6 (0.77%) patients untreated. Data for second-line and subsequent treatment rates have been taken from a variety of published sources”</p> <p>The model was based primarily on response rates and PFS times – due to a lack of long-term data it was assumed that OS was the same for all treatments. Originally median OS was used, but upon questioning from the AG, the manufacturer submitted the analysis based on mean OS. The model uses the individual patient data from CLL4 and puts this directly into the model until a patient is censored. An individual is censored when they enter second-line treatment, are lost to follow up or the follow up ended while still in the response to therapy or progression states. Once a patient becomes censored the model uses transition probabilities based on non-censored patients from the CLL4 trial, and on other studies (Table 32, p92), to estimate progression through the model. Since no previous data was reported to be available on the re-treatment response rate for patients treated with FC, the manufacturers have assumed a response rate equal to that observed at first-line and this appears a very strong assumption (given that other data used for re-treatment with Chl and F were lower than the estimates applied for first-line treatment). Furthermore, due to a lack of external data on the duration of response of re-treatment with the initial therapy, the model assumes that this duration is equal to the initial duration of response.</p> <p>The manufacturers assumed that overall survival was the same for all groups. This is implemented in the model by making the time from first progression to death shorter in patients who had received F or FC than it was for those who received Chl as first-line therapy. Consequently, the model assumes that any gain in median progression-free survival associated with F or FC was offset by an equivalent decrease in the median survival after final progression.”</p> <p>“The model uses data from the CLL4 study on initial response rates, duration of response and time between progression and re-treatment to estimate the relevant transition probabilities following first-line treatment. This assumes that the data from the CLL4 study is able to fully reflect the transitions of the patients that the decision problem is focused on. Furthermore, the model also assumes that transition probabilities between all states are constant over time. Given the aging of the cohort and the 20 year time horizon this assumption may not hold. Although the submission has varied the absolute transition probabilities in the sensitivity analysis it has not considered varying the transition probabilities over time and it is not clear what effect this would have on the results.”</p> <p>“The current approach to extrapolation assumes that the risk of particular transitions follow an exponential distribution (i.e. that the hazard is constant with respect to time). No supporting evidence is provided to justify this assumption. However, since the manufacturers had access to the patient-level data then statistical approaches using survival analysis could have been used to formally test this assumption. The disparity between the cost-effectiveness results at 5-years and at longer periods suggests that the main cost-effectiveness advantage is conferred in the period of extrapolation. Hence the approaches to extrapolation are likely to be central to the validity of the subsequent estimates of cost-effectiveness.”</p> <p>In reviewing the manufacturer’s transition calculations the ERG identified a number of potential</p>

	<p>extrapolation, then the subsequent findings must be interpreted accordingly. Given that FC has the highest initial response rate (and hence a larger proportion of this group continues to follow the transitions based on the CLL4 trial data), it is clear that the impact of assuming constant transition probabilities may be acting as a possible source of bias towards this group.”</p>	<p>sources of concern.</p> <ul style="list-style-type: none"> • The transition probabilities have been estimated by simply calculating the total number of exits of the state (not including the count of the number of patients censored) divided by the total amount of time spent in the state (including the time that those who were censored spent in the state). However, as the time spent in the state already includes the time experienced by censored patients, subsequent transition probabilities for this group need to be estimated conditional upon being censored. Since the appropriate conditional probabilities were not estimated, the subsequent transition probabilities for the censored/unobserved group are potentially underestimated. This would not necessarily cause bias in favour of any one treatment if the error impacted on all three treatments equally but since patients treated with F or FC remain in response to their first-line treatment longer than ChI there is potential for bias in favour of these treatments. • Another key concern exists with the approach taken to estimating transition probabilities from the CLL4 data, such that the method used to estimate the particular risks appear flawed. Two potential problems were identified by the ERG. Firstly, the transition probabilities themselves are based on time at risk for all possible transitions, rather than estimating time at risk for each possible transition separately. Secondly the probabilities of three transitions have been summed to estimate a single probability estimate for leaving the response state. In the manufacturer’s model this single probability is assumed to represent the probability of moving to the “progression” state but in fact this transition also includes the probabilities of dying and of patients moving to second-line treatment. This is in part due to the structural assumptions of the model which only allows patients to die from CLL after they have progressed through all possible lines of treatment. It would have been more appropriate to model the possible transitions from this state separately (i.e. to progression, to second-line treatment and to death) which would also allow the time patients are at risk for each possible transition to have been incorporated.” <p>“The ERG has a number of serious concerns about the approaches used in the manufacturer’s submission for the estimation of transition probabilities. Some of these relate to the structural assumptions in their model (e.g. patients can only make a transition from response to therapy to progression and not to any other states) and some are due to the statistical approaches used (e.g. not making the transition probabilities conditional upon being censored and not adjusting estimates for particular transitions to allow for time at risk of each particular event). To adequately address these concerns formal survival analysis methods could be employed (which would allow for the censoring issue to be dealt with correctly). However, then to incorporate the results into the model would require a major restructuring of the current model and hence this is beyond the scope of the ERG report.</p> <p>Formal survival analysis wasn’t used by the manufacturer.</p> <p>“The model also assumes that the duration of response for re-treatment with the same agent is equal to the duration of response to the initial treatment. This assumption has again been made as no further evidence was found during the manufacturer’s search... this assumption could be important as it is effectively double counts any initial treatment benefit, yet no evidence is provided to support such a claim and no attempt has been made to address the uncertainty surrounding this assumption.</p> <p>Transition probabilities for second and subsequent lines of treatment have been estimated from published median values (see Table 40, p118). They have assumed a constant relative risk of transition out of a state, and have used a standard exponential approach to calculate the cyclical transition probability.</p> <p>Overall Survival: “In the base-case of the manufacturer’s model it is assumed that overall survival is the same for all treatments. The manufacturer’s submission argues that, due to the limited follow-up data available, current CLL4-trial data are not mature enough to be able to demonstrate any mortality benefits with</p>
--	---	--

		<p>individual treatments. Consequently, the manufacturers appear to have taken a conservative approach of equalising survival across all treatments. They attempted to achieve this by “assuming that any gain in median progression-free survival associated with fludarabine or fludarabine with cyclophosphamide was offset by an equal decrease in median survival after final progression” (p94). The ERG identified a number of potential concerns regarding this assumption:</p> <ol style="list-style-type: none"> 1. This “conservative” assumption focuses on equalising median survival rather than mean survival. Within the current modelling approach, differences will still exist between treatments in terms of mean survival estimates (with estimates of mean survival highest for treatment with FC). As a result, any differences in mean survival will be reflected in the subsequent estimates of the ICER. (NB: Following correspondence with the manufacturer an addendum was submitted which presented the results of the analysis based on equalising mean survival). 2. There are also concerns that, by forcing people in the other treatment arms to spend longer in the final progression period (which has a low utility but still incurs a cost), the model may be potentially biasing the results in favour of the intervention with the longest original progression-free survival time. 3. The data from the CLL4 trial results actually show a higher mortality (although not statistically significant) in the F and FC treatment arms. Hence an analysis based on extrapolation of the CLL4 trial data itself could have altered the current cost-effectiveness estimates. Until more mature survival data are reported from the CLL4 trial it is unclear whether the current approach is actually conservative or not.” <p>Indeed, the ERG found that when they set utilities to 1.00 in the model the total number of life years gained were not the same – the discrepancy between using medians for OS biased the model in favour of Fludarabine - to equalise mean overall survival it is the mean gain in progression-free survival that needs to be offset by an equal decrease in mean survival after final progression</p>
Evidence synthesis (pool survival estimates?)	None.	<p>The ERG has several concerns about the approach used to implement data from other studies. In particular, the methods used to synthesise data from several sources for one parameter are a cause for concern. The submission has simply pooled the data from several sources for the response rates for second-line treatment with FC after F as first-line treatment and for F after Chl as first-line treatment. By taking absolute estimates from the studies, the benefits of randomisation are lost and the differences observed may simply be due to the different patient characteristics from the different studies. The ERG is, therefore, concerned that this approach could potentially affect the absolute cost-effectiveness estimates for both the fludarabine and chlorambucil first-line treatments and this may then impact upon the relative cost-effectiveness estimates for all treatments.”</p>
Survival model(s) fitted (Weibull, exponential etc)	Tested the fit of Weibull models.	<p>For a number of transitions formal survival analysis was not used, instead the manufacturer attempted to use empirical rates. For some transitions an exponential distribution was used.</p>
Independent survival models, or hazard ratio (proportional hazards) modelling	Fitted independently to response groups.	<p>In general, neither, since formal survival analysis was not undertaken. The ERG had some concern about the manufacturer not using relative treatment effects for indirect comparisons.</p>
Justification for survival model used?	Tested the suitability of the exponential assumption made by the manufacturer – it was not suitable.	<p>Not systematic.</p>
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Did not produce a novel model.	<p>The manufacturer’s submission is based on a ‘<i>de-novo</i>’ decision analytic Markov model to estimate the cost-effectiveness of treatment with (i) fludarabine monotherapy (F), (ii) fludarabine in combination with cyclophosphamide (FC) and (iii) chlorambucil (Chl). Cycle length is 28 days, and the model timeframe is 20 years.</p> <p>“There are 16 health states in the model, which can be separated into 5 different treatment states,</p>

		<p>5 treatment response states, 5 disease progression states and a death state (representing both CLL and non CLL mortality). Patients enter the model on initiation of first-line treatment and remain in the initial state for the period of time for which their first-line treatment continues. Patients are then divided between those who have a response of 12 months or more ("responders") and those who do not ("non-responders"), where a responder is a patient who has a response of 12 or more months. In subsequent cycles of the model, responders to first-line treatment remain in the "response" state or they experience disease progression and move into a "progression" state for a period of time before receiving their second-line chemotherapy. In accordance with CLL4 protocol, responders to first-line treatment are assumed to be re-treated with the same agent as first-line when their disease progresses. These patients then remain in the "re-treat" state while their treatment continues at which time they move to either the "response" state or directly to the "progression" state. Those patients that achieve a response to re-treatment will remain in the "response" state until they move into the "progression" state. Following a period in the "progression" state these patients then move into the "salvage" state where third-line therapy is initiated. Patients remain in this state for a number of cycles before moving into either the "progression" or "response" states as in a similar manner to that assumed for second-line treatment. Patients who respond to salvage therapy remain in the response state until they ultimately move to the final "progression" state. Once patients enter the "progression" state following third-line therapy they are assumed to be at a constant risk of death from CLL. Patients who do not achieve an initial 12 month duration of response to first-line treatment (non-responders) follow a similar path to the "responders" but the second-line treatment is not a repeat of the first-line therapy given (details of which second-line treatment is given are detailed in Table 32, p92). While mortality from CLL is only allowed once patients have progressed through the complete sequence of treatments (first-line, second-line/re-treatment and salvage), patients are allowed to make a transition to death due to non-CLL mortality from any state in the model."</p>
<p>Other issues noted (eg crossover)</p>	<p>The AG note: "The choice of second-line treatment has been modelled in a very rigid manner which may not reflect the variation in the use of second-line treatments in routine clinical practice. Indeed, patients on the CLL4 trial requiring second-line treatment were actually randomised to either treatment guided by the results of the DiSC assay or to treatment guided by protocol guidelines, which could result in any of the other 2 treatments or CHOP being used. As such it is clear that there is a number of alternative second-line treatment strategies that the manufacturers could have considered. Additional sensitivity analyses have been undertaken to examine two alternative sequences. An additional analysis ("FCR") considered the impact of assuming that patients who do not respond to FC at first-line receive a second-line of chemotherapy with fludarabine, chlorambucil and rituximab (FCR) before proceeding to salvage treatment. A second analysis ("C-FC") considered the use of FC instead of F as second-line therapy after patients fail to respond to chlorambucil monotherapy."</p> <p>Given the relatively long-term nature of the disease pathway, the modelled treatment sequences were important – the ERG felt that some alternatives (including possibly the optimal alternative) may have been left out. Crossover is an issue in so much as overall survival was not known and many different treatment sequences could be relevant. Evidence synthesis for future lines of therapy is also a key issue.</p>	

29. TA121: Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide, June 2007

Guidance: Temozolomide and carmustine implants have been appraised separately for the treatment of newly diagnosed high-grade glioma. On the basis of the evidence presented to the Committee, no recommendation can be made regarding the sequential use of these treatments for newly diagnosed high-grade glioma.

Temozolomide, within its licensed indications, is recommended as an option for the treatment of newly diagnosed glioblastoma multiforme (GBM) in patients with a World Health Organization (WHO) performance status of 0 or 1.

Carmustine implants, within their licensed indications, are recommended as an option for the treatment of newly diagnosed high-grade glioma only for patients in whom 90% or more of the tumour has been resected.

Treatment with carmustine implants should be provided only within specialist centres that in general conform to guidance in 'Improving outcomes for people with brain and other central nervous system tumours' (NICE cancer service guidance 2006; www.nice.org.uk/csgbraincns), and should be supervised by specialist neurosurgeons who spend at least 50% of their clinical programmed activities in neuro-oncological surgery. The specialists should also have access to:

- multidisciplinary teams to enable preoperative identification of patients in whom maximal resection is likely to be achievable
- magnetic resonance imaging (MRI) to enable preoperative identification of patients in whom maximal resection is likely to be possible, and
- image-directed technology, such as neuronavigation, for use intraoperatively to assist the achievement of maximal resection.

Carmustine implants are not recommended for the treatment of newly diagnosed high-grade glioma for patients in whom less than 90% of the tumour has been resected.

Source: Carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma, TA121, June 2007, <http://www.nice.org.uk/nicemedia/live/11620/34049/34049.pdf>, accessed 19/04/10

Garside R, Pitt M, Anderson R, Rogers G, Dyer M, Mealing S, Somerville M, Price A, Stein K. The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high grade glioma: A systematic review and economic evaluation, PenTAG, University of Southampton, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, September 2005, <http://www.nice.org.uk/nicemedia/live/11619/34040/34040.pdf>, accessed 19/04/10

PenTAG, Additional analysis submitted by assessment group in response to consultee and commentator comments, December 2005, <http://www.nice.org.uk/nicemedia/live/11619/34041/34041.pdf>, accessed 19/04/10

NICE, Final Appraisal Determination, Carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma, TA121, March 2007, <http://www.nice.org.uk/nicemedia/live/11619/34030/34030.pdf>, accessed 19/04/10

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The aim of the assessment group’s economic review was to separately assess the cost-utility of BCNU-W and TMZ as chemotherapy additions to radiotherapy (RT) and surgery for newly diagnosed patients with high grade gliomas who are suitable for surgery. No relevant full economic evaluations were found.</p> <p>In the AG model the same trial data as used by the manufacturers was used.</p> <p>Survival data was read from the published Kaplan-Meier survival curves of the largest RCTs for each chemotherapy treatment identified in the systematic review... The TMZ trial used also contains a Kaplan-Meier curve for progression-free survival”</p> <p>“To obtain cumulative survival probabilities for individual time intervals it was necessary to extract points from the curves manually. The transition probability at any point in time in the multi-state model is equivalent to the standard hazard rate function for a survival time distribution. If patient-level data were available, the relevant hazard function could be derived from the curve using a proportional hazard model. Since patient-level data is not available, it was necessary to approximate the Kaplan-Meier curve using a known distribution, in this case a Weibull distribution, which is both versatile and simple to implement. An approximate hazard function for the curve can then be derived. Transition probabilities can then be calculated using standard techniques.” – transition probabilities were time-dependent – formula given in the appendix to the report.</p> <p>“As no progression-free survival curve is presented in the BCNU-W trial, we have had to assume that progression from the “stable” to the “progressive” disease states in this model is a fixed variable, based on a constant rate. By contrast, the progression-free survival curve from the TMZ trial allow this to be modeled as a time dependent variable in the same way as overall survival.” – thus the BCNU-W part of the model includes some median data.</p>	<p>Link Pharma – BCNU-W Used data from the BCNU-W RCT by Westphal and colleagues. Note that median time to progression data was used.</p> <p>Schering-Plough – TMZ This was a trial-based study, using patient-specific cost and effectiveness data from the RCT, Stupp and colleagues 2005. The only use of modelling was for the statistical extrapolation of survival beyond 2 years. The AG state that the “submission has not fully described how survival was extrapolated beyond 2 years.” Means are used, and any additional analysis using the restricted means approach is presented for a 2-year analysis.</p>
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	<p>“Weibull curves were fitted to the overall survival curves from the main RCTs for BCNU-W and TMZ and to the progression-free survival curve in the TMZ trial. The quality of fit of the Weibull curves to the trial data was judged on two criteria:</p> <p>1. R2 should be as close to one as possible</p>	<p>Schering-Plough – TMZ It appears a generalised gamma was used for one modelled population, and a Weibull for another. However, due to CiC, and incomplete reporting by the manufacturer, not many details are available on the models.</p>

	<p>2. Median survival time predicted should match the trial data as closely as possible. The Weibull distribution is manipulated by adjusting the two defining curve parameters; the scale parameter (λ) and the shape parameter (γ). Best fit was used rather than constraining the fit to the trial medians. R2 values were very high for all curves (between 0.9852 and 0.9977), and median survival for the fitted curves was within 6% of the trial data for all curves (between 0.09% and 5.75%) Further details, and examples of the fitted curves are given in Appendix 12. Curves were only fitted to the first two years of data in order to help eliminate tail effects for survival curves.”</p> <p>Thus the assessment group only used data up to 2 years when fitting models.</p> <p>Little data is given on this, but the FAD states that the AGs method of fitting a model to PFS for TMZ was improved by fitting one Weibull for the first 12 months and a different Weibull for the second 12 months.</p> <p>For PFS in the BCNU-W analysis an exponential distribution was used.</p>	
Independent survival models, or hazard ratio (proportional hazards) modelling	Unclear, but seems likely to be independent.	Unclear.
Justification for survival model used?	Suggested lack of options due to not having patient-level data. Exponential necessary in some cases due to lack of PFS curve and thus reliance on a median.	The AG state: “the industry submission does not state the statistical parameters of the resultant fitted distributions or give an indication of how good a fit there was between the distributions and the trial data... Nevertheless, it is possible to fit a variety of Weibull curves to the 2-year survival data, each of which has an excellent fit (R2 all > 97%) and yet which also generate vastly different mean survival estimates. Also, given the uncertainty that generally surrounds the ‘tail’ of survival curves where, typically, small numbers are at risk, statisticians strongly warn about “overinterpretation of the right-hand part of the survival curve” (see Altman, 1991 p.386213). The fitting of standard distributions is one example of how such over-interpretation can occur. It is clear that, in the absence of a larger trial which follows up high grade glioma patients for three or four years, the estimates of mean extrapolated survival should be subject to extensive sensitivity analysis. This is not undertaken in the TMZ industry submission.”
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Markov model, with the following states: Surgery; Postoperative recovery; Radiotherapy; Stable disease; Progression; Death. Cycle length was one week and the duration was 5 years, after which almost all the cohort had died.	<p>Link Pharma – BCNU-W</p> <p>Model splits post surgery survival into two time periods – pre and post progression. The model is purely time-based, eg time spent pre progression + time spent post progression, each multiplied by utility scores. The assessment group state that whether the intervention can be classed as having a PFS advantage depends on which measure is used to assess progression, and whether there is an OS advantage depends on if the trial data is stratified by country. The AG believe that the manufacturer’s analysis was not ITT and was biased. Also, the manufacturer used median times until progression (although means were used for OS), which the assessment group states is incorrect for economic analyses.</p> <p>Schering-Plough – TMZ</p> <p>This was an EEACT. The only use of modelling was for the statistical extrapolation of survival beyond 2 years.</p>
Other issues noted (eg crossover)	<p>Treatment crossover / post progression treatments do seem to be an issue. The AG states: “the analyses presented in the industry submission in effect compare the costs and effects of a sequence of treatments given both initially and following tumour recurrence. Because of this it is highly plausible that both the incremental costs and incremental survival are partly driven by differing treatment choices during disease progression, rather than the choice of treatment when the gliomas were newly diagnosed. An alternative analysis of the cost-effectiveness of TMZ for newly diagnosed high grade glioma could assume that the effectiveness of treatments for newly diagnosed glioma is restricted to extending progression-free survival. Indeed there is no good evidence that TMZ, or any other chemotherapy treatment delivered as first-line therapy for newly diagnosed tumours, offers any benefit in slowing the rate of disease progression after recurrence.”</p> <p>And:</p>	

	<p>“No attempt has been made, either in relation to estimated costs or survival, to adjust for the fact that patients in the RT only (control) arm of the trial received substantially higher levels of salvage chemotherapy (especially TMZ).”</p> <p>The AG therefore conducted a reanalysis of the Schering model assuming only a PFS benefit.</p> <p>In the AG model, “the model does not contain a health state to allow for patients receiving subsequent surgery or chemotherapy after disease progression, this option is taken into account when evaluating the costs associated with the “progressive” state. In addition, as the transitions used are based on trial data, where a proportion of patients received second line therapy, the curves already incorporate any survival influence such treatment may cause.”</p> <p>So a sequencing approach is implicitly taken – the relevance of this depends on it the sequences are ‘realistic’. There were problems with this, with the AG having to perform additional analysis. Discussion centred around whether to cost treatments observed in the study, or treatments that would be more reasonable (eg assume same %s go on to receive further treatment), or whether to only include survival benefits up to PFS.</p> <p>One option for the analysis was said to be: “Keep the cost of 2nd line chemotherapy following radiotherapy-only as 70% getting PCV (i.e. current standard treatment), but also alter the effectiveness of the radiotherapy-only arm (because the Stupp <i>et al</i> survival estimate for control patients partly derives from 43% (72% x 60%) in the control arm getting TMZ on recurrence).” But this was deemed not to be feasible because: “we are unable to distinguish how much of the overall survival in the either arm of the Stupp trial is attributable to 1st line treatment and how much to different treatments at recurrence”</p> <p>Thus treatment crossover and methods for dealing with it are a definite issue in this appraisal.</p>
--	--

30. TA124: Lung cancer (non-small-cell) - pemetrexed, August 2007

Guidance: Pemetrexed is not recommended for the treatment of locally advanced or metastatic non-small-cell lung cancer.

People currently receiving pemetrexed should have the option to continue therapy until they and their clinicians consider it appropriate to stop.

Source: Pemetrexed for the treatment of non-small-cell lung cancer, TA124, August 2007, <http://www.nice.org.uk/nicemedia/live/11823/36170/36170.pdf>, accessed 20/04/10

Bagust A, Boland A, Dunder Y, Davis H, Dickson R, Green J, Hockenull J, Macbeth F, McLeod C, Proudlove C, Tudor Smith C, Stevenson J, Walley T. Pemetrexed for the treatment of relapsed non-small-cell lung cancer: ERG Report, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, September 2006 <http://www.nice.org.uk/nicemedia/live/11715/35188/35188.pdf>, accessed 20/04/10

Eli Lilly and Company Ltd, Clarification Response from Manufacturer, August 2006, <http://www.nice.org.uk/nicemedia/live/11715/36147/36147.pdf>, accessed 20/04/10

Eli Lilly and Company Ltd, STA Submission to NICE, June 2006, <http://www.nice.org.uk/nicemedia/live/11715/36146/36146.pdf>, accessed 20/04/10

Note: STA: Manufacturer submission available on NICE website. There was an appeal by the manufacturer, but this was dismissed.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	No evaluations comparing pemetrexed with docetaxel were found. The assessment group reviewed the manufacturer’s analysis.	<p>The economic model compares pemetrexed with docetaxel and includes data from nine randomised controlled trials including JME1 (pemetrexed versus docetaxel). An ICER of £16.8k is estimated. However, the ERG state: “However, a number of key assumptions and parameters in the model do not seem to be clinically and/or economically justified, particularly in terms of survival rates. For example, the company model assumes a survival benefit for pemetrexed compared to docetaxel. However, the ERG does not believe that this supposition is justified as it is based on flawed pooling methodology. When the more realistic assumption of equivalent survival is incorporated into the model, the ICER rises to £458,333 per QALY gained.”</p> <p>Clearly the survival estimates have a huge impact on the ICER. However, the ERG also state that when other inputs within the model are adjusted (eg drug acquisition costs) the ICER increases to £1.2 million.</p>

		<p>The ERG states: "Furthermore, the model does not allow for patients to die of anything other than cancer, or treatment related causes, which is an unrealistic assumption. In addition, patients cannot die in the first cycle of treatment which artificially inflates the survival benefit in both arms."</p> <p>Median values were used for OS, after pooling (inaccurately, as the comparisons are naïve and don't retain the randomised nature of the studies). It is not stated whether the pooled time to progression estimates are means or medians. These values were transformed into transition probabilities assuming a constant (exponential) probability per cycle.</p>
Evidence synthesis (pool survival estimates?)	<p>"In the company submission results were presented from an updated analysis of the one primary and six secondary outcomes from the JME1 trial. These confirm that there is no significant difference in the primary outcome (overall survival) between pemetrexed and docetaxel. A similar finding was also noted for five of the secondary outcomes:</p> <ul style="list-style-type: none"> - Progression-free survival - Time to progressive disease - Duration of tumour response - Duration of clinical benefit - Time to objective tumour response <p>Only one secondary measure (time to treatment failure) appears to show a small advantage for pemetrexed in median TTTF (2.3 vs. 2.1 months, $p = 0.046$). In terms of model states and events, this implies that there should be no differences in patient time spent in the three states (stable, response, progression) which govern the calculation of survival and state specific quality of life. The only possible difference implied by these results is that some docetaxel patients will discontinue active therapy earlier than those on pemetrexed, but with no impact on response, or the timing of progression or death. Thus if the small apparent difference in TTTF were to be allowed, its effect on the cost-effectiveness analysis would be solely that of reducing the mean number of treatment cycles (and therefore the cost) for docetaxel patients. However, by costing treatment in terms of the actual treatments given in the trial this effect is already accounted for.</p> <p>In the absence of differences in overall survival or time spent in health states, the only valid outcome differences are the utility effects of treatment-related adverse events. The overall utility gain claimed for pemetrexed over docetaxel has been re-estimated after applying a half-cycle correction (not used in the company model), and then disaggregated into components attributable to modelled survival gain, and treatment-related adverse events"</p> <p>When survival equivalence is assumed the utility benefit associated with pemetrexed falls from 0.07 to 0.003.</p>	<p>The ERG states: "The company attempted to estimate several efficacy inputs (overall survival (OS), time to disease progression (TTdP), and response rates) using clinical trial data. The company carried out both direct and indirect comparisons of clinical trial data (see Table 4-1). However, the only direct and reliable clinical evidence available, which is relevant to the reference case of this appraisal, is the JME1 trial of pemetrexed versus docetaxel."</p> <p>"The only direct and reliable evidence available which is relevant to the reference case of this appraisal is the JME1 trial of pemetrexed versus docetaxel. However, the discussion in section 3.2.4 highlights that the JME1 trial investigators failed to establish equivalence of effect or even non-inferiority with regard to overall survival of pemetrexed compared to docetaxel."</p>
Survival model(s) fitted (Weibull, exponential etc)	No models fitted.	None.
Independent survival models, or hazard ratio (proportional hazards) modelling	None.	None.
Justification for survival model used?	Investigated the justifications for the manufacturer's analysis, but no models were fitted.	None.

<p>Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)</p>	<p>The ERG critiqued the manufacturer's model: "However, we may instead choose to adopt a more pragmatic position: on the basis of the Kaplan-Meier analysis of overall survival for the JME1 trial we may accept that there is no realistic difference between the trial arms, and therefore base an analysis of costs and outcomes on the assumption of outcome equivalence. This then reduces the analysis to one which depends on a single trial, adjusted as appropriate to UK clinical practice and costs. Although this is an appealing option, it causes serious problems for the relevance and reliability of the submitted Markov model, which uses a series of intermediate states and differential transition rates to generate important gains in survival for pemetrexed relative to all alternative treatments. Clearly, if we accept that survival equivalence is itself a generous assumption, then the submitted model appears to have failed the primary validation test - to reproduce the single most important clinical outcome. The ERG have therefore concluded that it is unlikely that the submitted model, even with minor modifications and parameter changes, could be used as the basis for generating useable cost-effectiveness evidence. Indeed, a quite different model structure would be required to constrain survival to true equivalence between treatments, and this is beyond the scope of the ERG in preparing this assessment report."</p>	<p>Markov model with 4 main states: Response, Stable, Progressive disease, Death. Cycle length is 21 days and the model time frame is 3 years (the maximum life expectancy for the population).</p>
<p>Other issues noted (eg crossover)</p>	<p>Patients randomised to pemetrexed were allowed docetaxel after progression, but patients randomised to docetaxel were not allowed pemetrexed. The ERG state that this could have caused bias in favour of pemetrexed, but seem satisfied that in this case such bias was not present: "A number of patients in the pemetrexed arm received docetaxel following disease progression, whereas pemetrexed was not offered to patients progressing in the docetaxel arm. This could lead to bias in results for overall survival, however careful examination of additional details on patient disposition provided by the company have satisfied the ERG that no significant bias is detectable in the submitted trial findings for overall survival."</p> <p>This is based on further analysis requested of the manufacturer, which seems fairly basic, and may not really get to the bottom of whether crossover was important:</p> <p>"Overall survival and other efficacy endpoints were analyzed without consideration of potential effects of post-study-treatment anti-cancer therapy. However, an additional exploratory analyses were undertaken by Lilly to examine whether there was any evidence of one treatment arm or the other receiving a differential benefit due to post-study-treatment anti-cancer therapy. Our conclusion was that while it is possible that post-study-treatment therapies may have provided additional benefits for those patients who received them, there was no evidence of a differential advantage from these therapies gained by one study arm relative to the other.</p> <p>When considering the potential effect on overall survival of additional lines of therapies, the first consideration is whether there is a different outcome for those secondary efficacy measures that apply only to the study treatment period. Results for best tumour response and progression-free survival show very similar results to overall survival (in terms of the numerical comparisons between pemetrexed and docetaxel), which does not suggest any evidence of a differential benefit due to additional lines of therapy. If, for example, patients on the pemetrexed arm received additional benefit from cross-over to docetaxel, we would expect to see relatively better survival (at least numerically) on the pemetrexed arm --- some degree of improvement over the progression-free survival hazard ratio. Since we do not see any difference between the survival and progression-free survival hazard ratios, there is no evidence from these analyses of any systematic bias in the survival comparison due to additional lines of therapy.</p> <p>A second, exploratory analysis of post-progression survival was also conducted. With few exceptions, patients received additional therapies only after progressive disease. So if there was any differential survival benefit between arms due to the additional lines of therapy, we might expect to see some differences in the data for post-progression survival (Kaplan-Meier analysis of the time from progressive disease to the date of death, among all trial patients experiencing progressive disease). The analysis of post-progression survival in fact showed no evidence of such a differential effect."</p>	

31. TA129: Multiple myeloma - bortezomib, October 2007

Guidance: Bortezomib monotherapy is recommended as an option for the treatment of progressive multiple myeloma in people who are at first relapse having received one prior therapy and who have undergone, or are unsuitable for, bone marrow transplantation, under the following circumstances:

- the response to bortezomib is measured using serum M protein after a maximum of four cycles of treatment, and treatment is continued only in people who have a complete or partial response (that is, reduction in serum M protein of 50% or more or, where serum M protein is not measurable, an appropriate alternative biochemical measure of response) **and**
- the manufacturer rebates the full cost of bortezomib for people who, after a maximum of four cycles of treatment, have less than a partial response (as defined above).

People currently receiving bortezomib monotherapy who do not meet the criteria in paragraph 1.1 should have the option to continue therapy until they and their clinicians consider it appropriate to stop.

Source: Bortezomib monotherapy for relapsed multiple myeloma, TA129, October 2007, <http://www.nice.org.uk/nicemedia/live/11869/38001/38001.pdf>, accessed 20/04/10

Green C, Bryant J, Takeda A, Cooper K, Clegg A, Smith A, Stephens M. Evidence Review Group Report commissioned by the NHS R&D HTA Programme on behalf of NICE: Bortezomib for the treatment of multiple myeloma patients, Southampton Health Technology Assessments Centre, University of Southampton, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, April 2006 <http://www.nice.org.uk/nicemedia/live/11713/35150/35150.pdf>, accessed 20/04/10

NICE, Final Appraisal Determination, Bortezomib monotherapy for relapsed multiple myeloma, TA129, July 2007, <http://www.nice.org.uk/nicemedia/live/11713/37094/37094.pdf>, accessed 19/04/10

Janssen-Cilag Ltd, STA Submission to NICE: Velcade (Bortezomib) for the treatment of multiple myeloma patients at first relapse, July 2006, <http://www.nice.org.uk/nicemedia/live/11713/35151/35151.pdf>, accessed 20/04/10

Note: STA: Manufacturer submission available on NICE website (although appendices with further methods information are not available). There was an appeal in this appraisal and the appraisal committee was asked to assess the evidence again. Velcade was then recommended under a patient access scheme.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	The ERG states: "The model assumes that there is an overall survival hazard ratio of 0.42 (from APEX RCT) in the first year and 0.83 in years 2 and 3. The latter hazard ratio has not been justified. The manufacturer's submission assumes that there are independent benefits for TTP and OS. Given the workings of the model these benefits may not be independent, and it may be that the group of patients who have OS benefits will also have TTP benefits. Thus there may be some double counting for the effect of bortezomib. The submission states this not to be the case, but the ERG would like further clarification on this."	The model compares bortezomib with HDD. "The model uses (non-trial) observational data to predict the treatment experience of a cohort of patients treated with HDD. The patient group are defined as MM patients who have experienced a first relapse of MM treatment. Data from the RCT on HDD are deemed to be unavailable/inappropriate because of the early termination of the trial and subsequent inability to predict long-term outcomes and mortality data with HDD. The model uses data from the Mayo Observational Study (Kumar <i>et al</i> 2004) to model this baseline/comparator cohort. Data from the Mayo Study are from patients who have been treated with a range of different drugs, although few of these were treated with HDD. The submission states that 188 patients (32.5%) were treated at some point in their follow-up with VAD (combination of vincristine, adriamycin and dexamethasone), with 74 of these receiving VAD after their first relapse." "The clinical effectiveness data from the APEX RCT, showing a relative benefit in time to treatment progression (HR=0.56) and a relative benefit in overall survival (HR=0.42), are applied to the baseline prediction for HDD patients. The model adjusts the baseline transition rates (between health states) according to the hazard rates estimated in the APEX RCT (bortezomib vs. HDD). The model uses the comparative data to simulate the treatment effect from bortezomib, its impact on survival, and the subsequent cost-effectiveness of treatment. Treatment effect on adverse events is not included in the CEA. Bortezomib is assumed to have a treatment effect lasting for up to three years." Event Free Survival for each line of treatment appears to be based on median figures. This is not discussed by the ERG.
Evidence synthesis (pool survival estimates?)	No novel analyses.	Some use of external data.
Survival model(s) fitted (Weibull, exponential etc)	None.	None.
Independent survival models, or hazard ratio (proportional hazards) modelling	No novel analyses.	Proportional hazards modelling.
Justification for survival model used?	Not applicable.	The model was calibrated against observational and trial data to ensure reasonable OS was estimated.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	No novel model.	The CEA uses a decision-analytic model (quasi-Markov) to estimate the effect of treatment with bortezomib compared to HDD. "The model predicts treatment experience, rather than disease progression, modelling the flow of patients through a series of treatment regimens (from regimen 2 to 6). There are two general health states used in the model; these are 'on treatment regimen i', and 'death whilst on treatment

		<p>regimen i' ($i=1$ to 6). These health states apply to each of the potential treatment regimens (regimens 2 to 6), therefore there are 10 health states in which patients can arrive (5 regimens x 2 states). In any cycle patients may remain on that treatment regimen, progress to a new regimen or die on that regimen. The first and second cycle times are 3-months each, the third cycle time is 6 months, and all subsequent cycles are 1-year. There are 18 time cycles in the model, with all patients starting in the second regimen, giving the model a time horizon of 15-years (a lifetime horizon in this patient group)."</p>
<p>Other issues noted (eg crossover)</p>	<p>The model is based on treatment sequences rather than health states, and relies upon observational data for the baseline. Therefore there is likely to be some issues with different treatments being received by different groups. However, this forms the baseline, to which the treatment effect is applied. Thus it is important to know if there was any treatment crossover in the RCT from which the treatment effect was taken:</p> <p>"The model uses data from the study by Kumar <i>et al</i> (2004) (Mayo clinic data, Rochester USA), to predict the baseline disease progression for the comparator group, i.e. HDD treated patients. This assumes that the data from Kumar <i>et al</i> are able to reflect disease progression in the specified patients (controls). Whilst the Kumar <i>et al</i> study seems a good quality observational study, and there is an absence of alternative data sources available (see below), when applying this data in the context of the CEA presented there may be some areas of uncertainty. For example, the patients used from the Kumar <i>et al</i> study are (i) a subset of the Mayo patient data presented, (ii) this observational study reports data collected over a 13-year period (in a USA context), and patients may not have benefited from the latest treatment protocols, (iii) HDD was not one of the reported treatment regimens for the observational study, (iv) the observational data are not specific on which patients had what treatment and when, (v) that there are some differences in the APEX RCT and Mayo patient profiles, e.g. patients in APEX RCT are diagnosed approximately 5 years earlier than the Mayo patients. The Mayo study data show 355 persons receiving a 2nd regimen, the biggest group ($n=160$) getting combination alkylating agents, with 33 patients received VAD (where dexamethasone is expected to be the most active ingredient). In the dataset presented by Kumar <i>et al</i> 11 114 persons received VAD as 1st regimen. The suitability of this data is open to some judgement and interpretation."</p> <p>"The ERG suggest that the data used may predict a more severe disease progression/profile (e.g. time to progression may be shorter than expected, and mortality may be higher) than may be expected in a hypothetical cohort of patients treated with HDD in a RCT context (i.e. a direct comparison to the data applied for bortezomib from the APEX RCT). This suggestion is based on the issues discussed above. These issues could bias the estimates of treatment effect, given the model uses transit probabilities for the base case and adjusts these for treatment effect using hazard rates from the APEX RCT data. The submission does make an adjustment to the observational data to reflect the survival rate of HDD patients calculated from the APEX RCT (patients who had received only one prior therapy and then received HDD)."</p> <p>The manufacturer was asked to clarify issues around crossover in the APEX RCT. They do not fully answer the questions – they do not state if they accounted for crossover in their analysis, so it seems likely that they didn't. Information on the APEX trial suggested that crossover was allowed at 8.3 months, but other information suggests that at this point 44% of control patients had already crossed over.</p> <p>In addition, the manufacturer states: "The Independent Data Monitoring Committee terminated the APEX trial prematurely after 8.3 months follow-up, when the interim analysis showed superior efficacy benefit with VELCADE compared to HDD. Although ethically and clinically unavoidable, the early termination of the APEX trial and the subsequent cross-over of patients from the HDD arm to VELCADE treatment presents a number of methodological challenges in terms of quantifying the incremental health outcomes and associated costs for use in economic evaluation. Of primary importance for the economic model is the need to derive an accurate estimate of the expected lifetime survival gain with both VELCADE and HDD. The quantification of this survival gain directly from the APEX trial was not possible for two reasons: The first is that the early termination of the APEX trial resulted in considerable censoring which meant that direct observation of the long-term survival differences between VELCADE and HDD is not possible. The second is that patients within the HDD treatment arm were allowed to cross-over to receive VELCADE following early termination of the study. Within the APEX trial 60% of HDD patients crossed over to receive VELCADE. Therefore, it was necessary to identify sources of data from outside of the APEX trial to model the survival benefit of the comparator arm as well as using survival modelling techniques to estimate lifetime benefits and costs."</p> <p>This appears to be why the observational study was used to estimate long-term survival for HDD, and why PH modelling was used. HRs were taken from a study which published results of APEX up to the 8.3 month interim analysis – therefore beforemost crossover would have expected to have occurred. However, it seems likely that some crossover already had occurred, and it appears the analysis was not adjusted for this (this is confirmed by the manufacturer). Using the observational data was a method for adjusting long-term survival estimates for the control group and calibrating this with 1 year data from the RCT, avoiding the crossover problem. However, the observational data may not have been fully suitable for this because of differences in patient characteristics and treatment regimens. The Appraisal Committee demonstrated some concern over the suitability of the external data used, concurring with the ERG.</p>	

32. TA135: Mesothelioma - pemetrexed disodium, January 2008

Guidance: Pemetrexed is recommended as a treatment option for malignant pleural mesothelioma only in people who have a World Health Organization (WHO) performance status of 0 or 1, who are considered to have advanced disease and for whom surgical resection is considered inappropriate.

Patients currently receiving pemetrexed who do not fall into the patient population defined above should have the option to continue therapy until they and their clinicians consider it appropriate to stop.

Source: Pemetrexed for the treatment of malignant pleural mesothelioma, TA135, January 2008, <http://www.nice.org.uk/nicemedia/live/11909/38945/38945.pdf>, accessed 20/04/10

Dundar Y, Bagust A, Dickson R, Dodd S, Green J, Haycox A, Hill R, McLeod C, Walley T. Pemetrexed disodium for the treatment of malignant pleural mesothelioma: a systematic review and economic evaluation, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, December 2005, <http://www.nice.org.uk/nicemedia/live/11698/36841/36841.pdf>, accessed 20/04/10

Dundar Y, Bagust A, Dickson R, Dodd S, Green J, Haycox A, Hill R, McLeod C, Walley T. Pemetrexed disodium for the treatment of malignant pleural mesothelioma: a systematic review and economic evaluation: Addendum, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, March 2006, <http://www.nice.org.uk/nicemedia/live/11698/36840/36840.pdf>, accessed 20/04/10

Janssen-Cilag Ltd, STA Submission to NICE: Velcade (Bortezomib) for the treatment of multiple myeloma patients at first relapse, July 2006, <http://www.nice.org.uk/nicemedia/live/11713/35151/35151.pdf>, accessed 20/04/10

Note: STA: Only the executive summary of the manufacturer submission available on NICE website. There were two appeals. In the first, one of the appeal points from the manufacturer were that there had been a reliance on means rather than medians, even though means created uncertainty due to extrapolation. It was explained that means were required for economic evaluations and the appeal was dismissed on this point, but was upheld on others, which led to the topic being referred back to the AC. At the second appeal all grounds were dismissed.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The Assessment Group were denied access to the RCT patient-level data, despite repeated requests.</p> <p>In the model produced by the AG means are used, and it is stated that medians are not useful for economic evaluations, and that estimating means from medians is prone to error as the median tells us nothing about the tail of the distribution. Hence where data is incomplete it is better to fit a parametric model to estimate the mean.</p> <p>From the data supplied to the AG some approximate KM analyses were possible. It was found “that the K-M estimated means are systematically lower than the corresponding medians due to the truncation of the data required for estimation of the mean when not all patients have complete follow-up to death...Using the aggregated monthly data we estimated Weibull model parameters by Maximum Likelihood Estimation (MLE), and calculated the expected mean survival for each of the three populations (ITT, FS, FS/AD)... Comparison with the corresponding K-M results demonstrates:</p> <ul style="list-style-type: none"> • the extent to which K-M estimated means under represent true survival; and • the lack of precision of observed medians leading to unreliable estimates of survival gains between trial arms.” <p>“For the two remaining populations (FS/PS 0/1, FS/AD and PS 0/1) no aggregate data were provided, and so a different approach had to be adopted, based on the CSR K-M charts. This involved digitising the chart images as closely as possible, to provided approximations to the survival patterns in the trial. By calculating the total area under the curve (AUC) we obtained estimates which should correspond quite closely to the K-M mean estimates generated from the aggregate data for three populations.</p> <p>Comparing results in the first and second vertical segments of <i>Table 7A</i> indeed confirms this expectation. Establishing parameters for a Weibull model from the digitised K-M plots proved more problematic, since we had little information on which to judge how to weight the multiple observations underlying each point on a K-M plot. To address this problem we used point-wise standard errors from the</p>	<p>The Eli Lilly submission was split into two sections: “The submission was split into two sections each employing a separate economic model. The first model is based on trial data of pemetrexed plus cisplatin versus cisplatin. The second model was not based on any single trial but undertaken using an amalgamation of data from several published sources to estimate how pemetrexed plus cisplatin would compare with MVP, vinorelbine and active symptom control (ASC)”</p> <p>Model 1 The AG state “Life years gained were estimated using K-M survival curves of trial data and expressed in terms of both mean and median. However, only means will be considered in this discussion as medians are of limited economic importance.”</p> <p>Unclear how means were estimated, but clinical data relatively complete, seems likely to be AUC of KM.</p> <p>Regarding Model 2, the AG state: “There is a fundamental problem with the evidence provided to support outcome gains claimed in Model 2, which is highlighted by the following passage from the company submission: <i>“There have been few studies investigating the use of MVP, vinorelbine (+/- platinum) in MPM, however most are small, non-randomised phase II trials. There are no randomised controlled trials comparing chemotherapy to ASC. The patient population characteristics varied widely between studies that make comparison of agents problematic and hence inconclusive.”</i> Despite these limitations, the authors have assembled data apparently showing important survival gains for the pemetrexed plus cisplatin combination therapy, particularly in comparison to supportive care. Unfortunately the evidence base underpinning Model 2 is not credible since it is not founded upon direct or even indirect comparisons of RCTs, and there is no evidence to support comparability of the patient populations between the various studies quoted, nor with the EMPHACIS trial. The crucial issue is the extent of survival gain to be expected between pemetrexed plus cisplatin and the various comparators offered, and we have concluded that there is no objective basis on which to estimate such gains nor to assess the uncertainty associated with such estimates. Without these figures the Model 2 endeavour is fruitless, and therefore we have not pursued this approach any further.”</p> <p>In model 2, survival estimates were medians from a range of studies.</p>

	<p>approximate K-M analyses (i.e. from the first segment of <i>Table 7A</i>) and fitted polynomial functions of time to each population-arm so that we could obtain interpolated estimates of point standard errors for every point of the digitised K-M plot. This then facilitated the fitting of a Weibull survival model by weighted least squares, using the inverse of the standard error to weight each observation. In the case of the two populations without aggregate data, we used the FS polynomial functions to provide proxy weights. The results are shown in the final segment of <i>Table 7A</i> and graphically the fit between observational data and fitted models is shown in <i>Figures 7A-7E</i>. There is good correspondence between MLE estimates of mean expected survival, and those using weighted least squares and digitised data. It is also clear the extent to which projected mean survival estimates generally exceed those obtained by truncated observational data. A significant problem associated with the weighted least squares method is that it is not possible to estimate confidence ranges around the estimates directly. In the leftmost vertical section of <i>Table 7B</i> approximate confidence intervals have been derived by reference to the distribution of mean survival estimated by the MLE method.”</p> <p>Typically the extrapolated OS values were between 0.5-2 months longer than the KM based estimates, and thus the KM data was probably fairly complete.</p>	<p>Mean values for use in the cost-effectiveness analysis were derived by calculating a weighted average of reported medians and assuming the same mean to median ratio as that observed in the cisplatin only arm of the EMPHACIS trial.</p>
Evidence synthesis (pool survival estimates?)	<p>In an addendum to the AG report, the group drew upon SEER data to provide more information for their survival analysis. They stated that the long-term survival data showed that there was a small proportion of long-term survivors, showing that a model with a variable hazard, such as the Weibull (rather than the exponential), would be reasonable. No other possible models were suggested though. However the group conclude:</p> <p>“Consideration of the limited literature available from Europe and North America, does not offer a basis for estimating typical expected survival in MPM, nor for identifying an unambiguous set of prognostic indicators for better survival. Long-term time trends in survival may suggest some improvement in life expectancy at diagnosis, at least in men, but cannot rule out that this may be artefactual due to several confounding effects. However, the data do strongly suggest that despite the generally poor prospect, a small number of patients may survive for several years.”</p>	<p>Model 1 involved no synthesis; Model 2 used data from several studies.</p>
Survival model(s) fitted (Weibull, exponential etc)	<p>Weibull.</p>	<p>Not specified – likely to be AUC of KM for model 1, and no models fitted for model 2.</p>
Independent survival models, or hazard ratio (proportional hazards) modelling	<p>Independent models fitted to separate populations.</p>	<p>Not specified.</p>
Justification for survival model used?	<p>No others tried, but some justification for the approach was given: Exploratory analysis of suitable parametric survival models indicated that a constant hazard (exponential) model was inadequate to account for the observed data, but that a two-parameter Weibull model provided a robust fit to all patient populations.</p> <p>Some attempts were made to allow a weighting of the models fitted to the digitised data to take account of where the KM was informed by more data (see above).</p>	<p>No details.</p>
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>The Assessment Group developed a model comparing cisplatin and peremetrexed using a similar approach to the manufacturer, and the same RCT data. The group produced an equation based model, using mean survival as a key input.</p>	<p>“Model 1 is based on individual patient data (IPD) taken from the phase III trial of cisplatin versus peremetrexed plus cisplatin (only fully supplemented patients included) over a period of 29 months. The justification for cisplatin as a comparator is based on the assumption that cisplatin is likely to be at least as good as active symptom control (ASC), and at the time of trial design was considered the best available single agent, owing to no clear evidence of efficacy for either MVP or vinorelbine”</p>

Other issues noted (eg crossover)	There are no mentions of crossover.
-----------------------------------	-------------------------------------

33. TA137: Lymphoma (follicular non-Hodgkin's) - rituximab, February 2008

Guidance: Rituximab, within its marketing authorisation, in combination with chemotherapy, is recommended as an option for the induction of remission in people with relapsed stage III or IV follicular non-Hodgkin's lymphoma.

Rituximab monotherapy as maintenance therapy, within its marketing authorisation, is recommended as an option for the treatment of people with relapsed stage III or IV follicular non-Hodgkin's lymphoma in remission induced with chemotherapy with or without rituximab.

Rituximab monotherapy, within its marketing authorisation, is recommended as an option for the treatment of people with relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma, when all alternative treatment options have been exhausted (that is, if there is resistance to or intolerance of chemotherapy).

Source: Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma: Review of TA37, TA135, February 2008, <http://www.nice.org.uk/nicemedia/live/11923/39618/39618.pdf>, accessed 21/04/10

Bagust A, Boland A, Dickson R, Chu P, Hockenhull J, Davis H. Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma: ERG Report, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, August 2007, <http://www.nice.org.uk/nicemedia/live/11730/38878/38878.pdf>, accessed 21/04/10

Roche Products Ltd, Rituximab for the treatment of relapsed follicular lymphoma, STA Manufacturer Submission to NICE, June 2007. <http://www.nice.org.uk/nicemedia/live/11730/38897/38897.pdf>, accessed 22/04/10

Roche Products Ltd, Roche response to clarification letter, October 2007. <http://www.nice.org.uk/nicemedia/live/11730/38899/38899.pdf>, accessed 22/04/10

NICE, Final Appraisal Determination, Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma (review of technology appraisal guidance 37), TA137, January 2007, <http://www.nice.org.uk/nicemedia/live/11730/38825/38825.pdf>, accessed 22/04/10

Note: STA: The full manufacturer submission is available on NICE website.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The ERG discuss the survival analysis completed by the manufacturer at length, and re-run model 2 using new estimates. They state:</p> <p>“Survival analysis: The submitted models depend on the results of parametric survival modelling as the basis for estimating lifetime benefits from use of rituximab. The outcome gains shown in the submitted models are very sensitive to the way this modelling of the trial data is performed. There are several reasons to question the approach taken by the modellers:</p> <ol style="list-style-type: none"> 1) It is assumed that trend lines for PFS and OS based on trial observations will continue unchanged until all patients die. In view of the relatively long expected lifetime of this group of patients (7-10 years median survival from diagnosis), this seems to be a brave suggestion. In addition, the modellers acknowledge that after relapse, patients will undergo a series of further treatments, each of which may modify the prognosis of patients in different ways. 2) The MS provides ‘goodness of fit’ measures for a range of different types of statistical model, and selects Weibull as the base case function with Log-logistic as an alternative. However, it is far from clear that this choice can be justified from the analysis which suggests that other functions may be preferable for most of the comparisons being analysed. 3) All the fitted survival models assume that non-zero hazards occur in all comparisons immediately after randomization. This is true for patients who do not 	<p>“The MS presents the results of two sets of economic evaluations. The first compares the use of rituximab maintenance (following response to an induction therapy) versus observation only (no treatment until relapse). This is referred to as the maintenance 2-arm model. A three state transition model (progression free, progressive disease (PD) and death) is used to capture the costs and benefits of relapsed/refractory FL.</p> <p>The second model compares the use of rituximab maintenance therapy with observation only for patients responding to chemotherapy with or without rituximab and tests whether the use of rituximab as an induction therapy in addition to maintenance therapy is cost-effective. This is referred to as the induction plus maintenance 4-arm model. A five state transition model (progression free in the induction setting, progression free in the maintenance setting, progression free but not in the induction or maintenance setting, PD and death) captures the costs and benefits of relapsed/refractory FL.”</p> <p>Evidence from the EORTC trial is the principal source of clinical data used in the economic evaluations.</p> <p>Model 1</p> <ol style="list-style-type: none"> a) The transition from progression free to progressive disease is derived from the PFS observed in EORTC and the corresponding Weibull parametric extrapolation b) The transition from progression free to death is based on the overall survival observed in EORTC and the corresponding Weibull parametric extrapolation

<p>achieve a response to treatment where lack of response justifies exclusion from the second randomisation. However, for the other four groups, this assumption is false because the groups are selected on the basis of achieving a response to treatment which justifies their inclusion in the second randomization i.e. they are still alive and have not experienced disease progression when assessed for phase 2 of the trial. Thus for these four groups there is a protocol driven eventfree period equivalent to the time it takes to undergo six cycles of chemotherapy and a formal assessment prior to randomization of at least 126 days (the true mean duration of this period can only be obtained from analysis of patient-level data).</p> <p>The impact of the omission of this factor on the estimation of survival model parameters is profound - it alters the shape of the hazard function (changing the relative 'goodness of fit' of different statistical models) and also substantially alters the long-term estimated survival.</p> <p>4) In estimating model parameter values from the trial data, the modellers have calibrated groups of patients in pairs and assume that the treatments share common parameters, except for a 'treatment effect' parameter which is added to one of the common parameters to capture the influence of rituximab maintenance therapy. Thus they estimate only three parameters, instead of the four required to fit the two functions independently. This approach is justified on the grounds of making a proportional hazards assumption, but this assumption has not been substantiated by reference to the trial evidence. The ERG believes there are good grounds for questioning this strong assumption, which has the effect of over-riding potentially important differences in response patterns, and may exaggerate the long-term size of the apparent outcome benefits of rituximab.</p> <p>In view of all these problems, and because of the importance of survival modelling to the economic evaluation of rituximab, the ERG requested further information from the manufacturer about the disposition of patients and the mean time spent in each segment of the treatment pathway. In addition, the ERG requested access to a limited anonymized extract of patient-level outcome data from the trial in order to allow the ERG to explore alternative approaches to survival modelling which might overcome these difficulties. The manufacturer did not respond positively to either request.</p> <p>In the absence of other evidence, the ERG considers that it is prudent to give preference to the observed and reported evidence on PFS and OS, rather than to the manufacturer's projections. The ERG assumes that the observed effects are real and sustainable, but that no additional benefits accrue beyond the chosen cut-off point; for this we have used the same 1500 day limit as used by manufacturer's analysts... Although in general the K-M estimates are lower than the projections, the pairwise differences... are reduced to a much greater extent, illustrating the influence of joint parameter estimation on projection-based outcome gain results...</p> <p>The ERG has used the submitted 4-arm model to obtain cost-effectiveness results employing the K-M 1500-day outcome estimates, making appropriate adjustments to care costs (for routine maintenance and post-progression treatment). Table 5-17 compares these to the submitted manufacturer's base case results: in all cases the cost-effectiveness of use (or greater use) of rituximab worsens, though to different degrees, so that some strategies no longer appear to be cost-effective.</p> <p>Table 5-17: Effect of using restricted K-M outcome estimates on modelled costeffectiveness</p>	<p>c) The transition from progressive disease to death is based on the overall survival observed in EORTC and the corresponding Weibull parametric extrapolation</p> <p>Between 0-24 months: data from EORTC trial used +24 months: PD and mortality hazards from parametric curve fitting used.</p> <p>Hazards for PD and death for the rituximab maintenance group are assumed to be equivalent to those in the observation group after 5 years.</p> <p>Model 2 The transition from the induction setting to "progression free – not in the induction/maintenance settings" is based on results of EORTC. Those patients who complete induction therapy without progressive disease but who did not qualify for maintenance therapy according to the EORTC protocol will enter this health state. The transition from the induction setting to "progression free –in the maintenance setting" is based on results of EORTC. Those patients who qualified for maintenance therapy according to the EORTC protocol will enter this health state. The transition to progressive disease is based on the PFS and OS observed in EORTC. The transition to death is based on the overall survival observed in EORTC.</p> <p>Between 0-24 months: data from EORTC trial used +24 months: PD and mortality hazards from parametric curve fitting used.</p> <p>Hazards for PD and death for the rituximab maintenance group are assumed to be equivalent to those in the CHOP>O group after 5 years.</p> <p>The ERG state that: "The manufacturer extrapolates the K-M data (truncated at 1500 days [because of small event numbers and high uncertainty at the tails of the KM curve]) for progression free and OS from the EORTC trial. This was performed for the survival curves following (i) second randomisation in the EORTC trial (2-arm model) and (ii) first randomisation in the EORTC trial (4-arm model). For both models, the OS and PFS data used in the economic evaluations for each of the treatment groups are based on the fitted Weibull distributions. The Weibull curve was selected by the manufacturer on the basis of a series of good fit evaluations. The parametric curve fitting for each of the treatment groups implies different hazards across the treatment groups for the life time of the model. This is considered an unrealistic assumption by the manufacturer and so in the 2-arm model the hazards for the rituximab maintenance group are assumed to be equivalent to those in the observation group after 5 years and in the 4-arm model the hazards for the rituximab maintenance group are assumed to be equivalent to those in the CHOP>O group after 5 years."</p> <p>Sensitivity analysis was undertaken using the log-logistic distribution instead of the Weibull, which reduced the ICERs (£7.7k to £6k for model 1; £16.7k to £9.8k for model 2). SA was also completed around the duration of treatment effect. Assuming 2 years led to a substantial increase in the ICER (£7.7k to £18.1k for model 1; £16.7k to £36.5k for model 2) whereas assuming 30 years led to a reduction in the ICER (7.7k to £6.3k for model 1; £16.7k to £8.9k for model 2)</p> <p>The ERG state that they have "some concerns about the modelling of the survival data. The ERG was unable to overcome such concerns (e.g. by conducting PSA) as the manufacturer did not provide the requested additional information on the disposition of patients in the EORTC trial and the mean time spent in each segment of the treatment pathway." The ERGs concerns centre around the question of what is the most appropriate approach to the modelling of survival data.</p>
---	--

	<p>Comparison</p> <p>R-CHOP>R vs CHOP>R Model ICER KM ICER R-CHOP>R vs R-CHOP>O £16,749 £36,718 R-CHOP>R vs CHOP>O £11,904 £30,665 CHOP>R vs R-CHOP>O £11,910 £23,721 CHOP>R vs CHOP>O Dominant £73,140 R-CHOP>O vs CHOP>O £9,076 £13,895 R-CHOP>O vs CHOP>O £11,916 £19,657"</p> <p>This appears to be a type of restricted means approach.</p>	
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	A restricted means approach was taken in the absence of patient-level data.	Weibull. Log-logistic tested in sensitivity analysis. Data to which the models were fitted were truncated. Exponential, log normal and gompertz also considered.
Independent survival models, or hazard ratio (proportional hazards) modelling	Not applicable.	<p>Some use of proportional hazards modelling – only 3 parameters for 2 Weibull curves were estimated. The manufacturer states: "If the model assumption is of proportional hazard (ie. same shape) then the two curves for OS or PFS will differ only in the scale (location) parameter however their shape parameters will be the same. In this model the assumption of proportional hazard was maintained and thus only 3 parameters were estimated (1 shape, 2 scale) for PFS and 3 for OS."</p> <p>However, some information in the appraisal is also suggestive of independent modelling: The ERG state: "The parametric curve fitting for each of the treatment groups implies different hazards across the treatment groups for the life time of the model. This is considered an unrealistic assumption by the manufacturer" hence the manufacturer assumed the duration of treatment effect. The FAD confirms that such independent modelling was also carried out.</p>
Justification for survival model used?	It is interesting to note that in the appraisal more justification was given for model choice than in many other appraisals, but the ERG picked up on the issue more than in almost any other previous appraisal, and concluded that the model choices had not been appropriately justified – in response they chose to fit a model using a restricted means approach. The ERG question the analyses based upon the AIC and BIC, and also the proportional hazards assumption.	<p>In their submission, the manufacturer states: "The major distributions that have been proposed for modelling survival (or failure) times are the log-logistic, the log-normal, the exponential, the Weibull and the Gompertz. The Weibull distribution is a suitable distribution if events occur early in the follow-up period whereas the log normal and log logistic distributions have heavy right tails and are therefore suitable for situations in which events occur later in the follow-up period".</p> <p>They then present AIC and BIC statistics for each of these models. However, although the Weibull was chosen, frequently the logged models performed better based on these statistics, and in some scenarios the exponential also performed better. Differences in the statistics were relatively small.</p> <p>The manufacturer states: "Given the age and health status of the patients it was felt that the shape of the Weibull distribution was the most appropriate for this analysis. The distributions were fitted to the first 1500 days of the clinical trial period. The decision to go with the 1500 days truncation point was used because it is at this point where all curves were flattening out and thus might unduly influence the parameter estimates."</p> <p>This decision was made due to the long tails associated with logged models, as the manufacturer believed that these models were too optimistic. They reference one external data source to support their claim. They also state that it is generally best not to use a logged curve for PFS and a non-logged curve for OS as this could result in the impossible case of having more people in PFS than in OS.</p> <p>They also state: "Furthermore, the underlying assumption of the Kaplan-Meier is of proportional hazard and the Weibull is consistent with this assumption. The Log Logistic and Log Normal functions do not have this attribute of proportional hazards."</p> <p>Although it is not really stated why the proportional hazards assumption is or is not applicable.</p>

		<p>In the FAD, it is stated that when additional analysis was conducted fitting independent Weibull models instead of using the proportional hazards assumption, the ICER actually decreased slightly.</p> <p>The AC relied upon a combination of these analyses and was confident that the ICER was probably around £16k and thus acceptable.</p>
Type of economic model (Markov Model, Decision Tree, AUC, How many health states? etc)	Used the manufacturer's model.	<p>Model 1: A three state transition model (progression free, progressive disease (PD) and death) is used to capture the costs and benefits of relapsed/refractory FL.</p> <p>Model 2: A five state transition model (progression free in the induction setting, progression free in the maintenance setting, progression free but not in the induction or maintenance setting, PD and death).</p> <p>Both have monthly cycles and a 30 year time period in order to estimate lifetime costs and effects.</p>
Other issues noted (eg crossover)	<p>The ERG bring up an issue regarding survival analysis when survival is relatively long-term: "Another important issue to consider is the modifying role of subsequent treatments offered to patients. Since each treatment is liable to have a different mode of action and particular response profile, it cannot be concluded that parametric survival models calibrated only on the basis of within-trial data will remain valid when new regimens are introduced. This therefore calls into question the interpretation of long-term projections of benefit based on survival models. A more credible approach may involve limiting apparent gains only to the observation period prior to initiation of later treatments, though this is also not unambiguous in respect to inferences that can be drawn as to the relationship between observed improvements in PFS and potential gains in OS. It is quite possible for apparent gains in OS to be offset by later accelerated mortality in subsequent treatment phases."</p> <p>In addition, it appears that post-study treatments and crossover is important. The ERG state: "Post-progression treatment costs: Once patients have suffered PD, it is assumed that they will incur additional periods of active treatment at regular intervals (assumed to be two years on average). Although additional costs are included in the model, there is no opportunity for these treatments to have any effect on patient health outcomes, which are wholly determined by the survival models estimated from within-trial data. This might not be too great a problem if it were clear that the case-mix of patients suffering PD in the different treatment arms were equivalent (and therefore had similar prognoses). Unfortunately this is not the case, and therefore appears to be a modelling shortcoming, in that both cost and outcome effects of additional treatments are not estimated. The assumed interval between treatments corresponds approximately to the median progression free period for the whole trial dataset. However, the estimated weighted mean interval based on the manufacturer preferred Weibull models is in fact 3.00 years. For each treatment strategy, an average cost of post-progression treatments is calculated directly from the EORTC trial evidence of the distribution of treatments between ten regimens. Because the unit costs of individual treatments vary widely (from zero to £41,700), and the proportions of each are derived from very small numbers in the majority of cells, this approach to costing is liable to generate unwarranted apparent differences between strategies. An alternative method is to aggregate events into a small number of meaningful categories, and estimate joint averages where there is no strong evidence of significant differences. In this case, the largest category is 'chemotherapy'. For all patient groups, the post-progression use of chemotherapy appears to be inversely related to the use of regimens involving rituximab. Aggregating the latter into a single group, and creating an 'other treatments' group for all remaining events, a clear pattern emerges in which subsequent use of chemotherapy is greater for patients receiving rituximab (initiation and/or maintenance) during the trial, while the use of other treatments remains constant. We therefore assume that 25% of further therapy is attributable to 'other treatments' for all strategies. The remaining 75% is split between chemotherapy and rituximab-based treatments. Table 5-12 shows that 55% of patients not previously given rituximab will receive it postprogression whilst 45% will receive chemotherapy, falling to 25% of those previously given rituximab either for initiation or for maintenance (75% will receive chemotherapy), and 10% of those given rituximab in both trial phases (90% will receive chemotherapy). Weighted average unit costs were calculated directly for the three categories of therapy, and these were applied to the estimated proportions of therapies to generate an alternative set of average costs for each of the four treatment strategies."</p>	

34. TA145: Head and neck cancer - cetuximab, June 2008

Guidance: Cetuximab in combination with radiotherapy is recommended as a treatment option only for patients with locally advanced squamous cell cancer of the head and neck whose Karnofsky performance-status score is 90% or greater and for whom all forms of platinum-based chemoradiotherapy treatment are contraindicated.

Patients currently receiving cetuximab in combination with radiotherapy for the treatment of locally advanced squamous cell cancer of the head and neck who do not meet the criteria outlined above should have the option to continue therapy until they and their clinicians consider it appropriate to stop.

When using Karnofsky performance-status score, clinicians should be mindful of the need to secure equality of access to treatment for patients with disabilities. Clinicians should bear in mind that people with disabilities may have difficulties with activities of daily living that are unrelated to their prognosis with respect to cancer of the head and neck. In such cases clinicians should make appropriate judgements of performance status taking into account the person's usual functional capacity and requirement for assistance with activities of daily living.

Source: Cetuximab for the treatment of locally advanced squamous cell cancer of the head and neck, TA145, June 2008, <http://www.nice.org.uk/nicemedia/live/12006/40996/40996.pdf>, accessed 22/04/10

Griffin S, Walker S, Sculpher M, White S, Erhorn S, Brent S, Dyker A, Ferrie L, Gilfillan C, Horsley W, Macfarlane K, Thomas S. Cetuximab plus radiotherapy for the treatment of locally advanced squamous cell carcinoma of the head and neck: ERG Report, Centre for Health Economics, University of York and NHS Northern and Yorkshire Regional Drug and Therapeutics Centre, University of Newcastle, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, August 2007, <http://www.nice.org.uk/nicemedia/live/11697/34956/34956.pdf>, accessed 22/04/10

Merck Pharmaceuticals, Erbitux (cetuximab) for the treatment of locally advanced squamous cell carcinoma of the head and neck, STA Manufacturer Submission to NICE, August 2006. <http://www.nice.org.uk/nicemedia/live/11697/36792/36792.pdf>, accessed 22/04/10

Merck Pharmaceuticals, Erbitux (cetuximab) for the treatment of locally advanced squamous cell carcinoma of the head and neck, STA Manufacturer Submission to NICE Appendix 2, August 2006. <http://www.nice.org.uk/nicemedia/live/11697/36794/36794.pdf>, accessed 22/04/10

Note: STA: The full manufacturer submission including appendices is available on NICE website. There was an appeal on this appraisal. Many points were not upheld, but arguments against the use of carboplatin as a comparator, and related to patient comorbidities and performance status, were upheld. The result was a reanalysis of the evidence which relied upon sub-group analysis using the original economic model.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>Regarding the modelling of the 'cured' patients, the ERG noted: "The published meta-analysis and the use of UK life tables seem to be reasonable data sources. However, approximately 20% of the trial population were female, and so the use of male life-expectancy will underestimate the expected survival in the general population, and lead to a corresponding overestimate of the proportional hazard of death for patients who have experienced SCCHN. As the greatest proportion of cured patients is in the cetuximab plus radiotherapy group, this is unlikely to bias the model results in favour of cetuximab plus radiotherapy. Following a request by the ERG, the manufacturer provided an explanation of how the proportional hazard was calculated (see addendum "SCCHN NICE STA response letter for comments of 26th September 2006"); however, this has resulted in some confusion as it appears to suggest that the age-adjusted life expectancy has also been based on both male and female life tables. The ERG has been unable to replicate the calculation and, therefore, it remains unclear whether the possible bias described above will be an issue."</p> <p>Regarding the use of a random uniform probability combined with the cure rate probability, the ERG stated: " The use of a random uniform probability in this way assumes that nothing is known about the probability of each patient surviving beyond the observed censored time. This would contrast with other possible multivariate approaches that might incorporate the effect of patient covariates in the probability of survival, such as age, gender and co-morbidity. However, so long as the treatment effect of cetuximab does not differ according to these characteristics, the results of the economic model should not be biased. The use of a single draw from a random uniform probability could mean that, if the analysis were run again, the results would be quite different. In addition, the cure model predicts survival with uncertainty. The 95% confidence interval around treatment effect on the estimated cure fraction incorporated zero. This uncertainty could have been characterised in the economic evaluation by making use of the standard errors around predicted survival or the upper and lower confidence limits. The survival extrapolation is likely to be the main source of uncertainty in the economic evaluation but this is not reflected in the model results."</p>	<p>The manufacturer's submission is based on a <i>de-novo</i> economic evaluation to estimate the cost-effectiveness of treatment with (i) radiotherapy and (ii) cetuximab plus radiotherapy.</p> <p>Survival estimates were based on the one and only relevant RCT. The manufacturer extrapolated in order to estimate means.</p> <p>A cure model was used by the manufacturer, which basically allowed survival times to be estimated separately for those who had been cured and those who had not.</p> <p>"The manufacturers extrapolated censored survival times (i.e. in patients remaining alive at the end of the trial) using parametric survival models for progression-free and overall survival. To do this the manufacturers used a 'cure model', which allows a non-zero cure fraction. In other words, the survival model estimated the proportion of patients who were 'cured' (survival probability equal to 1) and who would never experience the event of interest (progression or death). This allowed the manufacturers separately to estimate the overall survival probability of cured and non-cured patients. The overall survival probability of cured patients was estimated from UK life tables together with an estimate of the proportional increase in mortality hazard for patients who have experienced LA SCCHN. The cure model predicts the progression-free or overall survival probability for the proportion of patients not cured... Survival times beyond censoring were imputed using the survival probabilities from the cure model corresponding to the censored time, multiplied by a random uniform probability. Predictions for censored overall survival were constrained to be at least as great as observed progression-free survival. Correspondingly, any predictions for progression-free survival that were greater than predicted overall survival were re-estimated. A single, deterministic imputation of progression-free and overall survival was calculated for each patient where necessary due to censoring."</p> <p>In more detail:</p> <p>"In order to estimate the overall survival probability of cured patients, the manufacturers used age-adjusted mortality risks for UK males (http://www.gad.gov.uk/Life_Tables/Interim_life_tables.htm) and applied a proportional hazard to account for the higher risk of death among patients who have experienced LA SCCHN in comparison to the general population. This proportional hazard was calculated by comparing the survival in a published meta-analysis of trials comparing radiotherapy</p>

		<p>to chemotherapy plus radiotherapy in patients with LA SCCHN to the survival probabilities calculated from UK life tables...</p> <p>The published meta-analysis provided survival curves that incorporated data up to 10 years follow-up. The manufacturer assumed that patients still alive after 5 years were cured. The clinical advisor to the ERG thought this to be a reasonable assumption. The manufacturers then estimated a straight line of 'best fit' between the published curves (loco-regional treatment plus chemotherapy compared to loco-regional treatment alone) and extended this line until it intercepted the x-axis (i.e. the survival probability was 0). The method by which this line was fitted to the published curves is not reported. The point of intercept for the fitted line was estimated to be 19 years. The mean hazard rate was then estimated by dividing the slope of the line by the survival probability in each year from 5 to 19 and pooling the results. The calculated mean hazard (0.1167) was divided by the estimated mean hazard in the general population (0.04188) to calculate a hazard ratio of 2.786. The initial survival time was set to equal the mean age (57 years) of the trial participants. For progression-free survival, the time to progression was imputed using the survival probabilities estimated in the cure model. The manufacturers state that the cure rate probability that corresponded as closely as possible to the observed censored time was multiplied by a random uniform probability. The survival time corresponding to this resulting probability was then taken as the imputed failure time. Adding this failure time to the censored time gave the estimated progression-free survival. A similar procedure was used to estimate overall survival, but using probabilities generated by the cure model for proportion of patients not cured and adjusted probabilities from UK life tables for the proportion cured."</p>
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)		<p>A log-normal distribution with a logistic link function was selected for the cure model. This is appropriate for characterising patterns of survival where the hazard is initially increasing, but then begins to decrease.</p> <p>Weibull and exponential models were also tested.</p>
Independent survival models, or hazard ratio (proportional hazards) modelling	<p>No novel models fitted.</p> <p>The ERG state that a sensitivity analysis was undertaken using a method which sounds like a restricted means approach: "A sensitivity analysis was conducted in which no extrapolation was performed. As greater proportions of patients receiving cetuximab plus radiotherapy were progression-free or alive at the end of the trial period, this analysis will underestimate the benefit of cetuximab plus radiotherapy. However, it does provide a useful extreme case scenario for comparison with the extrapolated results."</p> <p>Under the restricted means approach, the ICER increased from £6.4k to £20k.</p>	<p>Cure model fitted to patients according to response. For 'cured' background population survival rates were used combined with a proportional hazards assumption.</p>
Justification for survival model used?	<p>The ERG go into some details discussing the justification of the manufacturer's survival modelling technique. It is stated that: "The manufacturers explored a number of survival models for extrapolation. Survival was modelled using a Weibull distribution, which resulted in estimates more favourable to cetuximab plus radiotherapy in comparison to the cure model, and so the manufacturers state that their use of the cure model is conservative. In addition, the results of the cure model were compared to a simple extrapolation assuming an exponential survival distribution. The results of the simple extrapolation are described as very similar to the results of the cure model, but are not provided. This could potentially indicate that the cure model was poorly estimated on the overall survival data, and this model may have added little to a simple extrapolation assuming a constant hazard of death."</p>	<p>The manufacturer provided details of a Weibull model fitted to the data which showed that the cure model was conservative with regard to its cost-effectiveness results. They also note that the results of their cure model are similar to those obtained when an exponential model is fitted. This is claimed to act as a validation.</p> <p>The manufacturers also compared the Akaike Information Criterion (AIC), a statistic based on the log-likelihood, between models estimated using an exponential, Weibull, log-normal and log-logistic distribution, and found this to be lowest for the log-normal survival distribution.</p> <p>The manufacturer justifies using the lognormal for the cure model because the model hazards seem appropriate, and it has been used in the past.</p>

	<p>Also: "The shape parameter of the Weibull distribution for the estimation of overall survival was not significantly different from 1, indicating that an exponential distribution may equally be able to describe the survival data. The shape parameter for overall survival was estimated to be 0.93, indicating that the hazard was, if anything, slightly decreasing with time. The shape parameter of the Weibull distribution for progression-free survival was estimated to be 0.81, and was significantly different from 1, indicating that the hazard for progression-free survival was, on average, decreasing with time.</p> <p>The manufacturer states that the observed survival curves appeared concave, indicating that a log-logistic or log-normal model would be more appropriate than an exponential model that assumed a constant hazard, or a Weibull model that assumes a monotone hazard. The ERG considers this to be appropriate... Whilst the use of the cure model may have been conservative with respect to the use of a Weibull model, the choice of distribution for the cure model was not conservative. The log likelihood for each of the distributions tested within the cure model was lowest for the logistic distribution, but the log-normal distribution was selected. The manufacturers chose to use the log-normal distribution, as it resulted in the lowest cure fraction (estimated proportion of patients cured) compared to the Weibull, logistic, gamma or exponential. While this is true, it also resulted in the biggest difference in cure fraction between the treatment groups (11.7% in favour of cetuximab plus radiotherapy) compared to the alternative distributions (smallest difference 8.4%)."</p>	
<p>Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)</p>		<p>A type of state transition model is used, but the transitions are such that patients cannot return to a previous state, and so in a sense the model is survival based, based on the trial data and not a Markov Model:</p> <p>"Patients enter the model in the acute treatment stage at the beginning of the trial period. In the acute stage patients reside in the health state relating to the worst adverse event(s) they experience during treatment... Following the acute treatment stage, patients enter the locoregional control state and remain in this state until they experience disease progression at which point they enter the progressive disease state. These flows between states are uni-directional: patients are unable to return to the progressive disease or acute treatment states once they have left them. At any point in the model, patients may die and exit to the absorbing death state."</p> <p>Essentially, the evaluation is an EEACT, but uses extrapolation where necessary: "Disease progression is based on the individual patient data where it is recorded, and is imputed via the cure model where it is not."</p> <p>"Transitions between health states were based on actual observations rather than transition probabilities where such data was recorded. Where data were censored, transitions from health state J (locoregional control/stable disease) to health state K (progressive disease) and from either state to death was imputed via a statistical cure model."</p>
<p>Other issues noted (eg crossover)</p>	<p>Treatment crossover is not mentioned, it is only stated that salvage therapy and subsequent chemotherapy was balanced between the treatment groups.</p> <p>The use of a 'cure' model allows some use of external data (life tables) to estimate survival for a proportion of patients. [Note STATA command for the cure model is given in appendix 2 of the manufacturer's submission]</p> <p>An additional issue raised by the authors is that: "the approach undertaken within the model assumes that nothing is known about the probability of each patient surviving beyond the observed censored time. Multivariate approaches to extrapolation could have been used incorporating effects of extrapolation etc" – therefore if the treatment effect of cetuximab differed according to covariates then the results may be biased.</p> <p>It's worth noting that the ERG states that a number of the issues regarding the extrapolation were likely to be of low importance because even with no extrapolation (ie restricted means approach) the ICER was just below £20k – thus such an analysis was a helpful sensitivity analysis.</p>	

35. TA162: Lung cancer (non-small-cell) - erlotinib, November 2008

Guidance: Erlotinib is recommended, within its licensed indication, as an alternative to docetaxel as a second-line treatment option for patients with non-small-cell lung cancer (NSCLC) only on the basis that it is provided by the manufacturer at an overall treatment cost (including administration, adverse events and monitoring costs) equal to that of docetaxel.

The decision to use erlotinib or docetaxel (as outlined above) should be made after a discussion between the responsible clinician and the individual about the potential benefits and adverse effects of each treatment.

Erlotinib is not recommended for the second-line treatment of locally advanced or metastatic NSCLC in patients for whom docetaxel is unsuitable (that is, where there is intolerance of or contraindications to docetaxel) or for third-line treatment after docetaxel therapy.

People currently receiving treatment with erlotinib, but for whom treatment would not be recommended according to section 1.3, should have the option to continue treatment until they and their clinicians consider it appropriate to stop.

Source: Erlotinib for the treatment of non-small-cell lung cancer, TA162, November 2008, <http://www.nice.org.uk/nicemedia/live/11777/42657/42657.pdf>, accessed 23/04/10

Bagust A, Boland A, Dundar Y, Davis H, Dickson R, Green J, Hockenull J, Macbeth F, McLeod C, Proudlove C, Stevenson J, Walley T. Erlotinib for the treatment of relapsed non-small-cell lung cancer: ERG Report, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, September 2006, <http://www.nice.org.uk/nicemedia/live/11714/35177/35177.pdf>, accessed 23/04/10

Roche Products Ltd, Achieving clinical excellence in the treatment of relapsed non-small cell lung cancer, Tarceva (erlotinib) NICE STA Submission, May 2006. <http://www.nice.org.uk/nicemedia/live/11714/37396/37396.pdf>, accessed 23/04/10

NICE, Final Appraisal Determination, Erlotinib for the treatment of non-small-cell lung cancer, TA162, September 2008, <http://www.nice.org.uk/nicemedia/live/11714/42154/42154.pdf>, accessed 23/04/10

Note: STA: The full manufacturer submission including appendices is available on NICE website. There was an appeal on this appraisal. Several points referred to the rejection of the manufacturer's survival analysis and the acceptance of the ERGs analysis – these were rejected by the appeals panel. However one point which was upheld was to do with the ERG refusing to name the software they used to digitise the survival curves. This was named in the appeal hearing (it was techdig). Other non-survival based points were upheld.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The ERG state: "TAX317 is not the largest trial involving docetaxel, nor the most recent; hence the heavy reliance on this small trial does not seem justified." The ERG were also very concerned about the comparability of TAX317 and BR21 due to differences in patient characteristics. Therefore they had concerns over the indirect comparison.</p> <p>The ERG had concerns about the assumptions made about survival in the manufacturer's submission. They reassessed some figures, using a program to digitise KM curves and extrapolate to estimate means:</p> <p>"Overall survival (OS) for both docetaxel and erlotinib was assumed to be equivalent based on the mean overall survival (time to last observation) for erlotinib from the BR21 trial (9.03 months). At first sight this assumption appears reasonable as figures for mean OS are presented for erlotinib and docetaxel at two years (9.03 and 8.89, respectively), suggesting a small advantage for erlotinib. However, close re-examination of the Kaplan-Meier plot for docetaxel 75mg (TAX317) by the ERG leads to an estimated mean survival of 9.47 months, which closely matches the estimate of 9.48 months calculated by the investigators and reported by Leigh. At the same time point (19.3 months) the equivalent restricted mean survival (by AUC) for erlotinib is 8.59 months. If instead exponential survival curves are fitted to the Kaplan-Meier plots and projected to death, the estimated mean survival times are 11.2 months for docetaxel and 9.9 months for erlotinib. It is therefore far from clear that the assumption of survival equivalence is in fact</p>	<p>The economic model submitted in support of the company submission was a basic three state model comparing erlotinib with docetaxel, furnished with clinical data from TAX317 and the BR21 trial (erlotinib and docetaxel have not been compared in the same trial, therefore indirect comparisons were used to estimate mean PFS and OS).</p> <p>Kaplan-Meier data for PFS and OS is used directly in the model – life expectancy was short and so little extrapolation would have been required. In the base case OS was assumed to be equivalent.</p> <p>The indirect comparison was naïve – ie unadjusted.</p> <p>The manufacturer states: "As all patients in the docetaxel trial have died within the 2 year time horizon, the docetaxel mean reported by Holmes <i>et al</i> can be described as an "unrestricted mean. As Holmes states: "The longest survival time in the Shepherd study was 19.3 months" (p.582, Holmes <i>et al</i>, 2004)." And: "The erlotinib mean survival estimate however is a "restricted" mean as patients are still alive within the trial at 2 years. A longer follow-up of the BR21 trial or an extrapolation of the BR21 survival curves would lead to an increase in mean overall survival for erlotinib by increasing the area under the curve." However, this assumes that there was no censoring in the docetaxel study.</p>

	<p>conservative. This suspicion is reinforced by the reported median overall survival results: 7.5 months for docetaxel 75mg versus 6.7 months for erlotinib.</p> <p>In the model, docetaxel PFS was based on the estimate of mean treatment duration during the TAX317 trial (3.33 months), as data on docetaxel <i>mean</i> PFS was not available directly. This was compared to the mean PFS for erlotinib, estimated as 4.11 months (based on the proxy measure of mean treatment duration from the BR21 trial).</p> <p>However, data on docetaxel median time to progression (TTP, which is virtually equivalent to PFS) was available from TAX317, but was not mentioned in the submission. Unsurprisingly the median PFS (using TTP as a proxy) for docetaxel is greater than the median PFS for erlotinib (2.5 months versus 2.2 months). Furthermore, the JME1 trial estimates the median PFS of docetaxel as 2.9 months, which is greater than the 2.2 months reported for erlotinib.</p> <p>Once again, whilst it is appropriate to use means in economic analyses, the median should have been discussed especially since a poor proxy measure for PFS was used in its place. Furthermore, upon examination of the median PFS, it could be argued that docetaxel and erlotinib are at best equivalent, and that, based on the clinical data, docetaxel may be superior. In summary, the company's supposition that erlotinib is superior to docetaxel in terms of PFS seems overly generous. It is based on a proxy measure which is inherently flawed, and reverses the conclusion that might reasonably be drawn from the available clinical data."</p> <p>And:</p> <p>"Secondly, the assumption that mean overall survival is equivalent between erlotinib and docetaxel is not unequivocally demonstrated but relies on an indirect comparison between BR21 and the small underpowered TAX317 study. Furthermore, re-analysis of the Kaplan-Meier survival curves suggests that docetaxel may offer a survival advantage compared with erlotinib, which is also supported by data on median overall survival. This view is endorsed by the Australian Department of Health who reported that an indirect comparison of erlotinib versus docetaxel "...favoured docetaxel such that a statistically significant survival advantage for docetaxel could not be excluded".¹³</p> <p>Thirdly, the case for a progression-free survival benefit in patients treated with erlotinib compared with docetaxel is also based on an indirect comparison of BR21 and TAX317, and furthermore relies on the proxy measure of mean treatment duration. Using the proxy measure of median time to progression, estimates of progression-free survival appear to be greater for docetaxel patients than for erlotinib patients. Hence, it could be argued that in terms of progression-free survival docetaxel and erlotinib should be considered clinically equivalent at best. The company do not discuss this in their submission."</p>	<p>Means are used, but due to the lack of availability of some mean data a proxy mean variable has been used for one treatment (treatment duration rather than PFS) and so the ERG suggested that a median should have been considered in place of the proxy mean.</p>
<p>Evidence synthesis (pool survival estimates?)</p>		<p>In response to the ERG critique Roche performed a network meta analysis of OS, including erlotinib, docetaxel and other treatments, compared to placebo/BSC. They stated that this backed up their assumption of clinical equivalence, and that erlotinib may be better. However, in the FAD the Appraisal Committee state that the MTC was not robust – patient characteristics were likely to impact upon relative treatment effects, not all relevant trials were included, and some irrelevant comparators were included.</p>
<p>Survival model(s) fitted (Weibull, exponential)</p>	<p>The ERG completed a restricted means analysis so that the same time period is considered for both drugs, and also assessed the impact of extrapolating using</p>	<p>The manufacturer estimated AUC for the KM curves.</p>

etc)	exponential models.	
Independent survival models, or hazard ratio (proportional hazards) modelling	The exponential models appear to be fitted independently but this is not specified.	Not applicable.
Justification for survival model used?	No details for the choice of exponential models.	Assumed equivalence (attempted to justify with MTC).
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	Conducted an additional analysis based upon mean survival restricted to the same time point, but still based upon the manufacturer's model.	3-state Markov Model: PFS, Post-progression and Death. The model doesn't work like a standard MM though, since time spent in different states is based purely on KM data from trials, thus there are no transition probabilities. The ERG states: "The time horizon is two years, even though all patients in the TAX317 trial had already died by that time, which is in part explained by the small patient numbers in the trial (n = 55). The company believe that this will bias the indirect analysis in favour of docetaxel as 15% of patients in the erlotinib arm of the BR21 trial (n = 488) were still alive beyond the two year cut-off. The company argue that this is a conservative approach as, by excluding these patients from the analysis, the full health benefits of erlotinib are not realised."
Other issues noted (eg crossover)	It is stated that for the erlotinib study: "The study was a parallel-group study with no cross-over allowed. However significantly more placebo (18; 7%) than erlotinib recipients (8; 2%) received EGFR inhibitor therapy after study withdrawal. This would be expected to diminish the survival impact of erlotinib given on study." However "Two patients in the placebo arm inadvertently received active drug for a single 28 day period (the 2nd and 8th treatment periods) and are included with placebo patients for purposes of analysis." No adjustment was made for this. While it is possible that crossover is of less importance with very short survival, it still may have had an effect. No mention of crossover in the TAX317 study is made.	

36 and 37. TA169 and TA178: Renal cell carcinoma - sunitinib March 2009; Renal cell carcinoma August 2009

Guidance: Sunitinib is recommended as a first-line treatment option for people with advanced and/or metastatic renal cell carcinoma who are suitable for immunotherapy and have an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1.

When using ECOG performance status score, clinicians should be mindful of the need to secure equality of access to treatments for people with disabilities. Clinicians should bear in mind that people with disabilities may have difficulties with activities of daily living that are unrelated to the prognosis of renal cell carcinoma. In such cases clinicians should make appropriate judgements of performance status taking these considerations into account.

People who are currently being treated with sunitinib for advanced and/or metastatic renal cell carcinoma but who do not meet the criteria above should have the option to continue their therapy until they and their clinicians consider it appropriate to stop.

Bevacizumab, sorafenib and temsirolimus are not recommended as first-line treatment options for people with advanced and/or metastatic renal cell carcinoma.

Sorafenib and sunitinib are not recommended as second-line treatment options for people with advanced and/or metastatic renal cell carcinoma.

People who are currently being treated with bevacizumab (first-line), sorafenib (first- and second-line), sunitinib (second-line) and temsirolimus (first-line) for advanced and/or metastatic renal cell carcinoma should have the option to continue their therapy until they and their clinicians consider it appropriate to stop.

Source: Sunitinib for the first-line treatment of advanced and/or metastatic renal cell carcinoma, TA169, March 2009, <http://www.nice.org.uk/nicemedia/live/12143/43556/43556.pdf>, accessed 11/05/10

Bevacizumab (first-line), sorafenib (first- and second-line), sunitinib (second-line) and temsirolimus (first-line) for the treatment of advanced and/or metastatic renal cell carcinoma, TA178, August 2009, <http://www.nice.org.uk/nicemedia/live/12220/45232/45232.pdf>, accessed 12/05/10

Coon J, Hoyle M, Green C, Liu Z, Welch K, Moxham T, Stein K. Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: A systematic review and economic evaluation. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, May 2008, <http://www.nice.org.uk/nicemedia/live/11817/41488/41488.pdf>, accessed 11/05/10

Coon J, Hoyle M, Green C, Liu Z, Welch K, Moxham T, Stein K. Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: A systematic review and economic evaluation: Addendum to the report submitted on 2nd May 2008, Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, September 2008, <http://www.nice.org.uk/nicemedia/live/11817/43144/43144.pdf>, accessed 11/05/10

Abrams K, Palmer S, Wailoo A. Bevacizumab, sorafenib, sunitinib, and temsirolimus for renal cell carcinoma, Decision Support Unit, September 2008, <http://www.nice.org.uk/nicemedia/live/11817/43145/43145.pdf>, accessed 12/05/10

PenTAG, Drugs for renal cancer: PenTAG response to Consultee comments, June 2008, <http://www.nice.org.uk/nicemedia/live/11817/43147/43147.pdf>, accessed 12/05/10

NICE, Final Appraisal Determination, Sunitinib for the first-line treatment of advanced and/or metastatic renal cell carcinoma, TA169, February 2009, <http://www.nice.org.uk/nicemedia/live/11817/43174/43174.pdf>, accessed 12/05/10

NICE, Final Appraisal Determination, Renal cell carcinoma – bevacizumab, sprafenib, sunitinib and temsirolimus, TA178, April 2009, <http://www.nice.org.uk/nicemedia/live/11817/43918/43918.pdf>, accessed 12/05/10

Note: This was an MTA which was split into separate appraisals due to an appeal. There was only one assessment report, so both appraisals are considered in this section. The executive summaries of the manufacturer submissions are available on the NICE website. Sunitinib was recommended for 1st line use based on the end of life ruling, and an ICER of approx £50k. None of the other drugs were accepted. Bevacizumab was deemed to have a patient population of greater than 4,000 as it can be used for a number of cancers, but this was not considered to be the case for sunitinib (even though it too can be used for many cancers, in the sunitinib FAD the AC chose to only consider the particular licensed indication under scrutiny – possibly because RCC was sunitinib's first licence). There was an appeal, in which much of the discussion focussed on the population issue, but all appeals were dismissed.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>Weibull survival curves were fitted to the progression free and overall survival Kaplan-Meier curves from clinical trials for the baseline comparator. Relative measures of treatment effectiveness (hazard ratios) were then used to estimate the expected disease progression compared to baseline.</p> <p>An indirect comparison based on the Bucher method was used to compare the HRs for sunitinib and bevacizumab.</p> <p>Phase III data was available for bevacizumab and sunitinib for first line treatment, for temsirolimus for first-line treatment in poor prognosis patients, and for sorafenib for second line treatment. Phase II data was available for sunitinib for second line treatment.</p> <p>“In the survival analysis used to structure the model, for each baseline strategy/treatment a Weibull curve is derived to describe the number of patients alive over time (overall survival data) and another Weibull curve describes the number of patients in PFS over time. Weibull survival curves were fitted separately, corresponding to a chosen baseline treatment (i.e. IFN or BSC), to the PFS and OS Kaplan-Meier curves from the RCT judged most appropriate. For each treatment being compared to the baseline disease progression (e.g. sunitinib vs. IFN) the model uses relative measures of treatment effectiveness (hazard ratios) to estimate the expected disease progression compared to baseline. For each treatment (baseline and comparator), the number of patients in the PD health state at any time is calculated as the number alive minus the number in PFS health state, at that time.”</p> <p>Data from the bevacizumab trial were used for IFN for 1st line PFS and OS as it was more complete than the sunitinib trial, and published. Weibull models were fitted, but it is not clear how this was done – eg</p>	<p>Pfizer (sunitinib)</p> <p>Modelling uses survival analysis, employing clinical effectiveness data from a RCT (1st line) and other sources (2nd line), to model survival and disease progression over time. “In the CEA for first-line use, much of the data used is from the Phase III RCT of sunitinib versus IFN... For baseline disease progression (IFN alone), uses Weibull survival curves, modelled from trial data. To model differences between treatment (sunitinib) and controls, the analysis applies relative measures of treatment effectiveness (hazard ratios) from the RCT. In the sensitivity analysis the submission explores alternative methods for survival analysis, and the estimation of treatment effects.” This sensitivity analysis involved fitting independent Weibull models rather than relying on HRs (this alters the ICER by only approx £500), and another analysis involved fitting a BSC curve based on survival data from 3 independent trials, and then applying the HR to that. In one analysis the manufacturer used IFN OS data from the bevacizumab trial since longer term unconfounded data was available, and HRs were applied to that. However, the AG stated that if OS data is used from the bevacizumab trial PFS data also should be, for consistency. This is the approach used by the AG, and they note that this doubles the ICER from £28k to £56k due to PFS estimates.</p> <p>“In the analysis for first-line use, Pfizer assume that patients receive sunitinib or IFN until disease progression (PD state), and following progression patients receive BSC (second line drugs are not part of the analysis).”</p> <p>The submission presented two base case analyses, one using pre-planned interim analysis data, and the other using unplanned updated analysis data. The assessment group cautioned that the unplanned updated analysis data includes patients who have crossed over from IFN to sunitinib, with potential for confounding in the estimates of treatment effect (hazard ratios). For this reason the assessment group relied upon evidence from the pre-planned interim analysis.</p> <p>For the 2nd line model survival analysis is used to model disease progression, survival and treatment effect, with Weibull survival curves used to extrapolate from different (and independent) sources of data. Sunitinib survival is modelled based on a single arm phase II trial, whereas BSC survival is based upon two different methods – one pooled several different trials, and one was based upon SEER data. The AG note that this is likely to be biased because of the differences in patient characteristics.</p>

	<p>using tech dig. Data from the sunitinib trial were tested in sensitivity analysis.</p> <p>For the temsirolimus evaluation Weibull curves were fitted to empirical Kaplan-Meier data on PFS and OS for the patient group on IFN and HRs were applied to these for temsirolimus.</p> <p>For the sorafenib evaluation "For this question we identified data on sorafenib versus best supportive care (BSC) only. Whilst data were identified on sunitinib versus BSC in second line therapy it comes from two single-arm trials. We did not use this data to model cost-effectiveness due to methodological concerns. We modelled disease progression and cost-effectiveness for sorafenib compared to BSC using data from the RCT reported by Escudier and colleagues. We used data from this RCT for all patients in the trial, although we note that only 82% had been previously treated with immunotherapy." Weibull models were fitted to the control group data, and hazard ratios were applied for the treatment effect.</p>	<p>Roche (bevacizumab) The model uses RCT data. For baseline disease progression (IFN alone), a Gompertz model is fitted to the IFN data. To model differences between bevacizumab plus IFN and IFN, the analysis considers PFS by applying a Gompertz survival curve for bevacizumab plus IFN, modelled from trial data. For overall survival, modelling applies a relative measure of treatment effectiveness (hazard ratios) from the RCT to the baseline survival analysis. The submission explores alternative mathematical survival curves in sensitivity analyses. Following disease progression (PD health state) patients receive BSC, and are assumed to use second line drugs.</p> <p>The AG "highlight a concern over the clinical effectiveness data (HRs) used in the manufacturer analysis for overall survival (OS) and progression free survival (PFS). The analysis uses the HR for OS from unpublished data on what is classed a 'safety population' (not the ITT data), using the OS HR of 0.709. This differs from the OS HR of 0.75 from the ITT reported by Escudier and colleagues (2007). Where PenTAG have used the manufacturer model and applied the ITT data of 0.75 (HR for OS), the ICER increases from £75,000 to £87,400 per QALY. It is not clear why the manufacturer analysis uses data from the safety population (compared to ITT data). Again, with PFS we note that the model uses a HR of 0.609 (CI 0.508, 0.728) for PFS and that this is from a 'safety population' (stratified, by risk group) rather than the data reported in the RCT. The RCT reports a HR of 0.63 (95% CI 0.52, 0.75) in unstratified analysis. However in their model, a PFS HR is not explicitly applied, because PFS for both treatment arms is fitted to empirical trial data independently (we assume that this HR is implicit in the Kaplan-Meier data)." It is not clear exactly what the safety population is, although it seems likely to be those patients who actually received a dose of the randomised treatment.</p> <p>"Also, in sensitivity analysis the submission reports findings where cost-effectiveness has been assessed using a log-logistic model (instead of the Gompertz methods in the base case analysis), and PenTAG would question the appropriateness and prominence of this sensitivity analysis. In this case, the ICER falls greatly, from £75,000 to £40,000 per QALY, and at a willingness to pay threshold of £30,000 / QALY, bevacizumab plus IFN has a 9% probability of being cost-effective compared to IFN. However, Roche acknowledge that this ICER may be unrealistic because the log-logistic model results in an expected lifetime which may be unrealistically long (see Figure 7). We do not see the log-logistic method as a credible approach, i.e. we agree with Roche that it is unreasonable to use the log-logistic distribution to model PFS and OS in the sensitivity analysis because the tail of the distribution is too long."</p> <p>Wyeth (Temsirrolimus) "The model uses survival analysis, employing clinical effectiveness data from a single RCT to model survival and disease progression over time. The approach uses Weibull regression models, applied to PFS and OS data, to calculate the time dependent transition probabilities used to model disease progression, and cost-effectiveness."</p> <p>"The model is based on a set of time dependent transit probabilities, derived from individual patient-level data (not available to PenTAG) from the RCT by Hudes and colleagues (2007). We are unable to consider the derivation of these probabilities in any detail. Transit probabilities cover PFS to death, and PFS to post-progression. Thereafter the model makes assumptions over other transition probabilities. An assumption is made that the probability of transition from post-progression to death is equal to that for PFS to death. An assumption is made that the probability of transition from post-progression to post-progression (i.e. remaining in that state) is equal to that for PFS to post-progression. The rationale / support for these assumptions is/are not presented."</p> <p>"For each treatment, Weibull regression models are used to derive transition probabilities, with Weibull data fitted for transition from PFS to death, and from transition from PFS to post-progression. For subgroup analysis, the PFS and OS Weibull curves are unique for each patient subgroup: clear cell, non-clear cell, nephrectomy, non-nephrectomy."</p> <p>The AG state that they have a major concern over the Wyeth model, because the mean survival estimate for temsirolimus seems to be an overestimate. Although the medians are similar, the shape of the curves are</p>
--	--	---

		<p>very different to the KM curves.</p> <p>Bayer (Sorafenib) "The model uses survival analysis, applying data from the RCT reported by Escudier and colleagues, to model survival and disease progression over time. Data from the RCT are classed as mature for the PFS survival analysis, but immature (short follow-up) for the data on overall survival of patients. Therefore, whilst trial data (Kaplan-Meier) was used for PFS in both sorafenib and BSC, for the OS data trial data were extrapolated (using an exponential function) over time. The analysis uses survival data (empirical, or projected) for both sorafenib and BSC (to derive time dependent transition probabilities), and the model does not use relative measures of treatment effect (hazard ratios) to predict differences between treatment arms. In subgroup analyses, different methods were employed to model progression and treatment effect, adjusting baseline survival analysis using different data on median PFS and OS." Therefore there was some use of medians, restricted means and extrapolation.</p>
Evidence synthesis (pool survival estimates?)	PH modelling is used so that an indirect analysis of comparators that were not included in the same trial could be undertaken.	Analyses were mainly based on trial data in the different submissions, but there was some use of external data and some pooling of trials – particularly in the Pfizer analysis (see above for details).
Survival model(s) fitted (Weibull, exponential etc)	Weibull.	<p>Pfizer (sunitinib) Weibull. The assessment group state that "We have some concerns with the model used to estimate the cost-effectiveness of sunitinib for 1st line use. First, and a major concern, is that the Weibull curve fitted to trial data on progression free survival (PFS) for IFN is a poor fit to the empirical survival data. Figure 6 (page 109) shows that the Weibull curve fits the empirical data well up to about 0.5 years, but that thereafter the model predicts a much shorter tail (more rapid disease progression) than is shown by the actual PFS survival data. The manufacturer submission acknowledges that the curve "does not fit the latter proportion of the Kaplan-Meier data, and therefore the PFS benefit of IFN-α could be underestimated" (p58 of the industry submission). We suggest that the consequences of this poor fit are important, and in addition to the suggested underestimated benefit, the modelling creates an underestimate of the cost per QALY (due to incremental costs and effects associated with PFS)." "We have noted that the Pfizer survival analysis for PFS is heavily influenced by the first few data points in the Kaplan-Meier trial data. The submission has the curve fitted to multiple data points each month (and the transformation of the Weibull survival function S(t) for regression, $\ln(-\ln(S(t)))$ is very large and negative when S(t) is just below 1, i.e. for small time t). PenTAG suggest that the first few data points are outliers in the regression. When we fit a Weibull curve to fewer data points, in this case one data point per month, the fit to the actual data is much improved, because there are then no outliers in the regression, see Figure 6, page 109." Therefore the AG outline that one Weibull may not be sufficient, as the cumulative log hazard curve is not linear. However, their decision to fit a Weibull curve using one data point per month seems arbitrary. Also, they judge the fit of the curves purely by visual inspection, without considering that a poor fit to the tail of the KM curve may not be important due to the lack of data points in this part of the curve.</p> <p>Roche (bevacizumab) Gompertz, but other models (log-logistic) considered in sensitivity analysis.</p> <p>Wyeth (temsirolimus) Transition probabilities using Weibull models</p> <p>Bayer (sorafenib) Exponential for OS, but restricted means used for PFS.</p>
Independent survival models, or hazard ratio (proportional hazards) modelling	PH modelling used.	<p>Pfizer (sunitinib) PH modelling, but independent models tested in sensitivity analysis. The assessment report does not state whether any reference to the proportional hazards assumption was made by the manufacturer.</p>

		<p>Roche (bevacizumab) PH modelling for OS, but independent curves for PFS.</p> <p>Wyeth (temsirolimus) Independent</p> <p>Bayer (sorafenib) Independent</p>
Justification for survival model used?	<p>The AG make use of PH modelling to a very large extent, but do not discuss proportional hazards in their report at all. PH modelling was used due to there existing multiple comparators in separate trials, but for some of the evaluations this was not the case, so use of the HR was not necessary. The source of the HR was not considered, and there is no detailed discussion as to why Weibull models were used instead of other parametric models.</p> <p>The AG were asked about why Pfizer and the AG predicted different mean PFS despite using the same data. The AG explained this by saying that Pfizer used a different Weibull fit to the data (it remains unclear if the AG had access to the PLD). The AG argue that their model presents a closer fit to the 'data' but no tests are done, and this appears just to be based on a visual inspection. No discussion on whether censoring impacts upon the fitted curves is given – ie it may not be important to fit to the tail of the KM.</p> <p>The AG were also asked to clarify why the use of PH modelling for sunitinib was the best approach. The AG clarified that the HR they used was appropriate because PFS was unaffected by crossover, and that HRs were used to allow a 3-way comparison with bevacizumab. This is in response to Pfizer arguing that fitting curves independently for sunitinib more accurately reflects the data. The AG state that the choice between PH modelling and independent models is circumstance-dependent. They state that HR approaches may be better as they allow the use of one treatment effect, whereas independent curves fit to the KM curves including the tails, which can be uncertain [however, the models should allow for this].</p>	<p>Pfizer consider a number of different data sources for their survival modelling, and try PH and independent models. However, there are little details on why they decided to use Weibull models.</p> <p>Roche (bevacizumab) The use of independent models for PFS and PH modelling for OS is rationalised with reference to the incompleteness of the data. It is reported by the AG that the manufacturer stated that they tested several survival models, but that the Gompertz was chosen because it provided the best fit to the trial data. No further information on this is given. Roche and the AG agree that their sensitivity analysis involving a log-logistic survival model is unlikely to be appropriate due to the length of the tails. However, this assertion is not backed up by any data showing that the log-logistic model estimates overlong survival.</p>
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	<p>A decision analytic Markov-type model was developed to simulate disease progression and to estimate the cost-effectiveness of the drugs under consideration. The model had three health states: progression free survival, progressive disease and death and uses estimates of effectiveness, costs and health state utilities assigned to these states to model disease progression and cost-effectiveness over time in a cohort of patients. The model has a ten year time horizon and a 6-week model cycle. Future costs and benefits were discounted at 3.5% per annum.</p>	<p>Pfizer (sunitinib) Markov model to compare sunitinib at 1st line to IFN and at 2nd line to BSC. There are three health states: progression-free survival (PFS), progressive disease (PD) and death. The model uses a lifetime time horizon, and a short model cycle (first line 0.01 years [4-days] per cycle; second line variable cycle lengths, 1 – 10 weeks). Patients start in PFS in both models.</p> <p>Roche (bevacizumab) A Markov Model with three health states is used: progression free survival (PFS), progressed diseases (PD) and death. The model uses a lifetime time horizon, and a model cycle of 1 month. The model uses survival analysis, employing clinical effectiveness data from the RCT reported by Escudier and colleagues, to model survival and disease progression over time.</p> <p>Wyeth (temsirolimus) Markov model comprising three primary health states: progression free survival (PFS), post-progression and death. However, the PFS health state is sub-divided into 3 categories (sub-states), of complete/partial response, stable disease, and progressive disease (PD). The model uses a time horizon of three years, and a model cycle of 1 month. The model uses survival analysis, employing clinical effectiveness data from a</p>

		<p>single RCT to model survival and disease progression over time.</p> <p>Bayer (sorafenib) Markov Model with three health states: progression free survival (PFS), progressed disease (PD) and death. The model uses a 10-year time horizon, and a 1 month model cycle.</p>
<p>Other issues noted (eg crossover)</p>	<p>It is worthy of note that in their response to comments on the AG report regarding the use of 2nd line treatments in the AVOREN trial, which they used for their base IFN PFS and OS curves, the AG state: "PenTAG would note that whilst the published paper includes the statement that "Other neoplastic agents were allowed subsequent to progression or toxicity", we are unaware of any published evidence to suggest that TKIs or temsirolimus were used as second line therapies. We were therefore unable to adjust the IFN baseline overall survival data to reflect the use of second line treatment options."</p> <p>"In addition, had it been clearer that the treatment protocol in the AVOREN trial could potentially introduce bias, due to use of second line therapies, we are unable to identify a clear and robust method for making any adjustment to the data to consider such uncertainty."</p> <p>The Assessment group note that there is little data available to inform on the effects of both sunitinib and bevacizumab "on overall survival due to the early crossover of patients on control treatment following interim analyses; both interventions show some benefits on overall survival."</p> <p>Indeed "treatment crossover following interim analyses was permitted in all but one (temsirolimus vs. IFN) of the included trials resulting in confounding of overall survival data. There is therefore a large amount of uncertainty in the estimates of overall survival used in the assessment of clinical and cost-effectiveness."</p> <p>"Moreover, further analysis of these trials is unlikely to add significantly to this particular evidence base as treatment crossover has occurred following interim analyses." – long-term follow-up is not useful if crossover has occurred.</p> <p>And "There is evidence of confounding in at least one of the included trials; final analysis of overall survival in the TARGETs trial, (after 48% (n=216) patients in the placebo group had crossed over to sorafenib treatment) produced a hazard ratio of 0.88 which was not statistically significant. Further analysis in which data from the crossed over patients were censored, produced a hazard ratio of 0.78 (p=0.0287). Clearly the true effect of sorafenib in this trial lies somewhere between these two estimates."</p> <p>And "In the current evidence base there is large amount of uncertainty surrounding the estimates of overall survival, primarily due to early crossover of people receiving control treatment following interim analyses. It is unrealistic and perhaps unethical to expect that further randomised clinical trials would be performed using IFN or best supportive care as a comparator in these interventions that are now widely used in Europe and the US. As the interventions provide little possibility of a cure and in the absence of unconfounded estimates of overall survival from RCTs, further understanding of the impact of the interventions on health related quality of life during progression free survival and progressed disease would facilitate the decision making process for clinicians and patients."</p> <p>Pfizer crossover Crossover was an issue in the Pfizer trial – switching occurred after progression. But in addition to this, after the study finished 2nd line treatments were also given. In subsequent analysis the model was re-run using OS from a 'no post-study treatment' group (less than half the original randomised patient number), and also using data for this group for both PFS and OS. The AG state that patient characteristics look similar for this group and those who did go on to receive treatment, but this does not appear to have been tested, and it is noted that PFS differs for these patients, even before switching occurred. In particular, they state that it did not seem reasonable to use data from this group for PFS.</p> <p>In response to a request from NICE, the AG modelled using the ACs preferred assumptions. These used final ITT data for PFS, and for sunitinib OS, but only data for patients who had no post-progression treatments in the control arm for OS. Also, independent Weibull's were used. The AG were concerned with this because: "The Weibull curves used for PFS are both underestimating the proportions of people in PFS over time, when compared to the empirical Kaplan-Meier data. And secondly, PenTAG have concerns over the approach requested by NICE to estimating OS i.e. using survival curves from different patient groups for OS, (ITT OS curve for sunitinib and the 'no poststudy- treatment group' for IFN OS)." This reduced the ICER from £62k to £54k.</p> <p>The DSU were also asked to consider the new data supplied by Pfizer, which included more OS data, but this was confounded by crossover. In the interim data the OS HR was 0.65, and in the new data the HR was 0.82. This increased the ICER calculated by Pfizer from £29k to £72k. Pfizer presented an analysis where patients were censored at crossover, which gave a HR of 0.81, and therefore censoring made little difference compared to the ITT analysis. Pfizer claimed that the most appropriate method was to exclude all patients who crossed over, which gave a HR of 0.65. The DSU state that: "excluding patients who progress and therefore require 2nd line therapy (regardless of whether their demographics are similar or not to those who remain included) will almost certainly produce inappropriate results since their reason for exclusion is inextricably linked to outcome, i.e. death. A more appropriate strategy would be to censor at the time at which they began 2nd line therapy, though this should be undertaken with caution too.". The DSU also consider the fit of the Weibull models fitted independently and using HRs, and conclude that the independently fitted models seem to fit better – but again this seems to be based purely on a visual inspection. Also, the DSU note that any ICERs based on censoring or excluding IFN patients who subsequently received sunitinib, but including the full dataset for sunitinib (11% of whom went on to receive further sunitinib) will likely be underestimated due to the potential benefits of post-progression treatment in the sunitinib arm.</p>	

	<p>In summary, using the AG's preferred assumptions (ITT for PFS for sunitinib and IFN, ITT for sunitinib OS, and the no post study treatment group for IFN OS), the Pfizer ICER was £49k and the AG ICER was £54k. This was the exclusion technique, rather than the censoring technique. The no post study treatment estimate was not used for sunitinib OS because this led to an unfeasibly large increase in OS estimated for sunitinib – which raises questions about the validity of using this patient group for either treatment arm. Sunitinib was recommended based on the end of life ruling, with respect to these £49k - £54k ICERs.</p> <p>Roche Crossover</p> <p>Regarding the bevacizumab appraisal, there is a question as to the specific population used in the survival analysis – eg the 'safety population' was used in the manufacturer submission. The DSU report confirms that this is the population who received at least one dose of the randomised treatment. The DSU state that analysis should have been based on the ITT population, for randomisation reasons.</p> <p>In their response, Roche brought up the crossover issue. The DSU were asked to look at this. Roche argued that because the benefits of post-progression treatments will effect survival estimates, their costs should also be included (which Roche did), or the survival benefits should be factored out. Roche presented an analysis censoring all patients who received second line treatments, and the OS HR fell from 0.75 to 0.61, reducing the ICER from £171k to £101k according to the AG model. The DSU also tested a scenario where the OS HR was the same as the PFS HR (0.63) as it may be considered unlikely that a treatment has a bigger effect on OS than on PFS. This gave an ICER of £107k. The DSU state: "There is a difficulty that arises from these differences in approach. The ITT overall survival analysis respects the original trial randomisation whilst the censored analysis is based on particularly small numbers of patients. The patient groups are not entirely balanced in terms of their baseline characteristics between the censored treatment and control groups or between the censored groups and the ITT population, although it is difficult to assess whether these differences should be considered significant. Furthermore, there is a risk of unobserved differences between the treatment and control censored groups influencing the estimated treatment effect." The DSU state that provision of PLD would allow a much more rigorous analysis to take place. In the absence of this several censoring scenarios should be presented so that their impact can be more fully understood. Alternatively, they state that 2nd line treatment costs could be included in the AG model, but that this may involve modelling sequences not recommended in the NHS.</p> <p>As an additional issue, the assessment group note: "When modelling treatment of RCC there are methodological challenges when using summary data (survival analysis) from clinical trials, and research to explore the impact of using aggregated data compared to individual patient-level data would be helpful"</p> <p>The AG note that "The ICERs for both drugs are particularly sensitive to variations in the estimate of the hazard ratio (HR) for overall survival (OS) from the clinical effectiveness review. This is a particularly uncertain parameter in the modelling of disease progression and cost-effectiveness, with wide confidence intervals. The ICERs are less sensitive to changes in the estimates of clinical effectiveness against PFS, and are also seen to change in a counter intuitive manner. As would be reasonably expected, when the HR for OS is reduced (greater benefit), the ICER decreases. However, when the HR for PFS is reduced (greater benefit), the ICER increases. As shown in the tables and figures this is the case for both sunitinib and bevacizumab plus IFN. This result is due to the fact that the change in effect size (HR) retains a greater proportion of patients in PFS, which has a relatively high incremental cost (drug and drug administration costs). The incremental costs in PFS outweigh the survival and QALY gains when in PFS." Results were also particularly sensitive to the OS HR in the temsirolimus evaluation. Sensitivity analysis using upper and lower OS HR confidence limits was conducted. Often, even at the lower limits the drugs were not cost-effective based on the AG model.</p>
--	--

38. TA171: Multiple myeloma - lenalidomide, June 2009

Guidance: Lenalidomide in combination with dexamethasone is recommended, within its licensed indication, as an option for the treatment of multiple myeloma only in people who have received two or more prior therapies, with the following condition:

- The drug cost of lenalidomide (excluding any related costs) for people who remain on treatment for more than 26 cycles (each of 28 days; normally a period of 2 years) will be met by the manufacturer.

People currently receiving lenalidomide for the treatment of multiple myeloma, but who have not received two or more prior therapies, should have the option to continue therapy until they and their clinicians consider it appropriate to stop.

Source: Lenalidomide for the treatment of multiple myeloma in people who have received at least one prior therapy, TA171, June 2009, <http://www.nice.org.uk/nicemedia/live/11898/44812/44812.pdf>, accessed 12/05/10

Hoyle M, Rogers G, Garside R, Moxham T, Stein K. The clinical and cost-effectiveness of lenalidomide for multiple myeloma in people who have received at least one prior therapy: An evidence review of the submission from Celgene. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, September 2008, <http://www.nice.org.uk/nicemedia/live/11937/42423/42423.pdf>, accessed 12/05/10

Hoyle M, Rogers G, Garside R, Moxham T, Stein K. The clinical and cost-effectiveness of lenalidomide for multiple myeloma in people who have received at least one prior therapy: An evidence review of the submission from Celgene: Addendum to the report submitted on 1st September 2008. Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, January 2009, <http://www.nice.org.uk/nicemedia/live/11937/43015/43015.pdf>, accessed 12/05/10

NICE, Final Appraisal Determination, Lenalidomide for the treatment of multiple myeloma in people who have received at least one prior therapy, TA171, April 2009, <http://www.nice.org.uk/nicemedia/live/11937/43868/43868.pdf>, accessed 12/05/10

Note: This was an STA – the manufacturer’s full submission is available on the NICE website (though with a large amount of CiC data removed, and the appendices are not available, which include much of the survival modelling methods information). The treatment was recommended based on a patient access scheme and the end of life ruling.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
<p>Survival data used (patient-level data, or summary statistics – mean, median etc)</p>	<p>The AG compared the actual survival experience from the MM RCTs to the modelled TTP and OS survival. They found that “The modelled time to progression for Len/Dex and Dex monotherapy both appear reasonably consistent with the trial results...As expected, the modelled overall survival for Dex is far lower than experienced in the Len RCTs due to the adjustment of post progression survival to reflect experience in the MRC trials. However, importantly, the modelled overall survival for Len/Dex is clearly higher than experienced in the Len RCTs”</p> <p>The AG went on to demonstrate that the OS PLD used in the model was different to that reported from the MM RCTs – fewer patients were used, and calculated HRs were very different.</p> <p>Also: “in an attempt to correct for the extensive cross-over of patients from Dex to Len/Dex, a factor was added to the post-progression equation for Dex to calibrate the estimated Dex overall survival to that observed in the UK Medical Research Council (MRC) myeloma trials. Their approach is to take the clinical effectiveness of Len/Dex from one group of trials (MM-009 and MM-010) and the clinical effectiveness of Dex from a different group of trials (the MRC trials), with the Dex effectiveness adjusted for the patient characteristics of the patients in the MM09 and MM10 trials. Although Celgene adjusted for differences between the MM trials and MRC trials, there will inevitably be other factors which may be unbalanced between the modelled population of Len/Dex. For example, OS of Dex may have increased from the time of the MRC trials to the time of the MM trials. Celgene state that although the MRC trial data is rather old, with patients enrolling between 1980 and 1997, they believe that the data is still appropriate to the economic evaluation because the MRC data shows no trend for improvement in overall survival over time. However, on p29 of their report, Celgene note that there was a trend in the Mayo clinic data towards improved survival during 1995 to 2000, and a statistically significant improvement in survival from 2000 to 2006. Apparently, the trend to an improvement between 1995 and 2000 coincided with increased use of high dose therapy (with stem cell transplant), which likely contributed to this change. The significant improvement in survival observed between 2000 and 2006 is believed to be due to the introduction of novel therapies. This suggests that the overall survival of patients taking Dex today may be better than calculated from the MRC data. In this case, Len/Dex may actually be substantially worse value for money versus Dex than calculated by Celgene. Given these uncertainties in basing progression-free survival for Dex on the MRC data, it would be useful to populate the cost-effectiveness model with data for Dex taken from MM-009 and MM-010 with patients who crossed over to Len censored.”</p> <p>Therefore the AG state that a censoring analysis to account for crossover may be better than the approach using external data.</p>	<p>Because of the extensive crossover in the trials, effectiveness data from MM-009 and MM-010 was supplemented with long-term follow up based on database information from MRC trials.</p> <p>“Having established the baseline cohorts for the simulation by treatment and population group, the Celgene model goes on to estimate progression (and therefore the progression free survival period) for each patient. This is achieved using one of three approaches:</p> <p>(a) Where disease progression was observed in the real patient record ascribed to the hypothetical patient, and the treatment group of the real patient is the same as that being modelled, the time to progression is simply noted from the real patient record.</p> <p>(b) Where progression had not occurred in the patient record being used for simulation, this is predicted by assuming that time to progression follows a Weibull distribution. Time to progression is estimated for that patient based on a regression of patient characteristics and best treatment response achieved.</p> <p>(c) Where progression occurred in the real patient record, but the actual treatment taken was different from that of the simulated patient, the simulated patient’s time to progression is predicted as if they had been included in the treatment group of interest.</p> <p>For example, if the real patient progressed on Dex at time T1, this corresponds to a point (P1) on the TTP survival curve denoting the probability of progression at T1. It is assumed that, had the patient been taking Len/Dex, progression would have occurred at the same probability (P1), although since expected TTP is greater on len/Dex, this would correspond to a different time of progression (T2), where T2 is greater than T1. This T2 is calculated by solving the regression equation for patient characteristics on TTP for Len/Dex, using the real patient data. In approaches (b) and (c), and all approaches for bortezomib, the Celgene model uses a regression equation to estimate progression free survival based on best response achieved with individual patient-level covariates. In the case of bortezomib, data on the proportions of patients achieving each response is taken from the APEX trial. The progression free survival equation for bortezomib is adjusted to calibrate median progression free survival to the same value as shown in the APEX trial (30.3 weeks).”</p> <p>“Broadly, similar approaches to predict post-progression survival are taken by the Celgene model as for time of progression. Post-progression survival is modelled as an exponential function of a range of predictors. Time of death is then calculated as PFS plus PPS.</p> <p>However, crossovers from Dex to Len/Dex, which occurred in 47% of patients in MM-010 and MM-009, confounds the analysis of post progression survival for Dex. Therefore some adjustment is needed, specifically of the estimated PPS, and hence time to death in the Dex group. This issue is more important in post progression than progression-free survival because most (75%) patients crossed over at or after progression. In order to correct for the confounding effect of crossovers on Dex survival, the postprogression survival equation for Dex includes an adjustment factor which calibrates the Dex group’s overall survival to that shown in the UK Medical Research Council’s trials. Inherent in this approach is that Dex confers no less benefit as the range of chemotherapeutic agents included in the MRC trials. Analyses carried out by Celgene suggest that this is the case. The calibration of Dex survival is reported in Appendix 8 of the manufacturer’s</p>

	The AG present an analysis which shows that when crossover is not accounted for in any way the ICER increases from £25k to £79k.	<p>submission. Briefly, parametric survival analysis was used on the MRC data to derive an equation for time of death, including predictors of age, m-protein, beta-2M and time to progression with first line treatments. The values of these predictors were then set to the corresponding mean values in the MM010 and MM009 trials to estimate median overall survival for MM009 and MM010 under MRC conditions i.e. in the absence of Len treatment. Celgene justify using the MRC data to model postprogression survival of Dex as follows. First, the MRC trials represent the outcomes experienced by a large population (1,372 patients for overall survival) of UK patients on treatment with Dex for multiple myeloma. Second, Celgene state that although the MRC trial data is rather old, with patients enrolling between 1980 and 1997, they believe that the data is still appropriate for the economic evaluation of Len because the MRC data shows no trend for improvement in overall survival over time. Because the Celgene model aims to predict events at an individual patient-level, the preceding step is insufficient and it is necessary to adjust the Celgene PPS equation for Dex so that it might be used to estimate individual times of death. This was achieved by adding a factor to the Dex survival equation and iteratively varying this until the Celgene estimated median OS matched that for the MRC equation, as calculated using mortality predictors from MM009 and MM010. The use of MRC data is justified by Celgene on the grounds that it is: based on a large population (n=1,372) of UK patients. Although the data is now rather old, Celgene demonstrate no secular trend for improvement in overall survival, suggesting that the MRC data is still a good representation of overall survival at initiation of second and further lines of therapy.”</p> <p>The AG state that: “TTP and PPS for the model patients for both Len/Dex and the baseline treatments (Dex and bortezomib) were based on data from MM-009 and MM-010, and the APEX RCT, which is reasonable. The baseline PPS of Dex was based on data from MRC trials. We believe that would be better to fit mean OS, not median OS from the MRC data (see Section 5.3.3.2). Baseline PPS of Dex was not taken from the RCTs of Len/Dex v. Dex due to the large number of patients originally allocated to Dex who crossed over to Len. We have criticised this approach, see Section 5.3.3.2”</p>
Evidence synthesis (pool survival estimates?)	No novel evidence synthesis.	Evidence was taken from two RCTs, but was supplemented with evidence from an external trial from which patient-level data were obtained.
Survival model(s) fitted (Weibull, exponential etc)	Attempted to re-run the manufacturer model using means as the calibration factor rather than medians – this was achieved by assuming an exponential distribution (to convert the median to a mean).	Weibull and exponential.
Independent survival models, or hazard ratio (proportional hazards) modelling	No novel analysis.	Appear to be independent.
Justification for survival model used?	<p>There is little debate on the use of Weibull and exponential models. However, the ERG state: “The adjustment of the post-progression survival equations for Dex was based on matching the median overall survival between the MRC data, using patient characteristics of the Len RCTs, with the median overall survival of Dex in the model.</p> <p>However, given that the cost per QALY of Len/Dex v. Dex equals (mean costs in Len/Dex arm - mean costs in Dex arm) / (mean QALYs in Len/Dex arm - mean QALYs in Dex arm), we suggest that it is preferable to match the mean Dex overall survivals... By matching to the median OS, we are taking no account of the tail of the distribution, or more precisely, the curve beyond the 50th percentile. Indeed, this is evident from Figure 6, where we see that the tail of the curve used by Celgene is a very poor fit to the dotted line representing the exponential distribution. When we run the model with our fit to the mean OS, the ICER for 1 prior therapy increases only marginally, but the ICER for >1 prior therapy increases substantially, from £24,600 to £33,200 / QALY. In the thalidomide-exposed subgroup, the ICER for 1 prior therapy increases only marginally, and the ICER for >1 prior therapy</p>	There is little detail on why Weibull and exponential models were used. There is a considerable amount of debate on why the manufacturer chose to use median survival from the MRC trial to calibrate survival times, rather than the mean.

	<p>increases substantially, from £22,600 to £30,200 / QALY.”</p> <p>In addition, but not noted by the ERG, the adjustment made to OS estimates to account for crossover seems to suggest that the impact of the crossover was quite similar in magnitude to the impact of randomising to the intervention in the first place – ie the OS benefit of the intervention is roughly doubled – therefore the clinical validity of these estimates could be questioned. Is the crossover of a certain % of patients really likely to alter survival by such a large amount?</p> <p>There was some debate about the AGs use of an exponential model when calibrating to the mean – the AG justify this by stating that the manufacturer carried out analyses that produced Weibull models with shape parameters close to 1 – and hence an exponential is likely to be suitable (see final box in this table for more details).</p>	
<p>Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)</p>	<p>No novel model produced.</p>	<p>The manufacturer used discreet event simulation to model the cost-utility of lenalidomide/dexamethasone and dexamethasone alone in the five subgroups. Because of the extensive cross-over in the trials, effectiveness data from MM-009 and MM-010 was supplemented with long-term follow up based on database information from MRC trials.</p> <p>“Rather than the Markov approach often used in the assessment of cost-effectiveness of terminal cancer drugs and more frequently seen in NICE submissions, Celgene use a discrete event simulation model. The discrete event approach differs from the Markov approach in that there is no time-based “cycle” in which the model predicts the occurrence of transitions between specific health states. Rather, the occurrence of events is predicted by the model based on patient characteristics and treatments received. At the predicted occurrence of progression, the model then calculates the expected time of death. The key events in the Celgene model are progression and death, which define two corresponding periods – progression-free survival and post-progression survival.”</p>
<p>Other issues noted (eg crossover)</p>	<p>The AG state that a main limitation of the trial is the crossover: “The main threat to validity for the clinical effectiveness data is the high level of crossover in the trials, leading to a strong lenalidomide effect in the comparator arm. This is likely to underestimate treatment effect, especially for overall survival.”</p> <p>“Cost-effectiveness is extremely sensitive to the estimate of overall survival with dexamethasone alone. Because of crossover in the trials, survival for dexamethasone is taken from experience in MRC trials. This breaks randomisation and, since survival data is historical, may underestimate with survival with dexamethasone.”</p> <p>In addition: “The model predicts better overall survival for lenalidomide/ dexamethasone than shown in MM-009 and MM-010 and, if adjusted to better predict trial data, incremental costeffectiveness ratio increases for all comparators.”</p> <p>The AG note the particularly long extrapolation period required in this appraisal: “A 30-year time horizon was used in the model. As survival 30 years after starting treatment is negligible (when less than 2% of patients are still alive) the time horizon is effectively lifetime. Given that data for patients receiving Len/Dex is available for a median combined follow-up of only 31.3 months, by which time over half of patients taking Len/Dex are still alive, the 30-year time horizon represents a very large extrapolation. There is therefore a great deal of uncertainty in the survival times of patients in the model. Given that the cost-effectiveness of Len is strongly affected by survival times, this introduces considerable uncertainty in the estimates of cost-effectiveness.”</p> <p>In additional analysis there was substantial debate between the AG and the manufacturer regarding the method used to estimate OS:</p> <p>“1) Celgene state that “Fitting has to do with what is most justifiable in terms of reproducing the information as accurately as possible, not with the use of the fits afterwards.”</p> <p>There are an infinite number of ways of fitting to the MRC OS dexamethasone data: e.g. to the mean, 25th percentile, 40th percentile, 50th percentile (median), 75th percentile, minimize sum squares of differences. We believe that it is important to choose the method which is most appropriate for use in the cost-effectiveness model. In this case, cost-effectiveness is driven by mean dexamethasone overall survival, therefore, we believe that the model should fit to the mean dexamethasone overall survival MRC data. Indeed, we go further and say that it would have been preferable to design the model so that the model output gave a very close fit to the exponential curve for dexamethasone from the MRC data. However, this is not possible within the structure of the model because overall survival is constrained as the sum of two distributions: a Weibull distribution in PFS and an exponential distribution in PPS.</p> <p>...</p> <p>3) Celgene state: “Based on their Figure 6 (reproduced below) it appears that their calculations are incorrect as the curve representing our submitted model should cross the exponential curve from MRC exactly at the 50% survival point (i.e., the median) and it appears to cross at about 42% instead.”</p> <p>According to Celgene’s stated aim of fitting to the median, the curve should indeed cross at the 50% percentile. However, although Celgene stated that they fitted to the median for the >1 prior therapy subgroup, in fact, the median dexamethasone overall survival is 13.3 months in their model - greater than the median of 11.6 months to which they were attempting to fit.</p>	

Therefore, Celgene's model does not fit exactly to the median. This was Celgene's error, not ours.
For the 1 prior therapy subgroup, as Celgene stated, they fitted exactly to the median of 19.5 months.

4) Celgene state: "The exponential distribution from the MRC is not likely to be the true shape as it is well known that human mortality accelerates with time, requiring either a Weibull or Gompertz fit (Román *et al.* 2007; Jucket *et al.* 1993). This was not a concern for our approach as we are not using the MRC-derived shape in the model. The only purpose of the MRC analyses was to provide a calibration point that would allow adjustment of the equations in the model to remove the cross-over effects. By calibrating to the mean produced by the MRC curves, the ERG is taking the exponential shape to be the true function of OS in multiple myeloma."

Celgene found that the exponential distribution fitted the MRC adjusted overall survival for dexamethasone very well for both the 1 prior and >1 prior patient subgroups. Indeed, in their original report, p53 of Appendix 8, Celgene state "the shape parameter of the Weibull distribution does not improve the fit of the models. In fact, the estimates of the shape parameters were 1.01 (95% CI: 0.96 – 1.07) and 0.98 (95% CI: 0.89 – 1.08) in the one prior and multiple prior groups, respectively.", and the exponential distribution has a shape parameter of 1. Therefore, although we acknowledge that human mortality sometimes requires a Weibull or Gompertz fit, as Celgene state, this is clearly not the case in this instance.

5) Celgene state: "Indeed, the ERG themselves, when adjusting the Len/dex survival calibrated to the median not the mean (pg 82)".

We did indeed adjust Len/dex overall survival to the median, not the mean. However, if the mean overall survival for the Len/dex treatment arm from the MM RCTs had been available, we would have calibrated the model to the mean, not the median. In this instance, the Len/dex overall survival was immature, with approximately 50% of patients still alive at data cutoff in the published data. Therefore, we were forced to use a different method to calibrate the modelled Len/dex overall survival. We used the next best option, and fitted to the median. Furthermore, Celgene's modelled Len/dex OS departs from the RCT OS not just at the median, but rather (as stated in our report and as can be seen in Figure 5 in our report), at all points from the 100th percentile to the 50th percentile.

6) Celgene state: "Given that the true OS distributions are right-skewed (most of the deaths happen early), calibrating to the mean ignores where most of the known deaths actually occur and overemphasizes the tail of the distribution where there are fewer patients and much more uncertainty. Thus, the accuracy of predicted survival times in the known earlier parts of the curve would be compromised to gain better fit to the less well known, much more inaccurate, tail."

As mentioned above, cost-effectiveness is driven by mean (not median) overall survival. The mean, which is the area under the survival curve, can be heavily influenced by the tail of the distribution. Conversely, the median is completely independent of the tail. Therefore, it is important to take into account the shape of the tail of the Kaplan-Meier curve. By ignoring the tail, and concentrating solely on the median of the survival distribution, one effectively discards 50% of the available data. The Kaplan-Meier curve for dexamethasone OS from the MRC data was constructed from the experience of 375 patients for the >1 prior therapy subgroup (p52 Appendix 8 Celgene submission). This is a substantial number of patients, and therefore we expect the tail of the curve to be reasonably accurate. It is true that the tail of Kaplan-Meier curves may be more uncertain than the early part of the curve, but only when there is a large amount of censoring. However, we understand that there is virtually no censoring in the 375 patients: 354 (94.4%) of the 375 patients were recorded to have died (p52 Appendix 8 Celgene submission). Therefore, Celgene's assertion that the tail of the distribution is inherently inaccurate is unsustainable. In fact, there may be more uncertainty about the median than about the tail when there is very little censoring. Furthermore, as stated above, the exponential curve fitted the data well, which implies that the exponential curve also fits the tail well."

FAD stated: "The Committee noted that the trial results included a crossover effect and considered whether it was appropriate to use data from historical MRC trials to predict survival for people treated with dexamethasone in this population in the absence of an unbiased estimate from the trials of lenalidomide. The Committee was aware that the MRC data were derived from trials of agents in first-line therapy for multiple myeloma. Despite this, it accepted that these data represented the best available survival data for people with multiple myeloma to be used in extrapolation of overall survival in the current analysis. The Committee also noted that use of these data assumed that dexamethasone monotherapy was a suitable proxy (in the absence of more specific evidence) for all anti-myeloma therapies used in relapse. The Committee also considered that there was no evidence to indicate that the effectiveness of dexamethasone in relation to survival had changed over time since the MRC trials. It accepted the statements from the clinical specialists indicating that where improvements were noticed these were likely to be attributable to the use of the newer agents and stem-cell transplantation."

Also: "The Committee considered the ERG's exploratory reanalysis with an improved fit of the len/dex overall survival curve to the trial data and calibration of the dexamethasone overall survival curve to predict mean (and not median) overall survival based on a risk equation for survival derived from the MRC trials. The Committee considered that the ERG's approach to modelling overall survival in both the len/dex and dexamethasone arms was valid and resulted in more plausible estimates of cost-effectiveness than those presented by the manufacturer. The Committee noted that these adjustments to the modelling of survival may have different effects in different subgroups and that the ERG's adjustments had been made separately to subgroups defined according to number of prior therapies."

Calibration to mean or median from the MRC trials was an important issue. The AC concluded that calibration to the median led to an improvement in OS predicted by the model that was out of proportion considering the improvement seen in PFS. However they also believed this to be the case when OS was calibrated using the mean, but that this calibration was preferable.

39. TA172: Head and neck cancer (squamous cell carcinoma) - cetuximab, June 2009

Guidance: Cetuximab in combination with platinum-based chemotherapy is not recommended for the treatment of recurrent and/or metastatic squamous cell cancer of the head and neck.

People currently receiving cetuximab in combination with platinum-based chemotherapy for the treatment of recurrent and/or metastatic squamous cell cancer of the head and neck should have the option to continue treatment until they and their clinician consider it appropriate to stop.

Source: Cetuximab for the treatment of recurrent and/or metastatic squamous cell cancer of the head and neck, TA172, June 2009, <http://www.nice.org.uk/nicemedia/live/12179/44644/44644.pdf>, accessed 13/05/10

Greenhalgh J, Bagust A, Boland A, Fleeman N, McLeod C, Dundar Y, Proudlove C, Shaw R. Cetuximab for recurrent and/or metastatic squamous cell carcinoma of the head and neck (SCCHN): ERG Report, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, November 2008, <http://www.nice.org.uk/nicemedia/live/11987/43053/43053.pdf>, accessed 13/05/10

Merck Serono, Response to NICE STA Questions; Cetuximab for Recurrent or Metastatic SCCHN, 24th October 2008, <http://www.nice.org.uk/nicemedia/live/11987/42927/42927.pdf>, accessed 13/05/10

Note: This was an STA – the manufacturer’s full submission is available on the NICE website.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The AG states that: “Evidence from the EXTREME trial demonstrated the clinical effectiveness of cetuximab plus CTX over CTX for the trial period. The manufacturer failed to provide adequate information regarding projective modelling. As a result, the ERG was not able to test its concerns about some of the assumptions used by the manufacturer to extrapolate costs and benefits”</p> <p>The AG also questions whether modelling was required because the clinical data included OS data for 75-80% of patients (not clear if they are taking into account censoring here).</p>	<p>The clinical data used in the economic evaluation are generated from the EXTREME trial. Although the economic evaluation is trial-based, there is also a modelling component with regard to the extrapolation of health effects beyond the period of the trial (24 months).</p> <p>The distribution of patients over the three health states over time was imputed using Weibull survival models for both PFS and OS estimated using individual patient data (IPD) from the EXTREME trial. No transition probabilities were calculated to describe the distribution of health states over time, these being implicit within the parametric survival functions. The fitted (and extrapolated) Weibull survival curves describe the proportion of patients in each health state at the beginning of each three week cycle.</p> <p>As the trial data for OS and PFS were censored at 24 months and do not provide full information on the course of disease for patients still alive at that time, the manufacturer chose to extrapolate both outcomes beyond the trial period. In particular PFS curves were extrapolated to inform the transition from the stable/response health state to either progressive or death health states.</p>
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	A type of restricted means analysis was undertaken by the AG, although it appears that this was still based on the Weibull models fitted by the manufacturer, just truncated at 24 months. However, the manufacturer did provide the AG with patient-level data at a later date, and so it is possible that a restricted means analysis based on the KM curves was conducted.	Weibull, but log-normal and log logistic were also tested, and the Weibull was chosen based on it minimising the log likelihood.
Independent survival models, or hazard ratio (proportional hazards) modelling	Appear to be independent, based upon the manufacturer’s analysis.	Independent.
Justification for survival model used?	The AG state: “The MS did not provide an adequate explanation of why the Weibull function was chosen for all survival models in the base case, as well as in all six subgroup analyses. It is normal to carry out comparative model-fitting exercises for a range of candidate models, using objective criteria for assessing suitability on statistical grounds. Moreover, graphical evidence of the appropriateness of the fitted models was only provided in relation to the base case overall analysis. Further detailed information was requested by the ERG via the original letter of clarification and charts were then provided by the manufacturer showing the	The choice of the Weibull function was said to be based on two assessments: (i) goodness-of-fit and (ii) clinical expertise for the estimated values for time points after the evaluation period – the clinical validity aspect was on the basis that log normal and log-logistic models resulted in heavy tails which were deemed unrealistic. (note, as the AG point out, the KM curves do appear complete, although numbers at risk aren’t given).

	<p>Weibull model survival function superimposed on the Kaplan-Meier survival curve for each of the populations used in the submitted economic model. However, none of the requested supporting statistical information from the Kaplan-Meier analyses was made available by the manufacturer, notably the estimated mean survival with confidence limits, and the number of patient records included in each analysis. The manufacturer explained that the Weibull model was found to fit the available data slightly better than either the lognormal or the log-logistic functions for the overall trial population, and to provide more clinically realistic projected mean survival values. Visual examination of the subgroup charts suggests that there may be a systematic mismatch between the Weibull model and the observed data in the middle time period (150-400 days), especially for PFS, and that this may result in an overestimate of the mean expected PFS in both control and intervention arms, though it is not clear whether this would seriously impact on the incremental PFS. It is apparent that there are small numbers of patients within some of the subgroups and that model-fitting in these instances may not be reliable.”</p>	
<p>Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)</p>	<p>No novel model produced.</p>	<p>A two-arm state-transition Markov model was developed by the manufacturer to evaluate the cost-effectiveness of cetuximab plus CTX compared to CTX. The course of disease is reflected by three mutually exclusive health states (stable/response; progressive; death). Cycle length is 3 weeks. The clinical data used in the economic evaluation are generated from the EXTREME trial. Although the economic evaluation is trial-based, there is also a modelling component with regard to the extrapolation of health effects beyond the period of the trial (24 months). The economic evaluation adopts a lifetime horizon. The manufacturer reports an incremental cost-effectiveness ratio (ICER) of £121,367 per quality adjusted life year (QALY). The ERG made several corrections and/or adjustments to the model logic and parameter values. In general, the combined effect of ERG corrections and/or adjustments yielded less favourable economic results for cetuximab than described in the MS. The highest ICER estimated by the ERG for the amended base case was £208,266 per QALY</p>
<p>Other issues noted (eg crossover)</p>	<p>The AG state that: “The MS (pg74) states that “...the OS and PFS curves as observed in the trial were extrapolated by fitting 2-parameter Weibull survival curves to the empirical patient-level data. The scale and shape parameters of the Weibull distribution were estimated with least-square regression methods.” The ERG notes that these two statements appear to be contradictory, or at best confusing. From additional information provided by the manufacturer regarding model fitting, it seems that the Maximum Likelihood method was used for model fitting (the normal procedure when analysing IPD), rather than least squares minimisation (more commonly employed when only aggregate Kaplan-Meier analysis results are available).”</p> <p>The AG also question the use of extrapolation, which may be controversial: “the use of simulation derived from a single data source rather than employing the observed data directly is vulnerable to challenge, since the modelling process introduces additional uncertainty to that already inherent in the trial itself. Moreover, there is little to be gained by evidence synthesis in this situation since there is only one source of outcomes data. The potential problem with projective modelling of OS in such a case is that it is more likely to exaggerate health gains than to underestimate them, leading to an overoptimistic result. This is particularly relevant for late-stage disease where no claim is made to provide a cure, merely to modify the timing and process of progression. In such cases, benefit often takes the form of a limited period of reduced risk after which disease progression resumes as before, so that virtually all the outcome gain will have been realised well before the final patient dies.”</p> <p>“It can be argued that in some cases projection modelling of outcomes and costs may not be appropriate, especially where the Kaplan-Meier survival curves have converged closely, and there is no <i>a priori</i> reason to expect them to diverge significantly at a later time. Under such circumstances, truncating the analysis at the point when the trial was terminated may be considered necessary to avoid the risk of spurious artefactual differences arising from ill-advised projection. An analysis was carried out by the ERG using the submitted base case model to compare costs and outcomes at 24 months (end of the follow-up period). Both net discounted incremental costs per patient and incremental patient utility were considerably reduced (-£526 per patient and -0.029 QALYs per patient) resulting in a large increase in the estimated ICER from £121,367 to £147,817 per QALY gained. This indicates the sensitivity of cost-effectiveness estimates to assumptions concerning projection modelling.”</p> <p>An additional issue surrounded the modelling of PFS and OS and correlations: “Furthermore Weibull parameters for OS and PFS have been estimated separately from the same patient data. It is highly likely that OS and PFS will be strongly positively correlated (since PFS is part of OS), so that covariance between the two sets of Weibull parameter estimates cannot be ignored and, ideally, model parameters for PFS and OS should be jointly estimated. The manufacturer has confirmed that OS and PFS models were developed independently in all cases. The central estimate of the cost-effectiveness ratio may not be affected to any great degree by this problem, but any assessment of the associated uncertainty will not be trustworthy, since variance in the distribution of uncertainty is most probably overstated within the PSA.”</p> <p>Crossover was not mentioned, and the trial protocol seems to suggest that it would not have happened.</p>	

40. TA174: Leukaemia (chronic lymphocytic, first line) - rituximab, July 2009

Guidance: Rituximab in combination with fludarabine and cyclophosphamide is recommended as an option for the first-line treatment of chronic lymphocytic leukaemia in people for whom fludarabine in combination with cyclophosphamide is considered appropriate.

Rituximab in combination with chemotherapy agents other than fludarabine and cyclophosphamide is not recommended for the first-line treatment of chronic lymphocytic leukaemia.

Source: Rituximab for the first-line treatment of chronic lymphocytic leukaemia, TA174, July 2009, <http://www.nice.org.uk/nicemedia/live/11907/44906/44906.pdf>, accessed 13/05/10

Main C, Pitt M, Moxham T, Stein K. The clinical and cost-effectiveness of rituximab for the 1st line treatment of Chronic Lymphocytic Leukaemia: an evidence review of the submission from Roche, Peninsula Technology Assessment Group, Universities of Plymouth and Exeter, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, January 2009, <http://www.nice.org.uk/nicemedia/live/12039/43572/43572.pdf>, accessed 13/05/10

Roche Products Ltd, Rituximab for the first-line treatment of chronic lymphocytic leukaemia, NICE STA submission, November 2008, <http://www.nice.org.uk/nicemedia/live/12039/43581/43581.pdf>, accessed 13/05/10

Roche Products Ltd, Rituximab for the first-line treatment of chronic lymphocytic leukaemia, Clarification Letter, December 2008, <http://www.nice.org.uk/nicemedia/live/12039/43582/43582.pdf>, accessed 13/05/10

NICE, Final Appraisal Determination, Rituximab for the first-line treatment of chronic lymphocytic leukaemia, TA174, June 2009, <http://www.nice.org.uk/nicemedia/live/12039/44601/44601.pdf>, accessed 13/05/10

Note: This was an STA – the manufacturer's full submission is available on the NICE website.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	The AG reviewed the manufacturer's analysis – they did not use any additional data. They discussed several issues which are included in the final row of this table.	<p>Effectiveness parameters for the model are derived from the CLL-8 trial data and a multiple treatment comparison is used to derive a hazard ratio value for R-FC v. chlorambucil.</p> <p>“For the R-FC v FC comparison, the probability of patients remaining in the PFS state in each arm of the model is time-dependant and, in the base case, derived from a Weibull parameterisation of the PFS survival rates taken from the respective arms of the CLL-8 trial. The probability of death within the PFS state is calculated as the maximum value of two sources; 1) the monthly probability experienced by patients in the CLL-8 trial (0.001196 for R-FC and 0.001388 for FC) 2) the age specific background mortality rate taken from UK life tables.”</p> <p>So, a Weibull is used for PFS. For the transition to death from PFS the age specific background mortality rate is used to ensure that the mortality rate for patients in this state in the model does not fall below that of the average UK population which seems clinical implausible. The very small difference in monthly mortality rate from PFS between the two arms found in the CLL-8 contributes only a negligible benefit advantage to R-FC in the model.</p> <p>“The comparison of R-FC v. chlorambucil deploys the multiple treatment comparison... to derive an estimate for the hazard ratio of 0.24 between the arms in the model for transition from PFS to Progressed state. Mortality rates, both from the PFS state and the Progressed state, in this comparison use the equivalent values to the RF-C v. FC comparison.”</p> <p>Roche present a KM style curve showing post progression survival, which shows the curves very close together and regularly crossing – this is useful and is different from simply showing the OS curves.</p> <p>And for PP to death: “Patients in the Progressed state were modelled as a single population, with no distinction in the probability of death between arms of the model. The rationale given for this is that there was no significant difference found between arms in the CLL-8 trial. For this probability a constant value was calculated based on the inverse of the mean from the CLL-8 trial Kaplan-Meier.</p>

		This resulted in a monthly probability of dying in this state of 0.0405. This probability of death from the Progressed state is used throughout the time horizon of the model.”
Evidence synthesis (pool survival estimates?)	No novel synthesis.	MTC for one comparator.
Survival model(s) fitted (Weibull, exponential etc)	No novel models fitted, but the AG did suggest the manufacturer complete further analyses around scenarios relating to the duration of the treatment effect. In response, the manufacturer ran scenarios whereby the probability of death was equal in the Progressed health state no matter the initial treatment, and also an analysis whereby zero OS gain was assumed.	Weibull for PFS, but exponential, log logistic, log normal and gompertz tested. For the transition from progression to death the inverse of the mean from the KM was used – this was a constant hazard, which suggests an exponential distribution.
Independent survival models, or hazard ratio (proportional hazards) modelling	Not applicable.	PH for PFS (two different Weibull's but with the same shape), and a type of PH modelling used for the indirect comparison.
Justification for survival model used?	The AG state that: “Given the importance of this parameter [PFS] in the model, Roche assess a range of different curve parameterisations for PFS survival curve for goodness of fit. They demonstrate a clear rationale for the adoption of a Weibull fit in preference to several alternative parametric survival functions: Exponential, Log Logistic, Log Normal and Gompertz (p. 115-117 of the submission). In general, we are satisfied that Roche have used an appropriate method both (a) to extrapolate progression-free survival from the clinical data and (b) to determine the best curve parameterisation.” Therefore the AG accepted the Roche justifications, but these weren't discussed further.	The manufacturer's submission states: “each function was assessed for its goodness of fit to the data using Akaike (AIC) and Bayesian Information Criteria (BIC), the mean squared deviance and graphical inspection of fit (e.g., Martingale residuals) to the data before deciding on the final functional form. The parametric model structures assessed for goodness of fit to the data were: Log Logistic, Weibull, Log Normal, Gompertz and Exponential.”. It is also stated that: “The same shape Weibull function was found to be the best fit to the PFS data. Independently shaped parametric models are assessed whenever there is an indication that the shape of the treatment arms differ. There was no indication of differences in the shapes of the treatments and no violation of the underlying assumption of proportional hazards was noted in the diagnostics (e.g. Martingales) plots. Thus a same shape Weibull model was selected as the best fit parametric function to model the PFS data” Thus a type of PH modelling was used, and justified. And: “The decision for the Weibull function was based on the AIC / BIC for PFS and graphical inspection of the fit. Mean squared deviation (MSD) is also reported so that some assessment of goodness of fit can be assessed for the Gompertz function. The SAS institute is developing a procedure to assess the Gompertz function and report the value of the likelihood which can then be assessed for fit using AIC and BIC methods.”
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	No novel model produced.	Roche uses a Markov model, separating the disease process into the three states; Progression-Free Survival (PFS), Progressed, and Death, to analyse the cost-utility both of R-FC v. FC and R-FC v. chlorambucil. The modelling approach adopted by Roche uses a Markov state-transition cost-utility model implemented in Microsoft Excel. Cycle length was one month and a 'lifetime' time horizon (equating to 15 years after which less than 1.5% of patients survive). All patients are assumed to start in the PFS state (defined with reference to the CLL-8 study). After each cycle a patient in PFS will experience one of three outcomes; 1) remain in the PFS state, 2) move to the Progressed state, or 3) die. Once within the Progressed state patients will either remain in this state or die. Death is an absorbing state. There is no provision within the model for patients to move back to the PFS state after moving to the Progressed state. The Progressed state therefore acts to aggregate all events subsequent to first treatment relapse (except death). Estimated cost per QALY for the R-FC v. FC comparison in the base case is £13,189 and for R-FC v. chlorambucil comparison the base case is £6,422.
Other issues noted (eg crossover)	The AG state that a key area of uncertainty is: “It is unclear whether the observed treatment benefit for use of rituximab combination therapy for PFS is associated with longer-term gains in overall survival and how plausible it is to extrapolate any PFS benefits in the longer term.” As a check on their results the manufacturer considered a similar trial to that upon which their model was based: “Roche report external validation of their model with reference to a comparison with the MD Anderson study reported in Tam <i>et al</i> /2008. This study, in common with CLL-8, compared patients treated with R-FC v. FC. The Roche model was modified based on the data from this study, the main difference being the post-progression probability of death, which was four times lower in the MD Anderson study than that observed in the CLL-8 trial. Roche report that the incremental life years and QALYs predicted by their economic model is broadly consistent when populated with either of these data sets (i.e. MD Anderson study or CLL-8).” – thus the ICER remained about the same but the life expectancy was quite different.	

	<p>The AG note that the manufacturer shows that the change to the model that creates the biggest changes in the ICER are altering the PFS parametric model to a Gompertz or exponential (eg Weibull gives an ICER of £13k, and log normal, log logistic and exponential all range between £10k (exp) and £13k, but the Gompertz increases the ICER to £23k).</p> <p>The AG suggest that the manufacturer's model is too simplistic, due to the relatively long survival of patients and only 3 Markov states. This is a particular problem for the all-encompassing post progression state: "The aggregation is clinically unrealistic because patients will receive further treatment at progression which may then result in further periods of progression-free survival. The relapsing nature of CLL means that subsequent periods of progression are less likely to respond to further treatment, implying that later periods of progression in the course of disease are likely to be associated with higher disease-related mortality. This casts doubts over the simplifying assumption of a constant hazard of death after progression as modelled by Roche. This assumption was not confirmed through exploration of the CLL-8 dataset, as was done comprehensively for the modelling of PFS. The overall effect of the aggregated Progressed disease state and constant hazard of death from this state is to imply a correlation between PFS and OS which we do not believe has been empirically demonstrated."</p> <p>This continues: "Following queries on this subject by the ERG, Roche have carried out an analysis which differentially increases the probability of death for the R-FC arm for the Progressed state of the model. The outcome of this analysis shows the effect of removing the impact of OS advantage in the R-FC arm of the model. At the limit of this analysis (where no OS advantage is generated by the model structure) the ICER rises to £30,336. At this limit it should be noted that the ICER becomes very sensitive to the level of utilities used for the PFS and Progressed states of the model."</p> <p>The AG assessed the proportion of the QALY gain that was being derived from PFS and the proportion derived from OS. The majority was derived in OS. This was mainly because control patients were reaching progression before intervention patients, at which point the risk of death increased by a lot. Once in this state the risk of death was equal with treatment had been with the intervention or the control. Crossover is not mentioned, but this assumption rather than a reliance on OS data from the trial would seem to avoid any crossover issues. However, the AG were concerned about the relationship between PFS and OS that this modelling method resulted in. This led to the additional analysis by Roche demonstrating the ICER when all OS gains are removed (by increasing the risk of death upon progression by 315% in the intervention arm. The AG recalculated by reducing the the probability of death in the control arm to equalise OS, and came up with similar ICERs as Roche.</p> <p>Crossover onto the alternative treatment in the CLL8 trial was not allowed, but after 3 cycles any patient with stable or progressive disease could be treated with any other treatment deemed appropriate (rituximab containing regimens were allowed).</p> <p>In the FAD it was noted that: "The Committee noted that an interim analysis of the clinical trial results had demonstrated a statistically significant gain in overall survival but this gain had not been maintained during longer follow-up. The Committee accepted that crossover and subsequent lines of treatment in the trial made the overall survival benefit difficult to prove. The Committee heard expert opinion that the degree of response to treatment and the duration of progression-free survival were generally accepted as surrogates for overall survival. In addition, the Committee heard that cohort studies using historical controls had also shown survival benefits for people treated with rituximab-containing regimens, although results may have been influenced by changing clinical management, such as earlier identification of people with chronic lymphocytic leukaemia. On balance, the Committee was persuaded that the benefits observed in progression-free survival and response rate were likely to lead to a gain in overall survival, although currently this would be difficult to quantify."</p> <p>Note, Roche provided the data behind the KM curves in subsequent correspondence.</p>
--	--

41. TA176: Colorectal cancer (first line) - cetuximab, August 2009

Guidance: Cetuximab in combination with 5-fluorouracil (5-FU), folinic acid and oxaliplatin (FOLFOX), within its licensed indication, is recommended for the first-line treatment of metastatic colorectal cancer only when all of the following criteria are met:

- The primary colorectal tumour has been resected or is potentially operable.
- The metastatic disease is confined to the liver and is unresectable.
- The patient is fit enough to undergo surgery to resect the primary colorectal tumour and to undergo liver surgery if the metastases become resectable after treatment with cetuximab.
- The manufacturer rebates 16% of the amount of cetuximab used on a per patient basis.

Cetuximab in combination with 5-FU, folinic acid and irinotecan (FOLFIRI), within its licensed indication, is recommended for the first-line treatment of metastatic colorectal cancer only when all of the following criteria are met:

- The primary colorectal tumour has been resected or is potentially operable.
- The metastatic disease is confined to the liver and is unresectable.
- The patient is fit enough to undergo surgery to resect the primary colorectal tumour and to undergo liver surgery if the metastases become resectable after treatment with cetuximab.
- The patient is unable to tolerate or has contraindications to oxaliplatin.

Patients who meet the above criteria should receive treatment with cetuximab for no more than 16 weeks. At 16 weeks, treatment with cetuximab should stop and the patient should be assessed for resection of liver metastases.

People with metastatic colorectal cancer with metastatic disease confined to the liver who receive cetuximab should have their treatment managed only by multidisciplinary teams that involve highly specialised liver surgical services.

Source: Cetuximab for the first-line treatment of metastatic colorectal cancer, TA176, August 2009, <http://www.nice.org.uk/nicemedia/live/12216/45198/45198.pdf>, accessed 14/05/10

Meads C, Round J, Tubeuf S, Moore D, Pennant M, Bayliss S, McCabe C. Cetuximab for the first-line treatment of metastatic colorectal cancer, ERG Report, West Midlands Health Technology Assessment, University of Birmingham, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, July 2008, <http://www.nice.org.uk/nicemedia/live/11918/42075/42075.pdf>, accessed 14/05/10

Merck Serono Ltd. Single Technology Appraisal Submission: Erbitux (cetuximab) for the first-line treatment of metastatic colorectal cancer, NICE STA submission, September 2008, <http://www.nice.org.uk/nicemedia/live/11918/42095/42095.pdf>, accessed 14/05/10

Merck Serono Ltd. Response to NICE ACD: Cetuximab in the treatment of metastatic colorectal cancer, Response to NICE ACD, January 2009, <http://www.nice.org.uk/nicemedia/live/11918/42962/42962.pdf>, accessed 14/05/10

Note: This was an STA – the manufacturer’s full submission is available on the NICE website, but the health economics appendices, where much detail of the survival analysis is said to be, is not available.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The AG state: “The submission includes claims about recent advances in surgical technique involving the alternative of resection of liver metastases following chemotherapy. At 13 weeks in the model, some patients can be referred for curative-intent resection of liver metastases.” It was noted that this was very important for the economic analysis.</p> <p>Unusually the AG did not discuss in any detail the survival analyses included in the economic model.</p>	<p>The clinical effectiveness consisted of two unpublished trials of cetuximab plus chemotherapy vs chemotherapy alone in the first line treatment of metastatic colorectal cancer. The CRYSTAL trial enrolled 1217 patients with EGFR expressing mCRC and the combination chemotherapy was FOLFIRI. The OPUS trial enrolled 337 patients with previously untreated EGFR-expressing mCRC that were not resectable with curative intent and the chemotherapy used was FOLFOX4. Full follow up in the CRYSTAL trial, K-RAS wild type subgroup was given at 16 months and there were six patients remaining in the intervention arm and three in the control arm. For the OPUS trial K-RAS subgroup, the equivalent numbers at 12 months follow up were four patients in the intervention arm and two in the control arm.</p> <p>For each trial, a subgroup of liver metastases from the K-RAS wild type subgroup was also presented. For the CRYSTAL and OPUS trials respectively, the full ITT populations, K-RAS wild type subgroups and liver metastases subgroups of subgroups were 1198, 348, 67 and 336, 134 and 38. The economic model clinical effectiveness was based on the liver metastases subgroups of subgroups.</p> <p>Liver resection is key to the economic model – resection rates were very low in the key RCTs, but some other observational data and expert opinion were used to back up rate claims.</p> <p>The manufacturer states: “If a patient is appropriate for resection of liver metastases with curative intent then this surgery can be successful or unsuccessful. After successful curative surgery, patients are considered tumour-free and will have longer estimated mean life expectancy of 4.76 years, with the observed median survival time of 3.23 years (as estimated from Adam <i>et al</i> [2004]).”</p>
Evidence synthesis (pool survival estimates?)	None.	Two RCTs, but with two different combination treatments, so it doesn’t appear that synthesis of these data occurred.
Survival model(s) fitted (Weibull, exponential etc)	Unusually there is no discussion of survival analysis methods in the AG report. Based on titles of tables taken directly from the manufacturer’s submission it appears KM curves were used for the trial period, with Weibull extrapolations for the tails. However in some analyses it appears that a log normal extrapolation was used.	<p>According to the manufacturer’s submission: “The economic model simulates a sequence of treatment lines using the data available from literature and from a third-line cetuximab trial (Jonker <i>et al</i> 2007). The modelling of subsequent lines of treatment was used because the CRYSTAL trial has not yet generated mature overall survival data. In order to model the short and long-term outcomes of treatment in patients within the cost-effectiveness model, survival analyses have been undertaken as a means of estimating the survival benefits of cetuximab, as follows:</p> <ul style="list-style-type: none"> • The first line is based on the results of CRYSTAL and OPUS for 1st-line progression and death before progression • The second line is based on the results from Tournigand <i>et al</i> [2004] (for 2nd line progression-free survival) • The third line is based on the results from the Jonker <i>et al</i> 3rd-line trial (for 3rd line progression-free survival)” <p>The manufacturer’s submission makes clear that a Weibull model was used to model PFS from the CRYSTAL trial – it appears that this used a PH technique, although this is not clarified. For the OPUS trial a lognormal model was used.</p>

		<p>The range of survival models used in the model are not well explained, but the manufacturer states the type of model used for each transition;</p> <p>1st line PFS to death: Log normal (CRYSTAL) 1st line PFS to progressed: Weibull (CRYSTAL); Log normal (OPUS) 2nd line time to progression: Log normal (RCT by Tournigand <i>et al</i>) 3rd line time to progression: Weibull (RCT by Jonker <i>et al</i>) Post curative surgery survival: Log logistic (from trial by Adam <i>et al</i>) PFS following curative surgery: Log logistic (from trial by Adam <i>et al</i>)</p> <p>For each the 'most suitable' modelling technique was used, but this was not discussed. More detail on the survival methods were said to be in the health economic appendices of the manufacturer submission, but these are not available on the NICE website.</p> <p>As a sensitivity analysis, the manufacturer tested using the KM curves for the trial period, followed by the Weibull extrapolation after this in the model, rather than using the Weibull for the entire period. This was found to reduce the ICER from £69k to £54k for the folfiri model. Also, when a Weibull instead of the log normal was used for the Folfox model the ICER increased from £63k to £70k.</p>
Independent survival models, or hazard ratio (proportional hazards) modelling	None.	Appears to be PH modelling for 1 st line PFS. Appears likely that future transitions did not depend on initial treatment.
Justification for survival model used?	None.	Manufacturer states that the Weibull and log normal models fitted the respected data from the two trials best, but it is not stated how this was assessed.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	No novel model produced. The AG were concerned that the manufacturer extrapolated results to approximately 20 years, even though the survival rate at 5 years based on registry data is only about 12%. The AG re-ran the model with a 5-year time frame and the ICER increased from £69k for the folfiri combination and £63k for the folfox combination to £124k and £143k respectively.	The de novo model examined the cost-effectiveness of cetuximab in patients with mCRC that is EGFR positive, K-RAS wild type and with liver metastases. The model is a time dependent state transition (Markov) model with a cycle length of one week and a 23-year time horizon (1,200 cycles). The model is quite complex, modelling 1 st , 2 nd and 3 rd line treatment.
Other issues noted (eg crossover)	<p>Although not mentioned by the AG, crossover and subsequent lines of treatment were an issue in the study, as pointed out by the manufacturer. They state: "As patients would be treated on an individual basis with subsequent lines of chemotherapy following their study medication, and that this would include further exposure to cetuximab in both arms, so this would mask any overall survival benefit from treatment with cetuximab in the initial therapy, in terms of overall survival. As can be seen from the following table, nearly two thirds of patients received subsequent chemotherapy and approximately a quarter received EGFR antibody therapies subsequent to their study drug... Of interest is the observation that the use of EGFR antibody therapy in subsequent treatment of patients was greater in those who had received FOLFIRI alone during the study compared with those that had received cetuximab and FOLFIRI. This imbalance will have contributed to the observation that overall survival was not statistically significantly different between the two groups." This is with respect to the CRYSTAL study, but is also likely to be relevant for the OPUS study. The method of modelling 2nd and 3rd line treatments using other trials may help avoid this problem, but that is dependent on the studies used for those lines not including crossover – which may not be the case. The majority of debate on this Appraisal concerned liver resection rates. Very little emphasis (particularly in the AG report) was placed on the survival analysis methodology in any documents. However there was some discussion on amending PFS curves when a treatment stopping rule was introduced, based on a DSU report.</p>	

42. TA179: Gastrointestinal stromal tumours - sunitinib, September 2009

Guidance: Sunitinib is recommended, within its licensed indication, as a treatment option for people with unresectable and/or metastatic malignant gastrointestinal stromal tumours if:

- imatinib treatment has failed because of resistance or intolerance, and
- the drug cost of sunitinib (excluding any related costs) for the first treatment cycle will be met by the manufacturer.

The use of sunitinib should be supervised by cancer specialists with experience in treating people with unresectable and/or metastatic malignant gastrointestinal stromal tumours after failure of imatinib treatment because of resistance or intolerance.

Source: Sunitinib for the treatment of gastrointestinal stromal tumours, TA179, September 2009, <http://www.nice.org.uk/nicemedia/live/12233/45513/45513.pdf>, accessed 14/05/10

Bond M, Hoyle M, Moxham T, Napier M, Anderson R. The clinical and cost-effectiveness of sunitinib for the treatment of gastrointestinal stromal tumours: a critique of the submission from Pfizer, ERG Report, Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, January 2009, <http://www.nice.org.uk/nicemedia/live/12040/43430/43430.pdf>, accessed 14/05/10

NICE, Final Appraisal Determination, Sunitinib for the treatment of gastrointestinal stromal tumours, TA179, August 2009, <http://www.nice.org.uk/nicemedia/live/12040/45125/45125.pdf>, accessed 14/05/10

Sunitinib for GIST: additional notes for the ACD meeting from PenTAG, March 2009, <http://www.nice.org.uk/nicemedia/live/12040/43431/43431.pdf>, accessed 6/6/12

Note: This was an STA – the manufacturer’s full submission is available on the NICE website. The treatment was recommended due to the end of life ruling, on the basis of a most likely ICER of £32k.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	The AG reviewed the manufacturer’s analysis – they did not use any additional data. They discussed several issues particularly around the use of the RPSFTM method, which are included in the final row of this table.	<p>“The evidence for this submission is based on one RCT (Demetri <i>et al.</i>, 2006/08), which compares the effects of sunitinib with placebo for people with unresectable and/or metastatic GIST after failure of imatinib due to resistance or intolerance and with Eastern Cooperative Oncology Group (ECOG) progression status 0-1 (the most physically able).”</p> <p>“Treatment effectiveness is taken exclusively from the RCT of sunitinib versus BSC (Demetri <i>et al.</i> 2006). In all analyses, Pfizer consider PFS for sunitinib and BSC and overall survival for sunitinib based on the ITT data. For the base case Pfizer adjust the ITT overall survival for BSC using the RPSFT method... In a sensitivity analysis, they use the unadjusted ITT overall survival for BSC. Effectiveness data were not taken from the interim analysis of the RCT reported in Demetri <i>et al.</i> (2006)(Demetri <i>et al.</i> 2006). Instead, longer follow-up survival data from the RCT was used. These data were similar, but not exactly the same, as that reported in abstract form (Demetri <i>et al.</i> 2008). Follow up for the interim overall survival data is about 1 year, whereas follow up is about 4.5 years for the mature data used in Pfizer’s model. Weibull curves were fitted to the Kaplan-Meier data by regression, (see p64 of the Submission). Pfizer considered two methods for fitting the PFS and overall survival curves for sunitinib; A. Weibull curves for sunitinib and BSC were fitted independently. B. The Weibull curve for BSC was fitted, and the Weibull curve for sunitinib was calculated by multiplying the Weibull parameter λ for the BSC curve by the hazard ratio. Pfizer found that Method B did not give a good fit to the sunitinib Kaplan-Meier curve (Figures 13 and 14, Submission p64-65). Therefore, they used Method A in the base case analysis, and Method B in sensitivity analyses.”</p>
Evidence synthesis (pool survival estimates?)	None.	None.
Survival model(s) fitted (Weibull, exponential etc)	<p>As an additional analysis the AG ran the model assuming that there were no treatment benefits after disease progression – ie the probability of death is the same after progression no matter the original treatment. This led to an ICER of £47k, compared to the base case of £27.3k and the ITT analysis of £77k.</p> <p>The AG discuss the Weibull curve fitting, and conclude it is good. However, one data point per month is used rather than all of the data, to prevent early data from dominating the curve fit – the AG do not consider that this may not be ideal, since the data should drive the curve fit: “Given that cost-effectiveness is heavily influenced by clinical effectiveness, it is important that the Weibull survival curves have been fitted correctly to the Kaplan-Meier data. In short, we are satisfied that this is the case. The Weibull curves were fitted by linear regression to one data point per month to improve the fit to the actual data by preventing the first few data points in the trial data, at times less than one month, from dominating the fit (Figure 8, Figure 9).</p>	<p>Weibull models. The control group OS estimates are adjusted using the RPSFT method.</p> <p>Regarding the RPSFT, the following details are given: “To explore potential confounding influence of crossover, a <i>post-hoc</i> analysis of overall survival was recently published (Demetri <i>et al.</i>, 2008). This analysis was performed using rank preserved structural failure time model (RPSFT) method (Robins & Tsiatis 1991). The RPSFT method estimates the true treatment effect, even in the presence of non-random non-compliance, i.e. the effect that would be realised if all individuals complied with the treatment protocol to which they were assigned, while preserving the unbiased test of the null hypothesis available from the ITT analysis. (Submission p29)”</p> <p>“statistical expert advice (personal communication from Ian White, MRC Biostatistics Unit, University of Cambridge, 27th November 2008), confirms that in this situation the correct analytical approach is to use the RPSFT method used by Pfizer, as outlined below. It should be noted that this method was used post-hoc when the Kaplan-Meier ITT analysis had failed to show a benefit from sunitinib. In addition, although White had some oversight of the methods and commented that the results ‘look about right’, he did not actually conduct the analysis (and received no remuneration for his advice to Pfizer). Therefore, although it is the correct approach, we cannot be certain that the methods were applied correctly. In particular the</p>

	<p>We found similar ICERs when we used an alternative method of fitting Weibull curves to the Kaplan-Meier data by minimising the sums of squares. Fitting PFS data by minimising the sums of squares gives an ICER of £27,600 per QALY assuming the first cycle of sunitinib is free, and fitting overall survival data by sums of squares gives an ICER of £27,000 per QALY. These figures are very similar to Pfizer's base case of £27,400 per QALY assuming the first cycle of sunitinib is free."</p>	<p>RPSFT method:</p> <ul style="list-style-type: none"> • Estimates the times of death of patients randomised to placebo as if they had not crossed over to receive the intervention • Is based on the ITT population • Is a non-parametric model that produces a randomisation-based effect estimator" <p>"We used the method of Robins and Tsiatis (1991) which is the only method currently available in the literature that can correct for time-dependent treatment changes in survival data while respecting the randomisation. This method is based on the accelerated failure time model ... we estimated the hazard ratio for starting sunitinib compared to not starting sunitinib by running a Cox regression on the observed event times in the sunitinib arm and the estimated counterfactual values in the placebo arm. Because this procedure is based on the randomisation, it does not change the level of evidence against the null hypothesis. It does however change the estimated hazard ratio, bringing it further from the null, as would be expected from the fact that crossovers make the overall treatment experience of the two arms more similar. As a result, the 95% confidence interval is wide.</p> <p>The initial Hazard Ratios and 95% Confidence Intervals for all analyses utilizing the RPSFT method are derived from the Cox regression analysis as presented in the 2008 publication (Demetri <i>et al</i>, 2008). Advice received since the publication is that because this procedure is based on the randomisation, it does not change the level of evidence against the null hypothesis and therefore a different analytical approach needs to be used. Adopting this results in a wider 95% confidence interval and for transparency we have therefore also presented revised estimates from our updated analysis. It should also be noted that after review by an independent statistician Pfizer was made aware of a number of methodological issues with the original RPSFT analysis, we therefore took the opportunity of the availability of the final data to re-conduct the analysis. This updated analysis has been externally reviewed and approved. (Submission p31)"</p> <p>The AG make further comment on the method: "it is crucial that we have confidence that the RPSFT method has been correctly performed. From the Web of Knowledge database, we identified 68 papers that cite the original statistics paper that describes the RPSFT method(Robins & Tsiatis 1991). None of these papers are of cost-effectiveness studies. This suggests that the method has rarely, if ever, been used in cost-effectiveness models. But of course, this in itself does not mean that the method is inappropriate in this instance. As mentioned in the previous chapter, whilst an unpaid, independent statistician, Ian White (MRC Biostatistics Unit, Cambridge), who has published on methods of adjusting for patient cross-over (White 2005a;White 2006), has endorsed the use of the method in this application, he did not perform the calculations. Therefore, we cannot be completely certain that the method has been correctly implemented. Furthermore, the RPSFT analysis was unplanned (Submission p29). We do, however, have some weak evidence to suggest that the method has been applied correctly: the mean overall survival hazard ratio under the RPSFTM of 0.505 as estimated by Pfizer is similar to the mean overall survival hazard ratio of 0.49 for the interim ITT data, before patient cross-over. But of course the interim analysis is based on far less mature data than that used in the final analysis (on which the RPSFT method is based). On the other hand, we have a reason to question whether the method has been correctly implemented by Pfizer. Ian White (MRC Biostatistics Unit, Cambridge), advised Pfizer that the 95% confidence interval of the overall survival hazard ratio of 0.388 – 0.658 (mean 0.505) as originally calculated by Pfizer, was incorrect. Instead, Ian White states that the confidence interval should be 0.262 – 1.134 (Submission p39)."</p>
<p>Independent survival models, or hazard ratio (proportional hazards) modelling</p>	<p>As an additional analysis, the AG reran the model assuming an HR for OS of 1.0, which fits in to the 95% CI estimated by the RPSFT model. This increases the ICER to £230k. Assuming an HR of 0.262 (the lower 95% CI) the ICER falls to £18k. There is therefore considerable uncertainty.</p>	<p>Independent Weibull's were chosen. PH modelling was tested but these were stated not to give a good fit for sunitinib – although how the fit was measured was not explained in the AG report – it appears that this has been decided based on a visual inspection alone. Of particular note is that when PH modelling is used for each sunitinib curve the ICER falls to £15.5k from £27.3k.</p>
<p>Justification for survival model used?</p>	<p>The AG accept that the Weibull models have been chosen and fitted appropriately. However, this may not actually be the case, due to the manufacturer's fit to one data-point per month, which may be arbitrary and inappropriate. This practice is, however, accepted by the AG.</p>	<p>PH models did not fit so independent models were used. No detailed justification for the choice of Weibull's.</p> <p>RPSFT was used because a very high proportion of patients in the control arm switched on to sunitinib following disease progression. It is stated that this is the 'only' option that maintains the randomisation of</p>

	The AG have some concern about the use of the RPSFT, particularly that it has been implemented correctly. However they accept that it is an appropriate method to use.	the trial – however the IPE algorithm also does this. No mention is made of other complex methods for adjusting for crossover.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	No novel model produced.	<p>“The manufacturer used a Markov model, based on the renal cell carcinoma RCC model developed by PenTAG, to model the cost-effectiveness of sunitinib compared to best supportive care for GIST patients. This had a three state structure; progression free, progressive disease and death.”</p> <p>“For sunitinib and for BSC separately, a Weibull curve describes the number of patients alive over time (overall survival), and another Weibull curve describes the number of patients in PFS over time (Submission p64-5 Figs. 13 - 14). Fitting of these curves to trial data is described in the following section. For each treatment, the number of patients in PD at any time is calculated as the number alive minus the number in PFS at that time. The time horizon of the model is 6 years, and the model cycle length is 6 weeks (to reflect the duration of a cycle of treatment with sunitinib). A half-cycle correction is modelled.”</p>
Other issues noted (eg crossover)	<p>Crossover was a key issue. The AG state: “The blinded phase became open-label upon disease progression or at the time of interim analysis (54 weeks) when patients were allowed to cross-over from placebo to treatment group.” – 84% of the control group crossed over onto the intervention treatment. The AG state that the use of the RPSFT approach was unusual, but believed that it was a correct approach, but that they had not been able to check that it had been applied correctly.</p> <p>The manufacturer used the RPSFT method to correct for crossover. This had a significant effect on the economic analysis: “Pfizer’s base case analysis produced an ICER of £27,365 per QALY with the first cycle of treatment sunitinib <i>not costed</i>, and using effectiveness estimates from their Rank Preserved Structural Failure Time (RPSFT) analysis). When we included the cost of the first cycle of treatment we estimated that the value of the base case ICER was £32,636 per QALY, again using RPSFT effectiveness data. Their sensitivity analysis produced a range of ICERs from £15,536 per QALY to £59,002 per QALY. When a conventional method of unadjusted ITT analysis is used to calculate the base case ICER, values of £93,062 per QALY (first cycle costed) and £77,107 per QALY (first cycle free), are produced. However, this method does not account for the overestimated effectiveness results in the placebo arm due to crossovers; independent, expert statistical opinion favours the RPSFT method.”</p> <p>Regarding the use of RPSFT, the AG remarks that the usual approach would be to censor at crossover, and recognises (as stated by the manufacturer) that this approach would result in bias due to informative censoring (those who crossover will be different from those who don’t): “The use of the RPSFT method of analysis (instead of the conventional approach of censoring participants at the point of crossover) greatly affects the estimated cost-effectiveness of sunitinib for GIST. However, this is a common analysis issue in trials of cancer drugs that are found to be effective mid-trial, and the use of the RPSFT seems appropriate.”</p> <p>And: “In particular, we agree with their use of the RPSFT method, and with their method of fitting survival curves.”</p> <p>The FAD shows that the AC requested that the manufacturer also conduct an analysis censoring when patients crossover, and that this led to BSC dominating sunitinib. However little weight was given to this analysis because only 15 patients did not crossover.</p> <p>The FAD states: “The ERG highlighted that the hazard ratio for overall survival produced using the RPSFT method (0.505) was similar to the hazard ratio for overall survival produced at the interim ITT analyses before crossover had occurred (0.49). It stated that this strengthened the confidence it had in the results derived using the RPSFT method. Additionally, the ERG agreed with the manufacturer that censoring the participants at crossover in this instance was an unreliable method for controlling for crossover. The ERG also noted that the first 4 months of the overall survival curve for people who received best supportive care, who were then censored at the point at which they crossed over, was similar to the RPSFT overall survival curve. The ERG stated that this gave further credibility to the results derived using the RPSFT method because there would have been minimal censoring during the first 4 months.” The AC accepted the use of RPSFT.</p> <p>Although the RPSFT is discussed at length, there does not appear to be a full understanding of the method and its assumptions and limitations and implications. For example there is not detailed discussion of the common treatment effect assumption. Other methods, such as the IPCW are not mentioned.</p> <p>As an additional issue related to crossover, the AG note: “However, we caution that the base case ICERs may be slightly too low as Pfizer’s calculation does not include the cost of sunitinib in progressive disease for some patients randomised to sunitinib (54 patients in the sunitinib arm carried on with this treatment after disease progression), and who theoretically may have benefited.”</p>	

43. TA181: Lung cancer (non-small cell, first line treatment) - pemetrexed, September 2009

Guidance: Pemetrexed in combination with cisplatin is recommended as an option for the first-line treatment of patients with locally advanced or metastatic non-small-cell lung cancer (NSCLC) only if the histology of the tumour has been confirmed as adenocarcinoma or large-cell carcinoma.

People who are currently being treated with pemetrexed for NSCLC but who do not meet the criteria in 1.1 should have the option to continue their therapy until they and their clinicians consider it appropriate to stop.

Source: Pemetrexed for the first-line treatment of non-small-cell lung cancer, TA181, September 2009, <http://www.nice.org.uk/nicemedia/live/12243/45501/45501.pdf>, accessed 18/05/10

Fleeman N, Bagust A, McLeod C, Greenhalgh J, Boland A, Dundar Y, Dickson R, Tudor Smith C, Davis H, Green J, Pearson M. Pemetrexed for the first-line treatment of locally advanced or metastatic non-small cell lung cancer (NSCLC), ERG Report, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, February 2009, <http://www.nice.org.uk/nicemedia/live/12045/43758/43758.pdf>, accessed 18/05/10

Fleeman N, Bagust A, McLeod C, Greenhalgh J, Boland A, Dundar Y, Dickson R, Tudor Smith C, Davis H, Green J, Pearson M. Pemetrexed for the first-line treatment of locally advanced or metastatic non-small cell lung cancer (NSCLC), ERG Addendum, Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, June 2009, <http://www.nice.org.uk/nicemedia/live/12045/45084/45084.pdf>, accessed 18/05/10

Eli Lilly and Company Ltd, Pemetrexed first-line NSCLC non-squamous – NICE STA Submission, December 2008, <http://www.nice.org.uk/nicemedia/live/12045/43762/43762.pdf>, accessed 18/05/10

NICE, Final Appraisal Determination, Pemetrexed for the first-line treatment of non-small-cell lung cancer, TA181, July 2009, <http://www.nice.org.uk/nicemedia/live/12045/45085/45085.pdf>, accessed 18/05/10

Note: This was an STA – the manufacturer's full submission is available on the NICE website.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	The AG states: "As the model had a time horizon of up to six years - well beyond any of the trials - it was necessary to extrapolate the clinical results. The manufacturer did this by using median OS figures and converting them into transition probabilities (per cycle risk of death), assuming an exponential survival model. The ERG notes that the published survival curves do not appear to follow exponential trajectories. Furthermore, the fit of the modelled survival to the published curves is poor, leading to underestimation of survival in the model."	The clinical data used in the economic evaluation are primarily generated from the JMDB trial, with two further trials used to conduct indirect analyses. The model was extrapolated beyond the 30 month JMDB trial up to six years. To achieve this, the median values for OS observed in the JMDB trial were converted into a per cycle risk of death (transition probability). This per cycle risk of death was then used to extrapolate the data to six years.
Evidence synthesis (pool survival estimates?)	The AG state: "Only one relevant trial (JMDB) was identified which directly compared pemetrexed/cisplatin with any comparator of interest (gemcitabine/cisplatin). Indirect comparisons analysis was therefore undertaken by the manufacturer to attempt to compare the effects of pemetrexed/cisplatin with other comparators. Evidence from the indirect comparisons should be treated with caution as key comparators were excluded from the indirect comparisons analysis. In addition, the statistical approach employed to generate the findings is not considered to be the most optimal as calculations were based on median survival times and individual trial arm level data from within trials were compared, thus ignoring the benefits of randomisation."	An indirect comparison was run so that comparators not included in the JMDB trial could be included in the economic model.
Survival model(s) fitted (Weibull, exponential etc)	The AG states: "Overall survival and PFS are primary outcomes from the JDMB clinical trial, and therefore should be accurately replicated in the economic model for each of the trial sub-populations. However, the model allows direct comparison of model estimated OS with only trial OS data for a single (unspecified) sub-population and shows a poor fit to all sub-populations. No PFS trial data are provided for comparison. Moreover, Kaplan-Meier survival charts were not provided for all sub-populations in the MS. To rectify this omission, the ERG requested the relevant patient events and censored patients for each sub-population in the NICE clarification letter, to allow survival curves to be derived and compared to model estimates. The requested trial data were provided by the manufacturer in their response to the NICE letter of clarification. The ERG has generated Kaplan-Meier survival plots for both OS and PFS, and these are shown in Appendix 4 of the ERG report, together with the latest model estimates for comparison. The manufacturer's model appears to overestimate OS in both arms and almost all patient subgroups. For PFS, the model tends to produce under-estimates in the first six months and to over-estimate thereafter. In no instance can the fit of the model be considered a good fit to the JMDB data." The new approach adopted by the manufacturer (using Weibull models) is an improvement, but the AG state: "does not necessarily ensure that the new approach represents the most appropriate or credible means of estimation, since fitting a Weibull model to data for the whole	Exponential in the original analysis. In the revised analysis subsequent to the ACD the manufacturer made use of Weibull models in one analysis, and in the other presented a within trial analysis, presumably using a restricted means approach, although this is not discussed.

	<p>trial period places most weight on early and intermediate period events, and much less weight on the sparser events towards the end of the trial, potentially leading to systematic over- or under-estimation of survival towards the end of the trial.” – the AG go on to show that graphically the Weibull model is not a perfect fit to OS or PFS data (although it is improved compared to the exponential).</p> <p>The AG state that in their resubmission the manufacturer provided some PLD which allowed the AG to consider models for OS. They state the following: “The approach taken by the ERG to survival estimation was designed to make full use of the trial data and to minimise the contribution of trend projection beyond the available IPD. The area under the curve (AUC) was calculated from a Kaplan-Meier analysis from the start of the trial until the time when the last recorded event (death) occurred. Beyond that time, expected mean survival for patients still alive was estimated using a fitted survival model, calibrated from long-term trial data. The approach taken to projection was based on examination of the cumulative hazard function for each treatment duration subgroup. It was observed that standard survival functions (e.g. exponential, Weibull, Log normal, etc) were not generally compatible with the trial data across the whole range of observation. This is not unusual when treatment is of limited duration and would be expected to have a short-term effect of altering/delaying the normal course of the disease, after which the long-term progression pathway resumes. It was observed that in all cases at some time following the end of treatment the cumulative hazard function assumed a steady linear increase, indicative of a constant risk per unit of time. Therefore, for each patient subgroup an exponential function was fitted to the data from the point at which the long-term linear trend in the cumulative hazard became established. This survival function was then used to estimate the likely additional mean survival from the time of the last recorded death until the time horizon of the cost-effectiveness analysis.” le they used PLD as far as possible, before fitting an exponential model to the final portion. Re-estimations by the AG generally led to improved survival estimates and slightly decreased ICERs.</p>	
Independent survival models, or hazard ratio (proportional hazards) modelling	Kaplan-Meier combined with exponential models, independently fitted.	Some use of proportional hazards modelling for the indirect comparisons. But independent survival curves appear to have been fitted for the comparators included in the main RCT.
Justification for survival model used?	The AG stated that the manufacturer’s exponential models were a poor fit. The Weibull’s were better, but the AG was concerned that these may have placed too much weight on early and intermediate observations. Again, this view from the AG is a little concerning as the only way to avoid this is to exclude some data, which is arbitrary. The AG base their views on the manufacturer’s models mainly on visual inspection, but when coming up with their own approach they analyse the cumulative hazard function to determine which parametric model is likely to be appropriate – this leads to their Kaplan-Meier – exponential approach.	No detailed justification.
Type of economic model (Markov Model, Decision Tree, AUC, How many health states? etc)	<p>The AG state: “Due to problems with the manufacturer’s model, however, a total of three submissions and one addendum were provided (see section 5.5.1 for a detailed history of model versions). Our critique of the manufacturer’s economic evaluation is based on the third and final version of the MS (dated 21st January 2009) and the addendum (dated 23rd January 2009)”</p> <p>The AG state: “Examination of the third version of the economic model submitted to NICE and considered by the ERG showed that, although minor modifications had been made to correct some of the problems identified by the ERG with earlier versions, the underlying structural problem and logic errors had not been addressed and the model was still unable to replicate the response rates arising in the clinical trial. These serious flaws rendered it impossible for the ERG to provide reliable ICERs.”</p>	<p>The manufacturer developed a Markov model to evaluate the cost-effectiveness of pemetrexed/cisplatin compared to gemcitabine/cisplatin, docetaxel/cisplatin and gemcitabine/carboplatin. Although the economic evaluation is trial-based, there is also a modelling component to allow the extrapolation of health effects beyond the 30 month trial period up to 6 years. Cycle length was 3 weeks, and there were 4 health states: response, stable disease, progression and death.</p> <p>The ICERs estimated by the manufacturer’s model (third version) range from £8,056 to £33,065, depending on the comparator, the population and the application of a continuation rule.</p> <p>The NICE ACD concluded that the manufacturer’s economic model was not sufficient for them to make recommendations, and therefore the manufacturer made a re-</p>

	<p>"the chosen Markov model structure does not appear to be appropriate in this respect, since it imposes some strong constraints which make it impossible to replicate the data used to calibrate the model to an acceptable level of accuracy. In particular, the manufacturer's model assumes that death only occurs from the progressive disease state, and this dictates that no patients can die within the first cycle, and very few in the second cycle (about 1%). By contrast, the trial data indicate that 4-5% of patients were dead by the end of cycle 2. Furthermore all transition probabilities during the trial period are assumed to arise from constant risk processes (i.e. exponential survival distributions), without any justification. It is therefore unsurprising that the submitted model is unable to generate results consistent with the trial evidence, especially with respect to three primary clinical outcomes (OS, PFS and response rate"</p>	<p>submission. In their new analysis they included: "(1) 'in-trial' cost-effectiveness analysis using the individual patient survival outcomes (censored) and resource use events from the JMDB clinical trial database (2) The original submitted Markov model was modified to more accurately represent the outcomes of the JMDB trial using Weibull distributions, and to take into account concerns raised by the Committee ... in order to re-estimate the incremental cost-effectiveness of pemetrexed/cisplatin when compared to gemcitabine/cisplatin (modified Markov model) (3) Findings from the economic model used for the Pharmaceutical Benefits Advisory Committee (PBAC) HTA submission in Australia, which was based upon the patient-level data from the JMDB trial and used Weibull distributions to extrapolate survival "</p> <p>In the new submission, ICERs ranged from £24k to £45k.</p>
Other issues noted (eg crossover)	<p>Regarding crossover / post progression treatment, the AG states: "Second-line therapy is received by approximately 53% (pemetrexed/cisplatin) and 56% (gemcitabine/cisplatin) patients based on JMDB trial data. In the model, second-line treatment is a single state in which costs are incurred as a lump sum as the patients enter the state; no additional benefit is accrued and no utility value is attached. In relation to the issue of first-line and second-line chemotherapy the manufacturer states that: "It is not possible to disaggregate the effect of first-line therapy from second-line therapy in the overall efficacy results. Therefore, the simplifying assumption was made, that all second-line therapies have equivalent efficacy, safety and duration. Costs associated with docetaxel and erlotinib are assumed to be equal in the light of the FAD [final appraisal determination] for erlotinib which recommends erlotinib based on the premise that it has equivalent efficacy, and should therefore have equivalent cost, to docetaxel" (MS, pg67)."</p> <p>The AG asked the manufacturer about this, and concluded the following: "There is a high level of post-treatment which may impact on the results, particularly in a noninferiority trial. However, as the proportion of patients is relatively similar by treatment arm, the risk of bias is minimised. It is noted that it is patients in the gemcitabine arm who received significantly more therapy. This may suggest that this group of patients were fitter and so lived longer but the ERG does not believe this would significantly impact on the findings. Additional analysis undertaken for the EMEA which excluded patients who switched treatment indicate findings consistent with those when such patients were not excluded."</p>	

44. TA183: Cervical cancer (recurrent) - topotecan, October 2009

Guidance: Topotecan in combination with cisplatin is recommended as a treatment option for women with recurrent or stage IVB cervical cancer only if they have not previously received cisplatin.

Women who have previously received cisplatin and are currently being treated with topotecan in combination with cisplatin for recurrent and stage IVB cervical cancer should have the option to continue their therapy until they and their clinicians consider it appropriate to stop.

Source: Topotecan for the treatment of recurrent and stage IVB cervical cancer, TA183, October 2009, <http://www.nice.org.uk/nicemedia/live/12328/45867/45867.pdf>, accessed 18/05/10

Author information marked as commercial in confidence. Topotecan for the treatment of recurrent and stage IVB carcinoma of the cervix, ERG Report, Centre for Reviews and Dissemination, Centre for Health Economics, University of York, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, February 2009, <http://www.nice.org.uk/nicemedia/live/12034/44617/44617.pdf>, accessed 18/05/10

GlaxoSmithKline, Topotecan (Hycamtin) for the treatment of recurrent and stage IVB carcinoma of the cervix – NICE STA Submission, February 2009, <http://www.nice.org.uk/nicemedia/live/12034/44620/44620.pdf>, accessed 18/05/10

NICE, Final Appraisal Determination, Topotecan for the treatment of recurrent and stage IVB cervical cancer, TA181, July 2009, <http://www.nice.org.uk/nicemedia/live/12034/45432/45432.pdf>, accessed 18/05/10

Note: This was an STA – the manufacturer's full submission is available on the NICE website.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>Regarding the indirect comparison, the AG states: "A further potential issue with the indirect comparison is that it considered only a two year follow-up, while GOG-0179 reports patient-level data for three years – the third year of data from GOG-0179 was disregarded. The MS justifies this as follows: <i>"Only 24 months of follow-up data were available for GOG-0169. Therefore,</i></p>	<p>"The manufacturer's submission focused on direct evidence from trial GOG-0179, comparing topotecan plus cisplatin with cisplatin monotherapy, and indirect clinical evidence from trial GOG-0169 comparing topotecan plus cisplatin with paclitaxel plus cisplatin. A second direct comparison trial (GOG-0204) was mentioned in the manufacturer's submission, which compared four cisplatin-based combination therapies: topotecan plus cisplatin, paclitaxel plus cisplatin, gemcitabine plus</p>

	<p><i>although the direct analysis for GOG-0179 is conducted with 36 months of data, only data for the first 24 months are considered in the indirect analysis, for consistency with the GOG-0169 data.” (MS pp.91)</i></p> <p>This approach is likely to be conservative towards topotecan plus cisplatin based on the results of GOG-0169, as any additional survival benefit incurred after 24-months will not be reflected in the ICER estimates. However, the cost-effectiveness results compared to paclitaxel plus cisplatin will be optimistic towards topotecan plus cisplatin. An alternative approach would be to derive the hazard ratio for paclitaxel plus cisplatin versus cisplatin monotherapy from GOG-0169, and then apply this hazard ratio to all three years of data from GOG-0179. Although this assumes that the hazard ratio observed over a shorter period can be extrapolated to a longer follow-up period, this does seem to provide a more reasonable assumption than assuming that any additional benefits are not accrued over a longer time period. Following queries by the ERG, this approach was undertaken by the manufacturer in a revision to the model”.</p>	<p>cisplatin, and vinorelbine plus cisplatin.”</p> <p>The AG state that: “For the direct comparison, the time horizon (36 months of follow up period) was considered appropriate by the manufacturer given that the majority of patients in all treatment arms of the GOG-0179 trial had died and thus most of the costs and outcomes for the cohort had been incurred.” Thus it seems that no extrapolation was required.</p> <p>For the indirect comparison: “An indirect comparison between topotecan plus cisplatin and paclitaxel plus cisplatin was modelled using clinical effectiveness data from GOG-0179 and GOG-169. These two trials provide an indirect estimate with cisplatin monotherapy acting as a common comparator (Section 4.1.3).</p> <p>The hazard ratio for overall survival derived from GOG-0169 for paclitaxel plus cisplatin versus cisplatin monotherapy was applied to the first 24 months of overall survival data for cisplatin from GOG-0179 to estimate the overall survival for paclitaxel plus cisplatin. This provided the basis for an indirect comparison of topotecan plus cisplatin versus paclitaxel plus cisplatin. However, since the main publication of GOG-0169 did not actually report the hazard ratio, the manufacturer estimated the ratio (HR= 0.87 favouring paclitaxel plus cisplatin) from the survival curves using published methods.</p> <p>Although the manufacturer acknowledged that the GOG-0204 trial provided a direct comparison between topotecan plus cisplatin and paclitaxel plus cisplatin, a number of potential methodological limitations were identified, namely the early closure of the trial and the high proportion of patients with a good performance status. As a result, the comparison between topotecan and paclitaxel was presented as a separate sensitivity analysis. For this analysis the overall survival data for paclitaxel plus cisplatin was estimated from the hazard ratio for paclitaxel plus cisplatin versus topotecan plus cisplatin taken from GOG-0204 (1.255, favouring paclitaxel plus cisplatin). This hazard ratio was then applied to the first 24 months of overall survival data for topotecan plus cisplatin from GOG-0179.”</p>
Evidence synthesis (pool survival estimates?)	The AG noted that: “In terms of the indirect comparison, the ERG believes that a potentially relevant network of indirect evidence has not been fully explored (see Section 4.1); although the ERG does acknowledge that the quality of such evidence would be limited.”	The manufacturer presented an indirect comparison as a secondary analysis.
Survival model(s) fitted (Weibull, exponential etc)	No new models were fitted, although at the ERGs request the manufacturer presented an additional analysis applying HRs to longer time periods where appropriate.	It seems that for both the direct and indirect analyses no extrapolation was used and so no survival models were fitted. An HR was applied to the trial survival estimates for the indirect comparison.
Independent survival models, or hazard ratio (proportional hazards) modelling	The ERG suggested an extension to the PH modelling method (applying the HR to a longer time period where appropriate).	For the indirect comparison proportional hazards modelling was used, although there was debate over where the appropriate HR should be taken from: Key issues were...“and the appropriate source of the hazard ratio used to estimate survival for paclitaxel plus cisplatin – deriving this hazard ratio from GOG-0169 favours topotecan plus cisplatin, while deriving it from GOG-0204 favours paclitaxel plus cisplatin.”
Justification for survival model used?	No detailed discussion of the proportional hazards assumption.	No extrapolation, so no justifications were given. No consideration of proportional hazards despite the use of PH modelling.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	“The ERG made a number of revisions to this model, altering (among other things) the assumptions made over utility values, the costs of administering each treatment and the assumed number of vials of topotecan utilised per treatment cycle. Where the number of vials used was assumed to be minimised (maximised), the ERG found topotecan plus cisplatin to have an ICER versus cisplatin monotherapy of £26,778 (£34,327) in the cisplatin-naïve patient population and £58,872 (£73,833) in the full licensed population from GOG-0179”	The AG state: “The manufacturer’s submission included a non-standard economic analysis consisting of two parts. The primary economic evaluation was an individual patient-level data cost-utility analysis of topotecan in combination with cisplatin versus cisplatin monotherapy. In addition, a secondary modelled analysis using indirect evidence was undertaken to compare topotecan plus cisplatin versus paclitaxel plus cisplatin” “In the base-case direct comparison, the ICER of topotecan plus cisplatin versus cisplatin monotherapy was £17,974 per QALY in the main licensed population, £10,928 per QALY in the cisplatin-naïve population (excluding stage IVB patients) and £32,463 per QALY in sustained cisplatin-free interval (SCFI) patients.

		Results for the indirect comparison were only presented for a cisplatin-naïve population. In the base-case indirect comparison, paclitaxel plus cisplatin was dominated by topotecan plus cisplatin, which in turn had a cost-per-life-year-gained of £19,964 versus cisplatin monotherapy; where the hazard ratio used to calculate overall survival with paclitaxel plus cisplatin was taken from GOG-0204 (rather than derived from GOG-0169, as in the base-case), paclitaxel plus cisplatin was found to have a cost-per-life-year-gained of £982 versus topotecan plus cisplatin."
Other issues noted (eg crossover)	<p>Treatment crossover and post-progression treatments are not mentioned. This may be because no extrapolation was required and the analysis was purely trial based, although there is no discussion of which treatments may have been given after disease progression. The trial protocol included collecting data on any follow-up treatment given before disease progression though there is no discussion of this data. No data was due to be collected on follow-up treatments after progression. It seems likely that the severity of the disease limited follow-up treatment/post progression treatment.</p> <p>Survival analysis plays a relatively small part in the AG report in this Appraisal, in which there is more focus on costs and utilities. However the AG do make some interesting points regarding the possibility of informative censoring: "As mentioned by the MS (pp.98) a variant of the Lin method was applied in order to estimate the mean total cost per patient, which is the sum over the intervals of the Kaplan-Meier estimator of the probability of dying in an interval multiplied by the mean total costs of those who die in that interval. The Kaplan-Meier method is a non-parametric technique for estimating time-related events. It is a univariate analysis and is especially applicable when length of follow-up varies from patient to patient, taking into account those patients lost to follow-up or patients where the endpoint of analysis was not verified yet at end of follow-up. However, this method implies the strong assumption of non-informative or independent censoring, that is, a subject censored at time t can be considered completely interchangeable with any other subject (on the same treatment) who has also survived up to time t (the censoring could have happened to any comparable subject). This assumption is emphasized in the discussion section of Lin <i>et al</i> (1997): "The assumption of independent censoring requires some care. This assumption is clearly not satisfied if patients are withdrawn from the study for health- or cost-related reasons"... "One must carefully examine the independent censoring assumption before applying the proposed methodology." With respect to the intervention under analysis, informative censoring may be arising due to patients that dropped out were more likely to die sooner than similar non-censored subjects, that is, given that there is a relationship between their propensity to drop out and their survival, the factor that causes censoring is evidently related to that survival time. Also, those who were removed from the study due to adverse side-effects are subject to informative censoring. They are clearly very different from other patients who were still alive at the point they were removed; a patient without adverse symptoms would not have been censored at this time. In order to overcome the abovementioned methodological limitations, Willan <i>et al</i> (2005) proposed a method for estimating the difference in mean costs and the difference in effectiveness, together with their respective variances and covariance in the presence of dependent censoring. This method uses inverse-probability weighting for estimating the parameters required for performing a cost-effectiveness comparison of two groups when the measure of effectiveness is some function of survival and censoring is present. This methodology might have been more adequate in the current circumstances, where the probability of being observed may be estimated conditionally on a series of covariates. Again, as with the previous issue, although this appears to provide a more suitable analytic approach, it not clear that the approach employed would necessarily introduce any significant bias into the results."</p>	

45. TA184: Lung cancer (small-cell) - topotecan, November 2009

Guidance: Oral topotecan is recommended as an option only for people with relapsed small-cell lung cancer for whom:

- re-treatment with the first-line regimen is not considered appropriate **and**
- the combination of cyclophosphamide, doxorubicin and vincristine (CAV) is contraindicated (for details of the contraindications to CAV see the summary of product characteristics for each of the component drugs).

Intravenous topotecan is not recommended for people with relapsed small-cell lung cancer.

People with relapsed small-cell lung cancer currently receiving oral topotecan who do not meet the criteria specified above, or who are receiving intravenous topotecan should have the option to continue their treatment until they and their clinicians consider it appropriate to stop.

Source: Topotecan for the treatment of relapsed small-cell lung cancer, TA184, November 2009, <http://www.nice.org.uk/nicemedia/live/12348/46326/46326.pdf>, accessed 18/05/10

Loveman E, Jones J, Hartwell D, Bird A, Harris P, Welch K, Clegg A. Author information marked as commercial in confidence. The clinical and cost-effectiveness of topotecan for small cell lung cancer: a systematic review and economic evaluation, ERG Report, Southampton Health Technology Assessments Centre, University of Southampton, commissioned by the NHS R&D HTA Programme on behalf of the National Institute for Clinical Excellence, March 2009, <http://www.nice.org.uk/nicemedia/live/12021/44800/44800.pdf>, accessed 18/05/10

GlaxoSmithKline, Multiple technology appraisal of topotecan for the treatment of small cell lung cancer, December 2008, <http://www.nice.org.uk/nicemedia/live/12021/44797/44797.pdf>, accessed 19/05/10

Note: This was an MTA – the manufacturer’s full submission is available on the NICE website. Oral topotecan was recommended based on the End of Life ruling, based on an ICER of approx £34k.

	Assessment Group Model / Evidence Review Group Alterations	Manufacturer
Survival data used (patient-level data, or summary statistics – mean, median etc)	<p>The AG note a possible problem with the manufacturer’s survival analysis: “The time horizon used in the economic evaluation is the length of the trial. No additional modelling was undertaken to extend survival beyond the end of the trial. The MS reported that there were six remaining participants (three in topotecan group and three in the BSC alone group) who were still alive at the end of the trial and it was assumed all died the day after the end of the study. However, from the Kaplan-Meier plot of overall survival from the O’Brien and colleagues trial this does not appear to be the case. It appears that there are fewer survivors in the BSC arm than the three survivors reported in the MS. The reason for this discrepancy is unclear. Nevertheless, assuming that there are three survivors in each arm, based on the participant level data in the manufacturer’s model, this represents just over 4% of the population in each arm. There is a possibility that this could have underestimated the survival benefit for either arm of the trial.”</p>	<p>The economic model used PLD from one RCT. The stated aim of the analysis was to assess the cost-effectiveness of oral topotecan plus BSC against BSC alone in people with relapsed SCLC in whom treatment with IV chemotherapy is not considered appropriate. The cost-effectiveness analysis was based on participant level data from the O’Brien and colleagues RCT.</p> <p>“The economic model used the data from the trial up until the final assessment period, when six participants (three in the BSC group and three in the topotecan plus BSC group) were still alive. The model assumed that all surviving participants died the day after this final assessment. The participant level survival data was divided into 21-day periods to reflect the study cycles in the RCT.”</p> <p>Thus no extrapolation was used – this is a restricted mean type analysis.</p>
Evidence synthesis (pool survival estimates?)	Adjusted indirect comparison for intravenous topotecan.	None.
Survival model(s) fitted (Weibull, exponential etc)	<p>The AG model “builds upon the Kaplan-Meier curves for overall survival from the O’Brien and colleagues study for topotecan with BSC and BSC alone. These survival curves were scanned using TechDig software and then imported into Microsoft Excel. In both arms, some of the participants remained alive at the end of the trial. Therefore, the final portions of the survival curves were extrapolated using a regression analysis. A range of parametric survival functions were fit to the observed Kaplan-Meier estimates (full details are included in Appendix 9). The log-logistic survival function provided the best-fit to the observed Kaplan-Meier estimates and was used in the economic model.</p> <p>The extrapolated survival curves are given in Figure 2 and compared to the Kaplan-Meier survival estimates (details of the regression estimates are found in Appendix 9). These show a good fit to the overall survival curves. The most appropriate measure of overall survival is the mean rather than the median. Therefore, the associated mean survival times were estimated for the relevant survival curves.”</p> <p>The parametric models tested were the log-logistic and Weibull for OS, for time to progression an exponential was also considered.</p> <p>The AG compared the mean survival estimated based purely on the KM curves (truncated at the maximum observed survival for each arm in the RCT by O’Brien and colleagues) and from the log-logistic survival functions (extrapolated to a maximum duration of five years) were compared. These were very similar for BSC, but the mean OS was noticeably higher for the log-logistic curve than the KM curve for the intervention – thus a restricted means analysis would underestimate the treatment effect compared to the extrapolation approach.</p> <p>For time to progression, the KMs were not published for all treatments, and thus medians were transformed to means assuming an exponential distribution. In additional analysis based on data submitted by the manufacturer the AG were able to consider Weibull and Log-logistic models and show that the exponential model does not appear to reflect the trial data very well for the oral topotecan group.</p> <p>Sensitivity analysis was carried out extrapolating to 10 years, using a Weibull model instead of a log logistic, and truncating at the maximum follow-up from the trial. The base case ICER for</p>	No extrapolation – type of restricted mean analysis.

	oral topotecan was £34k, and this SA caused it to range between £33k (10 year) and £37k (Weibull).	
Independent survival models, or hazard ratio (proportional hazards) modelling	The AG also conducted a comparison for intravenous topotecan. "An adjusted indirect comparison was undertaken to assess the effect of intravenous topotecan on overall survival relative to best supportive care, using data from three RCTs included in the clinical-effectiveness review. The relative risk for overall survival with intravenous topotecan was 0.68 (95% CI 0.45 to 1.02) compared with best supportive care." Thus a PH type approach was used for the indirect comparison. For the trial based analysis appendix 9 seems to suggest that a PH approach was taken for OS. For time to progression it seems likely that independent exponential models were fitted. The supplementary analysis of the fit of exponential, Weibull and log-logistic models involves investigation of their fit only to the oral topotecan trial data.	None.
Justification for survival model used?	The log-logistic was picked based on a better fit to the data based on r2, and sum of residuals statistics, as well as plots of the log hazard KM data and visual inspection. In particular: "The Weibull survival functions are likely to underestimate survival probabilities at higher survival durations when compared with the Kaplan-Meier estimates. The modelled probability of survival at 100 weeks is very close to zero, for the Weibull survival function, whereas the Kaplan-Meier estimate is around 5%. In contrast the modelled probability of survival at 100 weeks, for the log-logistic survival function, is around 4%."	None.
Type of economic model (Markov Model, Decision Tree, AUC. How many health states? etc)	The AG "developed a new model to estimate the cost-effectiveness of topotecan as a second-line chemotherapy compared with BSC, in a cohort of adults with relapsed SCLC for whom re-treatment with the first line regimen was not considered appropriate." A survival model was built, with states of stable disease, progressive disease, and death. Mean OS was estimated based on the clinical trial data, and utility scores were applied to the mean OS estimates in order to estimate QALYs gained.	The model was trial-based. The AG state: "An independent economic model was developed to estimate the cost-effectiveness of topotecan (oral or IV) compared with Best Supportive Care (BSC) for patients with relapsed SCLC, for whom re-treatment with the first line regimen was not considered appropriate, from the perspective of the NHS and Personal Social Services (PSS). The model used survival analysis methods to derive estimates of mean survival for patients treated with topotecan or receiving BSC alone, which were combined with Quality of Life (QoL) weights to derive estimates of mean quality adjusted life expectancy for patients receiving BSC alone or topotecan and BSC. The model includes an estimate of time to disease progression for patients receiving topotecan, to take account of the reduction in QoL following disease progression." The manufacturer estimated an ICER for oral topotecan of £27k (QALY gain = 0.211), compared to the AG who estimated £34k (QALY gain = 0.1830)
Other issues noted (eg crossover)	Crossover is not mentioned as an issue in the appraisal. However it is stated in the review of the key RCT that: "However, 13 participants in each arm (18.3% BSC, 18.6% topotecan) received poststudy chemotherapy either alone or in combination with other therapy such as radiotherapy and surgery. In addition poststudy radiotherapy alone was received by 7 (10%) topotecan participants and 1 (1%) BSC participant."	

Appendix 2: Summary of survival analysis methods used in NICE appraisals of advanced and/or metastatic cancer treatments

Ref	Title of appraisal	Mean or median survival used in model?			Was extrapolation undertaken?	Which parametric models were considered?					Were parametric models justified by the analyst?	Was a restricted means analysis undertaken?	Was a hazard ratio (proportional hazards) modelling approach taken?	Was crossover an issue?	Was crossover adjusted for?	
		Mean	Median	Weibull		Gompertz	Exponential	Log-logistic	Log normal	Gamma						
TA3	Ovarian cancer - taxanes (replaced by TA55) May 2000	✓	✓	✓	?	?	?	?	?	?	x	?	x	✓	x	
Appraisal based on a review of studies, some used medians, some used means. In those that used means no description of how the mean was estimated was given, though one study did fit parametric curves. One study simply assumed average survival to be 50% longer for the active treatment. One only considered PFS. Crossover to alternate treatment was an issue in all of the 4 relevant trials but this was not discussed or adjusted for.																

		Thus OS was estimated, sometimes through extrapolation. Treatment crossover was an issue but was not discussed or adjusted for. Paclitaxel was recommended for use as a standard 1 st line therapy, and for patients and later lines who have not previously received paclitaxel (docetaxel was not licensed for ovarian cancer and so could not be recommended).																		
TA6	Breast cancer - taxanes (replaced by TA30) Jun 2000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	x	?	?	✓	x
		The HTA report does not make it clear whether means or medians were used in the economic evaluations that were reviewed, or how means were estimated if they were used. Details on the manufacturers submissions were removed from the report. The FAD is not available. Crossover was an issue but was not adjusted for.																		
		For one of the indications considered (paclitaxel second line) over half of the patients in the control group switched to paclitaxel. This was brought up in an Appeal, particularly because OS had been used as a key outcome measure in the guidance. but the point was not upheld after the Appeal Panel decided that the AC had outlined the difficulties associated with this and given it appropriate consideration. Initially 2 nd line paclitaxel was not recommended, but after appeal (at which an appeal was upheld regarding the interpretation of OS data for paclitaxel – but not related to crossover) this was recommended. Crossover to alternate treatment was also allowed in at least one of the docetaxel 2 nd line RCTs. No further details were given on this, but docetaxel was recommended at 2 nd line.																		
		For the paclitaxel first line indication two trials allowed crossover to alternate treatment and one recommended treatment with a separate treatment upon progression. It appears that this was not adjusted for, and neither paclitaxel or docetaxel (only abstract data were available for docetaxel) were recommended for general use at 1 st line. The Appeal document suggests that the first line economic evaluation focussed on PFS due to several OS confounding factors.																		
TA23	Brain cancer - temozolomide Apr 2001	x	✓	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
		The HTA group recognised the problems with using median data, and problems associated with follow-up schedules. Crossover was not mentioned. OS was estimated but not using extrapolation.																		
		Temozolomide was recommended for a subgroup of patients at the second-line stage, but not for first-line treatment.																		
TA25	Pancreatic cancer - gemcitabine May 2001	✓	x	x	x	x	x	x	x	x	x	x	✓	x	x	x	x	x	x	x
		The manufacturer used an AUC of KM curve approach, which the HTA group endorse. Medians are not used. The KM curves appear complete, but consideration is not given to the extent of censoring and no extrapolations are tested (the HTA group did not have PLD). Crossover is not mentioned. It is noted that in one reviewed study a Gompertz model had been used, which resulted in an increased survival gain estimate. Only OS is modelled, not PFS.																		
		Gemcitabine was recommended for a subgroup of patients for first-line treatment. It was not recommended as a second line treatment. However these were restrictions were due to no clinical evidence being available, rather than effectiveness/cost-effectiveness concerns.																		
TA26	Lung cancer - docetaxel, paclitaxel, gemcitabine and vinorelbine (updated by and incorporated into CG24 Lung cancer) Jun 2001	x	✓	x	x	x	x	x	x	x	x	x	x	x	x	x	x	✓/?	x	x
		Medians were used due to a reliance on trial summary statistics. It was noted that two of the trials for vinorelbine were crossover trials, but these were phase II trials and were not used in the economic analysis. One other vinorelbine trial was noted to not allow crossover. There is no mention of potential crossover in other trials. Given that median survival data were taken from 23 studies, and the focus was mainly on first-line treatment, it seems likely that crossover or post progression treatment would have occurred in some of the trials. Only OS was modelled, but without extrapolation.																		
		Gemcitabine, paclitaxel and vinorelbine (and each in combination with other treatments) were recommended for first-line treatment. Docetaxel was recommended for second line treatment as it was not licensed for 1 st line use.																		
TA28	Ovarian cancer - topotecan (replaced by TA91) Jul 2001	?	?	x	x	x	x	x	x	x	x	x	x	x	x	x	x	✓	✓	✓
		CiC resulted in few details of the economic evaluations conducted by the manufacturers being given in the HTA report. However one manufacturer used a cost minimisation approach (this was used inappropriately in a comparison of topotecan and caelyx). Whether means or medians were used in the other manufacturer's cost-effectiveness analysis of topotecan and paclitaxel, and whether PFS or OS (or both) were used in the economic analysis is not stated. The key topotecan versus paclitaxel trial was a crossover trial whereby patients with stable disease after 6 courses crossed over, but data in the appraisal was only used for the pre-crossover period of the trial. Patients who progressed during treatment were removed from the study and therefore could have received other treatments. This is not discussed in the economic section of the HTA report, and thus it is not clear if (and how) PFS and OS were estimated – potentially PFS may be unconfounded, but there will be a high degree of censoring due to the crossover, whereas OS will be confounded by post-progression treatments.																		
		Topotecan was recommended for a subgroup of patients at second-line. It was not licensed for 1 st line use. Details on the reasons for the ECOG and blockage restrictions in the recommendations are sparse as the FAD is not available, but this would appear to be on effectiveness groups.																		
TA29	Leukaemia (lymphocytic) - fludarabine (replaced by TA119) Sep 2001	?	?	x	x	x	x	x	x	x	x	x	x	x	x	x	x	✓/?	x	x
		The HTA group notes that it is not clear whether mean or median survival estimates were used. An independent model was not built and very few details on the manufacturer's submission are given. Crossover and post-progression treatments are not mentioned, but OS is relatively long, so it is likely to have been an issue, though the economic analysis focussed on time in remission rather than OS.																		

		Fludarabine was recommended for a subgroup of patients at 2 nd line. As it was only licensed as a second-line treatment, this represents a full recommendation.													
TA30	Breast cancer - taxanes (review)(replaced by CG81) Sep 2001	?	?	?	?	?	?	?	?	?	x	?	?	✓	x
		This was an update of TA6. No additional details were given and all details relating to the manufacturer submission were removed from the HTA report. The FAD was not available. There was new evidence relating to docetaxel for first-line treatment, but there was very little new second-line data. All data regarding the new docetaxel trial were removed due to CiC. Thus again crossover was an issue but was not discussed. The original guidance was not altered: Docetaxel and paclitaxel were recommended for second-line treatment, but not for first-line or adjuvant treatment.													
TA34	Breast cancer - trastuzumab Mar 2002	✓	✓	✓	x	x	x	x	x	x	x	x	?	✓	✓
		Comprehensive details of the manufacturer submission are not given, and ICERs quoted in the FAD differ to those in the HTA report. Medians and means are both mentioned in the HTA report with respect to the economic analysis, and the FAD mentions extrapolation. However no more details are given. Crossover was an issue in the key trial and in the manufacturer's initial analysis patients that crossed over were excluded, which the HTA group considered was likely to have caused bias and led to an analysis based on very low patient numbers, but may have been justified. The emphasis in the economic evaluation was on OS. Compared to the ITT analysis, the approach of excluding patients that crossed over led to a 10.9 increase in median survival gain (17.9 months compared to 7.0 months). The FAD reports that the manufacturer extrapolated survival to estimate mean OS for a case mix of patients that represented those who crossed over after progression, and subsequently they estimated a mean survival gain of 10 months, which led to an ICER of £37.5k. This analysis must have taken place after the writing of the HTA group report. The AC believed the survival advantage was likely to be greater than 10 months based on results of non-controlled studies. In addition the HTA group suggested that a lack of blinding may have caused bias – investigators were not blinded, but the central review group were. It was acknowledged that blinding in this situation is difficult, but non-blinding may cause biased administrations of co-interventions. The HTA group also suggested a HR should have been reported. Trastuzumab combined with paclitaxel was recommended as first-line treatment for a subgroup of patients. This represented a subgroup recommendation because there was clinical evidence for patients with characteristics other than those indicated by the recommendation (eg HER2 level 2+, whereas the recommendation is for HER2 level 3+, although the manufacturer's economic analysis only considered patients with HER2 3+) Trastuzumab monotherapy was recommended as third-line treatment for a subgroup of patients.													
TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (replaced by TA93) Mar 2002	✓	✓	✓	✓	✓	✓	x	x	x	✓	✓	x	✓	✓
		The HTA group used TechDig to estimate KM curves and extrapolate. Some models were compared, and the Weibull was chosen based on the sum of squared deviations. A restricted mean approach was also tested. Means and medians were discussed at length (manufacturers' used medians). The HTA group noted that post-progression treatment was a serious issue in the appraisal, with large proportions switching after progression in most of the trials. In some cases this was straightforward treatment crossover, but given the number of possible chemotherapy regimens in this area this was not always the case. Therefore they based their analysis on PFS. They considered that perhaps ideally sequences would be modelled, but little data on post progression treatments were available from trials. Post-progression treatment was also discussed in a later appeal hearing where the attributability of OS gains were questioned. Oxaliplatin with 5FU/FA was recommended for a subgroup of patients at first-line (this was a subgroup recommendation as further evidence on non-recommended subgroups was available). Irinotecan was recommended as a second-line treatment (this represents a subgroup as it was also considered at 1 st line). Raltitrexed was not recommended.													
TA37	Lymphoma (follicular non-Hodgkin's) - rituximab (replaced by TA137) Mar 2002	x	x	x	x	x	x	x	x	x	x	x	x	x	x
		Very little clinical data was available and a cost minimisation analysis was carried out. Crossover was not considered. Rituximab was recommended for last line use. This represents a subgroup as potentially rituximab could be used slightly earlier in the disease process (after 2 relapses)													
TA45	Ovarian cancer (advanced) - pegylated liposomal doxorubicin hydrochloride (replaced by TA91) Jul 2002	✓	x	✓	x	x	✓	x	x	x	x	x	x	✓/?	x
		The manufacturer carried out a CMA, but the HTA group incorporated survival in their analysis. Median survival for each treatment was transformed into an estimate of the mean assuming an exponential distribution and a constant hazard of death (justified due to a lack of PLD). OS was modelled. Crossover was not discussed. Given median PFS of around 16 weeks, and median OS of around 60 weeks, it seems likely that some type of post-progression / crossover treatment may have occurred. PLDH was recommended as a potential second-line treatment for a sub-group of patients. Although the recommendation was for a subgroup only (those with good performance status) this represents the population the drug had been tested in (and it is only licensed 2 nd line), so in this sense this is a full recommendation.													

TA50	Leukaemia (chronic myeloid) - imatinib (replaced by TA70) Oct 2002	x	x	✓	✓	x	x	x	x	x	x	x	x	✓/✓	x
		Survival rates were used in the manufacturer's analysis, as data only up to 1 year were available and OS was relatively long. After this time data for a different (similar?) treatment were used to estimate monthly survival rates using external data. Extrapolation using a Weibull model was also tested. The HTA group and the AC criticised the manufacturer's use of the external data, primarily because of the choice of trial, rather than the use of external data per se. Crossover was not mentioned, probably because the trial data used was short-term – in the long-term crossover and post-progression treatment may have been an issue (eg for the proxy external data). Using the external data from a separate treatment by definition means post-progression treatment is an issue, but not one that could be adjusted for. OS is estimated from the control using data from a different trial, without considering if crossover or post progression treatment occurred, although as OS was relatively long, post progression treatments were likely to have been given.													
		Imatinib was recommended in the chronic phase of disease if IFN was not appropriate/working. It was also recommended in the accelerated/blast phase if previous imatinib treatment had not been given. Given the license, this represented a full recommendation.													
TA54	Breast cancer - vinorelbine (replaced by CG81) Dec 2002	?	✓	x	x	x	x	x	x	x	x	x	x	x/?	x
		No new economic models were developed, instead previous models were reviewed. A number of these relied upon medians, and it is not clear if any used means. Often data was taken from individual arms of separate RCTs which was likely to have caused bias. The majority of the reviewed evaluations modelled based on either response rates or time until progression, although at least one review did include OS. Considering median survival was frequently more than 1 year, with time until progression around 6 months, it seems that post-progression treatment or crossover may have occurred.													
		Vinorelbine monotherapy was recommended as an option for second-line treatment, but not for first-line treatment. Combination therapy was not recommended due to a lack of data. This represented a subgroup recommendation because there was some (although limited) evidence of the use of vinorelbine monotherapy and combination therapy for 1 st and 2 nd line treatment.													
TA55	Ovarian cancer - paclitaxel (review) Jan 2003	✓	✓	✓	✓	x	x	x	x	x	x	✓	✓	✓	x
		Two new published models were reviewed. One based survival estimates (PFS and OS) on medians, the other used a restricted means analysis and fitted a Weibull model (which estimated a more favourable OS gain, more than halving the ICER). Crossover was an issue in the 4 relevant trials, but this was not discussed.													
		Paclitaxel combination therapy was recommended as a first-line treatment option, and for second-line treatment where it has not already been used at first-line.													
TA62	Breast cancer - capecitabine (replaced by CG81) May 2003	✓	✓	✓	x	x	✓	x	x	x	x	x	x	✓/✓	x
		For the monotherapy analysis the manufacturer used median PFS and OS from a number of trials (which may have caused bias). The HTA group re-did the analysis estimating means from the medians assuming an exponential distribution (due to lack of PLD). For combination therapy a single RCT was used and mean PFS and OS estimated (though the methods for this were not stated). The HTA group noted that ancillary treatment could have differed in the monotherapy trials, and in the combination therapy trial patients were withdrawn from the trial upon disease progression and hence post-progression or crossover treatment could have occurred. This was not discussed any further.													
		Capecitabine combined with docetaxel was recommended rather than docetaxel alone after anthracycline treatment. Capecitabine monotherapy was recommended after an anthracycline if a regimen including a taxane has not worked, and if capecitabine with docetaxel has not been taken. These represent full recommendations given the license.													
TA61	Colorectal cancer - capecitabine and tegafur uracil May 2003	✓	x	?	?	?	?	?	?	?	x	?	?	✓	x
		Several reviewed papers and the manufacturers' analyses used CMA. In some cases the HTA group believed this was reasonable, whereas it was not in others. The group estimated mean PFS and OS based on published KM curves, but the method of doing this was not stated. Post-progression treatments were an issue in the two key trials from which survival estimates were taken for UFT, with similar proportions in each trial arm (around 40% in one trial and around 50% in the other) receiving secondary chemotherapy. This was not discussed and adjustments were not made – data on the secondary chemotherapies were not collected. It seems likely that similar occurred in the capecitabine trials, although this is not discussed.													
		Both treatments were recommended as possible first-line treatments, representing full recommendations.													
TA65	Non-Hodgkin's lymphoma - rituximab Sep 2003	✓	x	x	x	x	x	x	x	x	x	x	✓	✓/✓	x
		The manufacturer and HTA group both used external PLD to estimate long-term PFS and OS by response category. The HR for each response category, and the response rates for each treatment, are then used from the clinical trial. The treatment effect was assumed to last 3 years and the model was run for 15 years, with mean survival being calculated using AUC. Therefore this was an interesting use of external data. Some problems exist due to the non-RCT nature of the data, including the likely confounding effect of alternative treatments in the long-term data. Applying response rates from the clinical trial to registry survival data for response groups does not avoid all potential crossover problems that may have occurred within the trial (though this is not discussed) because the OS HR could be confounded. However, the observational data may be confounded by differing post-progression treatments which could be an issue.													
		Rituximab combined with other treatments was recommended for stage II, III, and IV disease. Rituximab was also licensed for stage I, but there was not evidence for this group. Thus this was a full recommendation given the data available.													
TA70	Leukaemia (chronic myeloid) - imatinib Oct	✓	✓	✓	✓	x	✓	x	✓	✓	✓	x	✓	✓	✓
		The HTA group scanned KM curves from various different studies (OS was long) for different treatments using TechDig, and then fitted Weibull curves. For two treatments (imatinib and HU) survival curves were scanned and 'compared' in Stata, assuming an 'appropriate' distribution (Weibull, gamma, exponential, or log normal), and HRs compared to IFN were then used to estimate transition probabilities for													

	2003	<p>the model. The model was quite complex, and for two transition probabilities median survival had to be transformed to means under an exponential assumption. Also, for some transitions constant transitions were assumed, based on the literature. The manufacturer based their model on their trial for the initial 2 years, but then modelled long-term survival based on external trials and response rates. Crossover was an issue for the imatinib trial (0.7% crossed from imatinib to control, and 22.8% crossed from control to imatinib, due to intolerance; overall 2% and 58% had crossed over at 18 months), and the AC asked the group to also run their model using per protocol data rather than ITT data. The ICER reduced to £60k from £87k compared to HU (the treatment was already cost-effective compared to IFN).</p> <p>Imatinib was recommended as first-line treatment in the chronic, blast and accelerated phase. Although it may not have been CE compared to HU, it was against IFN, which was the main treatment used in the UK. This was a full recommendation.</p>														
TA86	Gastro-intestinal stromal tumours (GIST) - imatinib Oct 2004	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>x</td><td>✓</td><td>✓</td><td>✓</td> </tr> </table> <p>No head-to-head trials were available. For imatinib the manufacturer fitted exponential models to an imatinib trial, restricting data to 60 weeks for TTF, and 90 weeks for OS, due to heavy censoring and uncertainty after these time points. Sensitivity analysis suggested restricting the data made no difference to results. A similar method (using exponential curves fitted to data from an unpublished study) was used for IFN. The HTA group noted that the manufacturer had not calibrated the TTF and OS estimates, which led to unreasonable results. The HTA group therefore made amendments to the model to account for this. They also tested Weibull models. Crossover was an issue in the estimate of survival for the control group. The manufacturer included only patients who did not switch, whereas the HTA group tested including all patients but assuming the same risk of death upon progression for all patients. A further analysis by the DSU used data censoring when crossover occurred, which the AC thought was the least prone to bias. The manufacturer's method resulted in an ICER of £14k, the HTA group's method led to an ICER of £30k, and the DSU method led to an ICER of £32k.</p> <p>Imatinib was recommended as a first-line treatment, but should only be continued if a response is maintained. Full recommendation given license, but with cost agreement.</p>	✓	x	✓	✓	x	x	x	x	x	✓	x	✓	✓	✓
✓	x	✓	✓	x	x	x	x	x	✓	x	✓	✓	✓			
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review) May 2005	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>x</td><td>x</td><td>✓</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td> </tr> </table> <p>The HTA group stated that HRs for PFS and OS were the most accurate statistics for comparing treatments (proportional hazards are not discussed). HRs also could be meta analysed. Studies which quoted HRs were meta analysed (studies which did not were excluded as exponential assumptions about the medians would have to be made). However, an exponential transformation was used on a median to form an estimate of the mean for the base curve – this then causes all other curves to have exponential form. The group justified this as they had no PLD, and survival was short, so the form may not be of great importance. Two manufacturers used CMA, while the other appeared to have used a type of restricted mean approach estimating the AUC for KM curves from different studies, truncated at the same time point. The HTA group noted that this may have been biased due to different patient characteristics. Crossover was acknowledged to have occurred in at least two of the trials from which survival data was taken and included in the economic model, but no adjustments were made. In addition, the HTA group noted that third line drugs (the appraisal looked at second-line) and subsequent therapies may have confounded OS estimates. Thus this was an issue, but was not adjusted for.</p> <p>Paclitaxel in combination with carboplatin or cisplatin was recommended as a 2nd line (or subsequent) treatment option, as was paclitaxel alone and PLDH. Topotecan was recommended as a second-line (or subsequent) treatment for patients for whom PLDH and paclitaxel monotherapy are not appropriate. Paclitaxel was also licensed for use in 1st line but there did not appear to be evidence considered for this. The recommendations were for subgroups based on platinum sensitivity and resistance.</p>	✓	x	✓	x	x	✓	x	x	x	✓	✓	✓	✓	x
✓	x	✓	x	x	✓	x	x	x	✓	✓	✓	✓	x			
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (review) Aug 2005	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td> </tr> </table> <p>The HTA group modelled 1st line PFS, 2nd line PFS and OS, using sequencing data. KM curves were scanned using TechDig and the final portions of curves were extrapolated using Weibull models. Baseline curves were used and HRs (presumably from the Weibull model) were used to model the other treatments. Additional analyses using only the empirical KM curves were undertaken. For some treatments this actually led to slightly higher survival estimates. Log-cumulative hazard plots were constructed to test the appropriateness of the Weibull model, and it is noted that HR modelling implies proportional hazards (this is not tested). Fairly similar sequencing approaches using Weibull extrapolations were received by two manufacturers, though the data used in them was less complete. Crossover / post-progression treatments was an important issue, hence the use of sequencing modelling. This was a rare circumstance where a sequencing trial was available. The HTA group also state that where crossover is an issue and sequencing trials are not available using PFS as a measure is the only way to avoid confounding, but that there are problems with using PFS (schedule, surrogate). It was also noted that even in the sequencing trials further salvage therapy was sometimes used post the final progression.</p> <p>Irinotecan and oxaliplatin were both recommended in line with their licenses (irinotecan/5FU/FA 1st line, irinotecan alone subsequently; oxaliplatin/5FU/FA 1st or subsequent lines). Raltitrexed was not recommended (did not improve PFS or OS).</p>	✓	x	✓	✓	x	x	x	x	x	✓	✓	✓	✓	✓
✓	x	✓	✓	x	x	x	x	x	✓	✓	✓	✓	✓			
TA101	Prostate cancer (hormone-refractory) - docetaxel Jun 2006	<table border="1"> <tr> <td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td> </tr> </table> <p>The manufacturer completed two analyses – one used median OS and one used a Weibull extrapolation. This was justified by reference to the log cumulative hazard plot. The Assessment Group extended this, including more comparators. Indirect hazard ratios were estimated to allow the other comparators to be included. Crossover was an issue in one relevant trial of a comparator, but the HTA group noted that the treatment effect was actually stronger in this trial than in others, so no adjustments were made. Crossover was also an issue in at least two docetaxel trials, with patients switching between the different treatment arms. These were not discussed in the economic section. Crossover and alternative post-progression treatments were also allowed in other relevant trials, some consideration was given to the impact of this on costs, but this was not seen to make a significant difference and was not treated as an important issue.</p> <p>Docetaxel was recommended for a subgroup of patients, as this was the only population for which there was clinical evidence (despite the license being broader). Thus in this sense this was a full recommendation.</p>	✓	✓	✓	✓	x	x	x	x	x	✓	x	✓	✓	x
✓	✓	✓	✓	x	x	x	x	x	✓	x	✓	✓	x			
TA105	Colorectal cancer	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>x</td><td>x</td> </tr> </table>	✓	x	✓	x	x	x	x	x	x	x	x	✓	x	x
✓	x	✓	x	x	x	x	x	x	x	x	✓	x	x			

	- laparoscopic surgery (review) Aug 2006	The HTA group estimated recurrence and mortality rates for 6 month time periods from published survival curves. Rates were similar for each period, and so constant rates were used in the model. Relative risks were applied to these curves for the intervention. Crossover was not considered, probably due to the surgical nature of the intervention (assume no crossover). Laparoscopic surgery was recommended as an option.	✓	x	✓	✓	x	✓	✓	x	x	✓	x	x	✓/✓	?
TA110	Follicular lymphoma - rituximab Sep 2006	The manufacturer used a log-logistic extrapolation for PFS (a Weibull was also tested, increasing the ICER by £1k), and external registry data for the progressed state (using an exponential extrapolation). There was not an OS benefit in the RCT, and the HTA group suggested that the modelling technique suggested an implied OS benefit. They discuss the relationship between PFS and OS, but note that even when 0% of PFS benefit is extrapolated to OS (and using a Weibull for the registry data) the ICER remains below £20k. Crossover / post-progression treatment is not mentioned, but it may have been one reason that external data was used. Also OS is relatively long, so post-progression treatment seems likely. The HTA group show that a Weibull is a better fit than an Exponential for the external data (this increases the ICER), and note that the manufacturer does not consider the patient characteristics in the RCT and in the registry data. Rituximab was recommended in stage III and IV patients, in line with its license.	✓	x	✓	✓	x	✓	✓	x	x	✓	x	x	✓/✓	?
TA116	Breast cancer - gemcitabine Jan 2007	The manufacturer pooled absolute median PFS and OS times from different studies with no common comparator, and transformed the medians into means assuming an exponential distribution. The HTA group noted and illustrated that this was very prone to bias. The group re-estimated survival times using a parametric model (not stated which), and showed that the risk was non-constant – thus the exponential model was not reasonable. Post-progression treatments were costed in the manufacturer’s model, but it was not clear how this tallied with any post progression treatments received in the RCT. One relevant trial was noted to be affected by crossover (docetaxel patients switched onto docetaxel plus trastuzumab) and the OS for docetaxel from this study was above the 95% CI for docetaxel from pooled studies. Crossover and/or different post-progression treatments were also likely to have occurred in another study (the manufacturer’s key trial), but adjustments were not made. Gemcitabine combined with paclitaxel was recommended when docetaxel monotherapy or combined with capecitabine was also considered appropriate. This represented a subgroup/partial recommendation.	✓	x	✓	x	x	✓	x	x	x	✓	x	x	✓	x
TA118	Colorectal cancer (metastatic) - bevacizumab & cetuximab Jan 2007	One manufacturer fitted Weibull models to PFS and post progression, though for post progression the same risk of death was used for both groups to avoid crossover problems. Sensitivity analysis with an exponential was conducted, which reduced the ICER from £60k to £44k – the HTA group questioned this since the exponential is a special case of the Weibull [but when the number of parameters are considered it shouldn’t be assumed that the Weibull is better]. The HTA group chose instead to use post-progression data from the trials (with a Weibull extrapolation), because the trial seemed to show that crossover did not impact survival. Otherwise, the HTA group used TechDig to estimate KM curves and fitted Weibull models to the final portions. The other manufacturer extrapolated from KM curves at the point at which the intervention and comparator curves diverged (Gelber, 1993). For censored patients, OS was estimated by adding the additional survival beyond the censored timepoint estimated by the parametric curve. Given no comparator data for the intervention, HRs were used for a different treatment, from a different trial, which the HTA group deemed unreasonable. Crossover was an issue for bevacizumab, which the manufacturer attempted to adjust for, but the HTA group deemed this unnecessary. Continuation of bevacizumab treatments, as well as switching to a range of other post-progression treatments occurred in the bevacizumab trials (% that received different types of treatment were given). Sometimes these were off licence. The HTA group introduced 2 nd line PFS into the model based on sequences recommended by NICE, thus implicitly modelling sequences to cope with crossover / post-progression treatments. However, OS data was still used from the bevacizumab trial in which the confounding occurred. There is not a discussion on the crossover / post-progression treatment that may have occurred in the cetuximab trial. Neither bevacizumab or cetuximab were recommended.	✓	x	✓	✓	x	✓	x	x	x	✓	x	✓	✓	✓
TA119	Leukaemia (lymphocytic) - fludarabine Feb 2007	The manufacturer used RCT data to model PFS, but due to a lack of long-term data OS was assumed to be equal. Originally medians were used, but upon questioning from the HTA group means were estimated. PLD is used directly in the model for PFS, with OS estimates based on non-censored trial data and other sources used when patients are censored in the model. Patients in the trial were censored when they received 2 nd line treatment but 1 st and 2 nd line PFS is modelled and it is assumed that the response to the active treatment is the same at the 2 nd line as it was at the 1 st line, which was questioned by the HTA group, since the effectiveness of other treatments was assumed to decrease (based on pooled absolute estimates – which were also a cause for concern). Because OS was assumed to be equal, post progression was shorter for the active treatment. The model includes several states, and the transition probabilities are all constant over time – thus assuming an exponential distribution. The HTA group stated that this should have been justified. They obtained PLD and fitted Weibull distributions, to carry out a t-test to see if the shape parameter was equal to 1. They rejected the exponential confidently, as the shape was greater than 1, indicating an increasing hazard. The group also states that transition probabilities were estimated incorrectly as censoring is not taken into account (a formal survival analysis was not performed) and the risk of death is not correctly accounted for. Crossover and post-progression treatments were an issue in so much as OS was unknown and many different sequences could be relevant (including crossover). The HTA group felt that the sequence modelled may have been too rigid, and that other options should also have been modelled. Fludarabine monotherapy was not recommended. Fludarabine combination therapy could not be recommended due to a lack of licensing authorisation.	✓	✓	✓	✓	x	✓	x	x	x	✓	x	x	✓	✓
TA121	Glioma (newly diagnosed and	One manufacturer used median PFS and OS in their model. The other extrapolated beyond the trial period using a generalised gamma and Weibull, but the HTA group note that no justification for the models is given. They show that a number of Weibull models can be fit to the data that appear to have a good fit, but which give quite different mean survival estimates. The HTA group manually	✓	✓	✓	✓	x	x	x	x	✓	✓	✓	x	✓	✓

	high grade) - carmustine implants and temozolomide Jun 2007	<p>estimated the KM curve and fitted Weibull models wherever possible (in one circumstance a median had to be used). The fit of the models was tested by the R² statistic, and the closeness of the model estimates to the median trial estimates. Precedence was given to the R² statistic. Models were only fit to the first 2 years of data to eliminate tail effects, and it seems from the FAD that different models were used for the first 12 months and the second 12 months. The manufacturer also completed a restricted means analysis. Both crossover and different post-progression chemotherapies were important issues and the industry submissions did not adjust for these. The HTA group attempted to model sequences, but problems were that the survival benefits were confounded after PFS; which sequences were realistic; the uncontrolled nature of the sequences (ie what sequence to model); and not being able to determine how much of the survival benefit was associated with the different treatments – it was deemed not feasible to adjust effectiveness estimates because it was unknown how much of the OS was attributable to 1st and 2nd line treatments. The manufacturer of carmustine implants provided an analysis including post-progression treatment costs from the trial (including crossover) during consultation, whereas the HTA group produced a model costing post-progression treatments based on expert opinion/recommended treatment. The HTA group carried out these same analyses for TMZ (ie one analysis looking at within trial post-progression treatment, and one looking at expert opinion/recommended treatment).</p> <p>TMZ and carmustine implants were both recommended for certain patient subgroups.</p>	✓	✓	✓	×	×	✓	×	×	×	×	×	✓	×
TA124	Lung cancer (non- small-cell) - pemetrexed Aug 2007	<p>The manufacturer used pooled estimates of median OS, which the HTA group stated were likely to be biased. Median PFS estimates also seem likely to have been used. Medians were transformed into transition probabilities assuming a constant (exponential) probability per cycle. The HTA group argued that the clinical data showed no difference in any type of survival for the intervention. Crossover was highlighted as an issue by the HTA group (patients randomised to pemetrexed were allowed to switch to docetaxel, but patients randomised to docetaxel were not allowed to switch to pemetrexed) and post-progression treatment differences were also an issue, but the group were satisfied by simple analysis undertaken by the manufacturer that appeared to show that OS and PFS differences seemed similar, and no one treatment arm benefited more due to post-progression treatments received (the HR is no better for OS than it is for PFS). The manufacturer also showed that there was no difference in post-progression survival.</p> <p>Pemetrexed was not recommended.</p>	✓	×	×	×	×	×	×	×	×	×	✓	×	
TA129	Multiple myeloma - bortezomib Oct 2007	<p>The manufacturer used external observational data to estimate the survival experience of baseline patients, given the short follow-up of the key clinical trial. The HRs from the trial were applied for PFS and OS, assuming a treatment effect of 3 years. The HTA group state that a reduced OS HR is assumed in years 2 and 3, but this is not justified. The model was calibrated against observational and trial data. 5 treatment regimens are modelled in a sequence. The HTA group note problems with using observational data for the baseline, including patient characteristics, age of the data, and the order of sequences received.</p> <p>Crossover was an issue, which was a main reason for using the observational data for the control group, and HRs from 8.3 months of the clinical trial – before most (but not all) crossover would have occurred. Post progression treatment differences are likely to be an issue due to the observational data used, and the sequences modelled, but given that HR modelling is used it is most importance to clarify whether crossover occurred in the RCT from which the HRs are taken. In the trial patients were allowed to crossover after the 8.3 month interim analysis or upon disease progression. Data was only used up until the 8.3 month interim analysis, but by this time 44% of control patients had crossed over. Thus observational data was used to estimate long-term survival for the control group, but an unadjusted confounded HR was used to estimate OS for the intervention. Most of the discussion surrounded the possible bias associated with using the observational data which may have been for a more severe patient group, thus underestimating survival for the control, and possibly causing an overestimation of the effect of the intervention. However it was not considered that the OS HR used may have been biased against the intervention due to crossover.</p> <p>Bortezomib was recommended for its licensed subgroup of patients, which represents a full recommendation, but this is based on response and a rebate scheme.</p>	✓	×	×	×	×	×	×	×	×	✓	✓	✓	
TA135	Mesothelioma - pemetrexed disodium Jan 2008	<p>The HTA group did not have PLD from the key RCT and from the KM curve estimated restricted means and extrapolated using Weibull models after determining that exponential models were inadequate. TechDig was used, though weighting problems associated with this were noted, although weightings were interpolated. Estimates were compared to a restricted means analysis. Data from SEER were used to show that a small proportion of long-term survivors could be expected, hence explaining why the exponential would not be reasonable but a Weibull, with changing hazards, may be. The HTA group did note the drawbacks of relying on SEER data though, due to many possible confounding factors. The manufacturer presented two analyses. One appeared to use a restricted means approach based on the key RCT, whereas the other pooled absolute medians from a number of studies (without adjustment), and estimated means as a proportion of the average median based on the relationship seen in one trial. This was criticised by the HTA group. Crossover was not mentioned in the Appraisal, but further research shows that patients in the control arm were not allowed to move onto pemetrexed even after progression and the conclusion of the trial. However, 38% of the intervention group and 47% of the control group went on to receive other post progression chemotherapies, the most common being gemcitabine. This was not considered but could have effected OS estimates. Those who received post study chemotherapy had median OS of approximately 3 months longer in both treatment arms. This could have been due to patients living longer receiving extra treatment, or vice versa. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2503655/</p> <p>Pemetrexed was recommended for a subgroup of patients, which represented a partial recommendation as some groups for whom the treatment was licensed were not recommended for treatment.</p>	✓	✓	✓	×	✓	×	×	×	✓	✓	×	✓	×
TA137	Lymphoma (follicular non- Hodgkin's) -	<p>The manufacturer submitted two models, both of which used PLD from the RCT for 24 months, before using Weibull extrapolations for PFS and OS beyond the trial period. KM data was truncated at 1500 days for the model fitting due to the small event numbers and high uncertainty after this point. Sensitivity analysis using a log-logistic model was presented (which led to a lower ICER) but the Weibull was</p>	✓	×	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×

	rituximab Feb 2008	<p>chosen as the model of best fit (based upon AIC and BIC although the Weibull was not always best for these). Gompertz, exponential and log-normal were also considered, and the manufacturers stated that the Weibull is suitable when events happen early, whereas the logged models are more relevant when events occur late on. Proportional hazards are not tested, but the manufacturer states that the underlying assumption of the KM is proportional hazards, and thus because logged models are not PH models they are less desirable. In the FAD it is stated that SA was conducted fitting individual Weibull's and the ICER decreased. Some use of HR modelling was present – it was assumed that the Weibull's had the same shape – only the scale parameter was altered. The HTA group felt that in this case PH should have been justified, and it was not. Treatment effect was restricted to 5 years and was tested in SA. The HTA group believed model choice had not been justified sufficiently particularly as survival was long-term and future treatments were likely, and as an additional analysis carried out a restricted means analysis. Both crossover and post progression treatments were relevant, especially as survival was long-term. Many different treatments were given post progression, including rituximab containing regimens. The HTA group suggested that including treatment benefits only up until a new treatment is taken is one option to avoid problems, as how the previous treatment benefits after that is unknown. The manufacturer assumes that following progression patients will incur additional periods of active treatment (every 2 years on average) and allowed for costs of post progression treatments (though their methods for this was criticised by the HTA group), but no allowance was made for adjustments in survival, which were based solely on within trial data and extrapolation.</p> <p>Rituximab was recommended for induction and maintenance treatment and for treatment of relapsed patients when other chemotherapy options have been exhausted. This was not a full recommendation as some subgroups were specified (eg stage III/IV for two of the indications) and rituximab monotherapy could have been used slightly earlier than last line.</p>														
TA145	Head and neck cancer - cetuximab Jun 2008	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>?</td><td>x</td> </tr> </table> <p>A cure model was used to allow the manufacturer to estimate survival separately for those who had been cured and those who had not been cured. Survival for cured patients was based on adjusted life table figures. For non-cured patients the model estimated PFS and OS using a log-normal distribution with a logistic link function, which the manufacturer deemed appropriate because it allowed an initially increasing hazard which then decreases. Weibull and Exponential models were also tested, with the Weibull producing a lower ICER, and the exponential producing similar results. The HTA group suggested that this might suggest the cure model was little better than a simple exponential model. The manufacturer observed that survival curves were concave, backing up their choice of a log function rather than a Weibull or exponential, and the HTA group agreed with this. The AIC test was also used for the Weibull, exponential, log normal and log-logistic, and was lowest for the log normal. The log normal model also led to the smallest cure fraction, which was given as a conservative reason for its choice, but the HTA group noted that it also gave the biggest difference in cure fractions between treatment groups of all the models. The manufacturer appear to have completed a restricted means analysis as an extreme case sensitivity analysis, which was deemed useful by the HTA group because a greater proportion of active treatment patients remained non-progressed and alive at the end of follow-up, and hence if this analysis resulted in a cost-effective ICER it demonstrated that the extrapolation was not of great importance (this turned out to be the case). Crossover is not mentioned, it is just stated that salvage and subsequent therapy was well matched between the groups, thus this may not have been an important issue.</p> <p>Cetuximab was recommended for a specific patient subgroup.</p>	✓	x	✓	✓	✓	✓	✓	✓	x	✓	✓	x	?	x
✓	x	✓	✓	✓	✓	✓	✓	x	✓	✓	x	?	x			
TA162	Lung cancer (non-small-cell) - erlotinib Nov 2008	<table border="1"> <tr> <td>✓</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>✓</td><td>x</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>x</td><td>?</td><td>x</td> </tr> </table> <p>The manufacturer used KM data directly in the model. Survival was short so little extrapolation was deemed necessary. In the base case OS was assumed to be equivalent. Indirect comparisons were necessary and were unadjusted, which the HTA group were very concerned with. The manufacturer claimed that the mean survival for the new treatment was a restricted mean as follow-up was not quite complete, whereas the mean for the control was not a restricted mean as data from the relevant trial was complete. However, this would have assumed that there was no censoring in the docetaxel study. For one treatment a mean was not available and so a mean for treatment duration was used as a proxy for PFS. The HTA group suggested that an available median should have been used instead. In later analysis the manufacturer conducted a network meta analysis to try to prove OS equivalence, but the HTA group stated that the analysis was flawed due to the included studies. The HTA group performed an analysis restricting the means to the same time point for both key treatments, and also tested fitting exponential models. Some crossover occurred in the key erlotinib study even though it was not allowed in one key study – 7% in the placebo arm and 2% in the erlotinib arm received post study EGFR inhibitor therapy. There was no mention of crossover or post progression treatment in the docetaxel trial. Thus low proportions crossed over, and no adjustments were made.</p> <p>Erlotinib was recommended as a second-line treatment option alternative to docetaxel for a patient subgroup, under an equal costing agreement with the manufacturer. This represented a subgroup recommendation, since the product license was wider than the recommendation.</p>	✓	✓	✓	x	x	✓	x	x	x	x	✓	x	?	x
✓	✓	✓	x	x	✓	x	x	x	x	✓	x	?	x			
TA169	Renal cell carcinoma - sunitinib Mar 2009 and TA178 Renal cell carcinoma Aug 2009	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td> </tr> </table> <p>The HTA group fitted Weibull models to PFS and OS for the baseline comparator, and applied HRs to these to estimate survival for the interventions. The HTA group noted that although the manufacturer of bevacizumab stated that second line therapies were a confounder in their trial, there was not good evidence on this and thus OS estimates for the control group could not be adjusted. The HTA group state that treatment crossover was an issue for sunitinib, bevacizumab and sorafenib, and thus that the OS estimates for these drugs were highly uncertain. They state that longer term follow up of the relevant trials will not be useful due to the crossover. They present evidence of a censored analysis for sorafenib and show that the HR reduces from 0.88 to 0.78.</p> <p>For sunitinib, the manufacturer initially presented an analysis where a Weibull model was fitted to the control group PFS and OS, and the survival for sunitinib was estimated using HRs. In one analysis, the manufacturer used OS data for the control group from the bevacizumab trial, where unconfounded OS data was longer (this is a use of external data). The HTA group noted that if OS data for IFN were used from the bevacizumab trial then PFS data should also be taken from that trial. The model was re-run using OS data only from control patients who received no 2nd line therapy. This was in line with the AC's 'preferred assumptions'. The AC also preferred individual Weibull's to be fit rather than a HR approach. In a subsequent submission the manufacturer switched to using individual models rather than relying on the HR approach. Despite making use of HR modelling the proportional hazards assumption was not discussed by the HTA group. The HTA group were also asked to explain why their Weibull and the manufacturer's Weibull resulted in different mean estimates. The HTA group claimed that this was due to a different Weibull fit, and that their model represented a better fit (although this was</p>	✓	x	✓	✓	✓	✓	✓	x	x	✓	✓	✓	✓	✓
✓	x	✓	✓	✓	✓	✓	x	x	✓	✓	✓	✓	✓			

		<p>based only on a visual inspection). The HTA group used only 1 data point per month, to avoid early data points overly influencing the model fit [though this may be arguable – perhaps all data points should influence the model?]. The HTA group stated that this led to a model that fitted better to the end of the KM curve [but if this is very uncertain, is this reasonable?]. The DSU stated that individually fitted Weibull's seemed to give a better fit, although again this appeared to be based only on a visual inspection. In addition, the DSU note that the OS HRs differ markedly depending on the population used. Excluding patients who used 2nd line treatment the HR was 0.65, censoring those patients who switched onto sunitinib gave a HR of 0.81, and the ITT analysis gave 0.82. It is interesting that the censoring and exclusion approaches give such different results. The DSU preferred the censoring approach, but noted that even this should be undertaken with caution. The DSU also note that 11% of patients in the sunitinib arm went on to received more sunitinib after progression, but this was not taken into account in any economic analysis. Also of note, although the 2nd line exclusion approach was used for IFN OS estimates, it was not used for sunitinib OS estimates because it led to unfeasibly long OS estimates [those who did not receive further treatment had a better prognosis?]. There was less focus in the appraisal on the 2nd line sunitinib indication, but in their analysis, the manufacturer based control survival on SEER data in one analysis, and on pooled trial data in another, since no comparative trials were available. The HTA group noted that these unadjusted estimates were likely to be biased.</p> <p>For bevacizumab, the manufacturer fitted individual Gompertz models for PFS, but used a HR method for OS (rationalised by saying that PFS data were quite complete, whereas OS data were not). Alternative models were considered and justified using AIC, BIC and visual inspection. SA using a log-logistic model was presented, which reduced the ICER, but the HTA group deemed this unrealistic due to the length of survival estimated (although data was not used to back this up). The manufacturer reported that censoring all those who received 2nd line treatments reduced the HR from 0.75 to 0.61. The DSU also tested a scenario where the OS HR was the same as the PFS HR (0.63) as it may not be reasonable to estimate an OS HR that is lower than the PFS HR. The DSU state that this censoring leads to informative censoring, but that this could be analysed with access to PLD. In the absence of this, several censoring scenarios should be presented in order that an appraisal of the possible bias of the censoring could be made. The DSU state that the costs of 2nd line treatments could be included in the analysis, but this may result in modelling sequences not recommended by NICE.</p> <p>For temsirolimus individual Weibull models are fitted for PFS and OS. However, based on visual inspection the HTA group deemed these a poor fit to the data.</p> <p>For sorafenib the manufacturer deemed that PFS data was mature, while OS data was not. A restricted means approach was used for PFS, and OS was extrapolated using an exponential model.</p> <p>The HTA group also note that further research addressing the use of aggregate data rather than PLD would be useful.</p> <p>Thus it is clear that both crossover and post-progression treatments were important issues for this appraisal, and some attempts were made to adjust for these.</p> <p>Sunitinib was recommended as a first-line treatment option for a subgroup. This is narrower than the license, but similar to the population for which there was clinical evidence, so in this respect could be considered a full recommendation. None of the other treatments were recommended.</p>														
TA171	Multiple myeloma - lenalidomide Jun 2009	<table border="1" data-bbox="360 807 2163 831"> <tr> <td>✓</td> <td>✓</td> <td>✓</td> <td>✓</td> <td>x</td> <td>✓</td> <td>x</td> <td>x</td> <td>x</td> <td>✓</td> <td>x</td> <td>x</td> <td>✓</td> <td>✓</td> </tr> </table> <p>The manufacturer used a patient-level simulation model. For PFS, where disease progression was observed in the real patient record ascribed to the hypothetical patient the time to progression is taken from that record. Where progression had not been observed it was estimated using a Weibull model, with patient characteristics and best response taken into account in the model. Post progression survival was modelled in a similar way, except an exponential model was used with a range of predictors (a Weibull was tested, but did not improve the fit, and shape parameters were very close to 1). However around 50% of control group patients crossed over. 75% of the crossover occurred after progression. To correct for this post-progression crossover the post progression survival equation for the control treatment included a calibration factor which calibrated OS with that seen in previous MRC trials (use of external data). This assumes that the control (dexamethasone) has similar OS as the range of chemotherapies used in the MRC trials (data for 1372 patients). OS is relatively long for this disease (30-year time horizon). The manufacturer conducted analysis to suggest that this was the case. Parametric survival analysis was used on the MRC trial data to derive an equation for OS including a range of patient characteristic variables, the values of which were set to the means from the relevant intervention RCTs in order to estimate unconfounded median OS for these trials. The median for the intervention RCTs was estimated by matching to the median OS from the MRC trials. The manufacturer justifies the use of the RCT data because of the large patient numbers. Although the data is relatively old (patients enrolled between 1980 and 1997) they show that there is not trend for improved OS over time.</p> <p>The HTA group noted some problems with the manufacturer's analysis. Firstly, they noted that mean OS rather than median OS should have been used from the MRC data (there was debate between the manufacturer and the HTA group surrounding this, with the manufacturer arguing that calibrating to the mean would place too much weight on unknown event times at the end of the distribution, ignoring the events that occurred earlier. The HTA group stated the importance of means, and the importance of tails, hence why medians were not justified. This was particularly the case because the OS data from the MRC trials had very little censoring and 94% were said to have died). Secondly, they noted that there were likely to be important patient characteristics not reported in both the intervention RCTs and the MRC trials which could therefore not be included in the OS equations. Thirdly, they noted that data from the Mayo clinic show trends to improved survival between 1995 and 2006, thus suggesting the MRC trials may represent an underestimate of control OS.</p> <p>The HTA group therefore suggested that it would be valuable to re-run the model with censored crossover data, rather than relying on the external data. The HTA group noted that when crossover was not accounted for in any way the ICER increased from £25k to £79k. The HTA group noted that the control survival estimates were much lower than those seen in the relevant RCT, as expected due to the use of the MRC trials. However they also noted that OS for the intervention was much higher than in the RCT.</p>	✓	✓	✓	✓	x	✓	x	x	x	✓	x	x	✓	✓
✓	✓	✓	✓	x	✓	x	x	x	✓	x	x	✓	✓			

		<p>The HTA group showed that by matching OS medians the tail of the exponential distribution used is not taken into account, and thus the estimates used by the manufacturer fitted the exponential distribution poorly. The group showed that fitting to the mean increased the ICERs by approx £8k (to over £30k) for two of the subgroups.</p> <p>An additional note, not stated by the HTA group, is that the adjustment to OS estimates seem to suggest that the impact of adjusting for crossover is similar in magnitude to the impact of randomisation to the intervention group – the OS benefit is roughly doubled. Is this valid?</p> <p>The AC accepted the use of the MRC data, and the reasoning that there had been little improvement in OS over time. They preferred the HTA group’s method of calibrating to the mean and for estimating OS with the intervention. However, the AC also noted that calibrating to both the mean and the median led to an improvement in OS predicted by the model which was out of proportion given the improvement seen in PFS.</p> <p>Lenalidomide was recommended only as third line treatment, and with a costing agreement with the manufacturer. This was a partial recommendation since evidence was available for the use of the treatment after 1 prior treatment.</p>														
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab Jun 2009	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>✓</td><td>x</td> </tr> </table> <p>The manufacturer fitted individual Weibull models to the PLD for PFS and OS from the relevant trial. Log Normal and log-logistic models were also tested, but the Weibull was chosen based on the log likelihood statistic. The manufacturer also stated that the log models resulted in heavy tails which were clinically unrealistic. The HTA group questioned whether extrapolation was necessary since OS data was available for around 75-80% of patients (although it is not clear if this took into account censoring). The group also noted that although some justification for the Weibull models chosen was given for the base case analysis, similar justification was not given for sub-group analysis. The group noted that the Weibull models did not appear to match the KM curve well in the middle time period of the curve, and that some of the sub-group analyses seemingly involved very low patient numbers, potentially making model fitting unreliable. The HTA group conducted a type of restricted means analysis – still using the Weibull models, but truncated at 24 months. The group argued that in end-stage diseases extrapolating almost complete OS data is more likely to exaggerate health gains than underestimate them, because the intervention is not a cure and usually the benefit will simply reduce the risk for a limited period so that virtually all the benefit will have been gained before the final patient dies. They state that this is particularly the case as the KM curves had converged closely, and there is no apriori reason to expect them to diverge at a later time. Thus they deemed it most appropriate to truncate at 24 months in order to avoid any “risk of spurious artefactual differences arising from ill-advised projection”. This increased the ICER from £121k to £148k. The group also note that the Weibull’s for PFS and OS were estimated independently of each other. Because these are likely to be correlated the model parameters should have been jointly estimated – this is particularly important for PSA. Crossover was not mentioned and seemed unlikely to be an issue in the trial. However, further research showed that 6% of the control group received cetuximab after progression, hence crossover did occur. It was also noted that data on any post study chemotherapy given was collected in the trial, although data on this is not presented. Therefore post-progression treatments could be an issue. http://content.nejm.org/cgi/content/full/359/11/1116</p> <p>Cetuximab was not recommended.</p>	✓	x	✓	✓	x	x	✓	✓	x	✓	✓	x	✓	x
✓	x	✓	✓	x	x	✓	✓	x	✓	✓	x	✓	x			
TA174	Leukaemia (chronic lymphocytic, first line) - rituximab Jul 2009	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td><td>✓</td><td>x</td><td>✓</td><td>✓</td><td>✓</td> </tr> </table> <p>The manufacturer modelled PFS based on fitting individual Weibull models to the RCT data for the direct comparison. However the same shape parameter was used for the Weibull’s, and hence this was a type of HR modelling. Exponential, log normal, log-logistic and gompertz models were also tested. Death while in the PFS state is based on the larger of two probabilities in each cycle – probability from the RCT, and from the general population. The manufacturer presented a diagram of the post progression KM curves (not OS) which showed them to be very close together and crossing regularly – hence they modelled post progression as one population. A constant mortality rate (exponential) based on the inverse of the mean from the trial was used. For indirect comparisons HR modelling was used.</p> <p>The HTA group state that they were happy that the manufacturer had justified why the Weibull model was best, but do not discuss the justification. The manufacturer used AIC, BIC, mean squared deviance and graphical inspection of fit (eg martingale residuals). Mean squared deviance was used because SAS does not allow AIC and BIC to be estimated for the Gompertz model. They also stated that Weibulls with the same shape gave the best fit, as there was no indication that the shape associated with the two treatment arms differed. They stated that based on the Martingale plots there was no evidence of the proportional hazards assumption being violated. The HTA group stated that the major uncertainty was associated with the plausibility of extrapolating benefits in PFS to OS – the constant hazard after progression assumed a relationship between PFS and OS which the group believed had not been proven. Based on this the manufacturer produced a version of their model where no OS gain was assumed, causing an increase in the ICER to £30k. They also note that the choice of parametric model for PFS is important – the Weibull gave an ICER of £13k, log normal, log-logistic and exponential range between £10k and £13k, and the Gompertz causes an increase to £23k. The group noted that the manufacturer’s model was too simplistic, using just three health states, since further treatment will be received in reality. Direct crossover was not allowed in the trial, but patients with stable or progressive disease were allowed to switch onto any other treatment deemed appropriate, including rituximab containing regimens after 3 treatment cycles. Thus post-progression treatments / crossover were issues. The AC accepted that this made long-term OS difficult to prove. They were advised that gains in PFS were likely to lead to gains in OS. The post progression modelling used by the manufacturer is a way of avoiding the crossover issue.</p> <p>Rituximab combined with fludarabine and cyclophosphamide was recommended in situations where fludarabine and cyclophosphamide is considered appropriate. This is not quite a full recommendation as rituximab is licensed with any chemotherapy combination (though there is little evidence of this), and chlorambucil is still given to several patients with particular characteristics.</p>	✓	x	✓	✓	✓	✓	✓	✓	x	✓	x	✓	✓	✓
✓	x	✓	✓	✓	✓	✓	✓	x	✓	x	✓	✓	✓			
TA178	Renal cell carcinoma Aug 2009	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td> </tr> </table> <p>See TA 169.</p>	✓	x	✓	✓	✓	✓	✓	x	x	✓	✓	✓	✓	✓
✓	x	✓	✓	✓	✓	✓	x	x	✓	✓	✓	✓	✓			

TA176	Colorectal cancer (first line) - cetuximab Aug 2009	✓	x	✓	✓	x	x	✓	✓	x	x	✓	✓	✓	✓/?
<p>Small subgroups from the two relevant trials are used in the economic model (those with KRAS Wild-type who had liver metastases only). Liver resection was key for the model, although few resections occurred in the main trials. Resection rates were therefore assumed and survival for these patients was estimated based on external trial data (and using log logistic extrapolation). The manufacturer used the pivotal RCTs to estimate 1st line PFS (there were two RCTs testing cetuximab in combination with different chemotherapies, so these were not synthesised and two models were produced). 2nd and 3rd line treatment and time to progression were modelled based on external trials. For each line parametric models were used, with Weibull, log normal and log logistic curves all being used. It appears that a type of HR modelling had been used for PFS but this was not well explained. The ‘most suitable’ models were used, but this was not explained (this may have been in the appendices to the manufacturer’s submission, but these were not available on the NICE website). A sensitivity analysis included using the trial data for the trial period and only using the data from the parametric models after this period. This made a large difference to the ICER (£69k to £54k) for one of the models. This suggests that the parametric models did not fit the KM curve well, but this is not discussed. Alternative parametric models were also seen to alter the ICER significantly (eg using a Weibull instead of a log normal for PFS in the Folfox model increased the ICER from £63k to £70k).</p> <p>Unusually, there was very little discussion by the HTA group regarding the survival analysis conducted by the manufacturer. They did express concern with the 20 year model time line, considering survival at 5 years is around 12% based on registry data. They re-ran the model restricted to 5 years and the ICERs increased from £69k (folfiri model) and £63k (folfox model) to £124k and £143k respectively. The manufacturer noted the issue of crossover, but the HTA group did not. The manufacturer reported high levels of both post-progression crossover and subsequent therapies, with the use of subsequent EGFR antibody therapies higher in the control groups. Two-thirds received subsequent chemotherapy, and one-quarter received subsequent EGFR treatment. The manufacturer therefore states that OS estimates will be confounded. It is not discussed, but the treatment sequencing modelling approach may avoid this issue as only PFS from the key RCT is used, but only if crossover did not occur in the trials used to estimate longer term survival. In the previous NICE colorectal cancer appraisal that looked at treatment sequences, it was noted that salvage therapies were received even after the third-line, and thus post-progression treatment could confound OS estimates even using a sequencing approach.</p> <p>In a subsequent DSU report there was some discussion of amending PFS curves when applying a treatment stopping rule.</p> <p>Cetuximab was recommended for a patient subgroup, with a cost agreement with the manufacturer. Thus this was a partial recommendation.</p>															
TA179	Gastrointestinal stromal tumours - sunitinib Sep 2009	✓	x	✓	✓	x	x	x	x	x	x	✓	x	✓	✓
<p>The manufacturer fitted individual Weibull models to PFS and OS for the intervention and control. HR modelling was considered but judged (visually) that this did not lead to curves that fitted well for sunitinib, thus HR modelling was only used in sensitivity analysis (ICER £16k) (therefore there was justification for this, but not for the use of Weibulls). ITT data was used for PFS and OS for sunitinib, and for PFS for BSC. However for OS with BSC the data was adjusted using the RPSFT method due to treatment crossover (84% crossed over) (ICER was £27k). A full ITT analysis was presented as sensitivity analysis (£77k). The Weibull models were fitted using 1 data point per month to avoid early data points dominating the fit. The HTA group stated that this gave a good fit – there is not a discussion on not using all the trial data.</p> <p>While stating that a search of the literature suggested that the RPSFT method had not before been used in CEA, the HTA group believed that it was the correct method to use. They could not guarantee that it had been used correctly, but gave some assurances that it was a justified method based on external statistical advice, and that the results seemed reasonable as the HR was similar to the interim HR prior to crossover. The group agreed with the manufacturer that censoring at crossover involved informative censoring and thus was biased, even though it is the method commonly used in NICE appraisals. The AC requested this analysis, but little weight was given to it as only 15 patients did not crossover. The analysis led to BSC dominating sunitinib. It is interesting to note that the manufacturer described the RPSFT method as the only method available from the literature that can correct for time-dependent treatment changes in survival data while respecting the randomisation. The HTA group do not question this.</p> <p>Alternative analyses run by the HTA group included assuming there were no treatment benefits after progression – ie probability of death was the same after progression (£47k), OS HR of 1.0 (£230k), OS HR of 0.262 (£18k) (these were the CIs for OS using the RPSFT method).</p> <p>An additional factor noted by the HTA group was that 54 patients in the sunitinib continued using sunitinib after progression, but their survival was not adjusted – hence ICERs may be underestimates.</p> <p>Sunitinib was recommended as a treatment option for a subgroup if imatinib had failed, reflecting its license, but with a cost agreement with the manufacturer.</p>															
TA181	Lung cancer (non-small cell, first line treatment) - pemetrexed Sep 2009	✓	x	✓	✓	x	✓	x	x	x	✓	✓	✓	✓	x
<p>In the manufacturer’s original base case analysis 30 month RCT data was extrapolated to 6 years by taking the median OS from the trial and converting it into a per cycle risk of death (exponential). Indirect comparisons were also used but they used individual trial arm data, thus ignoring randomisation (though some HR modelling was also used). Subsequently the manufacturer submitted further versions of the model, one using a ‘within trial’ analysis – presumably using restricted means – and one using Weibull curves.</p> <p>The HTA group stated that the exponential based OS estimates did not seem to fit the published survival curves, leading to an underestimation of survival in the model. Both PFS and OS were modelled, but any comparison of the model to trial survival was only given for OS, and this was in insufficient detail. The HTA group requested trial data, which was provided, and demonstrated that both PFS fits and OS fits were stated to be poor (based on a visual inspection). The group stated that the Weibull models were an improvement, but state that the Weibull fit will place most weight on early and intermediate events, with less weight placed on sparser events towards the end of the trial, potentially leading to systematic over- or under-estimation of survival towards the end of the trial. They go on to show that</p>															

		<p>the Weibull is not a perfect fit to PFS or OS (though it is better than the exponential, and this is based only on a visual inspection). The HTA group reanalysed the data and used the area under the KM curve combined with expected mean survival for patients still alive using an exponential extrapolation in order to estimate total mean survival. They justified their approach to projection by examining the cumulative hazard function for each subgroup. They observed that models such as the exponential, Weibull, Log normal, were not compatible with the trial data across the whole range of observation, which they expected given that treatment is of limited duration and would be expected to have a short-term effect before the long-term disease progression pathway resumes. They observed that at some point following the end of treatment the cumulative hazard function assumed a steady linear increase, indicative of a constant risk per unit of time. Thus for each subgroup an exponential model was fitted to the data from the point at which the long-term linear trend in the cumulative hazard became established. This was used to estimate the likely additional mean survival from the time of the last recorded death until the time horizon of the economic analysis. This approach led to slightly improved survival estimates and slightly reduced ICERs.</p> <p>Crossover is not mentioned, but over half of patients received second-line therapy. The manufacturer costed this in their model but made no adjustments to survival. They stated that it was not possible to disaggregate survival benefits from first and second-line treatments and thus they assume that all second line treatments have equivalent efficacy. The HTA group concluded that as the proportion of patients that received post-treatment was similar in the treatment arms, it was unlikely to cause significant bias. They also referenced analyses by the EMEA which excluded patients who switched treatment and found that results were similar.</p> <p>Pemetrexed combined with cisplatin was recommended for a subgroup, reflecting a partial recommendation.</p>														
TA183	Cervical cancer (recurrent) - topotecan Oct 2009	<table border="1"> <tr> <td>✓</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>?</td><td>x</td> </tr> </table> <p>The manufacturer conducted both a direct and an indirect comparison. The direct comparison had a timeframe of 36 months, and the indirect comparison had a timeframe of 24 months. As most patients had died by these times the manufacturer did not extrapolate and thus a type of restricted means analysis was used. HR methods were used for the indirect comparison. Where 3 year data was available in one trial and only 2 year data in another trial the HR for the 3 year data was applied only to the 2 years of data from the reference trial. The HTA group suggested an alternative method, which was estimating the HR in the 2-year trial and using the 3-year trial as the reference, thus assuming that the 2-year HR was appropriate for a 3-year analysis. The manufacturer subsequently presented this analysis.</p> <p>There is no mention of crossover, perhaps because the analysis was purely trial based. Although there was also no discussion of post-progression treatments that may have been received even though the trial protocol stated that data on any follow-up treatment given before progression would be collected, while data on follow-up treatment after progression would not be. Possibly the severity of the disease limited the scope for switching/follow-up treatment. Further research did not shed light on this, although there is some suggestion that crossover could have occurred. http://jco.ascopubs.org/cgi/content/full/23/21/4626</p> <p>The HTA group consider the informative censoring that may have occurred in the trial. Patients removed from the trial due to side-effects are subject to informative censoring, and there may be a relationship between drop out and earlier death. They therefore discuss the method of inverse-probability weighting for estimating the parameters required for a CEA when dependent censoring is present, proposed by Willan <i>et al</i> (2005). The group state that this may have been a more suitable method for use in this appraisal, where the probability of being observed may be estimated conditionally on a series of covariates. However they note that it is not clear if the method used (KM method) caused any significant bias.</p> <p>Topotecan combined with cisplatin was recommended as an option for stage IVB patients who had not previously received cisplatin. This represented a partial recommendation since there was evidence on repeat treatment with cisplatin, and this was allowed by the license.</p>	✓	x	x	x	x	x	x	x	x	x	✓	✓	?	x
✓	x	x	x	x	x	x	x	x	x	✓	✓	?	x			
TA184	Lung cancer (small-cell) - topotecan Nov 2009	<table border="1"> <tr> <td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>✓</td><td>✓</td><td>x</td><td>x</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>x</td> </tr> </table> <p>The manufacturer estimated PFS and OS purely based on the trial data. At the final follow-up 3 patients in both the BSC and intervention groups were stated to be alive. They were therefore assumed to die the following day. A type of restricted means approach was thus taken. The HTA group stated that based on the KM curves it appeared fewer people remained alive in the control group than in the intervention group at the end of follow-up. If there were 3 patients in each arm that would represent 4% and therefore the HTA group stated that not extrapolating could underestimate the survival benefit for either arm of the trial. Thus, the HTA group reanalysed the KM data using TechDig and extrapolated the final portions using a log-logistic function (to a maximum of 5 years), as this was said to fit the KM curve best (based on R², the sum of residuals, plots of the log hazard trial data and visual inspection). Only the log-logistic and the Weibull were compared. The Weibull was said to underestimate survival over time, given that the Weibull estimated that close to 0% of patients would be alive at 100 weeks, whereas in the trial around 5% were alive. The log-logistic model estimated 4%. According to the HTA analysis, the restricted mean approach overestimated the ICER – extrapolation led to an increased survival benefit estimated for the intervention. In their analysis initially the HTA group had to estimate time to progression by transforming medians into means assuming an exponential distribution. However, the manufacturer submitted more data which allowed Weibull and log-logistic models to be considered. Sensitivity analysis was conducted using different distributions.</p> <p>Crossover is not mentioned. It is noted that around 18% of patients in each trial arm received post-study chemotherapy, and post-study radiotherapy was used in 10% of intervention patients and 1% of control patients. Therefore there could be an important difference in post-progression treatment.</p> <p>Oral topotecan was recommended as an option for people with relapsed small-cell lung cancer when retreatment with first-line treatment is deemed inappropriate, and where the CAV combination (cyclophosphamide, doxorubicin and vincristine) is contraindicated. This represented a partial recommendation.</p>	✓	x	✓	✓	x	✓	✓	x	x	✓	✓	✓	✓	x
✓	x	✓	✓	x	✓	✓	x	x	✓	✓	✓	✓	x			

Appendix 3: Summary of treatment crossover in NICE appraisals of advanced and/or metastatic cancer treatments

Ref	Title of appraisal	Did crossover occur?	Were post-prog treatments given?	Was OS estimated?	Was crossover addressed?	How was crossover addressed?	Was crossover an important issue?	Was the treatment recommended?	Crossover / post-progression tx an issue, and a negative recommendation
TA3	Ovarian cancer - taxanes (replaced by TA55)	✓	×	✓	×	-	-	✓ first choice	×
TA6	Breast cancer - taxanes (replaced by TA30) Indications: Pac, 1 st and 2 nd line Doce, 2 nd line	✓	✓	✓ second-line indication × first-line	× second-line ✓ first-line	Relied upon PFS for first-line	-	✓ second-line indication, as an option × first-line	✓ first-line paclitaxel
TA26	Lung cancer - docetaxel, paclitaxel, gemcitabine and vinorelbine (updated by and incorporated into CG24 Lung cancer)	Probable	Probable	✓	×	-	-	✓ as an option	×/option
TA28	Ovarian cancer - topotecan (replaced by TA91)	✓	×	Unclear	✓	Censoring (only used data pre-crossover)	-	✓ subgroup as an option	×/subgroup/option
TA29	Leukaemia (lymphocytic) - fludarabine (replaced by TA119)	Probable	Probable	×	✓/× (indirect control)	Time in remission	-	✓ first choice	×
TA30	Breast cancer - taxanes (review)(replaced by CG81) Pac, 1 st and 2 nd line Doce, 1 st and 2 nd line	✓	✓	✓ second-line indication Unclear first-line	× second-line Unclear first-line	Unclear	-	✓ second-line as an option × first-line	✓ first-line paclitaxel and docetaxel
TA34	Breast cancer - trastuzumab	✓	×	✓	✓	Exclusion and case-mix approach	✓	✓ subgroup as an option	×/subgroup/option
TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (replaced by TA93) Indications: Oxali 1 st line; Iri 2 nd line; ralti	✓	✓	×	✓	PFS	✓	✓ Oxali subgroups as an option ✓ Iri subgroups as an option × Ralti	×/subgroup/option ×/subgroup/option ✓ Ralti (but did not improve pfs or os)
TA45	Ovarian cancer (advanced) - pegylated liposomal doxorubicin hydrochloride (replaced by TA91)	Probable	Probable	✓	×	-	-	✓ as an option	×/option
TA50	Leukaemia (chronic myeloid) - imatinib (replaced by TA70)	Probable	Probable	✓	✓/× (indirect control)	Used external trial data for extrapolation	-	✓ as an option	×/option
TA54	Breast cancer - vinorelbine (replaced by CG81)	Probable	Probable	✓ in at least one model	✓/× (indirect control)	A number of models were based on PFS	-	✓ partial as an option	×/partial/option
TA55	Ovarian cancer - paclitaxel (review)	✓	×	✓	×	-	-	✓ as an option	×/option
TA62	Breast cancer - capecitabine (replaced by CG81)	Probable	✓	✓	×	-	-	✓ first choice for combination therapy indication, but as an option for monotherapy indication	×/option
TA61	Colorectal cancer - capecitabine and tegafur uracil	Probable	✓	✓	×	-	-	✓ as an option	×/option
TA65	Non-Hodgkin's lymphoma - rituximab	Probable in both trial and	Probable in both trial	✓	✓/× (indirect control)	Used external trial data for extrapolation, but HRs and	-	✓ first choice	×

		external	and external			long-term control may be confounded			
TA70	Leukaemia (chronic myeloid) - imatinib	✓	x	✓	✓	Per protocol	✓	✓ first choice for chronic phase, as an option for blast/accelerated phase	x/option
TA86	Gastro-intestinal stromal tumours (GIST) - imatinib	✓	x	✓	✓	Exclusion, censoring, ITT	-	✓ first choice, but with response restrictions	x response restrictions
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review)	✓	✓	✓	x	-	-	✓ partial as an option	x/partial/option
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (review)	✓	✓	✓	✓	Sequencing (but still issue for salvage)	✓	✓ for iri and oxali as options x for ralti (did not improve PFS or OS)	x/options for iri and oxali ✓ Ralti (but did not improve pfs or os)
TA101	Prostate cancer (hormone-refractory) - docetaxel	✓	✓	✓	✓ (costs only)	Inc some costs	x treated as a minor issue	✓ as an option	x/option
TA110	Follicular lymphoma - rituximab	Probable	Probable	✓	✓/x (indirect control)	Use of external data for post prog survival in both arms	-	✓ as an option	x/option
TA116	Breast cancer - gemcitabine	✓	Probable	✓	✓ (costs only)	Inc some costs	-	✓ partial as an option	x/partial/option
TA118	Colorectal cancer (metastatic) - bevacizumab & cetuximab	✓	✓	✓	✓ (bevacizumab)	Manufacturer assumed same risk of death upon progression for both groups. HTA group modelled sequences, but still used OS from trial	x, HTA group thought that the crossover did not appear to impact results, used trial data	x	✓
TA119	Leukaemia (lymphocytic) - fludarabine	✓	✓	✓	✓	Assumed equal OS, and limited modelling of sequences	-	x	✓
TA121	Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide	✓	✓	✓	✓	Inc some costs within attempts to model sequences, but could not allocate effect	-	✓ subgroup as an option	x subgroup/option
TA124	Lung cancer (non-small-cell) - pemetrexed	✓ (but in favour of new intervention)	✓	✓	x	Considered similarities between PPS for two arms	x deemed that no one arm benefited more from post prog tx	x	✓
TA129	Multiple myeloma - bortezomib	✓ (in trial)	✓ (obs data used)	✓	✓	Used data prior to interim analysis to estimate HRs, but xover had already occurred	x paid more attention to the use of the obs data	✓ as an option and with response and rebate scheme	x subgroup/option with response and rebate scheme
TA135	Mesothelioma - pemetrexed disodium	x	✓	✓	x	-	-	✓ partial as an option	x partial/option
TA137	Lymphoma (follicular non-Hodgkin's) - rituximab	x (although rituximab-containing regimens were allowed)	✓	✓	x	-	-	✓ partial as an option	x partial/option
TA145	Head and neck cancer - cetuximab	x	✓	✓	x	-	x noted that salvage and	✓ partial as an option	x partial/option

							subsequent tx were well matched,		
TA162	Lung cancer (non-small-cell) - erlotinib	✓	Probable	✓	×	-	×	✓ partial as an option with cost agreement	×
TA169 and TA 178	Renal cell carcinoma - sunitinib and Renal cell carcinoma	✓ for sunitinib, bevacizumab and sorafenib (including continuation of active tx – not controlled for)	✓	✓	✓ (suni and bev, not sorafenib)	Use of external trial data for longer unconfounded data (suni); Exclude any patients that received 2 nd line tx (suni and bev); Censoring patients that received 2 nd line tx (suni and bev)	✓	✓ sunitinib as an option ×	×
TA171	Multiple myeloma - lenalidomide	✓ (in trial)	✓ (obs data used)	✓	✓	PLD from MRC trials used to calibrate OS estimates for control	✓	✓ partial as an option with costing agreement	×
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab	✓	Probable	✓	×	-	×	×	✓
TA174	Leukaemia (chronic lymphocytic, first line) - rituximab	×	✓	✓	✓	Post-prog modelled as one population – same risk of death	-	✓ partial as an option	×
TA176	Colorectal cancer (first line) - cetuximab	✓	✓	✓	✓	Sequences and long-term survival modelled using other trials (but these may have also been confounded)	✓ noted as important by manufacturer, not noted by HTA group	✓ partial but as first choice, with costing agreement	×
TA179	Gastrointestinal stromal tumours - sunitinib	✓ (inc continued tx in intervention arm – not adjusted for)	×	✓	✓	RPSFT; censoring; no benefit after progression; CI OS HR analysis	✓ large degree	✓ as an option with costing agreement	×
TA181	Lung cancer (non-small cell, first line treatment) - pemetrexed	Possible	✓	✓	✓ (costs only)	Included costs of post prog tx	×	✓ partial as an option	×
TA183	Cervical cancer (recurrent) - topotecan	Possible	Probable	✓	×	-	-	✓ partial as an option	×
TA184	Lung cancer (small-cell) - topotecan	Possible	✓	✓	×	-	×	✓ partial as an option	×

Appendix 4: Evidence tables

Initial Search

Reference	Robins JM and Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics – Theory and Methods 1991;20;8:2609-2631
Origin	
Was the method developed specifically in	Yes. It was developed to deal with non-random non-compliance in a survival context.

the survival analysis context?	
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The rank preserving structural failure time models are the structural/strong version of accelerated failure time models with time-dependent covariates as introduced by Cox and Oakes (1984). The models are 'structural' because they directly model the survival times that would have been observed had, contrary to fact, all exchangeable subjects received the same treatment regime. They are 'rank preserving' because if individual i fails before j on one treatment regime, then i would also fail before j on any other regime.</p> <p>The RPSFTM method provides an estimate of the true treatment effect by relating an individual's observed event time to their counterfactual event time. The counterfactual event time is that which would have occurred if treatment crossover had not occurred. The method is best described through an example. Imagine an RCT with two arms, where A is the control arm and B is the intervention arm. Each patient (i) has an observed time to event (T_i) or censoring time (Δ_i). In addition, each patient also has a counterfactual time to event (U_i) which is the unobserved time to event that would have occurred if no treatment had been received (and, for the control group, if no crossover had occurred). U_i is only observed for patients in the control arm who do not switch treatment onto the intervention – that is it is only observed for patients who have received no treatment with the intervention whatsoever. The RPSFTM method makes the assumption that U_i is independent of the treatment group due to the balance provided by the randomisation process.</p> <p>The method splits the observed event time (T_i) for each patient into a two-stage process, that is the event time when the patient is on the control treatment (T_{Ai}), and the event time when the patient is on the intervention treatment (T_{Bi}). For patients who are randomised to the intervention treatment, and who do not switch onto the control treatment, T_{Ai} is equal to zero, and for patients randomised to the control group who do not switch onto the intervention T_{Bi} is equal to zero. However, for patients who crossover treatments both T_{Ai} and T_{Bi} will be greater than zero.</p> <p>The RPSFTM method relates T_i to the counterfactual event time (U_i) with the following causal model:</p> $U_i = T_{Ai} + e^{-\psi_0} T_{Bi} \quad (1)$ <p>In this model, e^{ψ_0} represents the acceleration factor associated with the intervention. The acceleration factor represents the amount by which the intervention increases the patient's expected time to event. If e^{ψ_0} is greater than 1 the intervention slows down the speed at which the patient moves towards the event of interest (deceleration) thus increasing the time to event, signifying a beneficial treatment. Conversely, if e^{ψ_0} is less than 1 the patient moves toward the event more speedily, decreasing the time to event and signifying a detrimental treatment.</p> <p>The value of ψ must be estimated, which allows the counterfactual event time (U_i) to be calculated. This is achieved by defining a binary process $X_i(t)$ which equals 1 when a patient is on the intervention treatment, and equals zero when the patient is on control treatment. Under this process, equation (1) can be rewritten as:</p> $U_i = \int_0^{T_i} \exp[\psi X_i(t)] dt \quad (2)$ <p>Solving equation (2) will provide several potential values for ψ, which can then be tested using the test statistic $Z(\psi)$, under the assumption that the distribution of the baseline lifetime U_i is independent of treatment group. Wald tests from parametric models, or non-parametric tests such as the logrank or Wilcoxon tests can be used to calculate $Z(\psi)$, and the value of ψ that is chosen for ψ_0 is that which results in $Z(\psi)=0$, as this represents the value for which the counterfactual event time is equal for the two treatment groups, which is a key assumption of the method.</p> <p>The model treats the potential untreated outcome as a baseline characteristic which must be balanced between randomised groups and thus the treatment effect estimation is based on finding a treatment effect which achieves a balance in the potential untreated outcome.</p>
What are the key assumptions of the method?	<p>The authors state that in general the survival differences that would have been observed had all subjects remained on protocol will not be identifiable without making various assumptions that are themselves non-identifiable, the simplest and least plausible of which is that subjects who comply with their assigned treatment protocol are comparable to those who fail to comply with regard to important prognostic factors. However, the RPSFTM method does not require this assumption.</p> <ul style="list-style-type: none"> i) The time an individual would fail at if never treated is not related to the treatment arm to which the individual is assigned – should be ok in RCTs. ii) Equal treatment effect no matter when treatment is received. iii) If an individual fails before another individual on one treatment regime, he will also fail before that other individual on all other treatment regimes. iv) There is a strong non-interaction assumption: eg if two individuals had identical observed failure times and treatment histories, it would be assumed that they would also have identical failure times if treatment had been withheld. Sometimes this might be biologically implausible.

What are the theoretical advantages and disadvantages associated with the method?	<p>Advantages:</p> <ul style="list-style-type: none"> i) Does not require compliers and non-compliers to be similar in terms of prognostic factors. ii) Maintains ITT randomisation (rather than comparing groups as treated). iii) White (2005) states that the main benefit of the RPSFTM method (other than the fact that it is a RBEE) is its flexibility. <p>Disadvantages:</p> <ul style="list-style-type: none"> i) The method assumes that the treatment effect associated with the intervention is equal no matter when it is received. If a patient at a later disease stage has a lower capacity to benefit from a treatment the method may over-compensate when estimating counterfactual survival times in the absence of treatment crossover, and thus the treatment effect associated with the new intervention may be overestimated ii) White (2005) suggests that it is possible to conduct a more powerful analysis than the RPSFTM approach, but that if a different method gave a significant result whereas the ITT analysis did not, it may receive little weight. In economic analysis, typically whether a treatment effect is statistically significant or not is of secondary importance, since an economic model will use the mean treatment effect statistic in tandem with the confidence or credible interval which would be used in probabilistic sensitivity analysis. However, this will be important for the level of uncertainty in the analysis. iii) The method as described by Robins and Tsiatis requires no censoring prior to the end of follow-up and no other missing data. White (1999) addresses this with a suggestion of how censoring can be accounted for (see above), but recensoring results in a loss of information. iv) White (1997) notes that the disadvantages of the RPSFTM method are that it assumes a correct model and offers no way to check the model, and it is computationally intensive.
What are the potential biases associated with the method?	These are mainly to do with the assumptions and disadvantages mentioned above. Potentially this could lead to an overestimation of the treatment benefit. In particular the equal treatment effect assumption. The method as originally developed by Robins and Tsiatis required no censoring prior to the end of follow-up, but White has extended the method to allow for this, but recensoring results in lost information.
Why might the method not be appropriate?	Due to the equal treatment effect assumption.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	It does not require compliers and non-compliers to be prognostically similar, and is randomisation based, which are advantages. It is an AFT model and is essentially a 'randomisation-based' version of a structural nested model.
Does the method represent an extension to another method?	It is a standalone method, although the authors note that it is a special case of the structural nested failure time models introduced by Robins in 1989 – these are also included in the review.
Application	
Is there a worked example in the survival setting?	No, only a theoretical example.
Is the example relevant?	Not applicable.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	Not applicable.
Other Issues	
Are there any other relevant characteristics associated with the method?	None not noted above.

Reference	Robins JM and Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. Journal of the American Statistical Association 1994; 89; 427: 737-749
Origin	
Was the method developed specifically in the survival analysis context?	Yes
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	

How does the method work?	This paper uses structural nested failure time models to re-estimate survival curves in the case where a confounding treatment was taken at differential rates in the two arms of a trial. The methods allowed an estimate of the treatment effect to be obtained as if no patients in either the treatment arm or the control arm had taken the confounding treatment. Therefore this may be a method that could be used for controlling for post-progression differential treatments, which may be useful for adjusting for treatment crossover, if the impact of the experimental treatment is assumed to be different when it is given after disease progression. The SNM used is similar to that described by Robins (1998), and the specific SNM methods used are identical to those described by Yamaguchi and Ohashi (2004). Since that paper was reviewed before this one, see the method described in the Yamaguchi and Ohashi (2004) data extraction table, which will not be reproduced here. This paper is included in the review because it preceded Yamaguchi and Ohashi's analysis.
What are the key assumptions of the method?	See Yamaguchi and Ohashi's SNM analysis.
What are the theoretical advantages and disadvantages associated with the method?	See Yamaguchi and Ohashi's SNM analysis.
What are the potential biases associated with the method?	See Yamaguchi and Ohashi's SNM analysis.
Why might the method not be appropriate?	See Yamaguchi and Ohashi's SNM analysis.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	It is an extension of Robins (1998), and is identical to Yamaguchi and Ohashi (2004).
Does the method represent an extension to another method?	It is an extension of Robins (1998).
Application	
Is there a worked example in the survival setting?	Yes, the methods are applied to an AIDS RCT.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	It is difficult to interpret the results, since we do not know the 'true' treatment effect. However, it is clear that the 'observational' analysis gives more efficient results than the 'randomised' analysis, with narrower confidence intervals.
Other Issues	
Are there any other relevant characteristics associated with the method?	No.

Reference	Law MG and Kaldor JM. Survival analyses of randomized clinical trials adjusted for patients who switch treatments. <i>Statistics in Medicine</i> 1996; 15:2069-2076
Origin	
Was the method developed specifically in the survival analysis context?	Yes, specifically for adjusting for treatment crossover.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	The method is based upon the Cox proportional hazards regression model, and is an adjusted hazard ratio approach. The method can be used to deal with switching both from control to experimental and vice-versa. The method works by splitting patients into 4 groups: AA, AB, BB, BA (where A is the control treatment and B is the experimental treatment). AA and BB represent patients who don't switch, whereas AB and BA represent switchers. Hazards are assumed to be proportional, and a Cox model is fitted to with a time-varying covariate for switching time.

	<p>Let n_A = patients randomised to A; n_B = patients randomised to B; R = HR of treatment B to A</p> <p>If no patients switched, the HRs would be: $\lambda(t)*R$ for B; $\lambda(t)$ for A</p> <p>And the hazard ratio R can be estimated using proportional hazards regression in its simplest two group form. Now, if patients switch treatment and instantaneously alter their HR by the factor of R or $1/R$ from that time onwards, the HR for the four groups of patients are: Group AA: $\lambda_{AA}(t)$ Group AB: $\lambda_{AB}(t)\exp(\beta\delta_i[t])$ Group BB: $\lambda_{BB}(t)\exp(\beta)$ Group BA: $\lambda_{BA}(t)\exp(\beta(1-\delta_i[t]))$ [1] Where $\exp(\beta)=R$ and $\delta_i[t]$ is a time dependent covariate defined to be zero up to the point of switching treatment in patient i, and to be 1 from that time point onwards.</p> <p>The authors state that this model cannot be specified in a Cox model because the underlying hazard rates in each group are different. It can't be assumed that the underlying hazard rates are the same because patients who switch may have a different prognosis – if this assumption was made (ie $\lambda(t)$ was the same for all groups) this would be equivalent to an 'as treated' analysis with treatment as a time dependent covariate.</p> <p>This issue is dealt with by assuming that the differing underlying hazard rates in the four groups of patients can be expressed as multiplicative factors acting on a single underlying hazard rate. The models in [1] then become: Group AA: $\lambda(t)\exp(\alpha_1)$ Group AB: $\lambda(t)\exp(\alpha_2 + \beta\delta_i[t])$ Group BB: $\lambda(t)\exp(b_1+\beta)$ Group BA: $\lambda(t)\exp(b_2 + \beta(1-\delta_i[t]))$ [2]</p> <p>The authors then simplify this by noting that at time $t=0$ the average hazard in each group must be the same baseline hazard as in the study population as a whole, so for the patients in treatment group A at randomization (time $t=0$): $\{\lambda(0)\exp(\alpha_1)\}^{n_{AA}} \{\lambda(0)\exp(\alpha_2)\}^{n_{AB}} = \{\lambda(0)\}^{n_A}$ with a similar expression for treatment group B.</p> <p>Solving these equations then gives: $\alpha_1 = -\alpha_2(n_{AB}/n_{AA})$ and $b_1 = -b_2(n_{BA}/n_{BB})$ which can then be substituted into [2], giving a model which can fitted as a Cox model. The authors then use two dummy variables. One equals 1 for group AB, $-(n_{AB}/n_{AA})$ for group AA, and 0 for groups BB and BA. And another that is 1 for group BA, $-(n_{BA}/n_{BB})$ for group BB, and 0 for groups AA and AB. Then the HR between the two treatments can be estimated using a time dependent covariate.</p> <p>The authors state the parameters α_1 and b_1 are included to allow an unbiased estimation of the HR between the two groups, but they can also be interpreted as indicating the extent to which patients who switch treatments have a different underlying prognosis relative to those who don't switch (within each randomised group). However, even if this difference is not statistically significant it is important to retain these parameters in the model to ensure unbiased estimates of the HR.</p>
<p>What are the key assumptions of the method?</p>	<p>For the method to result in reasonable estimates the assumptions required are much weaker than for an 'as treated' analysis. However important assumptions remain. These are:</p> <ol style="list-style-type: none"> i) The difference in prognosis between patients who do and do not switch can be modelled using proportional hazards. ii) Assume that patients who switch treatment instantaneously alter their HR by the factor R or $1/R$ from that time onwards (though if it was a placebo controlled trial and patients in the experimental group switched to placebo perhaps this should not be counted as a switch, as the treatment effect would be assumed to immediately disappear. The authors suggest that in such a scenario switching should only be considered from the control arm onto the active treatment). iii) The relative effectiveness of the two treatments is on average the same at the time of switching as at the time of randomisation – equal treatment effect (the authors acknowledge that this is a weakness, though they note that this will be relatively minor if there is a treatment effect, but it is relatively modest)
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>The method is relatively simple but has a number of important disadvantages.</p> <p>White responded to the paper with a response in the form of a letter in the journal. His chief concern was that patients are grouped based on future events, i.e. before switching occurs. Thus patients in the AB or BA groups are assumed to have a certain hazard function before they switch. However, patients in these groups cannot die before they switch treatment as otherwise they would be in groups AA or BB, and therefore in reality they have a hazard of zero up until the point at which they switch treatment. White believes that this is likely to bias the estimated HR towards the null.</p> <p>Also the method is not based on groups as randomised, and therefore the overall significance level is changed.</p> <p>The method is based upon proportional hazards, which cannot be true both for the ITT and the adjusted analysis.</p> <p>The authors note that the method probably would not be appropriate when switching was onto another treatment, other than the randomised treatments (though it could be used if it was assumed that the efficacy of the other treatment was similar to one of the randomised treatments, but this is not recommended).</p>
<p>What are the potential biases associated</p>	<p>The authors state that the assumptions they make may not hold exactly, but are unlikely to result in very significant bias. They also note that their approach may become unreliable if the</p>

with the method?	proportion of switchers is very high (if everyone switched the analysis could not be completed). However they are unsure what proportion of patients switching would lead to an unreliable analysis.
Why might the method not be appropriate?	<p>The authors recognise that the method might not be perfect, and that it should not replace the ITT analysis. However they think that it could replace the 'as treated' analysis in the presence of crossover.</p> <p>In his letter of response, White notes the dangers of grouping patients based upon future events, and the counterintuitive results of the method. He notes that the authors state that a standard 'as treated' analysis is biased because it assumes that the prognosis of switchers is similar to non-switchers, and that in their example this is demonstrated by better OS in patients who switched. White agrees with the premise regarding 'as treated' analyses, but states that in their example it is not surprising that switchers had longer OS because they had to survive long enough to receive the crossover treatment. He states that in this case the 'as treated' approach may actually be preferable, since it at least adjusts the HR in an expected way.</p> <p>In addition, White states that he tested the Law and Kaldor method in a simulation study and found that the adjusted treatment effect was hugely and significantly different from the true effect (1.48, 1.44-1.52 compared to a true effect of 1.0). He therefore concluded that the method is very prone to large biases due to resting on false assumptions. In their response, the authors stated that this was probably because approximately half of patients switched in White's simulation study.</p> <p>In their response to White's criticisms, the authors suggest that they believe that their method may be useful in cases where patients immediately switch treatment upon randomisation, in which case patients are not condition on future events. They state that this often occurs in screening and randomized consent trials – thus the method may not be appropriate for metastatic oncology drug trials.</p>
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	It is not based on the randomised population, and it is an adjusted HR approach rather than an AFT approach. It is distinct from other reviewed methods.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	Yes, an example is given using trial data for radical radiotherapy versus radical radiotherapy plus cystectomy in patients with invasive bladder cancer in which treatment crossover (both ways) occurred.
Is the example relevant?	Yes
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>Based on an ITT analysis the HR (using cox regression – data was relatively complete) for OS was 0.82 (0.61 to 1.11, p=0.20). Using the Law and Kaldor adjustment method the HR was 0.85 (0.59 to 1.23, p=0.39), whereas using a simple treatment group as a time dependent covariate approach gave a HR of 0.76 (0.66 to 1.03, p=0.074), and an analysis which censored at the time of switching led to an HR of 0.73 (0.53 to 1.02, p=0.063). The analysis identified that patients who were randomised to the control group but received the cystectomy had a particularly good prognosis relative to the other control group patients. Those randomised to the experimental treatment but received the control had a slightly better prognosis than other experimental group patients. Thus an assumption that switching patients are similar to non-switching patients would be clearly incorrect.</p> <p>The authors note that it is counter-intuitive for the adjusted estimate of the HR to be higher than the ITT HR, and they cannot explain the result. They state that it is probably because the ITT method and the Law and Kaldor method fit quite different models. They state that theoretically the proportional hazards assumption cannot hold for both models – in the Law and Kaldor method it is assumed that switching treatment has an instant multiplicative effect, and if this is true the hazard ratio for the ITT analysis can not be proportional – rather it would probably have to reduce over time as patients switched treatments. They state that practically both model types are likely to fit the data adequately (and they tested the proportional hazards assumption for both models by fitting an interaction between the treatment effect and the natural logarithm of time and conducting a likelihood ratio test. This gave non-significant results for both models (though it was more significant for the adjusted analysis – p=0.34 compared to p=0.84)). Despite this, the authors accept that the proportional hazards assumption cannot be exactly true for both analyses, and so there may be some bias in the results (though this is likely to be small) which may have led to the counter-intuitive results.</p> <p>In his letter of response, White suggests that the counter-intuitive result may be due to the model-type, in particular the fact that for patients in group AB, the hazard of death has to be zero until they switch treatment. White states that this will be likely to bias the estimate of β (the adjusted treatment effect) upwards (obviously – as patients can't die while on the control treatment). This bias will be counteracted by a similar bias in group BA, but it is unclear which bias will be larger. Hence this practical application shows that the method is likely to work poorly.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	None.
Reference	Mittlbock M and Whitehead J. The interpretation of clinical trials of immediate versus delayed therapy. Lifetime Data Analysis 1998;4:253-263
Origin	

Was the method developed specifically in the survival analysis context?	Yes, the method is developed to be used for the analysis of a small trial of delayed versus immediate treatment with Zidovudine for patients with HIV (the LRAZT trial). In the LRAZT trial the control group were scheduled to begin the active treatment upon disease progression, but some patients switched before progression, which may have confounded results. Mittlbock and Whitehead attempt to demonstrate a parametric method for dealing with this.
If not, what was the original context and how has the method been adapted?	NA
Theoretical Suitability	
How does the method work?	<p>The method attempts to fit a parametric model in which the dependence between switching and survival times is modelled. The authors note in the discussion that their parametric model is not put forward as a comprehensive alternative to the ITT analysis, and that it is a highly speculative method.</p> <p>The authors present their method primarily in the form of an exponential model, but note that the Weibull model would be more likely to provide a suitable alternative to the ITT analysis.</p> <p>T = survival time of a patient in the new therapy group S = time from randomisation until switching for patients in the control group U = latent variable representing the potential survival time of a patient in the control group if switching had not occurred U' = actual survival time for patient in the control group</p> <p>Eg, if a control patient switches, only the switching time (S) and U' are observed. It will be known that U is greater than S, but the value of U won't be known. For a patient who dies without switching U (=U' in this case) is known and it will be known that S is greater than U, but the value of S will be unknown.</p> <p>The simplest version of the model requires that T and U are exponentially distributed, with parameters denoted by μ and λ respectively.</p> <p>Because in the case study considered by the authors patients are expected to be switched onto the new treatment when their disease becomes symptomatic, but that in the trial some patients switched earlier due to other indicators, they decided to allow S to depend on U under the assumption that: $E(S U) = U/(\kappa - 1)$ where $\kappa (>1)$ is an unknown parameter. When $1 < \kappa \leq 2$ death is expected to occur before switching and where $\kappa > 2$ switching is expected to occur before death.</p> <p>The authors show that under an exponential parametric model the above dependence between S and U can be induced by making the inverse of switching time conditional on U follow the gamma distribution with parameters κ and U:</p> $f_{S U}(s U) = \frac{1}{\Gamma(\kappa)} U^\kappa s^{-(\kappa+1)} \exp\left(-\frac{U}{s}\right) \text{ where } \kappa \geq 1; U, s > 0$ <p>The marginal density function and corresponding expectation of S are:</p> $f_S(s) = \frac{\kappa\lambda}{(1+s\lambda)^{\kappa+1}} \text{ and } E(S) = \frac{U}{\kappa-1} = \frac{1}{\lambda(\kappa-1)}$ <p>Conditional on S, U follows a gamma distribution with parameters $(\kappa+1)$ and $(1+\lambda S)/S$ so that:</p> $f_{U S}(u S) = \frac{1}{\Gamma(\kappa+1)} [(1+\lambda S)/S]^{(\kappa+1)} u^\kappa \exp[-(1+\lambda S)u/S] \text{ and } E(U S) = \frac{(\kappa+1)S}{1+\lambda S}$ <p>The likelihood function based on n patients in the new treatment group with survival times t_1, \dots, t_n and censoring indicators $\delta_1, \dots, \delta_n$ is:</p> $L_1(\mu) = \prod_{j=1}^n \mu^{\delta_j} \exp(-\mu t_j) \text{ where } \delta_j = 1 \text{ if the } j^{\text{th}} \text{ patient has died and 0 otherwise.}$ <p>If for m patients in the control group, both switching times and untreated survival times were known, the full likelihood function would be:</p>

$$L_0(\lambda, \kappa) = \prod_{i=1}^m \lambda \exp(-\lambda u_i) \frac{1}{\Gamma(\kappa)} s_i^{-(\kappa+1)} u_i^\kappa \exp\left(-\frac{u_i}{s_i}\right)$$

However, in practice only one of S, U or the censoring time C are known and therefore the full likelihood function is not known. However, it can be split into three parts for the m_s patients who switched, the m_u patients who died and the remaining m_c censored patients who did neither.

The likelihood for the m_s patients who switched is:

$$L_{0,S}(\lambda, \kappa) = \prod_{i=1}^{m_s} \int_{s_i}^{\infty} f_{U,S}(u_i, s_i) du_i = \prod_{i=1}^{m_s} \int_{s_i}^{\infty} \lambda \exp(-\lambda u_i) \frac{1}{\Gamma(\kappa)} s_i^{-(\kappa+1)} u_i^\kappa \exp\left(-\frac{u_i}{s_i}\right) du_i$$

The likelihood for the m_u patients who died is:

$$L_{0,U}(\lambda, \kappa) = \prod_{i=1}^{m_u} \int_{u_i}^{\infty} f_{U,S}(u_i, s_i) ds_i = \prod_{i=1}^{m_u} \int_{u_i}^{\infty} \lambda \exp(-\lambda u_i) \frac{1}{\Gamma(\kappa)} s_i^{-(\kappa+1)} u_i^\kappa \exp\left(-\frac{u_i}{s_i}\right) ds_i$$

The likelihood for the m_c censored patients is:

$$L_{0,C}(\lambda, \kappa) = \prod_{i=1}^{m_c} \int_{c_i}^{\infty} \int_{c_i}^{\infty} f_{U,S}(u_i, s_i) ds_i du_i = \prod_{i=1}^{m_c} \int_{c_i}^{\infty} \int_{c_i}^{\infty} \lambda \exp(-\lambda u_i) \frac{1}{\Gamma(\kappa)} s_i^{-(\kappa+1)} u_i^\kappa \exp\left(-\frac{u_i}{s_i}\right) ds_i du_i$$

The full likelihood is therefore the product of four parts:

$$L(\mu, \lambda, \kappa) = L_1(\mu) L_{0,S}(\lambda, \kappa) L_{0,U}(\lambda, \kappa) L_{0,C}(\lambda, \kappa)$$

The authors state that the maximum likelihood of μ can be found directly from $L_1(\mu)$, where the value is $\hat{\mu} = d_1 / (t_1 + \dots + t_n)$ where d_1 is the number of deaths in the treatment group. However, they state that finding the maximum likelihood of λ and κ is more difficult, and in their example the authors stated that they used a NAG-subroutine to maximise the likelihood function.

The authors state that their method can be extended to non-constant hazards by allowing T and U to follow Weibull distributions. They state that in this case the maximum likelihood estimates for all parameters (μ, λ, κ and a shape parameter γ common to T and U – thus taking a HR modelling approach, assuming proportional hazards) have to be found using an optimisation procedure. They state that it is not possible to give closed expressions for $f_S(s)$ and $f_{U|S}(u|S)$ and their expected values.

The authors state that another characterisation of their models is apparent from consideration of the correlation between S and U, which for the Weibull model is:

$$\text{corr}(S, U) = \sqrt{\left(\frac{\kappa - 2}{\kappa - 1 + A^2}\right)}, \text{ if } \kappa > 2$$

where: $A = \frac{\sqrt{\Gamma(1+2/\gamma) - [\Gamma(1+1/\gamma)]^2}}{\Gamma(1+1/\gamma)}$ is the coefficient of variation U. In the exponential model $\gamma = 1, A = 1$ and $\text{corr}(S, U) = \sqrt{\left(\frac{\kappa - 2}{\kappa}\right)}, \text{ if } \kappa > 2$

So in the exponential model the ratio E(U)/E(S), which is equal to $(\kappa - 1)$, and $\text{corr}(S, U)$ depend on a single parameter. In the Weibull model E(U)/E(S) again is equal to $(\kappa - 1)$, but $\text{corr}(S, U)$ depends on the additional parameter $A \equiv A(\gamma)$.

What are the key assumptions of the method?

- i) It is assumed that T and U follow a parametric distribution – in the authors examples either an exponential or a Weibull.
- ii) A key assumption is that switching time (S) depends on untreated survival time (U), and the nature of this relationship must be assumed.
- iii) An exponential model assumes constant hazards, a Weibull assumes monotonic hazards.
- iv) Proportional hazards are assumed.

What are the theoretical advantages and disadvantages associated with the method?	<ul style="list-style-type: none"> i) There is no value for the parameter κ which corresponds to independence of survival and switching – thus the model can't be used to test whether such a dependence exists. ii) Due to the complicated form of the likelihood no simple methods for finding standard errors or confidence intervals for parameter estimates exist. iii) The exponential model is limited because the ratio $E(U)/E(S)$, which is equal to $(\kappa - 1)$, and $\text{corr}(S,U)$ depend on a single parameter.
What are the potential biases associated with the method?	The method is highly dependent on suitable parametric models being available, and the complicated nature of the approach may encourage the analyst not to consider all possible parametric models. The authors only attempt fitting exponential and Weibull models in their example. In addition the proportional hazards assumption may lead to the treatment effect being modelled inappropriately. The model is also highly dependent on the untreated survival time being linked to switching time.
Why might the method not be appropriate?	<p>The authors note that development of their methodology requires allowance for prognostic factors other than treatment because when delayed treatment is being considered it is likely to be most useful to provide guidance on when treatment should begin, based on specific indicators. It should be noted that the method is developed very much in the context of assessing the benefits of immediate versus delayed treatment in circumstances when delayed treatment is given too early in the control group. This is different from the treatment crossover problem as defined in this thesis – i.e. where immediate versus no treatment is the context. However, when the method was applied, it was in the context of switching as we define it in this thesis, therefore the method may be suitable.</p> <p>It is important that the complicated nature of the approach may encourage the analyst not to consider all possible parametric models, and proportional hazards are assumed which may often be inappropriate. The importance of this is demonstrated in the authors application of the method, where the results differentiate massively for the exponential and Weibull models.</p> <p>In addition, the authors themselves state that their method is not put forward as a comprehensive alternative to the ITT analysis, and that it is highly speculative. Thus it is likely to require further development and as it stands, does not appear to be a realistic relevant option for addressing treatment crossover.</p>
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	Mittlbock and Whitehead's method is fully parametric, unlike most other identified methods. It attempts to model counterfactual survival, like the RPSFTM and SNM methods. The RPSFTM method leads to the same level of significance as the ITT method, whereas Mittlbock and Whitehead attempt to gain some of the power lost as a result of switching in the ITT analysis, and hence has a different significance level.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	Yes, the authors apply both exponential and Weibull models to an HIV clinical trial data set. It is worth noting, however, that out of 993 patients there were only 27 deaths. 472 control group patients were included in the analysis, of which 323 switched treatment. The log-rank test based on the ITT analysis was non-significant. The authors state that at the extreme, it might be assumed that all switching patients would have died immediately if they had not switched, and a log-rank test comparing the survival times for the treatment group and the switching or death time for the control group was highly significant. The authors suggested that the true treatment effect was likely to lie between these two extremes. The authors analysed control patients and found that there was a significantly raised hazard of death after switching, which they suggest may indicate that patients switch when their disease deteriorates. The authors then fitted the exponential and Weibull models as described above.
Is the example relevant?	Yes – the analysis was done in the context of what would the results have been had no switching occurred, rather than if switching had occurred at the pre-specified time. Thus the example was relevant to our definition of switching. However there were very few deaths observed, which meant that a vast amount of extrapolation was required. Thus the results of the method was highly dependent on the parametric model chosen, and only exponential and Weibull models were tested.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	The exponential model resulted in survival times that seemed very unrealistic – the mean survival times were 129 months for the control group and 1230 for the treatment group, resulting in a very significant treatment benefit. The estimated switching times also did not fit the observed switching times very well. The Weibull model was an improvement, and results were more realistic – mean survival was 90 months for the control group and 182 months for the treatment group (still a significant difference), and the estimated switching times were more similar to the observed times (though the fit to the observed data based on a visual inspection was still not particularly good).
Other Issues	
Are there any other relevant characteristics associated with the method?	None.
Reference	White IR, Babiker AG, Walker S and Darbyshire JH. Randomisation-based methods for correcting for treatment changes: Examples from the Concorde trial. <i>Statistics in Medicine</i> 1999;18:2617-2634
Origin	

Was the method developed specifically in the survival analysis context?	Yes. The method focussed upon is Robins and Tsiatis' RPSFTM applied to the Concorde trial dataset. Although this paper is an application of a previously developed method it offers extensions to the RPSFTM method regarding censoring and non-constant treatment effects. The censoring and non-constant treatment effect extensions will provide the focus of this evidence table – the RPSFTM methodology will not be repeated.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>Censoring update:</p> <p>Because censoring is common in survival data it needs to be included in the RPSFTM model. C_i is defined as the time to administrative censoring (i.e. the end of follow-up in the trial), and the censoring time of the counterfactual event times is given by:</p> $D_i(\psi) = \int_0^{C_i} \exp[\psi X_i(t)] dt$ <p>However, because crossover may be related to prognosis, both X_i and D_i may also depend on prognosis and therefore the censoring of $U_i(\psi)$ may be informative, and thus including censoring in this way may result in biased estimates. White <i>et al</i> (1999) suggest that this bias could be avoided by recensoring counterfactual survival times so that the censoring time equals the minimum of the administrative censoring time (C_i) and $C_i \exp \psi$. Then the counterfactual survival time $U_i(\psi)$ for a patient is replaced by the censoring time of the counterfactual event times $D_i(\psi)$ if $D_i(\psi) < U_i(\psi)$.</p> <p>Non-constant treatment effect:</p> <p>The authors note that a key problem with the RPSFTM method is that it assumes a constant treatment effect no matter when the treatment was received. In the Concorde trial the treatment appeared not to have benefit in early disease, yet it is known that the treatment works in more progressed disease. The authors therefore attempted to extend the model to allow the treatment effect to depend on the CD4 count at which the new treatment was started (as CD4 indicates the progression of the disease). The median CD4 count (350) observed in the trial was used as a cut-off point for 'high' and 'low' CD4. The bivariate form of the model is shown below:</p> $U_i = \int_0^{T_i} \exp(\psi_H X_{Hi}(t) + \psi_L X_{Li}(t)) dt$ <p>where $X_{Hi}(t)$ is defined as 1 if the patient is on the new treatment and CD4 is greater than 350, and otherwise 0, and $X_{Li}(t)$ is defined as 1 if the patient is on the new treatment and CD4 is less than 350, and otherwise 0.</p> <p>Recensoring was conducted as for the simpler univariate model, but to estimate two parameters two tests of equality of U across the two arms are required. The authors used logrank and Gehan-Wilcoxon tests and by demanding that both test statistics are zero they state that they are (roughly) demanding that the rate of U be equal in the two arms both early and in late follow-up.</p>
What are the key assumptions of the method?	See RPSFTM method.
What are the theoretical advantages and disadvantages associated with the method?	See RPSFTM method. An additional disadvantage of recensoring is the associated loss of information.
What are the potential biases associated with the method?	See RPSFTM method.
Why might the method not be appropriate?	See RPSFTM method.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	This is an extension to the RPSFTM method.
Does the method represent an extension to another method?	Yes, the RPSFTM method.

Application	
Is there a worked example in the survival setting?	Yes, the RPSFTM method is applied to the Concorde trial. The authors also test to see if recensoring is necessary, and to see the results of their bivariate model. These will be focussed upon here.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>Censoring:</p> <p>The authors found that in the Concorde trial treatment history predicts end of follow-up on the unrecensored U-scale, but not on the recensored U-scale. The authors state that this is consistent with censoring being non-informative on the original scale, informative on the U-scale without recensoring, but non-informative when recensoring is undertaken. Thus, recensoring is required for valid inference from the Concorde trial data. The authors also completed a simulation analysis based upon the Concorde trial, and found that the recensored estimator was unbiased, but the unrecensored estimator was biased towards the null by a small but consistent amount. The authors conclude that in datasets similar to the Concorde trial failing to recensor incurs a small bias, but has little effect on the overall error of estimation. However, they recommend that recensoring should not be ignored because bias could be important.</p> <p>Bivariate model:</p> <p>The authors found that their estimates of treatment effect based upon the bivariate model had very wide confidence intervals, and that point estimators were hard to identify. They stated that the bivariate models were neither informative or robust in the Concorde dataset.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Reference	Robins JM and Finkelstein DM. Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. <i>Biometrics</i> 2000;56:779-788
Origin	
Was the method developed specifically in the survival analysis context?	Yes, the method is developed alongside an AIDS clinical trial where a new treatment was compared to a control with primary endpoint of time to pneumocystis pneumonia (PCP) and secondary endpoint OS. Upon PCP patients were allowed to cross over treatments. The new treatment resulted in a significant increase in time to PCP, but the OS difference was not significant. The authors suggested that this may have been partly due to the treatment crossover. Of the 310 patients in the trial, 94 died and 216 were censored (some due to being lost to follow-up and some due to remaining alive at the end of follow-up). Of the 94 deaths 21 were in patients who had crossed over. 37 occurred in subjects who stopped all prophylactic therapy, 21 for nonmedical reasons and 16 for medical reasons. The IPCW method developed provided a way of estimating the treatment effect had subjects neither stopped therapy nor crossed over onto alternative therapy.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The method involves artificially regarding subjects as dependently censored at the first time a subject stops therapy, switches therapy, or is lost to follow-up. Then the IPCW Kaplan-Meier estimator and log-rank test developed by Robins (1993) is applied, that adjusts for dependent censoring by using data collected on time-dependent risk factors for failure and censoring. The IPCW versions of these statistics differ from the ordinary versions in that, in calculating the contribution of a subject at risk at time t, (1) the subject is given a weight inversely proportional to an estimate of the conditional probability of having remained uncensored until time t, and (2) the estimate is based on the fit of a Cox proportional hazards model for censoring in which time-dependent prognostic factors for failure and censoring are entered as covariates. Therefore whether a patient stops treatment due to toxicity that cannot be treated with a palliative therapy can be taken into account – these patients are not treated as noncompliant.</p> <p>The authors complete 4 analyses, of which 2 are most relevant for this review. The first is basically an ITT analysis. The second regards subjects as dependently censored by the minimum of time to loss to follow-up and time to treatment crossover – thus only 73 of the 94 deaths are included as failures as these occurred before crossover. This analysis estimates what the survival curves would have been if crossover was not allowed. The third analysis regards subjects as dependently censored by the minimum of time to loss to follow-up, time to treatment crossover, and time to voluntarily stopping treatment for nonmedical reasons – thus only 52 of the 94 deaths are included as failures as these occurred before crossover and without treatment being voluntarily stopped. This analysis estimates what the survival curves would have been if crossover was not allowed and if patients could not voluntarily stop treatment without medical reason. The fourth analysis regards subjects as dependently censored by the minimum of time to loss to follow-up, time to treatment crossover, and time stopping treatment for any reason. This analysis estimates what the survival curves would have been if crossover was not allowed and if patients could not stop treatment for any reason. The second and third analyses are relevant for this review. The final analysis is arguably not appropriate because results should take into account patients having to stop treatment due to</p>

	<p>toxicity.</p> <p>The authors state that censoring based upon crossover must be regarded as dependent since crossover is likely to be linked in some way to prognosis. They state that Robins (1993) and Robins and Rotnitzky (1992) showed that if there is data on all time-dependent prognostic factors for mortality that independently predict censoring, then we can correct for the dependence between the censoring and failure by replacing the Kaplan-Meier estimator, log-rank test, and Cox partial likelihood estimator of the ratio of the treatment-arm-specific mortality rates by their IPCW versions. In the dataset used in the paper the authors note that there was extensive data on time-dependent prognostic factors.</p> <p>The authors first demonstrate a method for retaining only those covariates that are significant prognostic factors for mortality. They use a time-dependent Cox proportional hazards model for censoring to estimate the treatment-arm specific hazards of censoring conditional on time-dependent prognostic factors for failure. They note that for a time-dependent prognostic factor to cause selection bias or confounding it must generally be a prognostic factor both for failure and for censoring. The model specification implied that, conditional on their most recent values, past values of the potentially prognostic values do not predict failure at t. The authors only kept prognostic values significant at the p=0.12 level. However, when deciding which factors to retain it must also be considered whether the excluded variables have any influence on the future values of the retained variables – thus these factors must both not be an independent predictor of outcomes, or future values of retained variables. The authors tested this by fitting 3 separate models – for each retained variable they modelled the logit of the probability of change in that variable at each follow-up visit as a linear function of the treatment arm and the values of the excluded variables at the previous visit, to identify whether any of the excluded variables were predictors of future values of the retained variables. They noted that none were. The authors acknowledged that other approaches to dimension might be considered, including a formal approach to elicit a subjective prior distribution for the joint distribution of all unknown parameters in the model and carry out a full Bayesian analysis.</p> <p>The authors then introduce the IPCW versions of the KM and Cox partial likelihood estimators and the log-rank test. The first step is to specify a model for the hazard of censoring based upon the retained variables:</p> $\lambda_c \{ \bar{V}^*(t), Z, T > t \} = \lambda_{0z}(t) \exp \{ \alpha_z' V^*(t) \}$ <p>where $V^*(t)$ represents the vector of the retained variables and Z is treatment received. Because the baseline hazards $\lambda_{0z}(t)$ and α_z may depend upon treatment arm, the model must be fit to data from the two treatment arms separately. This is important because the reason for censoring might be quite different in the two arms.</p> <p>The IPCW KM estimator is given below. It differs from the ordinary KM estimator by weighting the contribution of a subject at risk at time t by the inverse of an estimate of the conditional probability of having remained uncensored until time t, based on the fit of the model above.</p> $\hat{S}_T(t z) = \prod_{\{i: X_i < t\}} \frac{1 - \tau_i \hat{W}_i(X_i) I(Z_i = z)}{\sum_{k=1}^n Y_k(X_i) \hat{W}_k(X_i) I(Z_k = z)}$ <p>where $X = \min(T, C)$, $Y(u) = I(X \geq u)$ is the at-risk indicator and $r = I(T = X)$ is the failure indicator that is 1 if the subject is a failure and 0 if the subject is censored, and W is the subject specific weight, $\hat{W}_i(t) = \hat{K}_i^0(t) / \hat{K}_i^{V^*}(t)$ where $\hat{K}_i^0(t)$ is the normal treatment arm specific Kaplan-Meier estimator of the probability of being uncensored by time t in treatment group Z_i. and $\hat{K}_i^{V^*}(t)$ is the Kaplan-Meier extended for censoring.</p> <p>The IPCW KM estimates the probability of surviving without failure until time t in the absence of censoring.</p> <p>The adjusted Cox HR is calculated by applying a Cox model to the data, including baseline covariates, with each individual's contribution to the partial likelihood weighted for each time point using the estimated inverse probability weights.</p> <p>The way that the weighting works is that if an individual k survives to time t with a conditional probability of $\hat{K}_k^{V^*}(X_i)$ of 0.25 of having avoided censoring until that time point, there are on average 3 other prognostically similar patients who would have survived until that time-point but would have been censored before that time point. Thus person k receives a weight of 4 (1/0.25) to represent the number of patients who would have been at risk at that time point if it were not for censoring.</p>
What are the key assumptions of the method?	A key assumption is that there must be no unmeasured confounders for censoring – the probability of censoring must be fully described by the data available on prognostic factors and the cause-specific hazard of censoring C must not depend upon any further variables that impact upon the possibly unobserved failure time T. However, in the presence of censoring it is not possible to directly test this. The authors state that if this assumption holds the IPCW method can fully correct for bias due to the dependent censoring attributable to the identified

	variables. They state that in practice this assumption may not fully hold, but if data has been collected on a number of variables it may be approximately true, in which case the IPCW method can reduce, if not totally eliminate, the bias due to dependent censoring.
What are the theoretical advantages and disadvantages associated with the method?	<ul style="list-style-type: none"> i) The method results in greater power to detect significant differences than the standard ITT analysis. ii) Data needs to have been collected on whether treatment changes were due to toxicity or not, if survival estimates are to be adjusted both for crossover and for voluntary treatment cessation. iii) There needs to be extensive data on time-dependent prognostic factors for mortality that independently predict censoring (crossover). iv) Must be able to specify accurately models both for the probability of crossover and survival. v) Does not work if all patients crossover.
What are the potential biases associated with the method?	The assumptions around data availability and model specification are strong, and the method may not be suitable if such data is not available – bias could result.
Why might the method not be appropriate?	The assumptions around data availability and model specification are strong, and the method may not be suitable if such data is not available – bias could result.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	Compared to other methods the IPCW requires more data to be available, but also offers the possibility of providing more ‘efficient’ and ‘powerful’ results, as the power is not the same as the ITT analysis. The IPCW method is a type of marginal structural model.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	Yes, based on the AIDS trial that the paper describes.
Is the example relevant?	Yes, the method is used to control for crossover in two relevant analyses.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>The authors found that 3 of the 6 prognostic factors for which they had data were relevant and significant. When the authors applied their method they found that the treatment effect associated with the new treatment increased compared to the ITT method, and confidence around the result also increased. The treatment effect is increased when crossover is controlled for, and is increased even further when patients who stopped treatment voluntarily for no medical reason are also controlled for. Evidence was also found that censoring had been dependent, since adding in more of the variables found to be important altered the treatment effect estimate, and when all these variables were excluded the estimate was biased towards the null. In addition, there was evidence that crossover and the cessation of treatment was linked to prognosis, as when this is controlled for the KM for the control group was decreased, whereas it remained about the same for the new treatment. This indicates that patients in the control group who stopped treatment or crossed over had a poor prognosis compared to those who were not censored, whereas in the treatment group censored and uncensored patients were similar and censoring was not dependent (as typically treatment was stopped due to toxicities that were not related to prognosis).</p> <p>The authors conclude that the IPCW KM and Cox partial likelihood estimators can be used to correct for bias due to non-compliance and dependent censoring when (most of) the non-compliance and dependent censoring can be explained by measured time-dependent prognostic factors and a (nearly) correct time-dependent Cox model for the hazard of censoring given these prognostic factors can be specified.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Reference	Branson M and Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. <i>Statistics in Medicine</i> 2002;21:2449-2463
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	The authors build upon the RPSFTM method by replacing the rank test-based estimation of ψ with a likelihood-based analysis. An iterative parameter estimation (IPE) algorithm is used whereby, using the notation used in the Robins and Tsiatis (1991) review above, e^ψ is initially estimated by comparing treatment arms as randomised using a parametric failure time model

	<p>in an ITT analysis. Any parametric failure time model could be used for this (Weibull, log-logistic, log normal, gamma). Given this initial estimate of ψ, the observed survival times of patients who switched from control onto active treatment are transformed (counterfactual survival times as if crossover hadn't occurred are estimated) using e^ψ and $U_i = T_{Ai} + e^{-\psi_0}T_{Bi}$ (The actual equation suggested by Branson and Whitehead is slightly different in terms but exactly equivalent). Then the treatment groups are compared again using a parametric failure time model, which will give an updated estimate for e^ψ. Again observed survival times of patients who switched from control onto active treatment are transformed, and groups are compared again, producing another updated estimate for e^ψ. This iterative process is continued until the new estimate for e^ψ is very close to the previous value, signifying that the process has converged. The authors suggest that convergence could be assumed when the new estimate for e^ψ is within 10^{-5} of the previous estimate, although they offer no rationale for this and so it may be considered to be arbitrary.</p> <p>If the transformed survival time for a crossover patient is greater than the administrative censoring time the patient is assumed to be censored and their survival time is replaced with the administrative censoring time. Thus recensoring is included in the method, but is restricted to the crossover patients, which is different from the recensoring approach developed by White (1999) in his extension of the RPSFTM method. The authors state that recensoring is only required if the new treatment is detrimental compared to the control, because if the treatment is detrimental the observed survival time for switchers will be shorter than the re-estimated counterfactual survival time and so re-estimated survival could be longer than the administrative censoring time. If the treatment is beneficial this cannot be the case, and so recensoring would not be necessary.</p> <p>It is assumed that survival times have a parametric form. The method does not require modelling of the distribution of a patient's switching time. The authors state that the method can be generalised to include other baseline covariates, and an example of this is presented and simulation studies run.</p> <p>The authors state that standard errors could be taken from the final regression in the IPE algorithm, or bootstrapping can be used. The authors state that the standard errors from the regression will be underestimates because the covariance matrix from the final iteration of the IPE algorithm does not take into account the fact that control arm patients have had their survival times transformed by the algorithm.</p> <p>Like Robins and Tsiatis, the authors recommend that the IPE estimate and CI is accompanied by the ITT p value.</p> <p>The authors also state that relaxation of the parametric assumption has been explored and seems possible, and that this would be the subject of a future publication – but no such paper has appeared.</p>
<p>What are the key assumptions of the method?</p>	<p>The additional assumption of the Branson and Whitehead method over that of Robins and Tsiatis is that survival times are assumed to have a particular parametric form. The Weibull distribution was used in the Branson and Whitehead paper but the authors recommend that if the method were to be applied to real data, a distribution is chosen which best fits the experimental data. It would be interesting to see how well the method performs under different parametric models.</p> <p>Other assumptions are similar to those required for the RPSFTM method. For example, because an accelerated failure time model is used it must be assumed that the treatment acts multiplicatively on a patient's survival time.</p>
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>The advantages and disadvantages are similar to the RPSFTM method. The authors claim that their requirement for less recensoring is an additional advantage. They also state that their method is an RBEE, and that their approach offers a systematic approach to estimating the treatment effect, which is not offered by the RPSFTM method. The additional disadvantage is the parametric assumption made in the estimation procedure.</p>
<p>What are the potential biases associated with the method?</p>	<p>These are mainly to do with the assumptions and disadvantages mentioned above. Potentially this could lead to an overestimation of the treatment benefit. In particular the equal treatment effect assumption. Further analysis (see below) suggests the censoring approach is unreasonable.</p>
<p>Why might the method not be appropriate?</p>	<p>Due to the equal treatment effect assumption.</p>
<p>How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?</p>	<p>Similar to the RPSFTM method.</p>
<p>Does the method represent an extension to another method?</p>	<p>Yes, it is an extension to the RPSFTM method. Robins and Tsiatis (1991) estimate e^ψ using a rank test statistic (such as the logrank statistic) and no systematic estimation algorithm exists for providing e^ψ. Branson and Whitehead state that a grid search over possible values is used to determine the appropriate point estimate, whereas their approach provides a systematic method for this.</p> <p>There are other small differences compared to the RPSFTM method. For example, Branson and Whitehead state that Robins and Tsiatis use a strong version of the accelerated failure time model, referred to 'structural' or 'causal' models. This is because they relate observed and counterfactual event times using the following equation: $U_i = e^{-\psi_0}T_{Bi}$, where the = sign denotes 'true equality'. Branson and Whitehead state that for their method, only 'equality in distribution' needs to be assumed, and so their method does not require the 'strong' version of the AFT model.</p>

Application	
Is there a worked example in the survival setting?	Yes, both a real life example and simulation studies. The simulation studies generated survival times using a Weibull distribution, set expected failure time to 500 days, and the data were positively skewed. The shape parameter was 1.5 and the scale parameter $e^{-\psi}$ was set such that $\psi = 6.3169$. Different switching patterns were considered, varying the proportion of patients that switched treatment and the proportion of time spent on the crossover treatment. Various treatment effects were considered. The real-world example was based on an application of the method to a lung cancer trial which compared immediate to delayed treatment.
Is the example relevant?	Yes
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	The simulation studies demonstrated that the IPE method consistently estimated the multiplicative effect of treatment in the presence of switching. Bias was low. In the real-world example the authors report that based upon the log-cumulative hazard plot they decided that the Weibull distribution was suitable for modelling survival. The ITT analysis gave a point estimate of the acceleration factor of 1.111, and the IPE method gave an estimate of 1.158. The RPSFTM method was also tested, and provided an estimate of 1.14.
Other Issues	
Are there any other relevant characteristics associated with the method?	Note that further analysis conducted by White demonstrated that the recensoring method advocated by Branson and Whitehead was unreasonable. White (2006) demonstrates that Branson and Whitehead's recensoring is insufficient because recensoring is required whether or not the new treatment is detrimental. Informative bias is likely even if the new treatment is effective, because counterfactual survival times are associated with patient prognosis. Patients who benefit from crossing over and have their survival time censored may have died if they had not crossed over. Because crossover is associated with prognosis, administrative censoring times are therefore also associated with prognosis and hence recensoring at the minimum possible censoring time ($C_i \exp \psi$) is required. White demonstrated that applying the IPE method without full recensoring will result in important bias when a relatively high (80%) proportion of patients have censored survival times. Therefore, if the IPE method is to be applied, full recensoring should be applied.

Reference	Walker AS, White IR and Babiker AG. Parametric randomisation-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. <i>Statistics in Medicine</i> 2004;23:571-590.
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	In the RPSFTM method and the IPE algorithm method ψ is chosen to balance the counterfactual event time, U , between treatment arms using a semi-parametric procedure. However these methods result in informative censoring (because individuals with the same C_i in the control group are more likely to be censored in the experimental group because the new treatment increases T_i , thus censoring is dependent on treatment received) and recensoring methods to adjust for this are associated with a loss of information and arbitrary differences from the results of ITT analysis because recensoring can lead to information from late in follow-up being omitted. This is a particular problem if there is treatment by time interaction. Walker <i>et al</i> present an extension to these semi-parametric methods which involve a bivariate parametric frailty model for time to treatment change and time to trial endpoint, which involves modelling the relationship between U and the treatment a patient actually receives Z . This avoids informative censoring, recensoring and loss of information. The authors state that using their method maximum likelihood does not incorporate the randomisation balance and is likely to be very sensitive to misspecification of the parametric model, and so they use estimating equations to force randomisation balance into the model. The model used is a joint parametric model for the counterfactual time U_i and time to initiation of the experimental treatment in the control arm Z_i . The density is $f(z)f(u z)$ and the censoring mechanism is non-informative whenever the censoring time on the U -scale is independent of U_i conditional on the time of treatment change Z_i . The model only considers switching from the control treatment to the experimental treatment. They assume that any changes in treatment in the experimental group are a consequence of using the experimental treatment and we do not want to correct for this. Consider a trial with control (A) and experimental (B) arms where some patients who are randomised to control crossover to the experimental treatment at some point during follow-up. Consider U_i as a patient's counterfactual event time and Z_i as the time at which they start receiving experimental treatment. The authors propose specifying a joint parametric model for U_i and Z_i which is made up of three parts: 1. A causal model relating U_i to a patient's observed failure time T_i . This is the AFT model of the RPSFTM method and the IPE algorithm method.

	$U_i = \begin{cases} Z_i + e^{\psi}(T_i - Z_i) & \text{if } R_i = 0 \text{ and } \Delta_i^Z = 1 \\ T_i & \text{if } R_i = 0 \text{ and } \Delta_i^Z = 0 \\ e^{\psi}T_i & \text{if } R_i = 1 \end{cases}$ <p>Where Δ_i^Z is a censoring indicator, R_i is the randomised group (0 for control, 1 for experimental), Z_i is the time at which switching occurred, T_i is the observed event time, and U_i is the counterfactual event time.</p> <p>2. A model for the association between U and Z. This is a bivariate frailty model. Either a positive stable or gamma frailty are suggested. The model is derived by assuming that U and Z are independent, conditional on a common frailty which has a distribution defined by whether a positive stable frailty or a gamma frailty is chosen. These models include a parameter Φ which describes the level of association between U and Z. A positive stable distribution leads to an association between U and Z that is initially strong but which washes out over time. The association modelled using a Gamma distribution with mean 1 and variance $\Phi-1$ does not change over time.</p> <p>3. Models for the marginal cumulative hazards for U and Z. $H_u(u)$ and $H_z(z)$ are the marginal cumulative hazards of U and Z respectively. The authors define Weibull and Gompertz models that could be used here, but state that extensions to piecewise exponential or other non-monotonic hazards are relatively easy.</p> <p>Weibull: $H_u(u) = \lambda_u u^{\gamma_u}$ $H_z(z) = \lambda_z z^{\gamma_z}$</p> <p>Gompertz: $H_u(u) = e^{\lambda_u}(e^{\gamma_u u} - 1)/\gamma_u$ $H_z(z) = e^{\lambda_z}(e^{\gamma_z z} - 1)/\gamma_z$</p> <p>The authors state that fitting this model using maximum-likelihood techniques would only ensure the original randomisation balance is preserved if all models are correctly specified, because the original randomisation balance has not been used in the estimation procedure: in particular estimates of U_i are not guaranteed to be balanced between randomised groups. This is ok if the model is correctly specified, but parameter estimates will be very sensitive to inaccuracies in the model specification. To deal with this, the authors suggest an alternative approach to maximum likelihood to estimate parameters. They use an augmented model to maintain the randomisation balance between groups which corresponds to the Cox model based test statistic in the semi-parametric approach of Robins and Tsiatis. The model has the form:</p> $H_u^*(u) = e^{\rho R} H_u(u)$ <p>An estimate of ψ can be found so that an estimate of ρ would be equal to zero, indicating there is no relationship between a patient's underlying survival time and the treatment arm they are randomised to so randomisation balance is maintained. The parameter estimates $\theta = (\lambda_u, \gamma_u, \lambda_z, \gamma_z, \phi)$ and ψ are found by finding the solution of the estimating equations:</p> $S_\theta(\theta, \psi) \stackrel{\text{def}}{=} \sum_i \frac{\partial}{\partial \theta} L_i^*(\theta, \psi, \rho) _{\rho=0} = 0$ $S_\rho(\theta, \psi) \stackrel{\text{def}}{=} \sum_i \frac{\partial}{\partial \rho} L_i^*(\theta, \psi, \rho) _{\rho=0} = 0$ <p>Where $L_i^*(\theta, \psi, \rho)$ is the full log likelihood for participant i for the augmented model.</p>
What are the key assumptions of the method?	<ul style="list-style-type: none"> i) Assume that U_i is a baseline variable identically distributed across the randomised groups. ii) Assumes any change in treatment in the experimental group is as a consequence of using the experimental treatment. iii) The censoring mechanism is non-informative whenever the censoring time on the U-scale is independent of U_i conditional on the time of treatment change Z_i. iv) Models must be accurately specified.
What are the theoretical advantages and disadvantages associated with the method?	<p>The method avoids recensoring, and so avoids loss of information and any sensitivity to treatment by time interaction. Compared with recensoring in a semi-parametric model the parametric model is comparable to the ITT in terms of numbers of events and time at risk, and so ensured robustness to small treatment by time interactions. It may also allow greater precision in circumstances when recensoring leads to a considerable loss of information.</p> <p>However the disadvantages are the parametric assumptions of the model. It is particularly sensitive to model misspecification. This is particularly important for the frailty model, which is hard to test and so the chance of misspecification is high.</p>
What are the potential biases associated with the method?	If the model is misspecified bias could result.

Why might the method not be appropriate?	Due to potential model misspecification.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	It extends the RPSFTM method by using a fully parametric process and joint modelling of U and Z so that recensoring can be avoided.
Does the method represent an extension to another method?	Yes, the RPSFTM method.
Application	
Is there a worked example in the survival setting?	<p>Yes, simulation studies are carried out, and the method is tried out on a real-world dataset.</p> <p>The simulation study was conducted to assess the performance of the estimation equation (EE) and maximum likelihood (ML) approaches for estimating ψ when frailty and marginal hazard functions were correctly specified and incorrectly specified. Different true frailty functions, strength of associations, treatment effect and % of switching were considered. The parametric joint model was compared to a parametric ITT analysis not based upon treatment received and a parametric model for U ignoring Z dependence.</p> <p>The real-world data was from the HIC Concorde trial in which 34% of control patients started taking the active treatment prior to disease progression (at which point they were scheduled to switch). The authors estimate time to progression if this crossover had not occurred.</p>
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>Simulations</p> <p>The authors found that when correctly specified models were used for the frailty distribution the ITT estimates of ψ were unbiased when $\psi = 0$ as expected, but were biased towards the null when $\psi \neq 0$, with bias increasing as the proportion of switching increased. The parametric model for U ignoring the informative censoring due to Z always produced biased estimates of ψ – the underlying link between poor prognosis and change of treatment leads to bias in the essentially observational effect of treatment received. The authors found that there was very little bias in the estimate of ψ using the joint model whether an ML or an EE approach is taken.</p> <p>When the frailty part of the parametric model was misspecified, the joint model using both the ML and EE still performed well when the switching % was 35% and censoring was 70%. Bias increased as the % of switching increased and the association between U and Z increased. Overall the EE version performed as well as or better than the ITT and ML models, and significantly better than the independent U model, when frailty was misspecified. The ML version was particularly sensitive to reducing the % of censoring. When censoring was low, the ML version performed more poorly than the ITT model. The EE version was not sensitive to the censoring %. Misspecifying both the U and the Z margin and the frailty distribution gave similar results to misspecifying frailty only, with EE performing as well as or better than ML. The authors placed emphasis on the frailty misspecification as the frailty is harder to measure in real life than the marginal hazards (as marginal hazards and potential distributions can be assessed by examining the observed distribution of Z censored by T, and information from non-parametric RBEEs could be used to assess possible distributions for the U margin).</p> <p>The authors conclusions based upon the simulation study were that the EE version was more robust to frailty misspecification than the ML version. EE generally overestimates the treatment effect when a gamma frailty model is used incorrectly (when the true frailty model is a positive stable model). Performance was poorest when this misspecification occurred, a large % of switching occurred, and censoring was moderate. The authors therefore suggest that in general it will be best to fit a positive stable frailty model using the EE approach (generally this led to small positive bias (underestimates of the treatment effect). They state that in general Weibull models provide a flexible parametric approach for the marginal hazard (although this can be analysed by looking at the T-censored data on Z).</p> <p>Real-world</p> <p>Using an AFT ITT model the estimate of ψ was -0.101 (not stat sig, $p=0.18$). Using the semi-parametric RPSFTM method this was -0.178 with the same significance level. The authors found that the best fitting joint model using the ML approach used a gamma frailty and Weibull margins for U and Z. However the ML approach was very sensitive, with estimates of ψ varying substantially when different parametric models were used. Using the EE approach the estimates of ψ were much less sensitive to different parametric models (which the authors state was reassuring since some model misspecification probably exists, but perhaps this is not very important). There was a gain in precision associated with both the EE and ML approaches compared to the semi-parametric approach, probably due to the loss in information from recensoring. The authors found that all the EE estimates of ψ were closer to the ITT estimate than the semi-parametric method. The authors state that this probably means that the effect of recensoring was greater than expected in the Concorde trial, probably because any suggestion of treatment effect is confined to early follow-up. The authors selected the EE estimate of ψ from the Weibull models that were best fitting for the ML approach, which gave an estimate of ψ of -0.112.</p>
Other Issues	

<p>Are there any other relevant characteristics associated with the method?</p>	<p>The authors state that gamma and positive stable frailties are extreme positive association models, and so it is possible that using an intermediate frailty such as the lognormal or inverse Gaussian frailty, would provide better performance under all kinds of frailty misspecifications.</p> <p>The authors recommend that in practical applications various frailty distributions and marginal hazard distributions are used for sensitivity analysis. They also suggest starting with the ML version, followed by the EE approach using the ML approach to define parametric model choice based upon which fitted best when the ML approach was used.</p> <p>Note that in subsequent research Morden <i>et al</i> (2011) found that the method was of limited practical use because it often failed to converge and often led to substantial overestimates of the treatment effect. In fact, it generally produced more bias than naive approaches (ITT, censor or exclude crossover patients). Hence in the interests of practicality (not including methods that have already been shown to perform poorly) it is reasonable to exclude this method from the simulation study included in this thesis.</p> <p>In discussion with the authors of this paper (White and Walker), they have noted that the method often does not work well, and they can not fully explain this, although it is likely to be due to its complexity and the requirement to accurately specify several complex models.</p>
---	--

<p>Reference</p>	<p>Tanaka Y, Matsuyama Y and Ohashi Y. Estimation of treatment effect adjusting for treatment changes using the intensity score method: Application to a large primary prevention study for coronary events (MEGA study). <i>Statistics in Medicine</i> 2008;27:1718-1733.</p>
<p>Origin</p>	
<p>Was the method developed specifically in the survival analysis context?</p>	<p>Yes, the authors extend a previous method so that it can be used for time-to-event outcomes.</p>
<p>If not, what was the original context and how has the method been adapted?</p>	<p>Originally the method was developed for continuous outcomes.</p>
<p>Theoretical Suitability</p>	
<p>How does the method work?</p>	<p>The authors state that the intensity score reflects the cumulative difference over time between treatment actually received and treatment predicted by prior observed medical history. The treatment effect in a structural nested mean model (as originally developed by Robins) can be obtained by regressing outcomes on the intensity score. The authors state that the method provides an easy implementation of g-estimation for the analysis of non-random non-compliance.</p> <p>The method was developed in the context of the MEGA study, which was a large RCT to evaluate the primary preventive effect of a statin against CHD. Patients were randomised to pravastatin plus diet, or diet alone. Many patients crossed over treatment during the study. Patients in the control group could be switched to the intervention if their cholesterol levels were not reduced, whereas patients in the intervention group could stop taking pravastatin when a reduction in cholesterol was observed. Over the 10 year trial period 21% in the control arm and 63% in the intervention arm switched treatment at some point.</p> <p>The authors set out a multiplicative structural nested mean model (SNMM). Each patient <i>i</i> is randomised to two treatments and receives a treatment at the start of each time <i>t</i> (<i>t</i>=0,..., <i>M</i>-1 where <i>t</i>=0 is the randomisation time and time of first treatment). However some patients fail to comply and receive the other treatment at each time <i>t</i>.</p> <p>The authors assume that there are repeated measures on treatment $S_i(t)$ (=1 for intervention, 0 for control), and covariates $L_i(t)$ at time <i>t</i>. $H_i(t)$ is the observed history of treatment and the covariates prior to time <i>t</i> ($H_i(t) = L_i(0), S_i(0), \dots, L_i(t-1), S_i(t-1), L_i(t)$ with $H_i(0) = L_i(0)$). $T_i(\bar{S}_i(t), 0)$ denotes the potential event times in response to the hypothetical treatments ($S_i(0), \dots, S_i(t), S_i(t+1) = 0, \dots, S_i(M-1) = 0$). Therefore this represents the event time that would have been observed if (possibly contrary to fact), the patient had his/her treatment history up until time <i>t</i>, but was switched to control treatment at <i>t</i>+1 and remained on that treatment until the event was observed.</p> <p>The authors present a multiplicative SNMM as introduced by Robins.</p> $\log E[T_i(\bar{S}_i(t), 0) H_i(t), S_i(t)] - \log E[T_i(\bar{S}_i(t-1), 0) H_i(t), S_i(t)] = \beta_0 S_i(t)$ <p>Where β_0 is the constant (across <i>t</i>) incremental causal effect of a final treatment $S_i(t)$ at time <i>t</i> on the potential outcome $T_i(\bar{S}_i(t-1), 0)$ following a patient's actual treatment through times 0,..., <i>t</i>-1 and control treatment after <i>t</i>-1. In this constant treatment effect model β_0 multiplied by <i>M</i> is the average causal treatment effect that would be realised if all patients complied with the treatment to which they were assigned. The authors state that Robins proposed g-estimation for the estimation of β_0 under the no unmeasured confounders assumption. The authors state that this assumption is unlikely to be precisely true, but given a rich collection of prognostic factors that influence a patients decision to comply at time <i>t</i> recorded in $H_i(t)$ it may be approximately true.</p>

The authors state that Brumback *et al* showed that the SNMM treatment effect (the g-estimator of β_0) can be obtained by the intensity score method, which involves regressing the outcomes on the cumulative intensity score. The authors use Brumback's results and use an AFT model to obtain β_0 from the multiplicative SNMM. They assume that T_i is subject to independent random censoring such as end-of-study censoring where for censored subjects T_i is the time to drop out or time until the end of study. They assume a exponential (log-linear) regression model:

$$\log T_i = \mu + \beta_i \sum_{t=0}^{M-1} \hat{I}_i(t) + \varepsilon_i$$

where μ is the intercept parameter, $I_i(t) = S_i(t) - E[S_i(t)|H_i(t)]$ is the intensity score at time t, and ε follows the extreme value distribution. For binary treatment $S_i(t)$ the time-dependent intensity score $I_i(t)$ measures departures of actual treatment from the propensity score $\Pr [S_i(t)|H_i(t)]$ developed by Rosenbaum and Rubin. The authors state that the propensity score is usually unknown and has to be estimated from the data, therefore they assume a parametric model for $\Pr [S_i(t)|H_i(t)]$ such as;

$$\text{logitPr}[S_i(t) = 1|H_i(t)] = \theta^T H_i(t)$$

They state that the intensity score at time t can then be estimated by

$$\hat{I}_i(t) = S_i(t) - E[S_i(t)|H_i(t); \hat{\theta}]$$

Where $\hat{\theta}$ is the MLE of θ under the parametric model. The authors assume that the intensity score does not equal 0 at any time t.

The authors state that the estimate for β_i in the parametric model above can be obtained via ordinary weighted least squares (WLS) regression. However, they state that the cumulative intensity score is generally uncorrelated with the cumulative propensity score, although $E[I_i(t)E(S_i(t)|H_i(t))] = 0$ for any t. Hence to obtain a consistent estimator of β_0 the correction term $N\beta_i C$ must be subtracted from the WLS estimating function.

where

$$C = \left(\frac{1}{N}\right) \left(\sum_{t=0}^{M-1} \hat{I}_i(t)\right) \omega_i \left(\sum_{t=0}^{M-1} E(S_i(t)|H_i(t); \hat{\theta})\right)$$

And

$$\omega_i = \exp(-\mu) \cdot T_i \cdot \exp(-\beta_i \sum \hat{I}_i(t))$$

Thus the authors state that the corrected estimating function for the exponential (log-linear) regression model is

$$U(\mu, \beta_i) \equiv \sum_i^N \left(-d_i \sum \hat{I}_i(t) + \sum \hat{I}_i(t) \exp[\log t_i - \mu - \beta_i \sum \hat{I}_i(t)]\right) - N\beta_i C = 0$$

Where d_i is the event indicator which is 1 if the subject failed and 0 if censored. The authors offer a proof to show that the correcting term is required. They also demonstrate how the asymptotic variance can be estimated for the estimates of the intercept, the coefficient of the intensity score and the parameter used to model the propensity score, using a sandwich estimator.

The authors go on to state that an extension of the multiplicative SNMM that they use is to allow the treatment effects to vary over time. Thus the multiplicative SNMM is altered to:

$$\log E[T_i(\bar{S}_i(t), 0)|H_i(t), S_i(t)] - \log E[T_i(\bar{S}_i(t-1), 0)|H_i(t), S_i(t)] = \beta_0(t) S_i(t)$$

Where $\beta_0(t)$ is the causal parameter at each time t. They state that because $\beta_0(t)$ is the incremental causal effect of a final treatment $S_i(t)$ at time t, the cumulative effect $\sum_{k=0}^t \beta_0(k)$ is the average causal treatment effect that would have been observed if all patients had continued to comply with the treatment assigned to them at time t. Assuming the consistency assumption and the no unmeasured confounders assumption the time-dependent causal parameters in the above model are consistently estimated using the following model:

$$\log T_i = \mu + \sum_{t=0}^{M-1} \beta_i(t) \hat{I}_i(t) + \varepsilon_i$$

In this case, the authors state that the following correction term must be subtracted from the WLS estimating function for the above model:

$$\sum_i [\hat{I}_i(t) \cdot \omega_i \cdot \sum_{t=0}^{M-1} \{\beta_i(t) E(S_i(t)|H_i(t); \hat{\theta})\}]$$

What are the key assumptions of the

i) Rubin's stable unit treatment value assumption – that is potential outcomes for patient i do not depend on the treatment received by any other patient

method?	<ul style="list-style-type: none"> ii) Potential outcomes for patient i satisfy the consistency assumption that links the potential outcomes with the observed outcomes T_i. iii) Censoring is independent and random. iv) The intensity score never =0 at any time point t. Therefore at no covariate levels is a patient certain to receive the treatment. v) No unmeasured confounders assumption.
What are the theoretical advantages and disadvantages associated with the method?	<ul style="list-style-type: none"> i) The authors state that censoring requires special care when using g-estimation based on a structural accelerated failure time model, while the intensity score approach can treat the censoring within the framework of standard regression models. ii) The authors state that the intensity score approach allows estimates of parameters within a structural nested mean model that allow the treatment effects to vary over time, which they state is difficult to achieve practically using the g-estimation technique. They state that this results in a model that is more robust to the estimation of dynamic sequential treatments conditional on past medical history. However, they state that this added robustness is compromised by the sacrifice of precision. They state that Brumback proposed the use of parametric constraints among the $\beta_i(t)$ to avoid sparse data problems. iii) The authors also suggest that their method is easy to apply. Put simply, they state that they compute propensity scores, derive intensity scores, and fit an ordinary regression model for any outcome variable (and subtract the correction term) to derive the estimator. <p>However, the authors also note that the RPSFTM (g-estimation) approach also has several advantages over their method:</p> <ul style="list-style-type: none"> i) The RPSFTM is a semi-parametric randomisation-based approach which preserves the validity of tests of the null hypothesis regardless of what determinants of outcome have influenced a patient's decision to comply. It leads to estimated effects that are only significant if the ITT estimate is significant. ii) For the intensity score approach the analyst must be able to specify a correct model for the conditional probability of treatment for each t up to the end of follow-up (though there will be an increase in power). The authors state that the no unmeasured confounders assumption is not testable. iii) The authors state that even when the no unmeasured confounders assumption approximately holds other strong modelling assumptions are required as there are many covariates in $H_i(t)$. It is unlikely that all will be precisely correct. However, in their application to real-world data all potentially important prognostic factors were included. They tested other models where they used variable selection procedures and included variables in different ways, and found that the intensity score estimates were insensitive to the prediction model conditional on the measured covariates. iv) The authors state that another advantage of the RPSFTM method is that it can be used where at each time point there is a possibility that patients with specific covariates are certain to receive the identical treatment. This may often be the case in the analyses considered in this thesis, when we are only considering switching from one treatment arm (switching may occur the other way, but we may not define it as switching and may not wish to control for it). The intensity score approach should not be used when switching was only observed in one group because the intensity score will be 0 at each time for patients in the complete compliance group – the authors state that in these circumstances the estimators will be biased for those with structural 0.
What are the potential biases associated with the method?	The substantial data requirements and the fitting of several models risks misspecification and associated bias. Also the method is only suitable if switching occurs in both treatment arms. And the method appears to correct treatment effect estimates if any treatment switch occurs, even if this was due to toxicity. This could lead to bias.
Why might the method not be appropriate?	The authors state that the intensity score method is only suitable if switching occurs from both arms of the trial.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The method is largely independent to the other methods reviewed, although its use of an AFT model is similar to SNMs, and the modelling of treatment and covariate history is similar to observational methods. The propensity / intensity score is similar to inverse probability of censoring / treatment weights.
Does the method represent an extension to another method?	The authors note that Brumback <i>et al</i> proposed the intensity score approach for the analysis of time-varying treatments in the presence of time-dependent confounding. They provided conditions under which the intensity score approach consistently estimates a treatment effect in a structural nested mean model, previously developed by Robins. The authors state that the intensity score approach was previously proposed for continuous outcomes (presumably by Brumback), and they extend this for time-to-event outcomes with censoring.
Application	
Is there a worked example in the survival setting?	<p>Yes, the authors present a simulation study and a real-world study comparing the intensity score method to the Robins and Tsiatis G-estimation RPSFTM method, an as-treated analysis and an ITT analysis. The simulation study and the real-world application involved varying levels of patients switching between two treatment groups (there was always switching in both groups).</p> <p>In the simulation study baseline covariates and potential failure times were simulated and these were related. The treatment actually received was assumed to be assessed at two points after baseline, at which points' crossover could occur. The observed failure time was calculated using a structural accelerated failure time model and censoring was included. In the</p>

	simulation study the constant treatment effect intensity score model was applied, rather than the time-dependent model. For the real-world study the parametric logistic regression model presented above was used and four time-dependent factors (lipid levels and treatment received before t) and 12 baseline factors were included as covariates $H_i(t)$. Missing data for the covariates were imputed. The authors tested fitting the constant treatment effect model and the time-dependent treatment effect model.
Is the example relevant?	Yes, although switching occurs in both directions – which is not the primary interest in our analysis of treatment switching (as we assume that a patient switching from intervention to control does not need to be corrected for).
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	The simulation study showed that the ITT and as-treated models were both largely biased towards the null. The RPSFTM method performed well with low bias and power similar to the ITT method (the authors note that although the RPSFTM method uses information on non-compliance it does not increase the power against the null hypothesis compared to the ITT approach). The intensity score approach had low bias, but this was slightly higher than the RPSFTM approach. It gave slightly higher power and narrower CIs than the RPSFTM method. For the real-world data both the constant treatment effect intensity score model, the time-dependent intensity score model and the RPSFTM method gave larger treatment effects than the ITT approach, which was expected. The results using the different adjustment techniques did differ. Generally the time-dependent intensity score model gave results closer to the ITT results. Whereas the RPSFTM method requires that no result that was not significant in the ITT analysis can become significant in the RPSFTM analysis (this is because the p value is maintained from the ITT analysis), in one case for one outcome the result became significant in the intensity score analysis whereas it was not in the ITT analysis. In general the CIs were widest for the time-dependent intensity score analysis – the authors state that this is likely to be due to the sparse data problems associated with estimating $\beta_I(t)$.
Other Issues	
Are there any other relevant characteristics associated with the method?	The authors note that their intensity score approach can also incorporate the IPCW method, and that this will be addressed in future work.

Included Studies: Secondary Search

Reference	Mark SD and Robins JM. A method for the analysis of randomized trials with compliance information – an application to the multiple risk factor intervention trial. <i>Controlled Clinical Trials</i> 1993;14;2:79-97
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	In this paper the authors apply the RPSFTM method to a clinical trial for smoking interventions, and also test the method in a simulation study in which compliance rates were varied. The paper briefly introduces a multivariate extension to the RPSFTM method, and thus is included in this review. However, the summary given of the multivariate extension is very brief, and hence the review presented here is also brief. The multivariate extension is described in the context of the MRFIT trial in which the effect of cigarette smoking on various health outcomes was studied. The goal of the authors' analysis was to estimate the effect of quitting smoking on time to death or first MI if everyone in the treated group had quit smoking, and everyone in the control group had continued to smoke. In reality some members of the treatment group did not quit, and some members of the control group did quit. The authors use the RPSFTM method. However, in their discussion they state that a multiparameter model could have been constructed to allow for the effect of quitting to depend on treatment group, or on a time-varying covariate. This would be similar to allowing for a non-constant treatment effect when correcting for treatment crossover. In the example given by the authors they state that if we thought that the benefit associated with quitting was greater in subjects with an elevated blood pressure, such a smoking-hypertension interaction could be accounted for by using a two-parameter RPSFTM such as: $U_i = \int_0^{T_i} \exp(\psi_{10}Q_i(u) + \psi_{20}Q_i(u)I_i(u))du$ Where $I_i(u)$ is an indicator variable for hypertension at time u. The authors state that the importance of this interaction could be tested by performing a Wald test of $\psi_{20} = 0$.
What are the key assumptions of the method?	See RPSFTM.

What are the theoretical advantages and disadvantages associated with the method?	See RPSFTM.
What are the potential biases associated with the method?	See RPSFTM.
Why might the method not be appropriate?	The authors state that although they suggest using the multivariate approach to test for effect modifiers, they are not certain how many parameters can be accommodated in real data under the constraints of a randomised analysis. They state that in the MRFIT trial noncompliance was severe enough that meaningful estimates of a two-parameter model were not possible.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	See RPSFTM.
Does the method represent an extension to another method?	Yes, the RPSFTM method.
Application	
Is there a worked example in the survival setting?	Yes, the MRFIT trial, but the multivariate approach did not work and the results were not presented.
Is the example relevant?	Not applicable.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	Not applicable.
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Reference	Mark SD, Robins JM. Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. <i>Statistics in Medicine</i> 1993;12:1605-1628
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	In this paper the authors apply the SNM method with g-estimation to a clinical trial for smoking interventions. The paper is slightly confusing because it refers to the SNM as a RPSFTM, which we refer to in this thesis as a randomisation based approach. This demonstrates the similarities between the RPSFTM and SNMs as the RPSFTM is a type of SNM. The SNM developed and applied by the authors is the same as that developed in the Robins (1998) paper on structural nested failure time models, which is reviewed in detail later in this appendix. Thus the details are not replicated here. The added interest of the paper is that it compares the SNM to a time-dependent Cox model and a model that only adjusts for baseline covariates, and finds that the SNM produces almost identical results, because (according to the authors) the time-dependent covariate was relatively unimportant. The paper also discusses how variables might be tested to see if they are time-dependent, and also briefly introduces a multivariate extension to the SNM method (which can be done in an identical way to the RPSFTM extension), but this is summarised very briefly and is not tested. Thus, no further details are given in this table.
What are the key assumptions of the method?	See SNM.
What are the theoretical advantages and disadvantages associated with the method?	See SNM.

What are the potential biases associated with the method?	See SNM.
Why might the method not be appropriate?	See SNM, and also added difficulties associated with attempting to estimate two different treatment effects reliant upon the no unmeasured confounders assumption. Given that in an RCT context we conclude that an SNM is likely to be best applied in a two-stage approach, a multivariate approach becomes less important.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	Interestingly, the authors find that the SNM provides very similar results in their example to a simple time-dependent Cox model, and also a model that only incorporates baseline covariates. Their other paper published in 1993, reviewed above, states that a RPSFTM approach gives similar results (although point estimates are different – quitting associated with a 53% increase in time to event in the observational analysis compared to 74% in the randomisation-based analysis. The authors note that the extra assumptions required by the observational analysis resulted in narrower confidence intervals which led to a statistically significant result. However, they note that the observational approach is prone to bias under the null hypothesis if models of the relationships between time-varying confounders and compliance and outcome are wrong, whereas the randomisation-based analysis is unbiased under the null hypothesis regardless of the effects that these confounders have on compliance.
Does the method represent an extension to another method?	Yes, the SNM method.
Application	
Is there a worked example in the survival setting?	Yes, the MRFIT trial, but the multivariate approach was not applied and the results were not presented.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach)?	The SNM approach led to identical results to a time-dependent Cox model, because the impact of the time-dependent covariate did not appear to be important. The point estimate differed to that resulting from a randomisation-based RPSFTM, and confidence intervals were narrower.
Other Issues	
Are there any other relevant characteristics associated with the method?	It is worthy of note that the trial to which the method was applied was very large – there were 12,866 participants. Hence it might be expected that observational methods can work well.

Reference	Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association 2001;96;454:440-448
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>In this paper the authors introduce marginal structural models for estimating the causal effect of time-dependent treatments. Marginal structural models are compared to structural nested models. The author states that the main difference between MSMs and SNMs is that SNMs model the magnitude of the effect of a treatment given at t as a function of the prognostic factor history up to t, whereas MSMs models the causal effect of treatment given at t only as a function of baseline prognostic factors. Both methods use the counterfactual modelling approach, and the MSM approach uses inverse probability of treatment weights. The emphasis of the paper is clearly on measuring the effect of a time-dependent treatment. Treatment crossover is not mentioned. Together with a paper by Robins (1999) this paper provides a full description of the use of MSMs for causal estimation.</p> <p>The authors describe a marginal structural Cox proportional hazards model for the analysis of an AIDS observational dataset. Their objective was to identify the causal effect of AZT treatment and prophylaxis treatment for pneumocystis carinii pneumonia. For their analysis they specified the following MSM:</p> $\lambda_{T_{\bar{a}}}(t V) = \lambda_0(t) \exp(\beta_1 a_1(t) + \beta_2 a_2(t) + \beta_3' V)$ <p>Where $T_{\bar{a}} = T_{\bar{a}_1, \bar{a}_2}$ is the patient's time to death if he or she had followed AZT treatment history \bar{a}_1 and prophylaxis treatment history \bar{a}_2. $\lambda_{T_{\bar{a}}}(t V)$ is the hazard of $T_{\bar{a}}$ at t conditional on having pretreatment variables V. $\lambda_0(t)$ is an unspecified baseline hazard function. $\exp(\beta_1)$ and $\exp(\beta_2)$ are the causal rate ratios for the effects of AZT and prophylaxis treatment.</p>

V is the vector of baseline regressors comprised of a selection of prognostic variables.

It is important to note that this model specifies the hazard of death at time t to depend only upon current treatment status and not previous treatment.

Hernan *et al* (2001) suggest that a standard time-dependent Cox model such as:

$$\lambda_T(t|\bar{A}_1, \bar{A}_2(t), V) = \lambda_0(t) \exp(\beta_1 A_1(t) + \beta_2 A_2(t) + \beta_3' V)$$

could be used to estimate the vector of β , if the giving of treatment at time t is completely random, or if the treatment decision only depended upon the history of treatment prior to t . In this case the β parameters will have causal interpretation, because the treatment decision is causally exogenous. Randomised treatments are causally exogenous, but we know that when unplanned treatment crossover occurs the choice of treatment received by potential crossover patients in the control group is not causally exogenous. The extent to which a treatment process is statistically non-exogenous can be estimated using the following equation:

$$W(t) = \prod_{k=0}^t f[A(k)|\bar{A}(k-1), \bar{L}(k)]/f[A(k)|\bar{A}(k-1), V]$$

Where the numerator is the probability that a patient received his or her own observed treatment at time k , $A(k)$ given his or her past treatment and prognostic factor history ($\bar{A}(k-1)$ and $L(k)$); and the denominator is the probability that the patient received his or her observed treatment conditional only upon past treatment history and baseline variables – not also conditional on prognostic factor history. If the treatment process is exogenous, this will equal 1 for all t . If the treatment process is not exogenous, the inverse probability of treatment weights (IPTW) method is used, whereby the time-dependent Cox model is weighted by applying the weight W^t to each patient for each time k . Essentially this means weighting by the inverse of a patient's probability of having his or her own observed treatment history. Using this weighting means that $\bar{L}(t)$ does not predict treatment at t given past treatment history, and thus a counterfactual pseudo population is created in which treatment is exogenous. Also, the causal effect of treatment is the same in the counterfactual pseudo population as it is in the original population, and so to estimate treatment effects we can conduct standard time-dependent Cox model analysis on the pseudo population. If the assumption of no unmeasured confounders holds – that is $L(t)$ includes all relevant time-dependent prognostic factors, then the weighted estimators will converge to values of β that can be appropriately interpreted as the causal effect of treatment history on the time to event.

The authors demonstrate how to incorporate censoring into the analysis. They allow for both administrative censoring and censoring due to drop-out. They adjust for this by defining that it is desired to estimate the effect of \bar{a} when $\bar{c} \equiv 0$, that is, a patient's failure time when treated with a certain regimen is estimated in the absence of censoring:

$$\lambda_{T_{\bar{a}, \bar{c}=0}}(t|V) = \lambda_0(t) \exp(\beta_1 a_1(t) + \beta_2 a_2(t) + \beta_3' V)$$

In order to obtain consistent estimates of β in this situation, it must be assumed that censoring is noninformative (ignorable) given treatment history and time-dependent covariate history, and thus there is no unmeasured confounding. To take censoring into account a patient who is alive and uncensored at time t is weighted by $W(t) \times W^\dagger(t)$, where:

$$W^\dagger(t) = \prod_{k=0}^t \frac{\Pr[C(k) = 0 | \bar{C}(k-1) = 0, \bar{A}(k-1), V, T > k]}{\Pr[C(k) = 0 | \bar{C}(k-1) = 0, \bar{A}(k-1), \bar{L}(k), T > k]}$$

Which is equivalent to the inverse of the ratio of a patient's probability of remaining uncensored up to time t divided by that same probability calculated as though the only determinants of censoring were past treatment history and V . Therefore, IPTW is complemented with inverse probability of censoring weights (IPCW) to account for censoring. $W(t)$ is also slightly amended, with $C(k)=0$ added to the conditioning events in the numerator and the denominator:

$$W(t) = \prod_{k=0}^t f[A(k)|\bar{A}(k-1), \bar{L}(k), C(k) = 0]/f[A(k)|\bar{A}(k-1), V, C(k) = 0]$$

Under these conditions, the denominator of $W(t) \times W^\dagger(t)$ is the probability that a patient has his or her own observed treatment and censoring history through time t .
135;145132;142132;142132;142132;142132;142132;142132;142132;142128;139129;140

The authors then go on to explain how $W(t)$ and $W^\dagger(t)$ are estimated. In the case where there are two treatment effects of interest a pooled logistic model for the binary responses for A_j and censoring are fitted, as follows:

	<p>Let:</p> $P_1(k) = \Pr[A_1(k) = 0 \bar{A}_1(k-1) = 0, \bar{A}_2(k-1), \bar{L}(k), \bar{C}(k) = 0, T > k]$ $P_2(k) = \Pr[A_2(k) = 0 \bar{A}_1, \bar{A}_2(k-1) = 0, \bar{L}(k), \bar{C}(k) = 0, T > k]$ <p>Which denote the conditional probability of remaining off treatments A_1 and A_2 at time k given past treatment and covariate history. These are equivalent to the numerator of $W(t)$. The denominator of $W(t)$ is addressed by defining $P_1^*(k)$ and $P_2^*(k)$ which are similar to $P_1(k)$ and $P_2(k)$ except $\bar{L}(k)$ is replaced by V.</p> <p>The authors then show that estimates $\hat{P}_j(k)$ of $P_1(k)$ and $P_2(k)$ can be obtained by fitting the following pooled logistic model for the binary response to A_j to patients at risk of initiating treatment j at time k.</p> $P_j(k) = \text{expit}\{\alpha_{j0}(k) + \alpha'_{j1} Q_j(k)\}$ <p>Where the $\alpha_{j0}(k)$ are time-unit specific (eg day, week, month) intercepts; $Q_1(k) = (A_2(k-1), L(k)', V)'$; and $Q_2(k) = (A_1(k), A_1(k-1), L(k)', V)'$; (note that expit is the inverse function of the logit).</p> <p>Using the same technique, but removing $L(k)$ from $Q_j(k)$ estimates can be estimated for $P_1^*(k)$ and $P_2^*(k)$. Then, the estimate of the numerator of $W(t)$ is $\bar{P}_{1i}(t)\bar{P}_{2i}(t)$, which is calculated by:</p> $\bar{P}_{ji}(t) = \prod_{m=0}^t \hat{P}_{ji}(u)$ <p>if the patient i did not start treatment A_j up to time t, and</p> $\bar{P}_{ji}(t) = [1 - \hat{P}_{ji}(k)] \prod_{m=0}^{k-1} \hat{P}_{ji}(u)$ <p>if the patient i started treatment with A_j at time k for $k \leq t$.</p> <p>These calculations highlight an additional assumption of the MSM – it is assumed that once a treatment has been started the patient does not stop taking it at any point.</p> <p>Estimates of the denominator of $W(t)$, $\bar{P}_{1i}^*(t)\bar{P}_{2i}^*(t)$ are obtained by simply replacing $\hat{P}_{ji}(k)$ with $\hat{P}_{ji}^*(k)$ in the above estimators. To estimate the censoring weight, $W^\dagger(t)$ the exact same methods are used to fit logistic models for the binary responses for $C(k)$. The authors state that Robins (1999) proved that these IPTW estimators will be consistent so long as the models for treatment initiation and censoring used in the numerators of $W(t)$ and $W^\dagger(t)$ are correctly specified. However, Hernan <i>et al</i> (2001) also note that to ensure that the IPTW estimate of β will be consistent in moderate sized samples it is necessary to ensure that the estimate of $W(t)$ is not overly variable, hence in their example they reduced the number of free parameters in the logistic model for $P_j(t)$ by not fitting a separate intercept $\alpha_{j0}(k)$ for every unit of time period k. In their example, k was measured in months and natural cubic splines with five knots at certain time points were used (effectively this would allow 5 different intercepts).</p> <p>The authors identify a slight complication with the MSM method – they note that while using the IPTW approach an ordinary time-dependent Cox model can be weighted to estimate consistent causal treatment effects, in actual fact few software programs allow for time-varying weights. To avoid this problem, they explain that a weighted pooled logistic regression model can be fitted, treating each person-month as an observation and allowing for a time-dependent intercept.</p> <p>The authors state that because weights are used for the IPTW estimation standard error estimates will be incorrect, and thus robust variance estimators – which will provide a conservative confidence interval – must be calculated.</p>
<p>What are the key assumptions of the method?</p>	<p>No unmeasured confounders. Once a treatment has been started the patient does not stop taking it at any point. Robins (1999) proved that IPTW estimators will be consistent so long as the models for treatment initiation and censoring used in the numerators of $W(t)$ and $W^\dagger(t)$ are correctly specified.</p>
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>The authors consider the major advantage of MSMs to be that they resemble standard models, since they take the form of Cox proportional hazard models. However, there are important advantages that SNMs hold over MSMs. Firstly, MSMs cannot be used if there are any particular covariate values that ensure that a patient will definitely receive a certain treatment at time k. In the context of treatment crossover this would be a problem if all patients in the control group switched treatment at the point of disease progression, and is an even more serious problem if we do not wish to control for any non-compliance in the group randomised to the experimental treatment. As discussed earlier, typically we do not wish to control for such non-compliance as it is likely to be due to medical reasons that are likely to also occur in the real-world. Hence it may only be possible to apply MSMs to the control arm of trials, which significantly reduces the amount of information which the model can use to make estimates for the treatment effect in crossover patients. This would be very likely to lead to estimating a different treatment effect in the control group than in the experimental group which may be considered an advantage from some perspectives, but such an analysis would certainly be open to question.</p>

	<p>Secondly, while the structure of MSMs makes them useful for considering interactions between treatment and baseline covariates, they are not as useful as SNMs for considering interactions between treatment and time-dependent covariates. Essentially SNMs model the magnitude of the treatment effect at t as a function of the prognostic factor history up to t, whereas MSMs measure the effect only as a function of baseline covariates, and weights to account for time-dependent covariates are estimated separately.</p> <p>On top of these issues with MSMs, the assumptions behind the approach are restrictive, as they are for SNMs that are not RBEEs. It must be assumed that the covariates included in the analysis are sufficient to adjust for both confounding and selection bias due to censoring – that is it is assumed that there are no unmeasured confounders and non-informative censoring. Also, the model specified for the effect of treatment on mortality must be correct, and the models used for initiation of treatment and for censoring must be correct. In defence of the problems associated with these strong assumptions the authors point out that when estimating the effect of a time-independent treatment using standard methods and observational data the same assumptions are made – there must be no unmeasured confounders, noninformative censoring and no model misspecification. However, as Robins (1999) states, that is why it is dangerous to draw causal inferences from observational datasets. This highlights the important advantage that the RBEE methods hold over observational SNMs and MSMs.</p>
What are the potential biases associated with the method?	Associated with the disadvantages and assumptions noted above – in particular the no unmeasured confounders assumption and modelling requirements.
Why might the method not be appropriate?	Associated with the disadvantages and assumptions noted above – in particular the method may not be fully applicable in an RCT context with treatment crossover.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	This method was developed as an alternative to SNMs – so it forms a separate type of method. It is based upon proportional hazards models rather than accelerated failure time models. The IPCW method is a type of MSM.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	The method is applied to an observational dataset, but the emphasis is very much on the theory.
Is the example relevant?	No.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach)?	Not applicable.
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Reference	Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. <i>Statistics in Medicine</i> 2000;19:1849-1864
Origin	
Was the method developed specifically in the survival analysis context?	No, but the method is extended for use in survival analysis.
If not, what was the original context and how has the method been adapted?	The method was developed specifically for estimating treatment effects in RCTs in the presence of non-compliance, but not in the context of survival analysis. However it was extended for use in survival analysis using the Cox proportional hazards regression model.
Theoretical Suitability	
How does the method work?	<p>The authors develop a method that they call the adjusted treatment received (ATR) method. Their basic idea is that they consider a situation in which the relationship between compliance and thus treatment received, and outcome is influenced by unobserved confounders. They show that the residual of the regression of the actual treatment indicator variable on the randomization arm indicator variable ‘intercepts’ the effect of such confounders, and they then include this residual in a multivariate analysis, in conjunction with the treatment indicator variable, which they believe should adjust for confounding. The residual measures the extent to which compliance splits the group into one with a good outcome and one with a bad (worse than the placebo group in the absence of treatment effect) outcome.</p> <p>The authors present examples from clinical trials in which non-compliers in the treatment group can be shown to have worse outcomes than a placebo control group. Assuming no placebo effect, the authors state that it would therefore be expected that compliers in the treatment group would have a better outcome than the placebo group, irrespective of treatment.</p>

	<p>Therefore such a comparison would be biased and this confounding needs to be taken into account in the analysis.</p> <p>The authors state that the confounders are usually unknown and standard methods to correct for bias cannot be used. However, they show that a variable (E) can be used in lieu of the real confounder(s) (C) to adjust for confounding in multivariate analysis in order to estimate the true causal effect.</p> <p>The authors set up their method by assuming:</p> <p>The randomization arm (R) has an effect on the treatment received (T), but this is also affected by C (which is a confounder as it also affects the outcome O), and variables N1, which do not affect O. The outcome O is affected by the treatment received (T), the confounder (C) and a group of other variables N2 (which do not affect T).</p> <p>The goal is to estimate the direct causal effect of T on O, which could be done by conditioning upon C, but this is not possible because C is unobserved. Instead, the authors state that it is possible to use any other variable that is not a descendent of T, and which blocks (or intercepts) every path (except the direct one) to O (this is called a 'back-door' variable, and such a variable avoids residual confounding (which occurs when C is modelled by a variable which does not block the path between C and T)). They define a variable E which blocks the path between C, N1 and T, which therefore meets these criteria. Nothing is known about E other than that together with R it fully explains T. We know that the effect of R on T can be measured by the fraction of patients who take the active drug in each trial arm, $E(T R)$. Then, if an additive effect between R and E is assumed, $T=E(T R)+E$, then E can be estimated by the residual of the regression of T on R. The residual is the difference between T and its expectation conditional on R. The value of E is then known for each individual in the trial. However, the authors state that if it is assumed that there is no switching in the placebo group, E will be identically 0 in this group, whereas in the treatment group while the mean of E will also be 0, the distribution will be different because E will be either positive or negative for each patient depending upon if they are a complier or a non-complier. This difference in distributions is unattractive because we assume complete comparability between the two groups due to randomisation.</p> <p>Instead, a multiplicative effect could be assumed: $T=E(T R)*E$. In this case: $E=0$ for the R=1 and T=0 group. This group comprises a fraction $1-E(T R)$ of the R=1 trial arm. $E=1/E(T R=1)$ for the R=1 and T=1 group. This group comprises a fraction $E(T R=1)$ of the R=1 trial arm. E is unknown or undetermined in the R=0 group. In this group the values of E can be estimated using mean imputation, setting $E = E(T R = 1) = 1$ in the R=0 group. This leads to the same values of E as under the additive model. The authors state that the multiplicative approach leads to E having the same distribution in the two trial arms.</p> <p>The causal effect of T on O can now be estimated by conditioning on E. This is due to the existence of the variable R, whose effect on O is via T – in this case R is an instrumental variable.</p> <p>The authors state that because E is a 'back-door' relative to (T,O) it can be treated as 'the' confounder, and it absorbs all of the non-treatment effects which may act as a confounder between treatment and outcome. Importantly, E must be uncorrelated (orthogonal) to R. If this were not the case, ITT would also be subject to confounding bias.</p> <p>One further assumption must be made to make the direct causal effect of T on O identifiable – the treatment effect does not depend on the level of the 'true' confounder C on a chosen scale (for example linear or logistic) of interest. C and T must be additive on that scale, and there must be no interaction – C is a confounder but not an effect modifier.</p> <p>Under this assumption we can then estimate the effect of T on O in a regression model by including E as an additional covariable in the regression. E absorbs the effect of C, such that conditional on E, C no longer has an effect on T. The authors state that it does not matter whether a linear regression, a logistic regression, a poisson regression or Cox's proportional hazards model is used, in all cases the treatment indicator variable can be used as a covariable in conjunction with E to correct for hidden confounders.</p> <p>The authors show that under a linear model relating O, T, C and R, the residual E can be used in lieu of C to estimate the causal effect of T on O, by using a linear regression of O on T and E. However they state that this is only approximately true under non-linear models. This is because in a linear model in the absence of a treatment effect the expected outcome is identical in the two treatment arms, and the mean value of E is 0 in both arms. However in a non-linear model this is no longer exactly true. E is 0 on average in both arms, but non-linearity might lead to different outcomes even in the absence of a treatment effect. The authors briefly suggest ways that could adjust for this, but state that unless E is very large, or the model has a very large curvature, the effects of non-linearity can be ignored. They also state that non-linear functions of the residual E may have a confounding impact on the relationship between T and O. They state that when patients can only switch in one direction (from intervention to placebo) only the residuals themselves need to be included as a covariable. However, when switching can occur in both directions and when it is suspected that in one arm patients with good prognosis comply whereas in the other patients with a poor prognosis comply, an R by E interaction may be included.</p> <p>The authors then consider survival analysis. They state that in this case compliance is not a 'yes' 'no' decision, but one that may be made after a prolonged period of compliance. They state that their method can be extended to such situations by taking as the regression of treatment group on randomisation arm $E(T R)$, the ratio of the life table estimate of the survival function of still being on treatment and alive and the life table estimate of being alive. They state that this regression, and thereby the residual E, then become time varying and can be</p>
--	---

	included (in conjunction with a time-varying treatment indicator) in a Cox proportional hazards model for time-varying covariates. They state that in this case (as whenever a Cox model is used), the proportional hazards assumption may be controversial. This is particularly important if late non-compliance is associated with good prognosis, and early non-compliance is associated with poor prognosis – in which case proportional hazards do not hold. Also, it must be assumed that there are no residual effects of the treatment – the benefit or negative effect of taking it disappears as soon as a patient stops taking it.
What are the key assumptions of the method?	<ul style="list-style-type: none"> i) Switching is only between the trial treatment arms. ii) There is no effect from R on O other than via T. This assumes that there are no placebo effects associated with being randomised to a particular treatment. iii) E is uncorrelated to R. iv) C is a confounder but not an effect modifier. v) In the survival context, proportional hazards are assumed. vi) The benefit/negative effect of a drug is lost as soon as a patient stops taking it. Any residual effect would invalidate the ATR method.
What are the theoretical advantages and disadvantages associated with the method?	The assumptions above are key – if they do not hold the method is not applicable. The method is randomisation based which is an advantage, but it has not been developed specifically for the crossover case as defined in this thesis, therefore its applicability may be questioned. It is also based upon a linear model, and therefore is only approximately true in non-linear applications. It also relies upon modelling different compliance types, which is not necessary for methods such as the RPSFTM
What are the potential biases associated with the method?	The model is only approximately true for non-linear models, and in the survival context is reliant on the proportional hazards assumption.
Why might the method not be appropriate?	<p>Although the method allows switching in both directions, the emphasis is firmly upon patients in the treatment group switching to the control treatment. The emphasis is also upon the control treatment being placebo. This does not reflect the switching from an active control treatment to the new intervention which is the primary problem associated with treatment switching. Also the method assumes no residual treatment effects and no examples are given of the method applied in the exact situation being studied in this review. Thus it is not certain that the method is entirely applicable to the crossover issue as defined here.</p> <p>Also, the authors state that their approach is different from one which includes compliance as a covariable. They state that in the case of a placebo control group in which there is no switching no imputations of results for potential non-compliers in the control group are needed – however again this is a case that is not relevant for the study of this thesis.</p>
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The authors state that their method is similar in spirit to the RPSFTM method. The RPSFTM considers hypothetical values for O in the absence of a treatment effect, which should be independent of the randomization arm indicator variable and then compare these hypothetical values to those that are actually observed. The ATR involves considering a covariate (E) that is unrelated to the randomisation arm indicator variable, and which accounts for differences in the outcome variable between compliance and randomisation groups in the absence of a treatment effect.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	<p>Yes, two examples using a logistic regression are given, and one example using a Cox proportional hazards model is given.</p> <p>The authors also present simulation studies based on a logistic model and a Cox proportional hazards model.</p>
Is the example relevant?	The Cox model example seems potentially relevant, but on closer inspection it is not particularly relevant. The example is of a trial of vitamin A supplementation, whereby patients in the intervention arm were given two doses. The authors use a Cox model with each dose treated as a time-dependent covariate. The residuals E1 and E2 (for each dose T1 and T2) are also fixed and time-dependent.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>The authors found that the first dose provided a 77% reduction in mortality risk, whereas the second dose has no significant effect.</p> <p>In the logistic simulations the authors found that their method gave estimates of the treatment effect that were closer to the truth than the ITT or as treated approach. However, the Cox model simulation was less successful. The simulation was very limited. 100 trials with 250 patients in each arm were simulated. It was assumed that there was no treatment effect, and the study duration was 12 time units, after which patients were censored. Non-compliance was homogeneously distributed over the 12 time units. Patients with non-compliance at time k had an exponential survival time with mean duration k – thus a very strong association between compliance and survival was assumed. The authors state that despite the non-linear nature of the model and an unrealistically strong effect of E, the method removed most of the spurious effect of T. However, the method was still significantly biased, which the authors put down to the data only meeting the proportional hazards assumption conditionally on some ‘frailty’ (in their example equal to the censoring time, but generally correlated with it) and violated the proportional hazards assumption unconditionally. They state that they do not know how to simulate such data, and that there is a clear need to further explore the ATR approach for clinical trials with survival end points.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Reference	Loeys T, Vansteelandt S, Goetghebeur E. Accounting for correlation and compliance in cluster randomized trials. <i>Statistics in Medicine</i> 2001;20:3753-3767
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>In this paper the authors argue that structural models such as the RPSFTM should not ignore clustering in the randomisation process, and they show how marginal modelling and random effects models can be used to allow structural models to account for clustering. Structural models rely heavily on the randomisation assumption, specifically on the creation of balanced groups and an ITT test of equal outcome distributions between randomised arms. However the authors state a number of problems with this assumption:</p> <ul style="list-style-type: none"> - False alarms of imbalance are easily generated when the clustered nature of randomisation is forgotten. - Clustering can cause differentially selective non-compliance when it is not expected – for example health experience of earlier cluster members can impact compliance of later cluster members. Small effects here, but large cluster sizes can yield substantial design effects, which can harm the validity of unadjusted statistical analysis. - In the context of infectious diseases, the assumption that the treatment of one person does not affect the outcome of another can be violated due to clustering. - In addition to this, clustering calls for corrections to standard errors to account for correlated outcomes. Robust standard errors of covariates, calculated under an independence working correlation GEE approach can dramatically change compared to naively calculated ones. <p>The authors state that ‘GEE’ methods can be used to calculate standard errors that account for correlated outcomes. Alternatively, random effects may model between-cluster heterogeneity explicitly.</p> <p>The authors consider these issues with regard to structural failure time models, and linear structural mean models.</p> <p>First, the authors consider how to correct the variance of the estimator of the marginal structural effect to allow for correlated treatment-free survival (thus assuming correlated counterfactual survival times) in a RPSFTM-type model (that does not include censoring). They state that traditionally (weighted) log-rank tests are used to check independence of the pseudo treatment-free outcomes and the randomisation group for a grid of acceleration factor values. The authors suggest two alternatives to this that can deal with correlated data. These allow for treatment-free responses that vary from cluster to cluster, but focus on a single average causal effect. They are both random effects models:</p> <ul style="list-style-type: none"> - Assume treatment-free survival times to follow a proportional hazards model conditional on a particular prognostic factor (in the example the factor was age as baseline risk was well known to depend on age): $\lambda(p A_{ij}) = \lambda_0(p)\exp(\beta A_{ij})$ <p>Where $\lambda(p A_{ij})$ denotes the hazard of treatment-free survival time P_{ij} conditional on age A_{ij}. The authors state that the effect of clustering can then be accounted for by using the robust variance estimator of Wei <i>et al</i> (1989). This is a marginal model approach.</p> <p>For this model, to construct estimating equations for the acceleration factor the authors propose for $V_{ij}(\psi)$ (pseudo treatment-free outcomes) estimation model:</p> $\lambda(v R_{ij}, A_{ij})\lambda_0(v)\exp(\beta(\psi)A_{ij} + \theta(\psi)R_{ij})$ - The second approach conditions on the cluster and assumes that a cluster-specific frailty captures the heterogeneity between clusters. This is a frailty approach: $\lambda(p A_{ij}, \omega_i) = \omega_i\lambda_{of}(p)\exp(\beta A_{ij})$ <p>They state that the most popular choice lets ω_i be gamma distributed over the clusters with mean 1 and variance σ^2. They state that diagnostics are available to assess the adequacy of the distributional assumption on the frailties using the observed data in the placebo arm. For this model, to construct estimating equations for the acceleration factor the authors propose for $V_{ij}(\psi)$ (pseudo treatment-free outcomes) estimation model:</p> $\lambda(v R_{ij}, A_{ij}, \omega_i) = \omega_i\lambda_{of}(v)\exp(\beta(\psi)A_{ij} + \theta(\psi)R_{ij})$ <p>$B(\psi)$ and $\theta(\psi)$ are unknown parameters which implicitly depend on ψ. At the true value of ψ_0 the estimating equations for both the approaches described above are independent of R_{ij} so that $\theta(\psi)$ should be zero. A point estimate of ψ is thus found as the ψ –value minimising the χ^2 value of the Wald test for $\theta(\psi)=0$. A test of $\theta(\psi)=0$ is called an auxiliary test.</p> <p>The authors then go on to consider that different clusters might experience different structural effects of treatment. They explore this by incorporating random treatment effects in a structural mean model framework.</p>

	<p>The authors assume that average causal effects are proportional to exposure, this time following the linear structural mean model for uncensored data:</p> $E\{T_{ij} - P_{ij} E_{ij}, b_i\} = (\psi_0 + b_i)E_{ij}$ <p>In this model $\omega_0 + b_i$ expresses how much the average causal effect of treatment is increased or decreased per unit increase in treatment exposure for members of the ith cluster. In contrast to the structural failure time model this model involves an equality in mean and not in distribution.</p> <p>The authors let b_i be independent equally distributed random draws with mean zero conditional on exposure. This results in ψ_0 being the population-averaged treatment effect. When ψ_0 is positive clusters with positive (negative) b_i experience larger (smaller) effects than can be expected in the general population. The variance σ_b^2 of b_i quantifies this heterogeneity. To estimate ψ_0 first pseudo treatment-free outcomes $V_{ij}(\psi) = T_{ij} - \psi E_{ij}$ are calculated for each individual. Under the randomisation assumption estimation proceeds testing to make sure there is no mean difference in pseudo treatment-free outcomes between arms. The authors go on to present an estimator for $\hat{\psi}$ when there is equal allocation to both treatments and zero experimental exposure in the control arm. As zero experimental exposure in the control arm is not relevant for the treatment crossover problem, this estimator and its variance are not presented in this review.</p>
What are the key assumptions of the method?	The method assumes that randomisation is clustered and that treatment-free outcomes are correlated. They develop approaches which allow the RPSFTM assumption that treatment-free outcomes are not correlated to be relaxed. Other assumptions of the RPSFTM method apply.
What are the theoretical advantages and disadvantages associated with the method?	The relaxation of the assumption that treatment-free outcomes are not correlated is an advantage, the disadvantage being the increased complexity of the model that this requires. The result seems to be that confidence intervals are wider and more robust, but may be unhelpfully wide in the case of the frailty approach when the frailty model is misspecified.
What are the potential biases associated with the method?	Model misspecification can lead to large losses in power if the frailty approach is taken.
Why might the method not be appropriate?	The method may be unnecessary if trial randomisation is not clustered. The structural mean model introduced is not developed in a context suitable for treatment crossover. Structural mean models are generally only suitable when there is no censoring.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The structural model section of the paper extends the RPSFTM method for cases when trial randomisation is clustered.
Does the method represent an extension to another method?	Yes, the RPSFTM.
Application	
Is there a worked example in the survival setting?	The authors went on to test the marginal and frailty structural failure time approaches compared to the naive standard estimation ignoring clustering, in simulation studies. They considered a scenario where there was no clustering effect, and various other scenarios regarding clustering. The authors also tested the different structural failure time approaches using a vitamin A trial.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	In the simulation studies of the structural failure time approaches the authors found that when there was no clustering effect the three approaches gave the same results. In all scenarios the mean treatment effect (but not the model-based variance) is the same for the naive approach and the marginal approach, but this differs for the frailty approach. When ω_i is drawn from a gamma frailty distribution the naive and marginal approaches lead to underestimates of the true conditional treatment effect but when ω_i is drawn from a positive stable distribution there were convergence problems for the gamma-frailty approach in 5% of simulations. Overall the authors conclude that the gain (loss) in power under a correct (misspecified) frailty model compared to the marginal approach is rather small (large) and hence the marginal approach might be preferred unless one is interested in estimating a cluster-specific baseline failure rate. In the Vitamin A application of the structural failure time approaches the authors found that the different auxiliary tests gave only very slightly different estimators of effect. Confidence intervals widened slightly for the marginal and frailty approaches compared to the naive approach, and in one case the frailty model led to an estimate being statistically non-significant when it was significant using the other two approaches. The authors stated that this may have been due to a loss of power for the frailty model due to a misspecification of the frailty distribution.
Other Issues	
Are there any other relevant characteristics associated with the	The structural mean model is not developed for treatment crossover in this paper. In the context of this thesis, where the focus is on metastatic cancer drugs and their RCTs, clustering is unlikely to be an issue.

method?	
Reference	<p>Yamaguchi T and Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part I: Structural nested models and marginal structural models to test and estimate treatment arm effects. <i>Statistics in Medicine</i> 2004;23:1991-2003</p> <p>And</p> <p>Yamaguchi T and Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. <i>Statistics in Medicine</i> 2004;23:2005-2022</p>
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>In this paper the authors discuss and examine the use of structural nested models and marginal structural models for adjusting for differential proportions of second-line treatments in clinical trials. This involves estimating the effects both of the initial treatment and the second-line treatment, as the paper does not consider crossover, but rather a separate second-line treatment. Therefore the method developed may be considered as not directly relevant for dealing with crossover. However, perhaps this is a good way for dealing with crossover, as the two treatments are considered separate, and have different effects. If we believe crossover patients are less likely to benefit from the experimental treatment, perhaps dealing with the drug as though it is a different treatment at first-line and second-line use is sensible as this allows a differential treatment effect to be modelled. Thus the paper is included in this review.</p> <p>The authors suggest two methods: structural nested models and marginal structural models to adjust for differential proportions of post-randomisation second-line treatment in cancer clinical trials, in the presence of time-dependent confounders that are themselves affected by previous treatment. The authors regard the second-line treatment as observational trials of these drugs.</p> <p>Structural nested models</p> <p>First, in the absence of censoring: R_i is the treatment arm indicator (1 if A, 0 if B). T_i is the time of death S_i is the time of second-line treatment initiation, if it is given, and the time of death otherwise σ_i is 1 if second line treatment is given ($S_i < T_i$), 0 otherwise.</p> <p>Four pre-treatment random variables, which due to randomisation are independent of the treatment arm indicator, are considered. Only one is observed for each patient, and the others are counterfactual or latent variables. They are: T_{i00} is the death time if B is allocated and second-line treatment is never received T_{i10} is death time if A is allocated and second-line treatment is never received $T_{i0s=h}$ is death time if B is allocated and second-line treatment history is h $T_{i1s=h}$ is death time if A is allocated and second-line treatment history is h</p> <p>The simple structural nested failure time model is:</p> $T_{i00} = \exp(\beta_r^* R_i) [S_i + \exp(\beta_s^*) (T_i - S_i)]$ <p>Which in terms of observable T_i is:</p> $T_i = S_i + \exp(-\beta_s^*) [\exp(-\beta_r^* R_i) T_{i00} - S_i]$ <p>To estimate parameters, the authors state that we define the observable random variable obtained by replacing (β_r^*, β_s^*) by (β_r, β_s)</p> $T_i(\beta_r, \beta_s) = \exp(\beta_r R_i) [S_i + \exp(\beta_s) (T_i - S_i)]$ <p>In the absence of censoring $T_i(\beta_r, \beta_s)$ is observed for each patient. The randomisation can then be used to estimate the parameters (T_{i00} is independent of R_i). Because two parameters</p>

are being estimated the authors use two weighted log-rank tests (log-rank score, Prentice-Wilcoxon score) to test the null hypothesis (T_{i00} is independent of R_i). If only the log-rank test was used there may be more than one set of values for (β_r, β_s) that meet the null hypothesis, which is known as the identifiability problem. The authors use:

$$Q_{LW}(\beta_r, \beta_s) = (S_L(\beta_r, \beta_s), S_w(\beta_r, \beta_s)) \sum_{LW} (\beta_r, \beta_s)^{-1} (S_L(\beta_r, \beta_s), S_w(\beta_r, \beta_s))'$$

Where $S_L(\beta_r, \beta_s)$ denotes log-rank score, $S_w(\beta_r, \beta_s)$ denotes Prentice-Wilcoxon score and $\sum_{LW}(\beta_r, \beta_s)$ their joint estimated covariance matrix applied to the data $\{R_i, T_i(\beta_r, \beta_s)\}, i = 1, \dots, n$. The value for the solution $Q_{LW}(\beta_r, \beta_s) = 0$ is the point estimate of (β_r^*, β_s^*) and the set that satisfies $Q_{LW}(\beta_r, \beta_s) < 6.0$ the 95th percentile of the χ_2^2 distribution are the joint confidence set.

In the presence of censoring a slightly different approach is taken.

C_i is the potential censoring time, which is the difference between i 's date of randomisation and the end of follow-up. We observe $R_i, C_i, X_i = \min(T_i, C_i)$ and $S_i^* = \min(S_i, X_i) = \min(S_i, T_i, C_i)$ instead of R_i, T_i, S_i .

Under the joint null hypothesis $(\beta_r^*, \beta_s^*) = (\beta_r, \beta_s), T_i(\beta_r, \beta_s)$ and C_i are jointly independent of R_i . But for censored patients we cannot observe $T_i(\beta_r, \beta_s)$. So, we define a new random variable $C_i(\beta_r, \beta_s)$:

$$C_i(\beta_r, \beta_s) = \min [\exp(\beta_r) \exp(\beta_s) C_i, \exp(\beta_r) C_i, \exp(\beta_s) C_i, C_i]$$

Which is independent of R_i under the null hypothesis. The authors state that because $C_i(\beta_r, \beta_s)$ is also a baseline variable any observable function of $\{T_i(\beta_r, \beta_s), C_i(\beta_r, \beta_s)\}$ is independent of R_i under the null hypothesis. Hence the authors define a new variable U_i :

$$U_i(\beta_r, \beta_s) = \min\{T_i(\beta_r, \beta_s), C_i(\beta_r, \beta_s)\}$$

Which is independent of R_i .

Let $\Delta_i(\beta_r, \beta_s) = 1$ if $T_i(\beta_r, \beta_s) < C_i(\beta_r, \beta_s)$ and 0 otherwise. Then log-rank and Prentice-Wilcoxon scores and their joint covariance matrix can be calculated using the data $\{R_i, U_i(\beta_r, \beta_s), \Delta_i(\beta_r, \beta_s)\}, i = 1, \dots, n$ to get the censored version of Q_{LW} . Under the null hypothesis Q_{LW} follows the distribution χ_2^2 .

To construct the confidence interval for β_r and β_s : For β_r we find a value of β_s that minimises $Q_{LW}(\beta_r, \beta_s)$ and refer to it as $\tilde{\beta}_s(\beta_r)$. If $\beta_r^* = \beta_r$ then $Q_{LW}(\beta_r, \tilde{\beta}_s(\beta_r))$ follows χ_2^2 distribution. Then the confidence interval for β_r^* is the set of β_r which satisfied $Q_{LW}(\beta_r, \tilde{\beta}_s(\beta_r)) < 3.84$. The confidence interval for β_s can be constructed in the same way.

Observational version:

The authors then state that by considering the second-line treatment as an observational study of the second-line treatment we can obtain a more precise estimate of β_r^* by first estimating β_s^* under the assumption of no unmeasured confounders, and by then taking into account the randomisation for first-line treatment. When β_s^* is estimated the randomisation is not taken into account. Based on the no unmeasured confounders assumption the decision as to whether to initiate second-line treatment is independent of T_{iR_i0} . Formalised, this is:

$$\lambda_{is}[t|R_i, \bar{L}_i(t), T_{iR_i0}] = \lambda_{is}[t|R_i, \bar{L}_i(t)]$$

λ_{is} is the patient i 's hazard of initiating second-line treatment at time t and $\bar{L}_i(t)$ is the vector of covariate history prior to time t including baseline variables. Any variables that may influence the initiation of second-line treatment should be included to satisfy the no unmeasured confounders assumption.

To estimate β_s^* the authors suggest modelling the above equation using a proportional hazards model with time-dependent covariates (a sufficient number of covariates, $\bar{L}_i^*(t)$, a subset of $\bar{L}_i(t)$ should be included so that the no unmeasured confounders assumption is approximately true). The model is:

$$\lambda_{is}[t|R_i = R, \bar{L}_i^*(t)] = \lambda_{0R}(t) \exp[\alpha_R^* \bar{L}_i^*(t)]$$

Where $\lambda_{0R}(t)$ is the arm-specific baseline hazard and α_R^* is the vector of arm-specific regression coefficients. That is, we fit this model separately for the two treatment arms, which is important because the reason for initiating second-line treatment might be quite different in the different arms.

The authors then state that in the absence of censoring they estimate β_s^* using the above model. Under the no unmeasured confounders assumption this model means that:

$$\lambda_{is}[t|R_i, \bar{L}_i^*(t), T_{iR_i0}] = \lambda_{is}[t|R_i, \bar{L}_i^*(t)]$$

And therefore if we include the term T_{iR_i0} (that is, $T_i(0, \beta_s^*)$) in the model, the coefficient of the term $\alpha^*(\beta_s^*)$ structurally equals 0. Hence the hypothesis that $\beta_s^* = \beta_s$ is equivalent to the hypothesis that $\alpha^*(\beta_s) = 0$. Based on this, a G-estimation procedure is used to estimate β_s^* . Counterfactual survival times are estimated for each value of the treatment effect within an SNM, and then a g-test is carried out using the above model of the hazard of treatment change. The procedure checks at each time t , for an association between initiation of second-line

treatment and $T_i(0, \beta_s)$ after adjusting time-dependent confounders and history of second-line treatment before t , but without adjusting for the covariates (confounders) and history of second-line treatment subsequent to t . They state that Robins (1992) has shown that this gives an asymptotically normal and unbiased estimator of β_s^* .

After this, the estimate of β_r^* is estimated using the 'observational' G-estimate of β_s^* . Under the null hypothesis that $\beta_r^* = 0$, T_{iR_i0} have the same distributions between treatment arms, so if we know the true value of β_s^* we can compute $T_i(0, \beta_s^*)$ and perform log-rank or Prentice-Wilcoxon tests to see if the distributions of T_{iR_i0} are similar. In fact, the authors acknowledge that we do not know β_s^* , but we have a consistent estimate $\hat{\beta}_s$. So we can substitute this for β_s^* . The authors' then state how β_r^* can be estimated. Under the hypothesis that $\beta_r^* = \beta_r$ we can apply log-rank or Prentice-Wilcoxon tests to the data $\{R_i, T_i(\beta_r, \hat{\beta}_s)\}$. They state that a 95% confidence interval consists of the values for which we fail to reject the hypothesis at the 5% level. The value for which the log-rank score of the Wald statistic equals 0 is the point estimate of β_r^* .

The authors then state that in the presence of censoring we cannot observe T_{iR_i0} that is, $T_i(0, \beta_s^*)$, for all patients. In this case, they state that the following can be used:

$$U_i(0, \beta_s^*) = \min\{C_i \exp(\beta_s^*), C_i, T_{iR_i0}\}$$

They state that under the assumption of no unmeasured confounders and because C_i is a component of $\bar{L}_i(t)$ the model in the presence of no censoring becomes:

$$\lambda_{is}[t|R_i, \bar{L}_i^*(t), U_i(0, \beta_s^*)] = \lambda_{is}[t|R_i, \bar{L}_i^*(t)]$$

The authors state that we can then use the same approach as in the absence of censoring to estimate β_s^* using $U_i(0, \beta_s^*)$ as a substitute for T_{iR_i0} . That is, β_s is set for a range of appropriate values and $U_i(0, \beta_s)$ is calculated, then the term for each of the values is added to the above model. A log-rank test or Wald test of the hypothesis that $\alpha^*(\beta_s)$, the regression coefficients of $U_i(0, \beta_s)$, equal 0, is completed at each repetition.

The authors then state that under the null hypothesis of $\beta_r^* = 0$, $U_i(0, \beta_s^*)$ and $\Delta_i(0, \beta_s^*)$ have the same distributions in the two treatment arms. Hence they state that we can use $\hat{\beta}_s$ and apply the log-rank or Prentice-Wilcoxon test to the data $\{R_i, U_i(0, \hat{\beta}_s), \Delta_i(0, \hat{\beta}_s)\}$ instead of $\{R_i, T_i(\beta_r, \hat{\beta}_s)\}$. They state that to estimate β_r^* the data $\{R_i, U_i(\beta_r, \hat{\beta}_s), \Delta_i(\beta_r, \hat{\beta}_s)\}$ can be used.

The authors state that using this technique, because we have used the estimated value for β_s^* the estimate of the variance of the log-rank or Prentice-Wilcoxon test statistic must be modified. The authors state that when they used a weighted log-rank test statistic with score $S(\beta_r, \hat{\beta}_s)$ they used the approximate variance with delta method:

$$\widehat{Var}[S(\beta_r, \hat{\beta}_s)] = \sum (\beta_r, \hat{\beta}_s) + \widehat{Var}_{corr}(\beta_r, \hat{\beta}_s)$$

Where $\sum(\beta_r, \hat{\beta}_s)$ is the estimated variance for a weighted log-rank test applied to the data $\{R_i, U_i(\beta_r, \hat{\beta}_s), \Delta_i(\beta_r, \hat{\beta}_s)\}$ and $\widehat{Var}(\beta_r, \hat{\beta}_s)$ represents the additional variability due to the estimation of β_s^* .

Marginal Structural Models

The authors state that unlike the usual time-dependent Cox model, a marginal structural Cox model can be used to obtain valid causal inference for the effect of time-varying treatment in the presence of time-dependent confounders which are themselves affected by previous treatment. The parameters of these models are estimated using IPTW.

The marginal structural Cox proportional hazards model considered is:

$$\lambda_{T_{\bar{a}}}(t|R, \bar{a}(t)) = \lambda_0(t) \exp[\beta_1 R + \beta_2 a(t)]$$

R is the treatment arm indicator

\bar{a} is the second-line treatment indicator

$T_{\bar{a}}$ is a counterfactual random variable reflecting a patient's time of death from randomisation if his/her history of second-line treatment had been \bar{a} , possibly contrary to what was observed. For those who receive \bar{a} until his/her death time $T_{\bar{a}} = T$.

Where $\lambda_{T_{\bar{a}}}(t|R, \bar{a}(t))$ is the hazard among patients with treatment arm indicator R in the source population, had all patients followed second-line treatment \bar{a} through time t and never been censored.

The authors state that the above model is the structural model for the marginal distribution of the counterfactual variable $T_{\bar{a}}$, and β_1 and β_2 are the causal log rate ratios for the treatment arm effect and the effect of second line treatment respectively.

$\exp(\beta_1)$ has a causal interpretation as a rate ratio of hazard at any time t , if possibly contrary to fact, all patients had been randomised to treatment A, compared to the hazard had all patients been randomised to treatment B.

$\exp(\beta_2)$ has a causal interpretation as a rate ratio of hazard at any time t , if possibly contrary to fact, all patients had been randomised to receive second-line treatment, compared to the

	<p>hazard had all patients been randomised not to receive second-line treatment</p> <p>β_1 and β_2 can be estimated using inverse probability of treatment and censoring weighted estimators. The Cox model is fitted with the contribution of patient i to a risk-set calculation at time t determined by the weight $W_i(t)W_i^*(t)$.</p> <p>$W_i(t)$ is the inverse of the probability of having patient i's observed history of second-line treatment. To estimate this, a Cox proportional hazards model is applied with the initiation of second-line treatment as an outcome.</p> <p>$W_i^*(t)$ is the inverse of the probability of patient i remaining uncensored. To estimate this, a Cox proportional hazards model is applied with the probability of remaining uncensored an outcome.</p> <p>The estimators based on these weights are known as inverse probability of treatment and censoring weighted estimators. The analysis using this weighting system creates what the authors call a 'pseudo-population' which consists of $W_i(t)W_i^*(t)$ copies for each patient i.</p> <p>The pseudo-population has two important properties: Firstly, time-dependent confounders do not predict the status of second-line treatment at time t given the history of second-line treatment. Secondly, the causal effect of second-line treatment in this pseudo population is the same as in the study population. The authors then state that because of this, in this population we can use the ordinary time-dependent Cox regression to obtain unbiased estimators .</p> <p>The authors state that in some cases there may be a few patients who have very high weights applied to them because some time-dependent prognostic factors are strongly associated with the initiation of second-line treatment, and thus these patients contribute a large amount to the pseudo population and dominate the weighting analysis. The authors state that this can lead to unstable parameter estimates with large variability, and to avoid this they describe how 'stabilized weights' can be estimated. The authors state that assuming no unmeasured confounders and no selection bias due to censoring, these stabilised weights can lead to consistent estimates of the causal parameters which are more efficient than the unstabilised weights. For stabilised weights to be estimated, the numerator of the non-stabilised weights $W_i(t)(W_i^*(t))$, that is 1, must be replaced by the probability that patient i had had his/her second-line history (here read censoring history for the censoring weight) through time t, conditional on his/her history of second-line treatment and baseline covariates but not adjusting for time-dependent prognostic factors. For these weights time-dependent prognostic factors are not included in the Cox models used to estimate the weights.</p> <p>Using this model the authors stated that individual weights induces within-patient correlation that must be taken into account in the calculation of the variance, for which the conservative robust variance estimator developed by Lin (1989) can be used. The authors state that this provides conservative confidence intervals for the causal parameters β_1 and β_2 because it does not account for the estimation of the weights.</p>
<p>What are the key assumptions of the method?</p>	<p>The simple structural nested failure time model assumes:</p> <ol style="list-style-type: none"> i) The effect of second-line treatment is multiplicative to the time $T_i - S_i$. It is assumed that once a patient starts second-line treatment they continue on it, and its effect is maintained, and it contracts or expands time to event by $\exp(\beta_s^*) (T_i - S_i)$. So $T_{iRi0} = \exp(\beta_s^*) (T_i - S_i)$ is patient i's death time if randomised to R_i and no second-line treatment is received. β_s^* quantifies the effect of second-line treatment. The authors state that if the time that second-line treatment is taken is known, the time patient i received the treatment can be modified in the model. ii) Once the effect of second-line treatment is removed the treatment arm effect is also multiplicative to T_{iRi0}. So $T_{i00} = \exp(\beta_r^*)T_{iRi0}$. The parameter β_r^* quantifies the direct effect of the treatment arm. iii) Also, censoring is only due to the administrative end of follow-up. iv) When the treatment effect of second-line treatment is estimated as though it was an observational study of this treatment, the no unmeasured confounders assumption is made. <p>Marginal structural Cox Proportional Hazards Model</p> <ol style="list-style-type: none"> i) No unmeasured confounders ii) No selection bias due to censoring iii) No patient stops second-line treatment once it has begun iv) The effect of second line treatment is maintained once it is initiated. v) Correct specification of models.
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>The authors consider the relative advantages and disadvantages of the SNMs and MSMs. They consider the more recognisable structure of the MSM, based upon the Cox model, to be an advantage. Linked to this, the hazard ratio measure that comes from the Cox model and the MSM is more understandable to non-statisticians.</p> <p>The authors consider that an important advantage of SNMs are that they can be used to measure the effect of dynamic treatment regimens. Where actual treatments depend upon a patient's clinical condition SNMs are useful, whereas MSMs are more suited to measure the effects of prespecified treatment regimens. For MSMs we have to make the assumption that</p>

	<p>no patient stops second-line treatment once it is initiated, and the effect of the second-line treatment is maintained once it is initiated, which may not be reasonable.</p> <p>The authors state that the most important disadvantage with the MSM, and also with the SNM using observational G-estimation, is the assumption of no unmeasured confounders, and the correct specification of models. There are many potentially important baseline variables and time-dependent confounders and model specification is difficult. The assumption of no unmeasured confounders is not testable using observed data.</p> <p>Additional points that the authors note are:</p> <ul style="list-style-type: none"> - MSMs cannot be used if there is any covariate level that makes a patient certain to receive treatment. - SNMs (apart from the SNM observational method) use only the randomisation to estimate parameters, and so do not suffer from the problem of needing to specify a model for initiation of the second-line treatment. The method is less efficient than SNMs with observational G-estimation, but model misspecification bias is avoided regarding the initiation of second-line treatment. Hence this is the method recommended by the authors.
What are the potential biases associated with the method?	MSMs in particular are reliant on several important assumptions, and require several models to be specified correctly. A lack of important data availability could cause bias.
Why might the method not be appropriate?	Again, a lack of important data availability could make the MSM method in particular (and the SNM method using observational G-estimation) inappropriate. Also, the methods developed in this paper are not specifically to deal with treatment crossover, rather they are to adjust for differential second-line treatments. However this type of method may be worthy of consideration if it is expected that the experimental treatment will have a different effect at second-line compared to first-line.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The SNM is similar to the RPSFTM method, but extended to deal with a different second-line treatment. The MSM method was originally developed for causal inference by Robins (1999) and Hernan <i>et al</i> (2001)
Does the method represent an extension to another method?	The methods extend previous SNMs and MSMs to estimate two treatment effect, although this is also done in other papers included in the review.
Application	
Is there a worked example in the survival setting?	In an accompanying paper, the authors test their methods using a non-small-cell lung cancer dataset, in which 45% of patients in the experimental arm and 61% of patients in the standard care arm received radiotherapy as second-line treatment.
Is the example relevant?	Although it is not treatment crossover, this method may be worthy of consideration if it is expected that an experimental treatment might have a different effect at second-line compared to at first-line.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>The authors compared 7 methods for estimating the treatment effects associated with the trial treatments (all results are for patients with stage III disease):</p> <ol style="list-style-type: none"> 1. Standard ITT: HR 1.24 (0.81-1.91) 2. Cox model using second-line treatment as a time-dependent covariate and a dummy variable for treatment arm effect: 1st line treatment HR 1.24 (0.80-1.91); 2nd line treatment HR 0.97 (0.62-1.52). However this method is prone to bias because there existed time-dependent covariates that were risk factors for death and also predicted second-line treatment. 3. Censor at initiation of second-line treatment: Results not presented, but non-significant. Biased because predictors of the initiation of second-line treatment were also predictors of survival. 4. SNM randomised based: AFT results 1st line treatment extends survival by -6.7% (-74% - 46%); 2nd line treatment extends survival by 63% (CIs not stated); If assume Weibull distribution this is equivalent to 1st line treatment HR 1.10 (CIs not stated); 2nd line treatment HR 0.50 (CIs not stated). 5. SNM observational. AFT results 1st line treatment extends survival by -9% (-39% - 42%); 2nd line treatment extends survival by 101% (2%-293%). Thus these estimates were more precise than the SNM randomised based analysis. 6. MSM with non-stabilised weights. Results not presented, but noted that these weights were very unstable. 7. MSM with stabilised weights. 1st line treatment HR 1.20 (0.64-2.28); 2nd line treatment HR 0.85 (0.44-1.62) <p>Thus, the authors noted that none of their methods allowed the active first-line treatment to appear significant, and none changed the point estimate associated with the treatment very much (although the estimated effect did move slightly in favour of the treatment). The authors infer that the impact of the differential administration of second-line treatment was not large enough to cause the ITT analysis to significantly underestimate the treatment effect associated with the first-line treatment.</p> <p>The authors also noted that they had incomplete information regarding prognostic variables once first-line treatment had been withdrawn, and data on second and later line treatments were variable. Hence the authors did not differentiate between second and third line treatments, and assumed that the treatment effect of these treatments was maintained until death once second-line treatment had been initiated. There were also problems with data on prognostic factors that influenced the initiation of second-line treatment, and where data was not available upon initiation last observed values were used. Hence the data-dependent methods such as MSM and SNM using observational G-estimation were not easy to implement.</p>

	The fully randomised RPSFTM approach did not successfully result in helpful and precise estimates of the two treatment effects – again showing (along with other reviewed papers) the problems associated with multiparameter RPSFTMs.
Other Issues	
Are there any other relevant characteristics associated with the method?	These methods are versions of the previously reviewed SNM, MSM and RPSFTM methods, applied in an RCT context. The methods themselves are not particularly novel, but as they are applied using variations and in an RCT context they are relevant for the review.

Reference	Huang X, Cormier JN, Pisters PWT. Estimation of the causal effects on survival of two-stage nonrandomized treatment sequences for recurrent disease. <i>Biometrics</i> 2006;62:901-909
Origin	
Was the method developed specifically in the survival analysis context?	The method was developed in the context of an observational study in which patients received (non-randomised) initial chemotherapy (some did not receive initial chemotherapy), and then upon disease recurrence there was an option for patients to receive salvage chemotherapy. Although the context is an observational study, the goal of the authors was to estimate the causal effects of both initial and salvage chemotherapy on OS, which might provide a useful technique that could be applied in an RCT context to correct for treatment crossover.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The aim of the authors is to estimate, for each treatment sequence, the overall potential mean lifetime and survival distribution that would have been observed if, contrary to fact, the entire cohort had received the same treatment sequence. They use a method whereby the potential outcome of a sequential treatment strategy is estimated by taking a weighted average of the observed outcomes. A patient who does not take the strategy is given a weight of 0, whereas a patient who follows the strategy receives a weight equal to the inverse of the probability of having the treatment pattern that is actually observed for him/her. This means that those patients who did not follow the treatment strategy in question are represented by those who did. To implement this IPTW method the probability of receiving a particular treatment is estimated by a logistic regression model. Patients with censored survival outcomes are represented by patients who are not censored using the IPCW method.</p> <p>The authors start by denoting four treatment strategies, where chemotherapy is denoted by A, and no chemotherapy is denoted by B: AA (chemo for initial treatment and salvage treatment) AB BA BB</p> <p>It is important to note that these are strategies, rather than treatment actually received. For example, if a patient has initial chemo and does not experience disease recurrence, their strategy may be AA or AB, and so data from that patient is used to evaluate both strategies AA and AB.</p> <p>The authors' note that in observational studies adjusting for covariates is important, as factors such as age could impact upon the treatment decision. It could be argued that this is also the case for non-randomised treatment crossover.</p> <p>Every patient has four potential survival times:</p> $T_{AA}, T_{AB}, T_{BA}, T_{BB}$ <p>Censoring time is C with $X = \min(T, C)$ Indicator function $I(\cdot)$ is $\Delta = I(T \leq C)$ and $Y(t) = I(X \geq t)$ R is the time to first recurrence, when recurrence doesn't occur $R = \infty$ H is a vector of summary variables of the prognostic information up to the time of recurrence. F_1 and F_2 denote initial and salvage treatments. Z_0 is a vector of baseline covariates excluding treatment indicators. For simplicity, the authors assume that C is independent of (T, Z_0, H) and the treatments received, but they show how this can be relaxed. The joint distribution of (Z_0, F_1, R, H, F_2, T) is assumed to be independent and identical among all n patients. $K(t) = P(C \geq t)$ and $L(t) = P(T \geq t)$ denote the Kaplan-Meier estimators $\hat{K}(t)$ and $\hat{L}(t)$ respectively.</p> <p>The authors note that the HR between treatment sequences is not common over time (consider AA versus AB), and so a standard Cox model is not appropriate. They choose to use the mean lifetime and survival distribution, restricted to a upper time limit τ. They state that τ should be chosen so that there is sufficient sample size for each treatment to obtain reliable estimates. For simplicity, they continue to use T for $T^{(\tau)} = \min(T, \tau)$ and μ for $\mu^{(\tau)} = E\{T^{(\tau)}\}$</p>

	<p>$P_1(\gamma_1; Z_1)$ and $P_2(\gamma_2; Z_2)$ are the probabilities of receiving therapy A in the initial and salvage treatments respectively, where Z_1 is a subset of Z_0 and Z_2 is a subset of $\{Z_0, F_1, R, H\}$. Z_1 and Z_2 are chosen by the data analyst using model selection techniques. The authors assume that there exists a small $\sigma > 0$ such that $\sigma < P_1(\gamma_1; Z_1) < 1 - \sigma$ and $\sigma < P_2(\gamma_2; Z_2) < 1 - \sigma$ for any bounded Z_1 and Z_2. Denote $\gamma = (\gamma_1', \gamma_2')$. Assume that P_1 and P_2 follow the logistic models $\text{logit}(P_i Z_i) = \gamma_i Z_i$, $i = 1, 2$. Note that these models imply that, conditional on information up to time t, the treatment decision at t is independent of the future potential outcome – this is the no unmeasured confounder assumption.</p> <p>For the sequence AA (other sequences have functions defined similarly) the authors define the weight functions: $W_1^{(AA)} = \frac{I(F_1=A)}{P_1(\gamma_1; Z_1)}$, $W_2^{(AA)} = \frac{I(F_2=A)}{P_2(\gamma_2; Z_2)}$ and denote $W_i^{(AA)} = W_{1i}^{(AA)} W_{2i}^{(AA)}$ for $i = 1, \dots, n$ If a patient receives initial treatment and has no recurrence the weight function $W_2^{(AA)} = W_2^{(AB)} = 1$ so his/her survival outcome is used in the estimation of the potential mean survival time for both AA and AB.</p> <p>$S_{AA}(t)$ is the potential survival function for treatment sequence AA. As the data is censored, the authors use the IPCW method to obtain an estimating equation for $S_{AA}(t)$:</p> $\sum_{i=1}^n \frac{\Delta_i}{K(X_i)} W_i^{(AA)} \{I(T_i > t) - S_{AA}(t)\} = 0$ <p>The authors state that replacing $W_i^{(AA)}$ and $K(X_i)$ with their estimators $\widehat{W}_i^{(AA)}$ and $\widehat{K}(X_i)$ gives the estimator:</p> $\widehat{S}_{AA}(t) = \left\{ \sum_{i=1}^n \frac{\Delta_i}{\widehat{K}(X_i)} \widehat{W}_i^{(AA)} \right\}^{-1} \sum_{i=1}^n \frac{\Delta_i}{\widehat{K}(X_i)} \widehat{W}_i^{(AA)} I(T_i > t)$ <p>The authors state that using the truncation point τ ensures that $\Delta/\widehat{K}(X)$ will not be infinity. The mean restricted lifetime under strategy AA, μ_{AA} can be estimated by:</p> $\sum_{i=1}^n \frac{\Delta_i}{\widehat{K}(X_i)} \widehat{W}_i^{(AA)} \{T_i - \mu_{AA}\} = 0$ <p>The authors state that variance formulas for $\widehat{\mu}_{AA}$ and $\widehat{S}_{AA}(t)$ can be obtained using Taylor expansions and life expectancies and survival probability distributions can be obtained for the other treatment sequences in the same way as for AA. The authors provide covariance matrices in their Appendix.</p> <p>For estimating covariate effects the authors suggest combining estimation equations, which can be done as long as it is assumed that covariates have the same effect on different treatment sequences. Assume that the life expectancy for patients with baseline covariate Z, if they follow strategy AA, is $\mu(\beta^{AA}, Z)$ where β^{AA} is an unknown parameter. Mean lifetime μ_{AA} can be obtained by integrating $\mu(\beta^{AA}, Z)$ over the distribution of Z. The authors state that a combined estimating equation for each treatment sequence can be obtained by including the four treatment strategy indicators and Z into \tilde{Z} and using a parameter $\tilde{\beta}$ to rewrite four separate estimating equations into one combined equation:</p> $U(\tilde{\beta}; \hat{\gamma}; \widehat{K}) = \sum_{S \in \{AA, AB, BA, BB\}} \sum_{i=1}^n \frac{\Delta_i}{\widehat{K}(X_i)} \widehat{W}_i^{(S)} \dot{\mu}^{(S)}(\tilde{\beta}, \tilde{Z}_i) \times \{T_i - \mu^{(S)}(\tilde{\beta}, \tilde{Z}_i)\}$ <p>The variance and covariance matrices for $\tilde{\beta}$ are given in the author's Appendix. The authors state that this combined estimating equation is more efficient than individual estimating equations for each treatment strategy. Interactions between covariates and treatment sequences can easily be incorporated.</p>
<p>What are the key assumptions of the method?</p>	<p>i) Stable unit treatment value assumption / consistency assumption. The outcome for a unit receiving treatment will be the same no matter what mechanism is used to assign the treatment and no matter what treatment other units receive.</p> <p>ii) No unmeasured confounders. Conditional upon information up to time point t, the treatment decision at time t is independent of the potential outcomes.</p> <p>iii) Correct model specifications.</p>
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>The method allows sequences of treatments to be considered. Treatment crossover may be perceived as a treatment sequence, whereby it is desirable to compare patients who receive the control followed by the new treatment, to patients who stay on the control, and to patients who just initially receive the new treatment. Isolating the effects of the different sequences in this way may allow a robust comparison of patients who were initially randomised to the treatment and patients who were randomised to the control and did not crossover. Considering</p>

	<p>sequences in this way does not constrain the second treatment (i.e. the crossover treatment) to having the same effect as the experimental treatment when given at first-line.</p> <p>However, the disadvantages of this method are that it is developed in an observational context, whereby the probability of receiving the treatments at first and second line is modelled based on probabilities associated with covariates. In RCTs, which are the focus of this thesis, we know who is going to receive the treatment at first-line based upon randomisation – we do not need to model this. Thus the model may not be appropriate in an RCT context. However – it may become relevant at the second-line context, where we do not know who is going to be crossed over from the control group to the experimental group – it is this that we are interested in. The IPCW approach takes into account the covariates of patients who crossover, and thus the method developed here may not add much to that approach in the context of RCTs.</p>
What are the potential biases associated with the method?	Potential biases are associated with the assumptions made (e.g. no unmeasured confounders). The approach is prone to bias based upon model misspecification.
Why might the method not be appropriate?	Because the method was developed in an observational trial context.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	It is most similar to the MSM/IPCW approach.
Does the method represent an extension to another method?	This is a standalone method, but could be said to extend the MSM/IPCW approach to deal with treatment sequences in the observational trial context. The key difference compared to the MSM method is that because sequences are compared the proportional hazards assumption is not appropriate, and so the authors use a restricted mean survival time estimate rather than a weighted Cox model in order to estimate the treatment effect. Given that in a treatment crossover context as defined in this thesis we are less interested in the effect of the whole sequence, and more interested in the true effect of the experimental treatment compared to the control, this is not a directly relevant extension.
Application	
Is there a worked example in the survival setting?	Yes, the author conducted a simulation study to assess the performance of their method with respect to estimating mean restricted lifetime and estimating the effects of a treatment sequence and covariates. In addition the authors applied their method to a real life observational trial of chemotherapy for soft tissue sarcoma.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>In the simulation study, the authors found that with regard to estimating mean restricted lifetime their method was very accurate with good coverage. However bias was introduced if the no unmeasured confounders assumption did not hold and this increased as the effect of Z, the vector of baseline covariates excluding treatment indicators, was increased. With regard to estimating the effects of a treatment sequence and covariates the method also gave answers very close to the truth, again with very good coverage.</p> <p>In the real-world data application the authors fitted their models, including several important prognostic factors to influence the probability of initially receiving chemotherapy, and added extra variables (such as time to recurrence) in their model estimating the probability of receiving salvage chemotherapy. They then used their method to estimate the covariate adjusted mean lifetime and the effects of covariates on the four treatment sequences. The effect of chemotherapy was not found to be significant, as the mean lifetime for strategy AA was not significantly different than strategy AB, BA or BB.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Included Studies: Citation Search

Reference	Shao J, Chang M, Chow SC. Statistical inference for cancer trials with treatment switching. <i>Statistics in Medicine</i> 2005;24:1783-1790
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	The method developed by the authors extends upon Branson and Whitehead's (B&W) method. Latent event times are used to model a patient's survival time that would have been observed if the patient had not switched treatment, conditional on the time of switching. The authors claim that the B&W method does not take into account the fact that a treatment switch is often (the authors claim perhaps always) based on prognosis and/or the investigator's assessment and medical knowledge. The authors claim that the survival time of a patient

who switched from control to treatment may be improved if switching is based on prognosis to optimally assign patients' treatments over time. The authors propose models that take this 'switching effect' into account. (Note – as remarked upon by White in his letter, the authors' reasoning is not always clear – it might be argued that the other methods implicitly do this anyway, and the IPE and RPSFTM methods allow for the fact that prognosis affects the probability of crossover).

The authors state that B&W's IPE algorithm is fully parametric (the survival time distribution is parametric, and a parametric model is imposed to relate the survival time distributions for two treatments). They state that they propose their latent event times model under this parametric setting. However, they state that besides the inclusion of the switching effect their model fitting procedure is a full parametric likelihood approach, which they state is more efficient and computationally simpler than the IPE method.

The authors also consider their method in a semi-parametric framework using a Cox model, which is parametric in modelling the relationship between two treatments, but otherwise is non-parametric. In this case instead of considering latent event times the authors consider latent hazard rates and incorporate the switching effect in latent hazard functions. They state that statistical inference can be carried out using Cox's regression with some additional covariates and parameters.

Parametric Approach and Latent Event Times

T_1, \dots, T_n are independent non-negative survival times

C_1, \dots, C_n are independent non-negative censoring times that are independent of survival times

Observations are $Y_i = \min(T_i, C_i)$

$\delta_i = 1$ if $T_i \leq C_i$

$\delta_i = 0$ if $T_i > C_i$

Assume treatment acts multiplicatively on survival time – thus an AFT model, magnitude of effect is $e^{-\beta}$, where β is unknown.

Assume the survival time distribution under control treatment has parametric form $F_\theta(t)$ where θ is an unknown parameter vector and $F_\theta(t)$ is a known distribution when θ is known.

$k_i = 1$ for the test treatment

$k_i = 0$ for the control treatment

This gives a survival time distribution of:

$$P(T_i \leq t) = F_\theta(e^{\beta k_i t}), t > 0 \quad [1]$$

If F_θ has a density f_θ , the density of T_i is $e^{\beta k_i t} f(e^{\beta k_i t}), t > 0$

Now, $S_i > 0$ denotes switching time. The authors allow switching either way. When there is no switch, the latent event time is the same as the observed time. For those who switch the latent event time \tilde{T}_i is an abstract quantity defined as survival time if the patient had not switched. B&W assumed the following model for a control patient who switched:

$$\tilde{T}_i \triangleq S_i + e^\beta (T_i - S_i)$$

Where \triangleq denotes equality in distribution. To allow switching in either direction, this can be altered to:

$$\tilde{T}_i \triangleq S_i + e^{\beta(1-2k_i)}(T_i - S_i)$$

Where k_i is the indicator of the original treatment.

The authors state that these models do not take into account the fact that treatment switching is often due to prognosis and judgement. A switching patient may do so because it is optimal for them, and it may result in a longer survival time than those patients who do not switch. They claim that ignoring this will bias results.

The authors model the effect of switching from control to treatment by $w_{0,\eta}(s) > 0$ and the effect of switching from treatment to control as $w_{1,\eta}(s) > 0$, and assume the following model, conditional on S_i

$$\tilde{T}_i \triangleq S_i + e^{\beta(1-2k_i)} w_{k_i,\eta}(S_i)(T_i - S_i) \quad [2]$$

η_i is an unknown parameter and $w_{k,\eta}(s)$ are known functions of the switching time s when η and k are given. The authors state that generally $w_{k,\eta}(s)$ will be nearly 1 when s is close to 0 (switching occurs very early).

Given [1] and [2], the survival times for switching patients, conditional on S_i are:

$$P(T_i \leq t) = P(\tilde{T}_i \leq S_i + e^{\beta(1-2k_i)} w_{k_i,\eta}(S_i)(T_i - S_i))$$

$$= F_{\theta}(e^{\beta k_i}[S_i + e^{\beta(1-2k_i)}w_{k_i,\eta}(S_i)(T_i - S_i)])$$

$$= F_{\theta}(e^{\beta k_i}S_i + e^{\beta(1-2k_i)}w_{k_i,\eta}(S_i)(T_i - S_i))$$

For $k_i = 0,1$

For patients who don't switch the distributions are $F_{\theta}(e^{\beta k_i}t), k_i = 0,1$

Now, assume F_{θ} has density f_{θ} , and for notation convenience define $S_i = \infty$ for patient i who never switches. Thus, the conditional likelihood function given S_i is:

$$L(\theta, \beta, \eta) = \prod_{i:S_i=\infty} [e^{\beta k_i} f_{\theta}(e^{\beta k_i} Y_i)]^{\delta_i} [1 - F_{\theta}(e^{\beta k_i} Y_i)]^{1-\delta_i}$$

$$\times \prod_{i:S_i=\infty} [e^{\beta(1-k_i)} w_{k_i,\eta}(S_i) f_{\theta}(e^{\beta k_i} S_i + e^{\beta(1-k_i)} w_{k_i,\eta}(S_i)(Y_i - S_i))]^{\delta_i}$$

$$\times [1 - F_{\theta}(e^{\beta k_i} S_i + e^{\beta(1-k_i)} w_{k_i,\eta}(S_i)(Y_i - S_i))]^{1-\delta_i}$$

Let $\gamma = (\theta, \beta, \eta)$. The authors state that the parameter vector γ can be estimated using the likelihood equation:

$$\frac{\partial \log L(\gamma)}{\partial \gamma} = 0 \quad [3]$$

They state that under usual regularity conditions on f_{θ} the estimated γ is asymptotically normal with mean vector γ and covariance matrix:

$$\left[E \frac{\partial^2 \log L(\gamma)}{\partial \gamma \partial \gamma'} \right]^{-1} \text{Var} \left[\frac{\partial \log L(\gamma)}{\partial \gamma} \right] \left[E \frac{\partial^2 \log L(\gamma)}{\partial \gamma \partial \gamma'} \right]^{-1} \quad [4]$$

Which the authors state can be estimated by substituting γ with its estimate. They state that statistical inference can be made based on this asymptotic result.

The authors state that they do not recommend using the IPE method for estimation. They state that if initial estimates of model parameters are obtained by solving the likelihood equation [3], then iteration does not increase the efficiency of the estimates and adds computational complexity. If initial estimates are not solutions of [3] then they are typically not efficient, and the estimates produced by IPE, if they converge, may not be as efficient as the solutions of [3].

Proportional Hazard model and latent hazard rate

The authors state that analysis based upon parametric models is not robust to model misspecification, and so the semi-parametric Cox proportional hazards model is useful. The authors set up the following model, where $F(t)$ is the survival time distribution and $f(t)$ is its density, and the hazard rate at time t is $\lambda(t) = f(t)/[1-F(t)]$:

Standard Cox model:

$$\lambda_{k_i}(t) = \lambda_0(t) e^{\beta k_i} \quad [5]$$

k_i is the treatment indicator

$\lambda_0(t)$ is left unspecified

The authors state that in a more general setting k_i can be replaced by a covariate vector, and β by a parameter vector.

If there is no treatment switch, β is obtained by maximising the partial likelihood function:

$$L(\beta) = \prod_i (e^{\beta k_i} / \sum_{j \in R_i} e^{\beta k_j})^{\delta_i}$$

Where R_i is the set of patients alive and under observation just before time T_i

When there is a treatment switch but the switching effect can be ignored (patients switch at random) [5] can be modified by replacing k_i by the time dependent covariate $k_i(t)$, where $0 \leq t < \infty$:

$$k_i(t) = \begin{cases} 1 - k_i, & t \geq S_i \\ k_i, & t < S_i \end{cases}$$

Where S_i is the switching time, and $S_i = \infty$ if the patient never switches. The authors state that this reduces to a special case of the proportional hazard model with time-dependent covariates.

However, to take into account the fact that the treatment switch may depend upon prognosis and/or the investigator's assessment the authors introduce the switching effect: $w_{k_i, \eta}(S_i)$ where η is an unknown parameter vector. The switching effect is then included in the proportional hazards model:

$$\lambda_{k_i}(t) = \lambda_0(t) e^{\beta k_i(t)} w_{k_i, \eta}(t, S_i) \quad [6]$$

Where,

$$w_{k_i, \eta}(t, S_i) = \begin{cases} w_{k_i, \eta}(S_i), & t \geq S_i \\ 1, & t < S_i \end{cases}$$

The authors state that this model can be referred to as a latent hazard rate model, since $\lambda_{k_i}(t)$ corresponds to a latent event time, and thus can be treated as a latent hazard rate.

Under the latent hazard rate model the partial likelihood is:

$$L(\beta, \eta) = \prod_i \left[\frac{e^{\beta k_i(T_i)} w_{k_i, \eta}(T_i, S_i)}{\sum_{j \in R_i} e^{\beta k_j(T_i)} w_{k_j, \eta}(T_i, S_i)} \right]^{\delta_i}$$

Estimators of β and η can be obtained by solving:

$$\frac{\partial \log L(\gamma)}{\partial \gamma} = 0$$

where $\gamma = (\beta, \eta)$. The authors state that under some regularity conditions these estimators are asymptotically normal with mean vector γ and covariance matrix given by [4], above. They state that statistical inference can be made based on this asymptotic result.

The authors state that if $\log w_{k, \eta}(s)$ is linear in η , eg $w_{k, \eta}(s) = e^{\eta k_i s + \eta k_i s^2}$, then model [6] is another special case of the standard proportional hazards model with time-dependent covariates, since the switching effect can be rewritten as:

$$w_{k, \eta}(t, S_i) = e^{\eta k_i S_i(t) + \eta k_i S_i^2(t)}$$

where:

$S_i(t) = \begin{cases} S_i, & t \geq S_i \\ 0, & t < S_i \end{cases}$ can be treated as another time-dependent covariate. Hence, model [6] is the proportional hazard model with time-dependent covariates $k_i(t)$, $S_i(t)$ and $S_i^2(t)$ in addition to the original time-dependent covariate k_i .

The authors state that the parameter vector is:

$$\gamma = (\beta, \eta_{0,0}, \eta_{0,1}, \eta_{1,0}, \eta_{1,1})$$

And it is estimated by solving:

$$\sum_i \delta_i \left(Z_{ii} - \frac{\sum_{j \in R_i} Z_{ij} e^{\gamma' Z_{ij}}}{\sum_{j \in R_i} e^{\gamma' Z_{ij}}} \right) = 0$$

Where:

$$Z_{ij} = (k_j(T_i), (1 - k_j)S_j(T_i), (1 - k_j)S_j^2(T_i), k_j S_j(T_i), k_j S_j^2(T_i))$$

The authors state that the resulting estimator, $\hat{\gamma}$ is asymptotically normal with mean γ and covariance matrix:

$$\hat{B}^{-1} \hat{A} \hat{B}^{-1} \text{ where:}$$

	$\hat{A} = \sum_i \delta_i \left(Z_{ii} - \frac{\sum_{j \in R_i} Z_{ij} e^{\hat{\gamma}' Z_{ij}}}{\sum_{j \in R_i} e^{\hat{\gamma}' Z_{ij}}} \right) \left(Z_{ii} - \frac{\sum_{j \in R_i} Z_{ij} e^{\hat{\gamma}' Z_{ij}}}{\sum_{j \in R_i} e^{\hat{\gamma}' Z_{ij}}} \right)'$ <p>And</p> $\hat{B} = \frac{\partial \log L(\gamma)}{\partial \gamma \partial \gamma'} \Big _{\gamma=\hat{\gamma}} = \sum_i \delta_i \left(\frac{\sum_{j \in R_i} Z_{ij} e^{\hat{\gamma}' Z_{ij}}}{\sum_{j \in R_i} e^{\hat{\gamma}' Z_{ij}}} \right) \left(\frac{\sum_{j \in R_i} Z_{ij} e^{\hat{\gamma}' Z_{ij}}}{\sum_{j \in R_i} e^{\hat{\gamma}' Z_{ij}}} \right)' - \sum_i \delta_i \frac{\sum_{j \in R_i} Z_{ij} Z_{ij}' e^{\hat{\gamma}' Z_{ij}}}{\sum_{j \in R_i} e^{\hat{\gamma}' Z_{ij}}}$ <p>The authors state that it is simple to extend this approach to the case where there are other time-independent and/or time-dependent covariates which can be added into model [6].</p> <p>In addition, the authors state that the latent event times approach and the latent hazard rate approach coincide when the survival distribution is exponential, ie $F(t) = 1 - e^{-t/\theta}$ with unknown parameter $\theta > 0$. For other types of distributions the methods are different.</p>
<p>What are the key assumptions of the method?</p>	<p><u>Parametric Approach and Latent Event Times</u></p> <ul style="list-style-type: none"> i) Censoring times are independent of survival times ii) Treatment acts multiplicatively on survival times – AFT model iii) Survival on control treatment has a parametric distribution which can be defined iv) Switching can occur either way and decisions are made on expert judgement/prognosis, which means that strategies are optimal and switching is beneficial <p><u>Proportional Hazard model and latent hazard rate</u></p> <ul style="list-style-type: none"> i) Assume proportional hazards ii) Correct model specification.
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>Theoretically, it seems beneficial to include ‘reasons’ for the treatment switch in the estimation of counterfactual survival times, but care needs to be taken with what this implies about the treatment effect – is it reasonable?</p> <p>The requirement of the proportional hazards assumption is a disadvantage with the semi-parametric approach. The reasoning of the authors is not always clear, and their methods have been severely criticised, suggesting that they are fundamentally flawed – see below (White letter).</p>
<p>What are the potential biases associated with the method?</p>	<p>The semi-parametric method relies upon the proportional hazards assumption, which is not discussed in any detail by the authors. The reasoning of the authors is not always clear, and their methods have been severely criticised, suggesting that they are fundamentally flawed – see below (White letter).</p>
<p>Why might the method not be appropriate?</p>	<p>It is debatable whether correcting for switching from the active treatment to the control treatment is required. It has been suggested by White that these methods are fundamentally flawed.</p>
<p>How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?</p>	<p>This method brings up the interesting issue of taking into account the fact that treatment switching is likely to be a strategic decision based upon prognosis and clinical expert opinion. The authors provide a method that they claim can take this into account, unlike (they claim) the B&W method that they reference (they do not reference any other crossover methods).</p> <p>An issue is that the authors also correct for switches from the active treatment to the control treatment, seemingly under the hypothesis that a treatment switch is in these patients best interests and thus even if the switch is to the control treatment this might inflate the treatment effect. It is debatable whether we want to adjust for this type of switching, as it may occur in the real-world (whereas switching from the control onto the treatment could not).</p> <p>Potentially the method could be applied, taking out the η terms for switching from the active treatment onto the control, which may make the method more efficient, and potentially more applicable to the treatment crossover problem as defined in this thesis.</p>
<p>Does the method represent an extension to another method?</p>	<p>The authors extend the B&W method, but then also develop a separate semi-parametric HR method.</p>
<p>Application</p>	
<p>Is there a worked example in the survival setting?</p>	<p>The authors carry out a simulation study where their semi-parametric Cox model method is compared to two others – one which ignores switching (exclude patients who switch, using standard Cox model) , and one which includes switching data but ignores the ‘switching effect’ (include all patients in a standard Cox model). The AFT version of their method is not tested in the simulation study.</p> <p>The authors simulate a trial with 300 patients in each arm, in which the survival time is generated by the exponential distribution with mean survival times 14.43 months and 21.65 months for the control and treatment arms. Censoring time is random and drawn from a uniform distribution on the interval 15-20 months, leading to a censoring % of 24.6% in the control group and 34.6% in the treatment group. Switch time is generated by an exponential distribution with mean 7.22 months for the control group and 10.82 months for the treatment group.</p>

	Switching effects, η , were chosen for each possible switch (control to treatment, treatment to control), and for if a switch is not chosen (stay on control, stay on treatment)
Is the example relevant?	Yes, although there is more switching from the treatment group to the control than vice-versa, which is perhaps not the type of switching that we wish to control for.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	The semi-parametric Cox method was seen to work well. Treatment effect β is estimated with less than 3% bias. β is better estimated than the η 's. The coverage probability of the asymptotic CI is close to the nominal level of 95%. The standard Cox model with all patients included results in very biased results, whereas the exclusion approach is quite close to the truth (though not as close as the authors method), but the standard deviation is larger, and so is less efficient. The authors estimate that their method results in an efficiency gain of about 15% under a switching rate of 67%, due to the lower SD. They state that this is smaller than they had hoped for, but that this is because the method involves estimating 4 additional η parameters. They suggest that even this 15% gain is equivalent to a 32% reduction in sample size.
Other Issues	
Are there any other relevant characteristics associated with the method?	<p>White (2006) in his response to Shao <i>et al</i>'s paper, states that it is confusing to suggest that the IPE (and RPSFTM) method does not allow for switching to be associated with prognosis, because in fact the RPSFTM and IPE methods make no assumptions around prognosis because instead their estimation is based upon the randomisation of the trial. It seems that the point Shao <i>et al</i> (2005) are attempting to address is that the treatment switching decision is specific to individual patient-level characteristics and prognosis, and that the effect of the treatment once switching has occurred will depend upon these variables. Hence, it seems that the true objective of Shao <i>et al</i>'s method is to allow treatment effect to vary between patients depending upon certain prognostic characteristics – in particular whether they were initially randomised to the treatment or whether they were switched on to the treatment after being randomised to the control group.</p> <p>However, for both methods, once the models have been fully specified the treatment effects are estimated using conditional likelihood functions, where the results are conditional on switching time. Thus the resultant treatment effect estimates are conditional upon switching time. White (2006) acknowledges this to be a serious problem with the methods, which leave them open to substantial selection bias. He states that the conditional likelihood approach is only valid if the treatment crossover observed is random, and thus ignorable and independent of prognosis. This is the same problem associated with the naive method of censoring crossover patients from a standard ITT analysis, which is very likely to lead to selection bias because typically crossover patients have a different prognosis (either better or worse) compared to non-crossover patients. Hence, the Shao <i>et al</i> methods are very likely to be biased.</p> <p>White also comments on the simulation study performed, and comments on the fact that Shao <i>et al</i> (2005) found that their methods performed well, with little bias. White points out that this was likely to be because in their simulation study they generated switching times independent of prognosis – making switching random and ignorable.</p>

Reference	Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. <i>Statistics in Medicine</i> 2006;25:3503-3517
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The method involves regarding censored observations as missing data. Other information collected about subjects may be associated with event times and can be treated as auxiliary variables that can be used to help recover some of the information lost due to censoring. Incorporating information from auxiliary variables can help reduce bias from dependent censoring, and can improve efficiency.</p> <p>The aim of the authors is to estimate the marginal survival distribution using longitudinal data, but without explicitly estimating the stochastic process for the longitudinal variable – which is what the IPCW method also does. Basically, this means developing a model for the censoring mechanism and using the results of this model to reweight the observations in an estimating equation. Under defined conditions consistent estimators are obtained even in the presence of dependent censoring. The authors state that for the IPCW approach the initial emphasis was on bias correction in the face of dependent censoring. In their approach, the emphasis is on using the information available in the data on the association between the auxiliary variables and the failure time to improve the efficiency of the estimate, while at the same time also attempting to minimise the impact of possible dependent censoring. Hence, they state that their method places more emphasis on modelling the failure time distribution than on the censoring time distribution.</p> <p>The method proposed by the authors involves using auxiliary variables to define a nearest neighbourhood of similar observations for each censored case, and then generate imputes from this set of neighbours.</p> <ol style="list-style-type: none"> 1. First, risk scores must be calculated: <p>t_1, \dots, t_n denote observed times for n subjects under study. $\delta_i = 1$ if t_i is an event time, and $=0$ if it is a censored time.</p>

$Z = (z_1, \dots, z_n)$ denotes the values of the auxiliary variables, and
 $Y = \{(t_1, \delta_1, z_1), \dots, (t_n, \delta_n, z_n)\}$

For each censored observation an imputing risk set is sought consisting of subjects who are similar to the censored case.

To define each imputing risk set, first the auxiliary variables are reduced to a scalar index (risk score), which provides an indicator of an individual's risk of disease or death (this is analogous to mean matching). The authors state that a time-independent proportional hazards regression model can be used to derive the risk scores, which summarises the relationship between the auxiliary variables and the failure time. The authors state that when this model is correctly specified the risk scores define an imputing risk set that can be used to improve efficiency when censoring is independent, and to reduce bias when censoring is dependent upon the auxiliary variables.

Bias can remain even if censoring is dependent on the auxiliary variables, if the model is misspecified. For this reason the authors also use a second proportional hazards model that calculates risk scores by summarising the association between the auxiliary variables and the censoring time (analogous to propensity score matching).

By combining risk scores based upon survival and censoring distributions the authors sought to study to what extent a double robustness property for model-misspecification could be established. They state that intuitively, if one of the models is correct, conditional on these two risk scores, event times are independent of censoring times. Hence, they state that within an imputing risk set that is defined using two risk scores, the event times are independent of censoring times.

Both of the above models use auxiliary variables as covariates, and so each risk score is a linear combination of Z . The authors propose that when auxiliary variables are time-independent a subset (e.g. baseline and latest observed) of the measurements of the variables up to the censored time could be used. Once covariates are chosen risk scores are $\widehat{RS}_f = \widehat{\beta}_f Z$ and $\widehat{RS}_c = \widehat{\beta}_c Z$ where $\widehat{\beta}_f$ denotes the estimates of the parameters of the failure time model, and $\widehat{\beta}_c$ denotes the estimates of the parameters for the censoring model. For time-independent auxiliary variables these models only need to be fitted once. However, for time-dependent auxiliary variables we need to fit these models for every censored observation to the data of those at risk at the censoring time using the currently available auxiliary variables as fixed covariates. Each risk score is then centred and scaled and denoted as: $RS_f^* = \{\widehat{\beta}_f Z - \text{mean}(\widehat{\beta}_f Z) / SD(\widehat{\beta}_f Z)\}$ and $RS_c^* = \{\widehat{\beta}_c Z - \text{mean}(\widehat{\beta}_c Z) / SD(\widehat{\beta}_c Z)\}$ respectively.

2. The Imputing risk set is then defined.

The authors state that the distance between subjects j and k is defined as:

$$d(j, k) = \sqrt{w_f \{RS_f^*(j) - RS_f^*(k)\}^2 + w_c \{RS_c^*(j) - RS_c^*(k)\}^2}$$

Where w_c and w_f are non-negative weights that sum to 1. For each censored subject j this distance is then employed to define a set of nearest neighbours. This neighbourhood, $R(j^+, NN)$ consists of NN subjects who have longer survival time than the censoring time of j , and a small distance from the censored subject j . For example the neighbourhood may consist of the 10 subjects with the 10 nearest distances from subject j , who have a longer survival time than the censoring time of subject j . If the number of individuals still at risk is less than NN , then they are all included in the imputing risk set. The authors investigated the effect of the size of the neighbourhood in a simulation study.

3. Imputation schemes.

The authors then introduce non-parametric imputation schemes that can be used to impute values once the imputing risk set $R(j^+, NN)$ has been defined. They state that each procedure can be repeated a number of times to obtain multiple imputed datasets for use in estimation. The final estimate is the average of all the repeats, and the final variance is the sum of a between-imputation and a within-imputation component. The authors state that in their case, the estimate from each imputed data set is a KM estimate, and the within-imputation variance is based on Greenwood's formula. The authors state that under certain settings the variance estimator may overestimate uncertainty and may not be consistent, but that they will monitor this through comparing estimates of SE and empirical SD. The imputation schemes considered by the authors are:

- Risk set imputation (RSI) (analogous to hot-deck imputation)
- Kaplan-Meier imputation
- Bootstrap imputation procedure, which can be used to supplement the above two methods to incorporate more fully the uncertainty in the imputes.

The authors focussed on the Kaplan-Meier imputation (KMI) technique.

They show that with a large number of imputes, the KMI survival estimates will on average reproduce the weighted Kaplan-Meier (WKM) survival estimates over the range that a weighted Kaplan-Meier (WKM) estimate is defined (it is defined up until the time at which the longest time is censored among patients with Z of interest). The authors state that the WKM estimator has been shown to be consistent, if the event time and the censoring time are independent conditional on Z . On the other hand, the RSI method will not reproduce WKM estimates, and

	<p>tends to impute censored times more often than the KMI approach. However, the authors also state that in complex situations where there are multiple categorical covariates, multiple continuous covariates, or time-dependent covariates, the WKM may not be defined and when it is the KMI method will not necessarily reproduce the WKM estimate.</p> <p>The authors give a proof that when using the KMI method, if one of the two working models is correct, T and C are independent conditional on the two risk scores. Following on from this the authors state the proof that if one of the two models is correctly specified, the KMI method for estimating the distribution of T will have small bias in large samples for values of t prior to the first censored value in the imputed data sets. Thus the KMI method has a large sample double robustness property. However, in small samples the nearest neighbourhood approach could lead to bias even if the models are correctly specified (and especially when they are misspecified – particularly the failure-time model).</p>
What are the key assumptions of the method?	All important covariates need to be identified, and data must be available for them. Models must be correctly specified.
What are the theoretical advantages and disadvantages associated with the method?	<p>If models are miss-specified and censoring is dependent on the auxiliary variables bias can remain.</p> <p>The approach has low reliance on statistical models, as they are only used to identify a nearest neighbourhood.</p>
What are the potential biases associated with the method?	<p>The reason that some bias remains with this method is largely due to sample size. The nearest neighbourhood may contain some observations that are not close enough to the target value, so the remnants of dependent censoring remain within the neighbourhood. The method will be harder to implement when there are a large number of covariates, as nearest neighbourhoods are more difficult to define.</p> <p>Also, despite the double robustness property the method is less good when either (particularly the failure time) model is miss-specified.</p>
Why might the method not be appropriate?	<p>Data availability is very important, and data on all important prognostic covariates may not be available. Sample size is important and the success of the method depends upon the 'nearness' of the neighbourhood, which will diminish at the tail of the survival distribution. In an RCT context where patient numbers are usually much smaller than in observational datasets, this could be a serious problem.</p> <p>The authors suggest that their method could potentially be improved by weighting different observations within the nearest neighbourhood based upon nearness, or by using a different number for NN depending upon the time of the censored observation.</p>
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The method is similar to the IPCW method, as it uses covariate information essentially in an attempt to control for dependent censoring.
Does the method represent an extension to another method?	No, but the method is similar in basis to the IPCW method.
Application	
Is there a worked example in the survival setting?	<p>The authors apply their method to clinical data from an AIDS trial to test their method (CD4 counts are used as covariates). They include the IPCW in their study for comparison. The % of patients that were censored were quite similar in both arms of the trial, as were the % subject to administrative censoring (95% vs 90%). Hence in this case the authors expect little bias due to dependent censoring, and instead hope to see some gain in efficiency from using auxiliary variables.</p> <p>In addition the authors complete simulation studies comparing their multiple imputation method to the IPCW method. They consider binary and multiple time-independent auxiliary variables, and time-dependent auxiliary variables. They investigated the effects of the censoring mechanism, model misspecification for calculating risk scores, and the weights and the size of the nearest neighbourhood (NN) on survival estimates.</p>
Is the example relevant?	<p>The real-world example is not directly relevant for a switching analysis, and is focussed more on efficiency than on adjusting for bias caused by dependent censoring.</p> <p>The simulation study is not specific to treatment crossover, but it is relevant for considering the performance of the method.</p>
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>The KMI approach supplemented with bootstrapping delivered very similar results to the partially observed results, but SE estimates are reduced, indicating increased efficiency. The authors state that the approach recovers about 70% of lost information for the control group, and about 33% for the intervention group. The results are similar for the IPCW approach, but the SEs are inbetween the KMI and partially observed (standard) results.</p> <p>Based upon the simulation studies for time-independent covariates it seems that if both models within the KMI approach are correctly specified the approach (with bootstrapping) gives estimates of the treatment effect that are similar in bias levels to the IPCW approach, and SDs are a little smaller. When either model is incorrect, there is more bias in the KMI approach than in the IPCW approach, particularly when the failure time model is incorrect (although SDs remain smaller, and the difference in the mean estimates of the different methods are small). In all cases the KMI method is better than the partially observed (standard) method. The method improves slowly as sample size increases (but some bias still exists even when sample size is 2000).</p>

	<p>The best choice of NN appears to be 10. Changing the weights associated with the failure time and censoring time models is beneficial when the weight associated with the model with the lower weight is miss-specified.</p> <p>For time-dependent covariates the KMI approach with bootstrapping appears to perform better than the IPCW approach in the presence of dependent censoring – with lower bias and similar SD (though some bias remains with both methods). The authors suggest that this may be because in the IPCW approach the estimator depends crucially upon the coefficients fit to the model of the censoring data, and the authors found that in their simulation these coefficients were attenuated towards zero.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	<p>Robins and Finkelstein (2000) note that a key problem with an AV approach is that when there are a range of auxiliary variables conditional modelling is required for the estimated event time and the process by which this is affected by the auxiliary variables. If the models used to capture these relationships are miss-specified the resulting treatment effect estimates can be biased and, for example, inconsistent estimates of the survival curve can be produced even when censoring is independent and an unweighted Kaplan-Meier would have been consistent. Robins and Finkelstein (2000) state that this problem can be avoided by creating a pseudo population using the IPCW approach rather than an AV approach. The IPCW approach requires fewer modelling assumptions than the AV approach. Thus, it appears reasonable to conclude that the IPCW approach supersedes AV approaches.</p>

Reference	Hsu C-H, Taylor JMG. A robust weighted Kaplan-Meier approach for data with dependent censoring using linear combinations of prognostic covariates. <i>Statistics in Medicine</i> 2010;29:2215-2223
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The method is an extension of their 2006 approach, included above. They stated that although the multiple imputation method has advantages, the adequacy of the imputation will depend upon the availability of ‘donor’ observations, which may be few in number towards the end of the survival distribution. Therefore their 2006 method could result in large variation in the tail of the distribution. Thus, the authors suggest an alternative approach that does not rely on multiple imputation. Instead the two risk scores are derived as above (one for the association between the covariates and event times, and one for the association between covariates and censoring times – to account for dependent censoring), and then these risk scores are categorised into groups to define homogenous risk groups. The Weighted Kaplan-Meier approach is then applied to re-estimate survival times for censored observations. The objective of the approach is to allow the WKM method to be able to deal with multiple prognostic variables and dependent censoring, without the data availability problems associated with the multiple imputation method.</p> <p>Hsu and Taylor’s 2010 approach is similar to their 2006 approach insofar as proportional hazards models are fitted to observed failure times and observed censoring times respectively, and then the risk score is calculated for each of these models. Their approach is also similar in that they then propose that these scores should be standardised by subtracting the mean and dividing by the standard deviation. However, after this point the methods diverge, and Hsu and Taylor’s 2010 approach does not use multiple imputation to impute values for the censored observations. Instead they use the risk scores to apply the WKM approach. They state that the two risk scores could be categorised separately and then the categorised risk scores could be jointly used to define risk groups. However, because the risk scores could be highly correlated and because separate categorisation could lead to groups with small numbers of observations, they suggest that principal component analysis should be performed on the two standardised risk scores to generate two orthogonal components, which can then be categorised to define the risk groups. WKM can then be performed on the defined risk groups. The orthogonal components described are linear combinations of the two risk scores. The authors state that these two components can be categorised separately based upon their percentiles into $I * J$ groups, where I is the number of categories in the first component, and J is the number of categories in the second component. The WKM estimator can then be derived for these categorised groups.</p>
What are the key assumptions of the method?	Similar assumptions as their 2006 approach, but without the multiple imputation. Risk scores must be modelled accurately, but the double-robustness property holds. Sample size and covariate groups must be adequate to allow sufficient population of risk groups. All prognostic covariates must be included.
What are the theoretical advantages and disadvantages associated with the method?	Compared to the standard WKM methods developed by Murray and Tsiatis and others, Hsu and Taylor’s 2010 approach has the advantage of being suitable for extending the WKM to a situation in which there are multiple prognostic covariates and dependent censoring, and the prognostic covariates can be categorical or continuous. Compared to the Hsu <i>et al</i> 2006 approach the method is less reliant on data availability and it similarly has less reliance on statistical models than earlier approaches as models are only used to identify risk groups. The method also ensures that censoring is independent conditional on the two orthogonal components, if at least one of the proportional hazards models is specified correctly (in a similar way as censoring being independent conditional upon the risk scores in the 2006 approach, given at least one of the proportional hazards was specified correctly). For both the 2006 and the 2010 methods this is assured if risk scores are categorical. If they are continuous a small bias could remain, as censoring in each categorised group will be close to, but not exactly, independent. In their 2010 paper, Hsu and Taylor state that their method, which applies the WKM approach using risk scores from two proportional hazards models to define risk groups can produce reasonable survival estimates even when the true failure time and censoring time models are AFT models, because the risk scores from the PH models are rescaled, and the definition of the risk groups is dependent on the relative magnitude of regression coefficients from the risk scores, not their absolute values. The authors state that these relative

	magnitudes are robust to the link misspecification, because it has been shown that the estimates of the regression coefficients of a PH model are consistent in that they are proportional to the regression coefficient of an AFT model when the true model is an AFT model. This suggests that the 2010 method is reasonable even if the risk score PH models are misspecified – although it seems likely that some bias may remain if this was the case.
What are the potential biases associated with the method?	If models are miss-specified, link functions are miss-specified (even despite the double-robustness property), sample sizes are too small, and too many covariates exist, bias may remain.
Why might the method not be appropriate?	For the above reasons. This may be a particular problem in RCTs, where patient numbers are often relatively small.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The method is an extension of the authors 2006 method, and also of the AV approach developed by Murray and Tsiatis (1996) and others. It is quite similar in concept to the IPCW/MSM method.
Does the method represent an extension to another method?	Yes, as above.
Application	
Is there a worked example in the survival setting?	Yes – real-world data and simulation studies.
Is the example relevant?	Yes, although they are not about treatment crossover – rather they are about estimating survival in the presence of dependent censoring.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	Hsu and Taylor (2010) applied their approach to a real-world prostate cancer data set in which dependent censoring occurred, and also conducted simulation studies. They compared their approach to their 2006 method, IPCW and a partially observed analysis. In the real-world analysis they found that both their methods and the IPCW approach led to higher survival estimates and generally lower standard errors over time than the partially observed analysis. The authors' 2006 approach led to higher standard errors in the tail of the distribution than all other methods (probably due to data scarcity). Towards the tail of the distribution the IPCW approach led to slightly lower survival estimates and higher standard errors than the authors' 2010 method. In their simulation studies, the authors found that the IPCW approach estimated the true effect with very little bias when all prognostic covariates were included in the analysis. When all prognostic covariates were not included (for example, when 3 of 5 covariates were included) the method was less successful, producing estimates approximately half way between the unadjusted partially observed results and the true results. The authors' 2010 approach produces similar mean squared error results as the IPCW approach whether one or both of the proportional hazards models are specified correctly, although their estimate of the true effect is less accurate than the IPCW that includes all confounders. When the proportional hazards models are miss-specified the authors assumed that only 3 of 5 covariates were included, and in these circumstances the results of the method were more accurate than the IPCW that only included 3 of 5 confounders. The authors' 2006 approach led to quite similar results as their 2010 approach, although estimates were generally slightly more biased. The authors also considered scenarios when the true link functions were AFT models rather than PH models, and found that all methods were prone to bias. The IPCW approach seemed to underestimate the true value, whereas the WKM and multiple imputation approaches overestimated the true value to a similar extent. The authors 2010 approach improved slightly as sample size increase, which was not the case for the other methods. The IPCW approach had poorer coverage than the other methods.
Other Issues	
Are there any other relevant characteristics associated with the method?	<p>Hsu and Taylor (2010) conclude that the IPCW approach can produce reasonable survival estimates when the true censoring time model is a PH model, and all prognostic covariates are included. On the other hand their 2010 WKM approach can provide reasonable survival estimates and is robust to misspecification of either one of the risk score PH models, and is robust to misspecification of the link functions, and was generally less biased than their 2006 approach. This interpretation seems slightly biased in favour of the authors' methods, since their results demonstrate that an IPCW approach that includes all relevant confounders and that has a correctly specified link function produces estimates of the true value that are closer than those obtained by a correctly specified WKM approach. There appears to generally be more bias in the WKM approach than the correctly specified IPCW approach and there is in fact bias with the WKM approach when the link function is miss-specified (as there is also with the IPCW approach).</p> <p>The authors suggest that the double robustness property of their approach means that their approach is likely to be free from bias when link functions are miss-specified, whereas the IPCW approach does not have this property and will be biased when the censoring time model is miss-specified. However, although the IPCW approach was biased in simulations testing such a scenario, and had particularly poor coverage, the WKM approaches also exhibited bias in these scenarios. The authors suggest that this is likely to be due to sample size, as with small numbers risk groups will contain some individuals whose risks are not particularly similar to the rest of the group. They state that this will be a larger problem as the number of covariates, and the range within covariates, increase. The authors also accept in their discussion that the double robustness property should not be relied upon as a substitute for correctly specifying models.</p> <p>Robins and Finkelstein (2000) note that a key problem with an AV approach is that when there are a range of auxiliary variables conditional modelling is required for the estimated event time and the process by which this is affected by the auxiliary variables. If the models used to capture these relationships are miss-specified the resulting treatment effect estimates can be biased and, for example, inconsistent estimates of the survival curve can be produced even when censoring is independent and an unweighted Kaplan-Meier would have been consistent. Robins and Finkelstein (2000) state that this problem can be avoided by creating a pseudo population using the IPCW approach rather than an AV approach. The IPCW approach requires fewer modelling assumptions than the AV approach. Thus, it appears reasonable to conclude that the IPCW approach supersedes AV approaches.</p>

Reference	Lee MLT, Chang M, Whitmore GA. A threshold regression mixture model for assessing treatment efficacy in a multiple myeloma clinical trial. <i>Journal of Biopharmaceutical Statistics</i> 2008;6:1136-1149
Origin	
Was the method developed specifically in the survival analysis context?	Yes. The paper analyses data from a multiple myeloma trial in which treatment switching occurred based upon patient response. Patients in the control arm generally switched onto the novel treatment (Velcade) after disease progression, whereas patients in the Velcade arm could receive other treatments after progression, which were assumed to be the same as the initial control treatment.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The authors use threshold regression methods and first-hitting-time survival models. The authors propose a model that differentiates the rate of disease progression before and after treatment switching, which they say can be used to model clinical trials with this crossover design.</p> <p>The authors state that they use a flexible mixture of threshold regression models – specifically a mixture of two first-hitting time (FHT) distributions for an underlying Wiener process representing patient health status. The mixture model includes a composite time scale that differentiates between the rate of disease progression during the interval from randomisation to switching and the interval from switching to death. The authors claim that the model realistically captures the nonlinear complex nature of the survival process.</p> <p>FHT A FHT is where the time-to-event data is interpreted as the elapsed time from initial observation until the sample path of a parent stochastic process first encounters a boundary set B. For example, a cancer progresses towards a critical state that proves fatal – the elapsed time from initial diagnosis to the critical state is an FHT. The authors state that many different mathematical forms may be used for the parent stochastic process that describes the progress of the disease, and for the boundary set that triggers the hitting time. They state that the flexibility in choosing the parent process and boundary set give FHT models great scope for describing real medical and health situations.</p> <p>The authors state that often the underlying parent stochastic process and boundary set cannot be observed, but time-to-event and covariate data do provide some insight into the nature of these latent or unobservable characteristics of an FHT model. They state that the latency of the postulated health status process may seem to be a vague element in the model, but that clinicians are well able to assess the health status of their patients. The latent health status process in the model can be considered as a one-dimensional index of how the patient's status changes through time under treatment. The authors state that linear combinations of observable covariate processes can emulate a latent patient health status process and allow it to be monitored, and they use baseline beta-2 microglobulin levels for this in their case study.</p> <p>The authors use a Wiener diffusion process to describe the health status of a subject with multiple myeloma, which allows a meandering path to be taken as health status gradually declines over time. The Wiener process has mean μ and variance σ^2. The authors set B as 0 so that death is triggered when a patient's health status first reaches 0. It is assumed that the initial level of the process at time zero is positive. The FHT is denoted by S and it is assumed that S has an inverse Gaussian distribution.</p> <p>The authors state that there is no guarantee that a parent stochastic process will reach a boundary set, and the possibility that it does not can be described as the cure rate of the FHT, denoted by $P(\text{cure})$. In a Wiener process this probability is positive if μ is positive. The probability is given by:</p> $P(\text{cure}) = 1 - \exp\left(-\frac{2x_0\mu}{\sigma^2}\right)$ <p>The authors state that because their study includes a latent health status process, the health status scale has an arbitrary unit of measurement and therefore there is one redundant parameter. They state that they choose to set the variance parameter of the Wiener diffusion process σ^2 to 1 (unity), the implication of which is that both the initial health state level x_0 and the mean parameter μ of the health status process will be measured in units of the standard deviation of the process.</p> <p>TR Threshold regression (TR) models are parametric models that do not generally possess the proportional hazards property, although variants exist that can incorporate this if deemed appropriate. TR models require an explicit formulation of the underlying health process and the triggers that define event times.</p> <p>The authors state that FHTs usually include regression structures in order to capture the reality of practical applications, and to account for variability in the data and to sharpen statistical inferences. An FHT model with a regression structure is referred to as a TR model, where threshold refers to the fact that the FHT occurs when the underlying process reaches a threshold state within a boundary set.</p>

The authors state that in TR, parameters of underlying processes, boundary sets and time scales are connected to linear combinations of covariates using suitable regression link functions, such as:

$$g_{\ell}(\ell_i) = z_i\beta$$

Where g_{ℓ} is the link function; ℓ_i is the value of parameter ℓ for individual i ; $z_i = (1, z_{i1}, \dots, z_{ik})$ is the covariate vector of individual i (with a leading unit to include an intercept term); and β is the associated vector of regression coefficients. The authors state that the mathematical form of the link function must be suited to the application, and generally it will be chosen to map the parameter space into the real line. In their case study the initial health status parameter x_0 was linked to a linear combination of covariates using the a natural logarithmic link function $\ln(x_0)$, whereas the mean parameter μ was simply given an identity link. The authors note that as with conventional regression analysis, TR requires judicious choices of the link function for each parameter, the list of covariates entering the regression function, and the exact mathematical forms of the covariates in the regression function.

Note that the authors transform calendar times to a composite time scale with form $r = \alpha t_1 + t_2$ where t_1 and t_2 correspond to duration to PD and duration post PD respectively. α is the ratio of the rate of progression on the primary therapy relative to that on the alternate therapy. r denotes the composite time scale.

Distributions

The authors state that the FHT distribution for a boundary by a Wiener diffusion model follows an inverse Gaussian survival distribution. This distribution depends on the initial health status level x_0 and the mean and variance parameters (μ and σ^2) of the underlying Wiener process.

The authors state that the p.d.f. for the FHT (defined in the composite time scale r , rather than calendar time) is:

$$f(r | \mu, \sigma^2, x_0) = \frac{x_0}{\sqrt{2\pi\sigma^2r^3}} \exp\left[-\frac{(x_0 + \mu r)^2}{2\sigma^2r}\right], \text{ for } -\infty < \mu < \infty, \sigma^2 > 0, x_0 > 0$$

If $\mu > 0$ the FHT is not certain to occur and the p.d.f. is improper – the probability that the boundary will not be reached is given by the cure rate above.

The c.d.f. is:

$$F(r | \mu, \sigma^2, x_0) = \Phi\left[-\frac{(\mu r + x_0)}{\sqrt{\sigma^2r}}\right] + \exp\left(-\frac{2x_0\mu}{\sigma^2}\right) \Phi\left[\frac{\mu r - x_0}{\sqrt{\sigma^2r}}\right]$$

Where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. The corresponding survival function is given by:

$$\bar{F}(r | \mu, \sigma^2, x_0) = 1 - F(r | \mu, \sigma^2, x_0)$$

The authors state that the p.d.f. and the c.d.f. show that μ , x_0 and σ^2 are not mutually estimable from censored survival data, and thus one of these parameters must be fixed. Arbitrarily, they set σ^2 to equal 1.

Mixture Model

The authors state that although the inverse Gaussian FHT model is a plausible parametric model for the survival distribution of a single subject, the survival curves for their case study suggested that the survival patterns across subjects in the study may be a mixture of two different inverse Gaussian distributions, with an initial plateau of low-risk and then a secondary slackening of risk after approximately one year being demonstrated. Thus, they used a mixture model, denoted by \bar{G} :

$$\bar{G}(r) = p\bar{F}_1(r) + (1-p)\bar{F}_2(r)$$

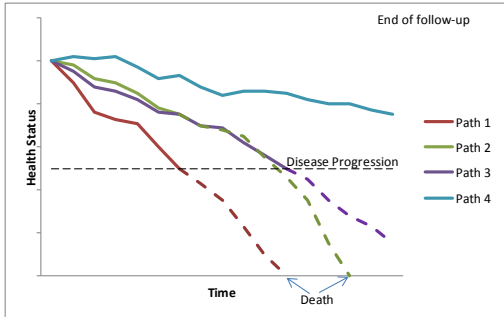
Where the mixing parameter is the proportion p . The $\bar{F}_j(r), j = 1, 2$ are the respective component survival functions of the mixture. Note it is assumed that each component survival function has a fixed mean parameter.

Statistical Inference

Each component of the survival function of the mixture model has its own p.d.f. and c.d.f., its own initial health status x_{0j} and mean parameter μ_j for $j=1,2$. The authors denote the vector of parameters of the mixture model by θ . This includes all the regression coefficients for parameters $p, \mu_1, x_{01}, \mu_2, x_{02}$ and α .

r_i denotes the composite time for patient i . It is the time of death for a dying patient, and a right-censored time for a surviving patient. Each dying patient contributes probability density $g(r_i | \theta)$ to the sample likelihood function for $i=1, \dots, n_1$ where $g(r | \theta)$ is the mixture p.d.f. corresponding to the mixture model $\bar{G}(r)$ above, and n_1 is the number of dying patients. Each surviving patient contributes survival probability $\bar{G}(r_i | \theta)$ to the sample likelihood function for $i=n_1 + 1, \dots, n_1 + n_0$ where $\bar{G}(r_i | \theta)$ is the mixture survival function in $\bar{G}(r)$ and n_0 is the number of patients who survive until the end of the study.

	<p>It is assumed that censoring is uninformative. The sum of $n = n_1 + n_0$ is the total number of patients in the trial. Therefore, the authors state that the sample log-likelihood function to be maximised has the form:</p> $\ln L(\theta) = \sum_{i=1}^{n_1} \ln g(r_i \theta) + \sum_{i=n_1+1}^{n_1+n_0} \ln \bar{G}(r_i \theta)$ <p>The authors state that they use the numerical gradient optimisation routine 'lf' in Stata to find the maximum likelihood estimate of the regression coefficient vector θ. They state that this optimisation routine requires only the sample log-likelihood function given above.</p>
What are the key assumptions of the method?	<p>It seems to be assumed that the novel treatment only impacts on rate of progression in the pre disease progression period.</p> <p>The σ^2 parameter of the Wiener distribution is arbitrarily set to equal 1.</p> <p>Seems to assume that all patients switch treatment after progression.</p> <p>Assumes uninformative censoring.</p> <p>Assumes that covariates are available in order to classify initial health status for each patient, and that the disease progression process can be appropriately modelled.</p>
What are the theoretical advantages and disadvantages associated with the method?	<p>Advantages</p> <p>Proportional hazards are not assumed.</p> <p>A flexible modelling approach is used.</p> <p>Requires an explicit formulation of the underlying health process to be defined (the authors give this as an advantage).</p> <p>Authors state that their method controls for response-adaptive switching.</p> <p>Does not assume an equal treatment effect.</p> <p>Disadvantages</p> <p>It seems to be assumed that the novel treatment only impacts on rate of progression in the pre disease progression period.</p> <p>The method seems to correct for whether switching occurred before or at disease progression, but it does not consider no switching at any point compared to switching at any point.</p> <p>The σ^2 parameter of the Wiener distribution is arbitrarily set to equal 1.</p> <p>Seems to assume that all patients switch treatment after progression.</p> <p>Requires an explicit formulation of the underlying health process to be defined – requires data on covariates for this.</p>
What are the potential biases associated with the method?	<p>See above disadvantages and below reasons why the method may not be appropriate. In particular it is essentially a method for extrapolation that happens to have been developed alongside a trial that incorporated treatment crossover. However, the type of crossover present is not exactly that investigated in this thesis and so the method is not directly relevant for the crossover problem as defined in this thesis.</p>
Why might the method not be appropriate?	<p>It seems to be assumed that all patients will switch treatment after PD, and this is a key part of the model. However, we are interested in cases where only a proportion of patients switch – it is not clear whether this is an important problem that would make the method inappropriate for the treatment crossover problem as defined in this thesis – but certainly the method would require adaptation.</p> <p>σ^2 is arbitrarily set to equal 1 – it is not clear whether this is an important problem.</p> <p>The aim is primarily to estimate μ of the Wiener process, which is equivalent to mean survival. In their model, the authors allow μ to depend on a treatment indicator variable for the assigned primary treatment, and PD, an indicator variable defining whether disease progression is observed under the primary therapy (i.e. this will be censored if switching occurs prior to PD. It is observed if PD occurs while on the primary therapy, and if switching occurs after PD. An interaction term between these two variables is also included. Covariates are also included for variables deemed important indicators of health status – age, baseline beta-2 microglobulin, and number of previous treatments.</p> <p>This set-up defines 4 treatment regimens: 1. Velcade assigned and taken until PD, at which point switching occurs. 2. Velcade assigned and switching occurs before PD. 3. Control assigned and taken until PD, at which point switching occurs. 4. Control assigned and switching occurs before PD.</p> <p>Thus, the model allows mean survival to be estimated for the above regimens (i.e. can see the effect of switching prior to PD rather than waiting to switch at PD), controlling for the baseline covariates. However, this does not seem like a suitable set-up to control for post PD switching – i.e. what would have happened if control patients did not switch after PD? For this it may be more appropriate to assign variables for whether switching was observed or not, and whether it happened before or after PD.</p> <p>Also, the way the $r = \alpha t_1 + t_2$ composite time scale is set up (with t_1 corresponding to pre progression time and t_2 corresponding to post-progression, it seems to be assumed that the primary therapy will only affect the rate of progression in the pre-progression period.</p>

	Also, the model seems to attempt to correct treatment effect estimates for pre progression switching controlling for health status indicators. However only three baseline indicators are used, which might not be appropriate.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The authors state that treatment switching as defined in this section is response-adaptive switching rather than random switching. They claim that the method used by Branson and Whitehead (2002) uses latent event times to model the survival time that would have been achieved had switching not occurred, but that this method does not take into account the fact that a treatment switch is often based on the observed effect of the current treatment. They claim that Branson and Whitehead's method is therefore for random treatment switching with a constant latent hazard rate over time (this seems unfair – the IPE method doesn't require any assumptions to be made about the prognosis of switchers – thus it is not only appropriate for cases where switching is random, and time-dependent confounders (possibly such as treatment response) are not a problem – provided randomisation is adequate and there is a common treatment effect). They claim that in fact even with random switching the hazard rate will increase after the switch, and the later the switch occurs, the larger the hazard rate after the switch. The authors claim that their method allows this problem to be addressed, as they deal with response-adaptive switching.
Does the method represent an extension to another method?	No, the method is quite separate from all other methods reviewed.
Application	
Is there a worked example in the survival setting?	<p>Yes, the authors give a case study using data from a multiple myeloma trial.</p> <p>The basic set up of the FHT model used in the case study given by the authors is shown in the diagram below reproduced from the paper, where the horizontal axis is time, and the vertical axis is health status.</p>  <p>Path 1 is a patient who experiences disease progression (PD) under the assigned therapy, and then dies before the end of follow-up. The dashed line denotes that after PD treatment was switched. Path 2 is a patient who switched treatments before PD, and also dies before the end of follow-up (thus PD under the assigned treatment is censored, but death is known). Path 3 is a patient who switches at PD, but is alive at the end of follow-up (and thus is censored for death). Path 4 is a patient who does not switch treatment and remains progression free at the end of follow-up (and thus is censored for PD and death). It is assumed that progression always precedes death.</p> <p>For the application of their method the authors set up an assigned treatment variable; an indicator variable for whether PD was observed under the primary therapy; an interaction term representing the product of the previous two variables; an indicator for the previous number of treatments received; a variable for the baseline level of beta-2 microglobulin; and a variable for baseline age. The four possible combinations of the treatment variable and the PD variable define four treatment regimens for patients: Active and PD observed under assigned treatment; Active and PD not observed (censored) under assigned treatment; Control and PD observed under assigned treatment; Control and PD not observed (censored) under assigned treatment.</p> <p>There are two time measurements for each patient – duration of PD and duration post PD. It is assumed that after PD treatment crossover occurs and so these measurements represent time on primary therapy and time on the alternate therapy. The sum of these two times equals total survival, be it censored or not.</p> <p>The authors chose a log-link function for the initial health status parameters x_{01} and x_{02}, and for the composite time parameter α. An identity link function was used for the mean parameters μ_1 and μ_2, and a logit link function for the mixture probability parameter p.</p> <p>Based on a Cox proportional hazards model of the Velcade data, the treatment indicator gave a hazard ratio of 0.7259 (p-value = 0.010), and the Kaplan-Meier showed that Velcade increases survival time when it is the primary therapy. However, the authors state that this method offers no insight to the source of nature of the benefit.</p>

	For their method, the authors presented the regression coefficients for each of their variables, which were exact maximum likelihood estimates. Standard errors were also presented. Initial health status was made to depend on the covariates for previous treatment, baseline beta-2 microglobulin and baseline age. The mean parameter covariates were made to depend on assigned treatment and the indicator variable for whether PD was observed on the assigned treatment, and the interaction term of these two covariates. These covariates define the treatment regimen. The mixing parameter p and composite time parameter α are not made to depend on any covariates.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>The authors found that the first component of the mixture model ($j=1$) represented an estimated 14% of the survival probability, which was significantly greater than 0. Thus there was evidence that the two-component mixture model offered a significantly larger likelihood than an unmixed model would have. In this first component only age was a significant indicator on the initial health status parameter. The mean parameter for the first period μ_1 showed a strong treatment effect favouring Velcade as the primary therapy. Taking into account the interaction parameter between assigned treatment and whether PD was observed on the assigned treatment, the benefit of the treatment regimen is enhanced (it is this interaction parameter which appears to attempt to correct for crossover).</p> <p>The second component of the model is the majority component of the mixture. Initial health status is significantly negatively affected by baseline beta-2 microglobulin and by the previous treatment covariate. In this component of the model the treatment regimen has no effect on the mean parameter μ_2, although survival is affected by whether disease progression is observed or not, suggesting that patients die more quickly when PD is observed while patients are on their assigned treatment.</p> <p>The authors present the estimated mean parameters for each of the model components for all 4 combinations of assigned treatment group and observed PD on assigned treatment. The mean estimates for μ_1 are positive (extend survival) for both Velcade regimens (i.e. whether or not PD was observed on the assigned treatment). However, the authors do not discuss the fact that estimates of μ_2 are negative for both Velcade regimens. All mean estimates are negative for the control treatment.</p> <p>The authors use their results to estimate cure rates, which were very low.</p> <p>They also consider the α parameter, the rate of PD prior to the therapeutic switching event PD. The point estimate of α signalled that health status declined slowly until the switch to alternative therapy. The authors rationalise this by suggesting that treatment switching will only occur when the patient's progression accelerates significantly under the primary therapy.</p> <p>The authors state that their results were broadly in line with the results from the Cox model. The Cox model found a benefit associated with Velcade, and the coefficients of the covariates are broadly in line. However, the authors state that their TR method provides more subtle insights into the source and nature of the comparative benefits of Velcade.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	This method is essentially a novel method for modelling survival from a clinical trial, and incorporates switching because the trial it is applied to involved switching. Although it seems that the method could be used to attempt to estimate survival benefits adjusting for crossover, this is not the focus of the paper, and it appears likely that the method would need to be adjusted in order to do this. Before using this method to adjust for crossover, ideally it should be accepted for use as a method for modelling survival, but at present the method has not been used other than by the authors.

Reference	Witteaman JCM, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MAJ, Hofman A, Robins JM. G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham heart study. American Journal of Epidemiology 1998;146;4:390-401
Origin	
Was the method developed specifically in the survival analysis context?	Yes. The authors use the G-estimation approach to estimate the causal effect of isolated systolic hypertension (ISH) on cardiovascular death, using data from the Framingham heart study.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The authors use g-estimation with an SNM to adjust for confounding caused by a time-dependent covariate (arterial thickness) that is a risk factor, and is influenced by previous ISH, and influences future ISH. The method used is an SNM as described by Robins (1998) in his paper on structural nested accelerated failure time, reviewed later in this appendix. Therefore most of the details are not replicated here. The reason for including this paper is that it claims to extend the approach for a case where censoring is by competing risks. However, this approach is using IPCW, which is also described in detail elsewhere in this appendix (see Robins and Finkelstein, 2000), and in fact this method combined with an SNM is also described by Robins (1998). Hence the methodology is not replicated in this table.</p> <p>Separately from this, the authors provide a clear explanation of administrative censoring in an SNM context. In a carefully planned RCT, administrative censoring (that is, censoring at the end of follow-up) is likely (and some censoring due to loss-to-follow up may occur), but this should be uninformative (that is, random and independent of the counterfactual survival time). In this case, censoring (C_i) can be dealt with relatively simply in an SNM. In the g-test model the counterfactual survival time for a given value of ψ ($U_i(\psi)$) is replaced with an indicator</p>

	<p>variable ($\Delta_i(\psi)$) that equals 1 if the event (for example, death) would have been observed if an individual i had been continuously exposed or unexposed to treatment, and 0 otherwise.</p> <p>$\Delta_i(\psi) = 1$ if $U_i(\psi) < C_i(\psi)$ and $\Delta_i(\psi) = 0$ if $U_i(\psi) \geq C_i(\psi)$</p> <p>where $C_i(\psi) = C_i$ if $\psi \geq 0$ and $C_i(\psi) = C_i \exp(\psi)$ if $\psi < 0$. The indicator variable $\Delta_i(\psi)$ will be observed for all patients, and will equal zero for all patients who had censored survival times. It may also be zero for some patients who had observed survival times, depending upon the size of the treatment effect ψ. This allows the g-test to be conducted, but with counterfactual survival time indicated by $\Delta_i(\psi)$ taking into account censoring, instead of simply being $U_i(\psi)$. This is reasonable because for treatment received to be independent of U_i conditional on past treatment and covariate history, it must also be independent of Δ_i because it is a function of U_i and C_i, which itself is essentially a baseline covariate (since it represents the end of follow-up time).</p> <p>The authors evaluate time-dependent confounders using time-dependent Cox models to examine whether covariate status before time t was related to the hazards of cardiovascular death at t, conditional on past ISH and baseline covariate status. Then a logistic model, pooled over all visits k, was used to examine whether time-dependent covariates predicted ISH at k, conditional on past ISH levels and baseline covariate status. Then it was examined whether the occurrence of remaining covariates was predicted by ISH, conditional on baseline covariates and the absence of that covariate occurrence at $k-1$. This determined covariates that were independent risk factors for CV death, predictors of future ISH, and predicted by previous ISH.</p>
What are the key assumptions of the method?	See SNM (and for competing risks censoring, IPCW).
What are the theoretical advantages and disadvantages associated with the method?	See SNM (and for competing risks censoring, IPCW).
What are the potential biases associated with the method?	See SNM (and for competing risks censoring, IPCW).
Why might the method not be appropriate?	See SNM (and for competing risks censoring, IPCW).
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The approach taken is an SNM, combined with IPCW.
Does the method represent an extension to another method?	SNMs are extended for censoring by competing risks.
Application	
Is there a worked example in the survival setting?	Yes.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	The authors fitted three alternative models, including different combinations of baseline and time-varying covariates. They found that the point estimate and confidence intervals associated with the treatment effect were almost identical for each model, suggesting that the time-dependent confounding in the study was small. However, the authors state that this finding is based on several assumptions – in particular, that the model linking observed and counterfactual survival time is correct; there are no unmeasured confounders for the risk of death and that predict the probability of ISH at time k ; and that there are no unmeasured confounders so that censoring by competing risks is independent of the time a subject would have died from CV disease had censoring been presented. It is also noted that the recensoring method could cause problems in smaller data sets.
Other Issues	
Are there any other relevant characteristics associated with the method?	The cohort in the trial was made up of 4404 subjects with available exposure and covariate information in the relevant time-frame, and thus it might be expected that observational methods can work adequately.

Included Papers: Reference Search

Reference	Robins JM. Structural Nested Failure Time Models. In "Survival Analysis" Section Eds.: Anderson PK, Keiding N, The Encyclopedia of Biostatistics, Eds: Armitage P, Colton T. Chichester, UK: John Wiley & Sons, 1998:4372-4389
Origin	
Was the method developed specifically in	Yes.

the survival analysis context?	
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>This paper provides a summary of what structural nested failure time models are, and how they can be used to estimate regime specific survival curves. This paper is essentially a review paper since Robins published a number of papers on SNMs prior to this paper. However, since the methodology discussed is similar but was developed over time in each paper, this review paper is included here, as it summarises the SNM methodology.</p> <p>Structural nested failure time models are causal models which estimate the effect of a time-dependent treatment or exposure on a survival time outcome in the presence of time-dependent confounding covariates. The models work by mapping a subject's observed failure time to the failure time that would have occurred if, possibly contrary to fact, treatment had been withheld, taking into account observed treatment, a patient's confounder history, and an unknown causal parameter. The causal parameter can be identified if the treatment at any time point is randomly assigned conditional on past treatment and confounder history. This requires the assumption that, conditional on the past treatment and confounder covariates, the treatment assigned is ignorable in that there are no other factors that influence treatment assignment that could cause selection bias – i.e. there are no unknown or unobserved confounders. In the context of treatment crossover this means that if there is a reason why some control patients are given the new treatment and others are not, this reason must be explained by covariates included in the SNM – hence it is possible to control for the fact that crossover patients may have a systematically better or worse prognosis than non-crossover patients. If this is the case, g-estimation can be used to provide robust semiparametric estimators of the causal parameter.</p> <p>Robins (1998) states that a specific example of an SNM is the strong version of the accelerated failure time model of Cox and Oakes (1984), which assumes that:</p> $U_i = \int_0^{T_i} \exp[\psi A_i(t)] dt$ <p>Where U is the counterfactual survival time, which is a known function of observed survival time (T), observed treatment (A), and an unknown parameter ψ. If observed treatment was zero – i.e. the treatment of interest was not given – then $U=T$. In addition, if the unknown parameter $\psi=0$, then U will always equal T. If ψ does not equal zero, then for a constantly treated patient $T = e^{-\psi}U$, hence the patient's untreated survival time is expanded or contracted by the factor $e^{-\psi}$.</p> <p>Robins (1998) demonstrates that ψ can be estimated using g-estimation, making use of the assumption that there are no unobserved confounders. For each potential value of ψ the counterfactual survival time can be estimated for each patient because all parameters making up its function are known (observed survival time, observed treatment, and ψ). Thus, Robins (1998) state that a grid-search can be undertaken to identify an unbiased estimate of the true value of ψ. This is achieved by estimating the counterfactual survival time for each potential value of ψ and then performing a g-test. A g-test first models the hazard of the treatment process as a function of the observed survival time and past treatment and covariate history, and then tests whether a coefficient (say θ) of a function that incorporates treatment and covariate history up until time t, and the estimated counterfactual survival time for a given value of ψ is significant. Provided the no observed confounders assumption holds, the value of ψ that is chosen is the one that results in $\theta = 0$, as this will be true when ψ is such that the hazard of a treatment change does not depend upon the counterfactual survival time, given treatment and covariate history. The model used for the g-test is a time-dependent Cox proportional hazards model for the hazard of treatment change:</p> $\lambda_0(t) \exp[\alpha'W(t)]$ <p>Where $W(t)$ is a known vector valued function of treatment history and covariate history up until time t, α is an unknown parameter vector, and $\lambda_0(t)$ is an unspecified baseline hazard function. To conduct the g-test the term $\theta Q(t, \psi)$ is added to $\alpha'W(t)$ in the above model, where $Q(t, \psi)$ is a function (which Robins states can be chosen by the analyst) of treatment and covariate history up until time t and the estimated counterfactual survival time for a given value of ψ. It is the value of ψ that results in a Cox partial likelihood score test (g-test) of the hypothesis $\theta = 0$ in this model that provides a consistent and asymptotically normal estimator of ψ_0. Given the no unobserved confounders assumptions holds, the Cox model of the hazard of treatment change is correct, and the SNM is correct. The confidence interval for ψ_0 is given by the values of ψ that result in the g-test not being rejected at the 0.05 level.</p> <p>Robins (1998) states that the method of g-estimation can be extended to the case of a multi-parameter deterministic SNM. Such a model could take the form:</p> $U_i(\psi) = \int_0^T \exp[\psi_1 A(t) + \psi_2 L^*(t)A(t)] dt$ <p>Where $L^*(t)$ is a known function of the covariate history, and where if the true value of ψ_2 is not zero, the magnitude of the treatment effect depends upon the patient's time-dependent</p>

covariate history. Conducting g-estimation and obtaining an estimate of the parameter vector $\psi = (\psi_1, \psi_2)'$ in this case is achieved by choosing $Q(t, \psi)$ to be a known vector-valued function of $\dim\psi$ and θ to be a $\dim\psi$ valued parameter with $\dim\psi$ the dimension of the vector ψ . For the context of treatment crossover this may be useful if it is deemed realistic that the treatment effect will be different for patients who crossover onto the treatment once their disease has progressed compared to those who are initially randomised to the treatment. A multivariate SNM would allow disease progression to be included in the modelled covariate history.

Robins (1998) briefly considers the situation in which, in the context of an RCT, a patient is randomised to the experimental treatment, yet possibly non-random non-compliance occurs and the dose taken is different, yet it is desirable to know the treatment effect of the prescribed dose. This could be likened to a situation in which treatment crossover occurs, but it is desired to estimate the treatment effect in the absence of such crossover. In such circumstances, Robins states that g-estimation can be used along with an SNM as described above, with indicators for randomised treatment and treatment actually taken. We assume that the prescribed treatment is independent of the counterfactual survival time given treatment and covariate history, rather than the actual treatment received, and hence the prescribed treatment acts as an instrumental variable. Although this method is suitable for this case, Robins states that alternative rank estimation procedures, such as the RPSFTM are also available.

It is vital that the covariates included in the model cover all variables that may impact upon treatment received and survival. This type of SNM is designed for observational studies and is not specifically designed for treatment crossover and so does not use the randomisation of an RCT in any way.

Censoring in SNMs

Censoring is a problem for the SNM methodology described above because it means that the counterfactual survival time can only be estimated for a proportion of patients (those who are not censored), as for some patients survival time is unobserved. Robins (1998) considers censoring by end-of-follow-up, and censoring due to loss-to-follow-up or competing risks, and develops the SNM to account for these.

For censoring due to end-of-follow-up the censoring time C can be dealt with as a time-independent covariate that is known for all patients at $t = 0$. The data that is available for each patient becomes:

$$[X = \min(T, C), \bar{A}(X), \bar{L}(X)]$$

Where $\bar{A}(X)$ is the treatment history and $\bar{L}(X)$ is the covariate history. If the treatment has a non-zero effect (ie $\psi_0 \neq 0$) a new random variable $X^*(\psi)$ cannot be used to replace counterfactual survival times used in the g-estimation process because if there is a treatment effect the chance of being censored at any time t will be altered and so the treatment history is not independent of $X^*(\psi)$. Because of this, Robins introduces an alternative method for taking this censoring into account, by defining two new variables, $X(t, \psi)$ and $\Delta(t, \psi)$. These must be observed for all subjects (unlike T and U) and must be independent of $A(t)$ unlike $X^*(\psi)$. When these variables have been defined, ψ_0 can then be estimated using the g-estimation process described above, except now, when the term $\theta Q(t, \psi)$ is added to $\alpha'W(t)$ in the Cox model, $Q(t, \psi)$ is now a function of treatment and covariate history up until time t , $X(t, \psi)$ and $\Delta(t, \psi)$. Thus, $X(t, \psi)$ and $\Delta(t, \psi)$ replace the estimate of the counterfactual survival time in the model and the value of ψ that results in a Cox partial likelihood score test (g-test) of the hypothesis $\theta = 0$ provides the estimate of ψ_0 , given the no unobserved confounders assumptions holds, the Cox model of the hazard of treatment change is correct, and the SNM is correct.

For a deterministic SNM such as that described here:

$$X(t, \psi) = \min\{H(\psi), C(t, \psi)\}, \quad \text{and} \quad \Delta(t, \psi) = I\{X(t, \psi) < C(t, \psi)\}$$

Where $H(\psi)$ is the estimate of the counterfactual survival time and $C(t, \psi) \equiv C - t + \int_0^t \exp\{\psi A(t)\} dt$ when $\psi \geq 0$ (i.e. treatment reduces survival) and $C(t, \psi) = \int_0^t \exp\{\psi A(t)\} dt + C - t \exp\{\psi\}$ if $\psi < 0$ (i.e. treatment increases survival). Given the definition of the indicator function $\Delta t, \psi$, if $\Delta t, \psi = 0$ an individual is ψ -censored, but if $\Delta t, \psi \neq 0$ some failures will be ψ censored while others will not be, since the observed failure time may differ from the estimated counterfactual survival time for a given value of ψ . Replacing the $H(\psi)$ in $Q(t, \psi)$ with $X(t, \psi)$ and $\Delta(t, \psi)$ is suitable for the estimation procedure because $X(t, \psi)$ and $\Delta(t, \psi)$ are observable and for ψ_0 at time t they are only functions of treatment history up until time t , the counterfactual survival time and C .

The model is further complicated when censoring by loss-to-follow-up or competing risks is also considered. In this case, in addition to the censoring by end-of-follow-up C , there is also censoring due to Q , which is the minimum of time to loss-to-follow-up or to a competing risk event (from here on referred to as "loss-to-follow-up"). Now, the data available are $X^* = \min(T, C, Q) = \min(X, Q)$, $\tau = I(X^* \neq Q)$, $\bar{A}(X^*)$, $\bar{L}(X^*)$. Hence $\tau = 1$ if and only if a patient's survival time was observed, or if it was censored due to reaching the end of follow-up. When censoring due to loss-to-follow-up is adjusted for in an SNM framework, an additional assumption is made – that there is enough data on a range of time-dependent and time-independent covariates such that there are no unmeasured confounders for censoring due to loss-to-follow-up, and hence, given the data available on the past, censoring by Q is ignorable. Given this assumption, an inverse probability of censoring weights (IPCW) procedure is then used in a weighted g-estimation process to estimate ψ . First, the probability $K(X)$ of a patient surviving to $X = \min(T, C)$ is estimated, and then this probability is used as an inverse weight in the g-estimation procedure. This is achieved by fitting a Cox proportional hazard model for

the hazard $\lambda_Q(t|\bar{A}(t^-), \bar{L}(t^-), X > t)$, which is the hazard of being censored at time t where $\bar{A}(t^-)$ is the treatment history up to (but not including) time t , and $\bar{L}(t^-)$ is the vector of various time-dependent and time-independent covariates up to (but not including) time t . The Cox model is:

$$\lambda_{0Q}(t) \exp\{\alpha^* W^*(t)\}$$

Where $W^*(t)$ is a known vector valued function of $\bar{A}(t^-)$ and $\bar{L}(t^-)$, α^* is the vector of unknown parameters and $\lambda_{0Q}(t)$ is an unspecified baseline hazard. This model is then used to estimate the Cox baseline hazard estimator at each time Q_j at which any patient j experienced censoring due to loss-to-follow-up, i.e.:

$$\hat{\lambda}_Q(Q_j) = 1 / \sum_{i=1}^n \{\exp[\hat{\alpha}^* W_i^*(Q_j)] I(X_i^* \geq Q_j)\}$$

Which is the Cox baseline hazard estimator of $Q_{0Q}(Q_j)$, ie the hazard across all patients of being censored due to loss-to-follow-up at time Q_j . Then, the probability $K(X)$ of surviving to $X = \min(T, C)$ is estimated by multiplying together the estimated conditional probabilities of not (hence the '1 -' in the estimator below) experiencing censoring due to loss-to-follow-up using the time-dependent Cox-model version of the Kaplan-Meier estimator, over the times at which there is some risk of censoring due to loss-to-follow-up ($Q_j < X$):

$$\hat{K}(X) = \prod_{\{j: Q_j \leq X, \tau_j = 0\}} \{1 - \lambda_Q(Q_j) \exp[\hat{\alpha}^* W^*(Q_j)]\}$$

Robins notes that this probability depends upon the patient's treatment and covariate history through $W^*(t)$.

Once this process is complete, to estimate ψ an inverse probability of censoring weights method is used within the g-estimation process. Through the above steps $\hat{K}(X)$ is estimated for each patient whose survival time is observed or who reach the end of follow-up (those patients with $\tau = 1$). In the g-estimation procedure previously the function $\theta Q(t, \psi)$ was added to $\alpha^* W(t)$ in the Cox model, $Q(t, \psi)$ being the function of treatment and covariate history up until time t , $X(t, \psi)$ and $\Delta(t, \psi)$. Now, $Q(t, \psi)$ is replaced by $Q^*(t, \psi) \equiv Q(t, \psi) / \hat{K}(X)$ for those with $\tau = 1$ (those who did not experience censoring due to loss-to-follow-up), and by $Q^*(t, \psi) = 0$ for those with $\tau = 0$ (those who did experience censoring due to loss-to-follow-up). Then the g-estimation procedure is completed and the value of ψ that results in a Cox partial likelihood score test (g-test) of the hypothesis $\theta = 0$ provides the estimate of ψ_0 , given the no unobserved confounders assumptions holds, the Cox model of the hazard of treatment change is correct, and the SNM is correct, and that there are no unmeasured confounders for censoring.

Robins gives a helpful intuitive explanation of why this inverse probability of censoring weights method gives a consistent estimator of ψ_0 , given that the Cox model for the hazard of censoring is correct, and given that there are no unmeasured confounders for censoring due to loss-to-follow-up or competing risk. For each patient for whom X is observed who has a cumulative probability (from the Kaplan-Meier estimator) of, for example $\hat{K}(X) = 0.25$ of avoiding censoring due to loss-to-follow-up or competing risks, there would on average have been three other patients who had similar treatment and covariate history and would have had a similar value of X , but who were censored due to loss-to-follow-up or competing risks prior to X , under our earlier assumption that there are no unmeasured confounders that might impact treatment and survival. Hence, to account for this censoring, the patient for whom X is observed is given a weight of 4 in the g-estimation procedure, which is achieved by multiplying their covariate $\Delta(t, \psi)$ by 4. This accounts for the three patients for whom X was not observed and $Q^*(t, \psi)$ was set to zero. The resulting estimate of ψ is a consistent asymptotically normal estimate, but the previously described method for estimating confidence intervals is not valid. However, Robins (1998) states that if the extended Cox model for the treatment process is fitted using a program that computes "robust variances" the resulting intervals are guaranteed to be conservative, and thus reasonable for use. This means that in large samples the nominal 95% confidence intervals are guaranteed to cover ψ_0 at least 95% of the time, and 0.05 level g-tests will reject the null hypothesis no more than 5% of the time.

It might be considered that the above methods could be used to adjust for treatment crossover, perhaps by censoring patients who crossover and classing this as competing risk, or censoring due to loss-to-follow-up. Robins (1998) briefly considers this, and states that if a reasonably large proportion of patients did not receive the treatment that we wish to control for (in our case, if a reasonable proportion of control group patients did not crossover) – Robins suggests at least 30% - then patients who crossover could be censored at the point of crossover, and then Q could be the minimum of time to censoring due to loss-to-follow-up, time to censoring due to competing risks, and time to crossing over. Then the above methods could be used to estimate ψ .

What are the key assumptions of the method?	No unmeasured confounders. Correctly specified SNM. If required to adjust for potentially informative censoring, IPCW assumptions are also required.
What are the theoretical advantages and disadvantages associated with the	A key disadvantage is that it is very data intensive. The no unmeasured confounders assumption is hard to justify.

method?	
What are the potential biases associated with the method?	If there are any unmeasured confounders the method will result in bias.
Why might the method not be appropriate?	<p>Because often data may not be available on all potentially prognostic variables. In addition, the method may be difficult to apply in the RCT treatment crossover context that we are interested in. The method could not be applied simply to an RCT dataset, because it would be inappropriate to attempt to model the treatment process (that is, the treatment received by patients over time) when patients are randomised to treatment groups – patients randomised to the control group would not receive the intervention (until crossover was allowed) irrespective of their covariates, and similarly patients in the intervention group would remain in that group (though they may discontinue treatment) irrespective of their covariates. In this situation, attempting to model the treatment process based upon observed covariates would be counter-intuitive. Even if this were possible, if the method were applied to the whole trial population in order to estimate a causal effect of the treatment the result would be an average effect based upon all patients who took the treatment. It is arguable how useful this is from the health economist's decision problem perspective, because what is desired is an estimate of the treatment effect in patients initially randomised to the intervention group – not an average effect incorporating this group and patients who received this treatment later on, once their disease had progressed.</p> <p>However, these issues do not mean that the SNM method cannot be useful in an RCT context, given the decision problem faced in the economic evaluation. In the context of an RCT, the control group after the point at which treatment crossover becomes possible could be treated as an observational dataset. The SNM method could then be applied to this dataset to estimate the treatment effect specific to control group (crossover) patients. Given the resulting treatment effect estimate (in terms of an acceleration factor – working on the time scale) the survival times of control group crossover patients could be adjusted to estimate counterfactual survival times had crossover not occurred. This is similar to the approach taken by Robins and Greenland (1994) and Yamaguchi and Ohashi (2004)</p>
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	The method influences and borrows from several others – IPCW, and the SNMs used by Robins and Greenland (1994) and Yamaguchi and Ohashi (2004), as well as the RPSFTM which is used by many of the papers included in this review.
Does the method represent an extension to another method?	No, although for informative censoring it combines with the IPCW method.
Application	
Is there a worked example in the survival setting?	No.
Is the example relevant?	Not applicable.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	Not applicable.
Other Issues	
Are there any other relevant characteristics associated with the method?	Not applicable.

Reference	Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. Biometrics 1996;52:137-151
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>In this paper the authors use a time-dependent prognostic covariate to help in the estimation of OS for patients whose survival time is censored. This is therefore an extrapolation method. Treatment crossover is not mentioned but this method could be useful.</p> <p>Murray and Tsiatis (1996) describe a method whereby crossover patients are treated as censored (as in the IPCW approach) and a pseudo population is created by imputing values for these patients using an Auxiliary Variable (AV). If there is only one AV the method can be conducted without additional modelling assumptions, or if categorical combinations of key covariates</p>

	<p>can be defined. However if this is not the case and there are several time-varying AVs (which may be likely) conditional modelling of event times and the process by which these are effected by the AVs is required. Robins and Finkelstein (2000) note that in this case the extra models can cause biased estimates if they are not specified correctly. The IPCW approach requires less modelling assumptions in this scenario and so may be preferable.</p> <p>The WKM approach described by Murray and Tsiatis (1996) weights survival conditional upon having a particular value of one particular auxiliary variable by the probability of having that covariate value. The approach is fully non-parametric but can only incorporate one auxiliary variable.</p> <p>The method is adapted by the authors for the case when the AV is time dependent.</p>
What are the key assumptions of the method?	There is only one AV and that AV is suitable for re-estimating event times for censored patients, or relevant covariates can be categorised without the need for further conditional modelling.
What are the theoretical advantages and disadvantages associated with the method?	It is quite unlikely that there will be one AV suitable for re-estimating event times for censored patients. If there is only one AV the method may be accurate and relatively free of modelling assumptions.
What are the potential biases associated with the method?	If there is more than one AV the method may be inaccurate.
Why might the method not be appropriate?	Because there is likely to be more than one AV. Also, it was not developed specifically for the treatment crossover situation.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	In contrast, IPCW can incorporate a range of variables, and thus is more likely to satisfy the no unmeasured confounders assumption. Robins and Finkelstein (2000) state that when several covariates are included the AV approach requires models for the conditional hazard of the event time and the process by which this is effected by the chosen auxiliary variables – and if these models are misspecified survival estimates can be biased even if there was no selection bias (in which case a standard Kaplan-Meier estimator would have been suitable). The IPCW method, in which weighting is based upon the inverse of the estimate of the conditional probability that a subject is uncensored through time t given the auxiliary data available, will not be biased in such a situation, given sequentially ignorable censoring. Thus, Robins and Finkelstein (2000) note that while extended versions of the WKM can be more efficient than the IPCW method when models are correctly specified, the IPCW approach is preferable. This provides some interesting context for the Hsu <i>et al</i> (2006) paper, that attempts to develop a method that can reproduce WKM estimates (in situations where the WKM is correctly specified), extend the method so that more than one auxiliary variable can be included, while using a non-parametric approach so that there is not an over-reliance on parametric assumptions.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	Yes, the method is applied to an AIDS trial and simulation studies are run.
Is the example relevant?	The example was not relevant for treatment crossover.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	The method allowed increased precision when censoring occurred.
Other Issues	
Are there any other relevant characteristics associated with the method?	Robins and Finkelstein (2000) note that a key problem with an AV approach is that when there are a range of auxiliary variables conditional modelling is required for the estimated event time and the process by which this is affected by the auxiliary variables. If the models used to capture these relationships are miss-specified the resulting treatment effect estimates can be biased and, for example, inconsistent estimates of the survival curve can be produced even when censoring is independent and an unweighted Kaplan-Meier would have been consistent. Robins and Finkelstein (2000) state that this problem can be avoided by creating a pseudo population using the IPCW approach rather than an AV approach. The IPCW approach requires fewer modelling assumptions than the AV approach. Thus, it appears reasonable to conclude that the IPCW approach supersedes AV approaches.

Reference	Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: “Statistical models in epidemiology: The environment and clinical trials” Eds: Halloran ME, Berry D. IMA Volume 116. NY: Springer-Verlag 1999: pp.95-131
Origin	
Was the method developed specifically in the survival analysis context?	Yes.
If not, what was the original context and	Not applicable.

how has the method been adapted?	
Theoretical Suitability	
How does the method work?	<p>In this paper the author introduces marginal structural models for estimating the causal effect of time-dependent treatments. Marginal structural models are compared to structural nested models which were developed by the same author. The author states that the main difference between MSMs and SNMs is that SNMs model the magnitude of the effect of a treatment given t as a function of the prognostic factor history up to t, whereas the MSMs models the causal effect of treatment given t only as a function of baseline prognostic factors. Both methods use the counterfactual modelling approach, and the MSM approach uses inverse probability of treatment weights. The emphasis of the paper is clearly on measuring the effect of a time-dependent treatment. Treatment crossover is not mentioned. Together with a paper by Hernan <i>et al</i> (2001) this paper provides a full description of the use of MSMs for causal estimation.</p> <p>The authors describe a marginal structural Cox proportional hazards model for the analysis of an AIDS RCT. Their objective was to identify the joint causal effect of AZT treatment and AP treatment for pneumocystis carinii pneumonia. For their analysis they specified the following MSM:</p> $\lambda_{T_{\bar{a}}}(t V^{\dagger}) = \lambda_0(t) \exp(\beta'_1 V^{\dagger} + \beta_2 \alpha(t))$ <p>Where $T_{\bar{a}}$ is the patient's time to death if he or she had followed AP treatment history \bar{a}. $\lambda_{T_{\bar{a}}}(t V)$ is the hazard of $T_{\bar{a}}$ at t conditional on V^{\dagger}. $\lambda_0(t)$ is an unspecified baseline hazard function. $\exp(\beta'_1)$ and $\exp(\beta_2)$ are the causal rate ratios. V^{\dagger} is the vector of baseline regressors comprised of a selection of prognostic variables, and includes a randomisation group indicator.</p> <p>It is important to note that this model specifies the hazard of death at time t to depend only upon current treatment status and not previous treatment.</p> <p>The authors suggest that a standard time-dependent Cox model could be used to estimate the vector of β, if the giving of treatment at time t is completely random, or if the treatment decision only depended upon the history of treatment prior to t. In this case the β parameters will have causal interpretation, because the treatment decision is causally exogenous. They state that randomised treatments are causally exogenous, but we know that when unplanned treatment crossover occurs the choice of treatment received by potential crossover patients in the control group is not causally exogenous. Robins (1999) shows that the extent to which a treatment process is statistically non-exogenous can be estimated using the following equation:</p> $W(t) = \prod_{k=0}^t f[A(k) \bar{A}(k-1), \bar{L}(k)] / f[A(k) \bar{A}(k-1), V]$ <p>Where the numerator is the probability that a patient received his or her own observed treatment at time k, $A(k)$ given his or her past treatment and prognostic factor history ($\bar{A}(k-1)$ and $L(k)$); and the denominator is the probability that the patient received his or her observed treatment conditional only upon past treatment history and baseline variables – not also conditional on prognostic factor history. If the treatment process is exogenous, this will equal 1 for all t. If the treatment process is not exogenous, the inverse probability of treatment weights (IPTW) method is used, whereby the time-dependent Cox model is weighted by applying the weight W^{\dagger} to each patient for each time k. Essentially this means weighting by the inverse of a patient's probability of having his or her own observed treatment history. Using this weighting means that $\bar{L}(t)$ does not predict treatment at t given past treatment history, and thus the authors state that a counterfactual pseudo population has been created in which treatment is exogenous. Also, the causal effect of treatment is the same in the counterfactual pseudo population as it is in the original population, and so to estimate treatment effects we can conduct standard time-dependent Cox model analysis on the pseudo population. Thus, Robins (1999) show that if the assumption of no unmeasured confounders holds – that is $L(t)$ includes all relevant time-dependent prognostic factors, then the weighted estimators will converge to values of β that can be appropriately interpreted as the causal effect of treatment history on the time to event.</p> <p>The authors demonstrate how to incorporate censoring into the analysis. They allow for both administrative censoring and censoring due to drop-out. They adjust for this by defining that it is desired to estimate the effect of \bar{a} when $\bar{c} \equiv 0$, that is, a patient's failure time when treated with a certain regimen is estimated in the absence of censoring:</p> $\lambda_{T_{\bar{a}, \bar{c}=0}}(t V^{\dagger}) = \lambda_0(t) \exp\{r[\bar{a}_1(t^-), t, V^{\dagger}; \beta_0]\}$ <p>Where t^- denotes treatment history.</p> <p>In order to obtain consistent estimates of β in this situation, it must be assumed that censoring is noninformative (ignorable) given treatment history and time-dependent covariate history, and thus there is no unmeasured confounding. To take censoring into account a patient who is alive and uncensored at time t is weighted by $W(t) \times W^{\dagger}(t)$, where:</p>

	$W^\dagger(t) = \prod_{k=0}^t \frac{\Pr[C(k) = 0 \bar{C}(k-1) = 0, \bar{A}(k-1), V, T > k]}{\Pr[C(k) = 0 \bar{C}(k-1) = 0, \bar{A}(k-1), \bar{L}(k), T > k]}$ <p>Which is equivalent to the inverse of the ratio of a patient's probability of remaining uncensored up to time t divided by that same probability calculated as though the only determinants of censoring were past treatment history and V. Therefore, IPTW is complemented with inverse probability of censoring weights (IPCW) to account for censoring. $W(t)$ is also slightly amended, with $C(k)=0$ added to the conditioning events in the numerator and the denominator:</p> $W(t) = \prod_{k=0}^t f[A(k) \bar{A}(k-1), \bar{L}(k), C(k) = 0] / f[A(k) \bar{A}(k-1), V, C(k) = 0]$ <p>Under these conditions, the denominator of $W(t) \times W^\dagger(t)$ is the probability that a patient has his or her own observed treatment and censoring history through time t.</p> <p>The authors then go on to explain how $W(t)$ and $W^\dagger(t)$ are estimated. In the case where there are two treatment effects of interest a pooled logistic model for the binary responses for A_j and censoring are fitted.</p> <p>Robins (1999) proved that IPTW estimators will be consistent so long as the models for treatment initiation and censoring used in the numerators of $W(t)$ and $W^\dagger(t)$ are correctly specified.</p> <p>The authors state that because weights are used for the IPTW estimation standard error estimates will be incorrect, and thus robust variance estimators – which will provide a conservative confidence interval – must be calculated.</p>
<p>What are the key assumptions of the method?</p>	<p>No unmeasured confounders. Once a treatment has been started the patient does not stop taking it at any point. Robins (1999) proved that IPTW estimators will be consistent so long as the models for treatment initiation and censoring used in the numerators of $W(t)$ and $W^\dagger(t)$ are correctly specified.</p>
<p>What are the theoretical advantages and disadvantages associated with the method?</p>	<p>The authors consider the major advantage of MSMs to be that they resemble standard models, since they take the form of Cox proportional hazard models. However, there are important advantages that SNMs hold over MSMs. Firstly, MSMs cannot be used if there are any particular covariate values that ensure that a patient will definitely receive a certain treatment at time k. In the context of treatment crossover this would be a problem if all patients in the control group switched treatment at the point of disease progression, and is an even more serious problem if we do not wish to control for any non-compliance in the group randomised to the experimental treatment. As discussed earlier, typically we do not wish to control for such non-compliance as it is likely to be due to medical reasons that are likely to also occur in the real-world. Hence it may only be possible to apply MSMs to the control arm of trials.</p> <p>Secondly, while the structure of MSMs makes them useful for considering interactions between treatment and baseline covariates, they are not as useful as SNMs for considering interactions between treatment and time-dependent covariates. Essentially SNMs model the magnitude of the treatment effect at t as a function of the prognostic factor history up to t, whereas MSMs measure the effect only as a function of baseline covariates, and take into account time dependent covariates through weighting.</p> <p>On top of these issues with MSMs, the assumptions behind the approach are restrictive, as they are for SNMs that are not RBEEs. It must be assumed that the covariates included in the analysis are sufficient to adjust for both confounding and selection bias due to censoring – that is it is assumed that there are no unmeasured confounders and non-informative censoring. Also, the model specified for the effect of treatment on mortality must be correct, and the models used for initiation of treatment and for censoring must be correct.</p>
<p>What are the potential biases associated with the method?</p>	<p>Associated with the disadvantages and assumptions noted above.</p>
<p>Why might the method not be appropriate?</p>	<p>A standard MSM estimates an average treatment effect across all patients who took the treatment. In the context of treatment crossover that would include both experimental group patients and crossover patients. This is problematic from the economic evaluation decision problem, as we are interested in the treatment effect specific to patients initially randomised to the experimental treatment. In addition, as for the SNM method it is problematic to apply the MSM to an RCT context. The MSM relies on being able to model the treatment process, yet when patients are randomised to treatment groups attempting to model the probability of treatment received based upon observed covariates is counter-intuitive. A similar approach to that taken for the SNM could be applied (that is a two-stage approach of estimating the treatment effect first in crossover patients, and then in the experimental group), whereby the MSM is applied only to patients in the control group after the time-point at which treatment crossover becomes possible. This would result in a treatment effect estimate specific to crossover patients. However, the SNM approach uses accelerated failure time models and produces an acceleration factor which works on the time-scale, allowing observed survival times of crossover patients to be 'shrunk' in order to arrive at an estimated counterfactual dataset. Conversely, the MSM produces a hazard ratio, which works on the hazard scale rather than the time scale. Therefore survival times of crossover patients cannot be shrunk in the same way and there is no obvious way to estimate the counterfactual dataset. For these reasons, neither a standard MSM or a 'two-stage' MSM represent a suitable approach for addressing the treatment crossover problem in the context of an RCT, given the economic evaluation decision problem.</p>

	However, the inverse probability of censoring weights (IPCW) method is a type of MSM which is directly relevant for our context, as it attempts to estimate the treatment effect specifically for the experimental group. The method can be used to address any type of informative censoring and basically represents a method for improving upon the naive censoring approach. Instead of simply censoring patients the covariates of censored patients are taken into account in an attempt to remove selection bias. In the context of treatment crossover, the method involves artificially regarding subjects as dependently censored at the time crossover occurs. When the IPCW method is used and the cause of informative censoring is treatment crossover, past treatment history is removed from the weighting model because as soon as crossover occurs the individual is censored, and weights are only applied to patients in the control group – therefore previous treatment has no impact on the crossover decision.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	This method was developed as an alternative to SNMs – so it forms a separate type of method. It is based upon proportional hazards models rather than accelerated failure time models. The methods described are similar to those described by Hernan <i>et al</i> and Yamaguchi and Ohashi, also included in this review.
Does the method represent an extension to another method?	No.
Application	
Is there a worked example in the survival setting?	The method is applied to an RCT dataset, but little details of this are given, the emphasis is very much on the theory.
Is the example relevant?	Not applicable.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	Not applicable.
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Included Papers: Expert Advice

Reference	White IR, Walker S, Babiker AG. strbee: Randomization-based efficacy estimator. The Stata Journal 2[2], 140-150. 2002.
Origin	
Was the method developed specifically in the survival analysis context?	Yes. This paper demonstrates how to apply an RPSFTM using the statistical computer package STATA. This paper was included in the review because it extends previous research on the RPSFTM method in that it presents computer programming code demonstrating how the method could be implemented in practice – hence the authors describe <i>how</i> to apply these methods, rather than simply reporting an <i>application</i> of these methods (which would have resulted in exclusion). Hence in this table the focus is on the STATA code presented, rather than the RPSFTM method itself (details on this are available from the Robins and Tsiatis (1991) evidence table, above).
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	<p>The authors develop the strbee command to estimate the treatment effect in the presence of treatment crossover in the context of an RCT, using an RPSFTM.</p> <p>The syntax for the approach is as follows:</p> <pre>strbee treatvar [if exp] [in range] [,test(logrank wilcoxon cox weibull exponential) xo0(timevar eventvar) xol(timevar eventvar) endstudy(varname) savedta(filename[, append replace]) psimin(#) psimax(#) psistep(#) tol(#) noci trace list graph level(#) graph_options]</pre> <p><i>Treatvar</i> is the treatment arm, which must have values 0 and 1.</p> <p><i>Test</i> specifies the test to use when comparing counterfactual survival in the two treatment arms</p> <p><i>xo0(timevar eventvar)</i> is the time of crossover and indicator of crossover from treatment arm 0 to treatment arm 1. If it is not specified it is assumed no such crossover occurred.</p> <p><i>xol(timevar eventvar)</i> is the time of crossover and indicator of crossover from treatment arm 1 to treatment arm 0. If it is not specified it is assumed no such crossover occurred.</p> <p>In this thesis we are only considering treatment crossover from the control treatment to the experimental treatment, since this is the type of crossover that is most common and which is most likely to cause bias. Also, crossover in the other direction may occur due to failure of the new treatment which we do not wish to adjust for.</p> <p><i>endstudy(varname)</i> this is the potential censoring time, specified for every patient (whether they were censored or not). For individuals who are censored due to competing risks or loss to follow-up the potential censoring time should be set to the time at which they were actually censored. For all other individuals the potential censoring time should equal the length</p>

	<p>of the study minus the time at which the individual entered the study. For example, if the study length was 3 years and the individual entered the study 3 months after the study began that individual's endstudy value is 33 months.</p> <p><code>psimin(#)</code> and <code>psimax(#)</code> specify the extreme values for the treatment effect parameter ψ. The default values are -1 and 1.</p> <p><code>psistep(#)</code> this specifies the step size between the extreme values for a grid search for ψ. If 0 is chosen an interval bisection approach is taken for ψ and confidence intervals. The authors note that if the test statistic is nondecreasing in ψ the interval bisection approach can give incorrect answers.</p> <p><code>tol(#)</code> this is the convergence criterion – the program continues searching for ψ until each solution differs by less than 10^{-tol}. The default is <code>tol(3)</code>.</p> <p><code>noci</code> if this is included the program does not search for confidence limits.</p> <p><code>level(#)</code> this designates the confidence interval level. The default is <code>level(95)</code>.</p> <p><code>savedta(filename[, append replace])</code> this designates where to save the results, and whether results should append or replace existing results.</p> <p><code>trace</code> this provides extra details on recensoring and ψ and the test statistic at each iteration. It also saves a data file with created variables included, including counterfactual survival. This is useful if, for instance, the analyst wishes to go on to fit a parametric model to the re-estimated survival times for the control group.</p> <p><code>list</code> this lists the values of ψ and the test statistic.</p> <p><code>graph</code> this graphs the test statistic against ψ.</p>
What are the key assumptions of the method?	See RPSFTM, with censoring as specified by White <i>et al</i> (1999).
What are the theoretical advantages and disadvantages associated with the method?	See RPSFTM, with censoring as specified by White <i>et al</i> (1999).
What are the potential biases associated with the method?	See RPSFTM, with censoring as specified by White <i>et al</i> (1999).
Why might the method not be appropriate?	See RPSFTM, with censoring as specified by White <i>et al</i> (1999).
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	See RPSFTM, with censoring as specified by White <i>et al</i> (1999).
Does the method represent an extension to another method?	See RPSFTM, with censoring as specified by White <i>et al</i> (1999).
Application	
Is there a worked example in the survival setting?	Yes, the authors illustrate strbee applied to an HIV RCT testing zidovudine treatment, in which some control group patients crossed over.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach)?	The ITT analysis gave an estimate of ψ of -0.147 (-0.302 – 0.002). The strbee analysis gave -0.181 (-0.350 – 0.002) (maintaining the p value).
Other Issues	
Are there any other relevant characteristics associated with the method?	White has made more recent versions of strbee available on his web-page (http://www.mrc-bsu.cam.ac.uk/Software/stata.html#Software_IW). These allow the treatment effect to be interpreted as a hazard ratio, and allow covariates to be included in the analysis. An option has also been added that allows estimation using the IPE algorithm.

Reference	Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. The Stata Journal 2[2], 164-182. 2002.
Origin	
Was the method developed specifically in the survival analysis context?	Yes. This paper demonstrates how to apply an SNM using the statistical computer package STATA. This paper was included in the review because it extends previous research on the SNM method in that it presents computer programming code demonstrating how the method could be implemented in practice – hence the authors describe <i>how</i> to apply these methods, rather than simply reporting an <i>application</i> of these methods (which would have resulted in exclusion). Hence in this table the focus is on the STATA code presented, rather than the SNM method itself (details on this are available from the Robins (1998) evidence table, above).
If not, what was the original context and	Not applicable.

how has the method been adapted?	
Theoretical Suitability	
How does the method work?	<p>Sterne and Tilling (2002) developed the <code>stgest</code> command to operationalise a SNM to estimate the effect of a time-varying exposure variable, <i>expvar</i>, on survival, accounting for potential confounding caused by a list of time-varying and baseline covariates. These are specified in <i>confvars</i>, and the lagged or baseline effects (for example, the values of variables 3 months previous to the current time could also effect the treatment process) of these can also be included using the <code>lagconf()</code> and <code>baseconf()</code> options within the command.</p> <p>The syntax for the approach is as follows:</p> <pre>Stgest expvar confvars, visit(varname) [lasttime(varname) range(numlist) step(#) tol(#) lagconf(varlist) firstvis(#) baseconf(varlist) pnotcens(varname) idcens(varname) saveres(filename) replace detail round(#)] Makelag varlist, firstvis(#) visit(varname) Makebase varlist, firstvis(#) visit(varname) Gesttowb</pre> <p>The options within the command are as follows:</p> <p><code>visit(varname)</code> specifies the variable identifying the measurement occasion.</p> <p><code>lasttime(varname)</code> this is included for censoring purposes, and specifies the last time at which follow-up would have occurred for each individual had they not experienced the outcome event.</p> <p><code>range(numlist)</code> this represents the range of estimates to be considered for the causal parameter. The default in the <code>stgest</code> program is -5 to 5. Unless the <code>step()</code> option is specified the program uses an interval bisection approach to estimate the causal parameter and 95% confidence intervals.</p> <p><code>step(#)</code> if this is specified a grid search rather than interval bisection will be used to estimate the causal parameter.</p> <p><code>tol(#)</code> the interval bisection search when successive values of the causal parameter differ by less than 10^{-tol}. The default is 3.</p> <p><code>lagconf(varlist)</code> this is a list of the variables whose lagged confounding effect must be included in the analysis. The lagged value is the value from the previous visit (day). The <code>stgest</code> program creates new variables for these in the dataset.</p> <p><code>firstvis(#)</code> this is the number of the first measurement occasion after which outcome events contribute to the analysis. By default this is the minimum value of <code>visit()</code>, but if lagged confounders are used this must be at least one greater than the minimum value of <code>visit()</code></p> <p><code>baseconf(varlist)</code> this is a list of variables whose baseline confounding should be controlled for in the analysis. The program creates new variables for these in the dataset.</p> <p><code>pnotcens(varname)</code> this is a variable containing the cumulative probability of remaining uncensored by competing risks to the end of follow-up for each individual. Thus the IPCWs must be calculated prior to running <code>stgest</code>.</p> <p><code>idcens(varname)</code> this must be specified if <code>pnotcens</code> is specified, and indicates whether the individual was censored due to a competing risk. Including this indicator allows robust standard errors to be used.</p> <p><code>saveres(filename)</code> this allows results (z statistic for each value in <code>range()</code>) to be saved.</p> <p><code>replace</code> this allows new results to replace old results.</p> <p><code>detail</code> this means that detail from each regression model fitted at each iteration is displayed</p> <p><code>round(#)</code> Sterne and Tilling state that this variable is rarely required, but can be used when there are problems in creating the indicator variable used in the logistic regression of exposure on counterfactual failure time allowing for censoring.</p> <ul style="list-style-type: none"> Deriving censoring weights in the presence of informative censoring, for use with <code>pnotcens</code> <p>First the probability of being censored at each examination is modelled. If lagged variables are being included, the first examination is not included in this analysis. Outcomes are included for each relevant event, e.g.: no censoring, death due to cancer, death from another cause, lost to follow-up. A separate model is used for the final examination as there is no loss to follow-up after this stage. The interest is in the probability of censoring due to competing risks – for example this may be the probability of censoring due to death from another cause and due to loss to follow-up – i.e. everything except the occurrence of the event of interest:</p> <pre>Mlogit cens [covariates] if [examination numbers] Predict pcens2 if e(sample), outcome(2) (option p assumed; predicted probability) [this is the probability of censoring due to death from another cause. Repeat this for all relevant outcomes, e.g. pcens3 for censoring due to loss to follow-up] Gen pcens=pcens2+pcens3 [i.e. add probabilities of all relevant outcomes] Mlogit cens [covariates] if [final examination number] Predict pcens4 if e(sample), outcome(2) (option p assumed; predicted probability) [this is the probability of censoring due to death from another cause – in the final period censoring due to loss to follow-up is not included]</pre>

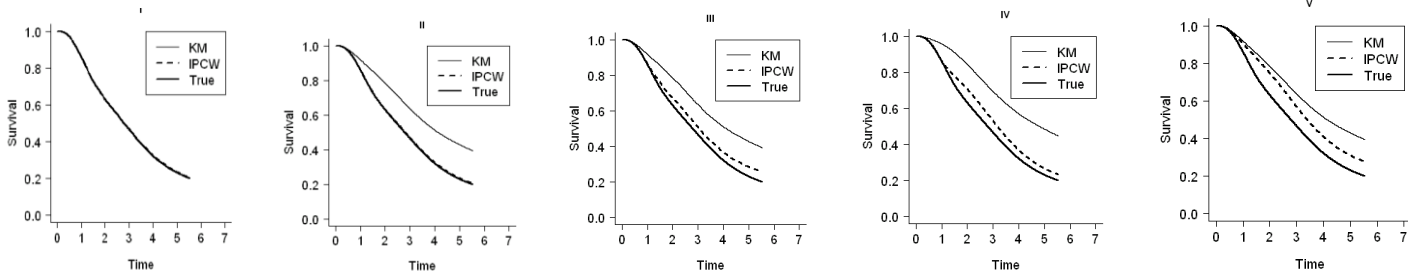
	<p>Replace <code>pcens=pcens4</code> if <code>phase==4</code> <code>replace pcens=0 if phase==1</code> [this represents setting the probability of censoring equal to 0 in the first examination period where it is assumed censoring cannot happen as lagged variables are being used (this does not have to be the case if lagged variables are not used), and setting <code>pcens</code> equal to the probabilities derived from the final examination period model for that time period] The above estimated probabilities for each examination are then used to estimate the probability of remaining uncensored for each individual to the end of the final examination period: <code>Gen lpnocens=log(1-pcens)</code> <code>egen sumpnoc=sum(lpnocens), by(id)</code> <code>Gen pnotcens=exp(sumpnoc)</code> <code>label var pnotcens "cumulative probability not censored"</code> This is then used within <code>stgest</code>, and <code>idcens</code> is used to indicate whether each individual is censored before the end of the study.</p>
What are the key assumptions of the method?	See SNM, and also IPCW for competing risks.
What are the theoretical advantages and disadvantages associated with the method?	See SNM, and also IPCW for competing risks.
What are the potential biases associated with the method?	See SNM, and also IPCW for competing risks.
Why might the method not be appropriate?	See SNM, and also IPCW for competing risks.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	See SNM, and also IPCW for competing risks.
Does the method represent an extension to another method?	See SNM, and also IPCW for competing risks.
Application	
Is there a worked example in the survival setting?	Yes. The authors apply the <code>stgest</code> command to estimate the effect of smoking on rates of heart disease based upon a longitudinal study. Data on 1756 subjects was used.
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach)?	A standard Cox model controlling for baseline covariates gave a HR of 1.61 (1.23 – 2.09). A time-dependent Cox model controlling for baseline and time-varying covariates gave a HR of 1.06 (0.71 – 1.58). <code>Stgest</code> gave an acceleration factor of 0.757 (0.67 – 0.97), which using the <code>gesttowb</code> command is approximately equivalent (using the shape parameter from a Weibull model fitted controlling for the baseline and time-dependent covariates, but not time-dependent confounding) to a HR of 1.38 (1.04 – 1.60). When censoring by competing risks was taken into account the acceleration factor from <code>stgest</code> was 0.733 (0.43 – 1.1), approximately equivalent to a HR of 1.43. (0.91 – 2.65).
Other Issues	
Are there any other relevant characteristics associated with the method?	Note, the way that the authors transform the <code>stgest</code> acceleration factor into a HR is prone to error, because the Weibull shape parameter in an analysis that does not account for time-dependent confounding may be different from the true shape parameter.

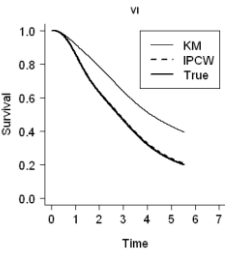
Reference	Young JG, Hernan MA, Picciotto S, Robins JM. Relation between three classes of structural models for the effect of a time-varying exposure on survival. <i>Lifetime Data Analysis</i> 16, 71-84. 2010
Origin	
Was the method developed specifically in the survival analysis context?	Yes. The paper does not present a novel method, but discusses a data generation mechanism that approximately satisfies a MSM and an SNM. Based upon this the authors discuss the assumptions that must hold in order for these methods to be unbiased. This offers useful insight on the practicalities of these methods and thus the paper was included in the review. This table focuses upon the findings of the paper in terms of when these methods cannot be expected to be unbiased, due to their limiting assumptions.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	Under the data generation mechanism developed by the authors several assumptions are made. Importantly, it is assumed that counterfactual survival times follow an exponential distribution, which allows treated survival times to follow a simply defined accelerated failure time model and a simple MSM, and the ψ from the MSM will equal the ψ from the AFT. Data

	were also generated under the assumption that the event rate was low in each interval – the authors referred to this as the “rare disease” assumption.
What are the key assumptions of the method?	Not applicable.
What are the theoretical advantages and disadvantages associated with the method?	Not applicable.
What are the potential biases associated with the method?	Not applicable.
Why might the method not be appropriate?	Not applicable.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	Not applicable.
Does the method represent an extension to another method?	Not applicable.
Application	
Is there a worked example in the survival setting?	Yes. The authors simulate data using their data generation mechanism. They simulated 1000 samples with 2500 subjects in each, with 10 observation times.
Is the example relevant?	Yes, although this is simulated data.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach)?	<p>The SNM and MSM models generally produced low bias, although bias of approximately 4% was produced by the SNM. However, bias increased for the MSM when the survival data was generated with a Weibull model rather than an exponential model (although this increase was not observed for the SNM, probably because the data were still generated under an SNM). For the MSM, bias increased to about 20% in this circumstance. The authors state that this is because when the exponential assumption is violated the data were no longer generated under a Cox MSM, while the SNM conditions were still met.</p> <p>The authors also found that when they violated another of their assumptions – the “rare disease” assumption, that is that event rates are very rare in each interval – the MSM method again is associated with bias. The authors state that in theory this should not be the case, but because it is common to apply an MSM using weighted logistic regression, there is a requirement that the rare disease condition holds in each time interval.</p> <p>The authors also found that when data were generated that met the requirements of SNMs and MSMs, the MSM had similar or less bias and smaller variance.</p>
Other Issues	
Are there any other relevant characteristics associated with the method?	None.

Reference	Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias. <i>American Journal of Epidemiology</i> 173[5], 569-577. 2011.
Origin	
Was the method developed specifically in the survival analysis context?	Yes. The authors address the limitations of the IPCW method. Although they do not present a novel method they provide useful equations for the survivor function that are not provided by Robins and Finkelstein (2000), and also present a simulation study assessing the limitations associated with the IPCW method. Therefore the paper was included in the review. The specific IPCW theory is not replicated here, as it is presented in the Robins and Finkelstein (2000) evidence table presented elsewhere in this appendix.
If not, what was the original context and how has the method been adapted?	Not applicable.
Theoretical Suitability	
How does the method work?	The authors note that the IPCW method is reliant on the “exchangeability” assumption and correct model specification. They state that “exchangeability” assumption is also referred to as the “no unmeasured confounders” assumption and depends upon three conditions: Firstly, all common predictors must be appropriately measured and accounted for in the analysis. Secondly there must be a sufficient number of participants under follow-up at all relevant times – among those at risk and those under follow-up there must be a nonzero probability of not being censored for every combination of values observed for the common predictor histories at each time point; Thirdly the common predictors cannot be deterministic or nearly

	<p>deterministic in relation to both the outcome of interest and the artificial censoring mechanism among patients over time.</p> <p>The authors state that small sample size or deterministic common predictors violate the exchangeability assumption because the outcomes that are observed among the participants who are not artificially censored in this context will likely not be representative of the unobserved outcomes among the artificially censored participants even if one appropriately measured and accounted for all common predictors. Small sample sizes, highly stratified data due to numerous common predictors, and continuous common predictors can also result in random nonpositivity. Smoothing via parametric models minimizes random nonpositivity due to continuous common predictors (7). Deterministic or nearly deterministic common predictors can result in strong induced selection bias with artificial censoring and, in turn, systematic nonpositivity.</p> <p>The authors state that when estimated weights are extreme in value of do not have a mean close to 1 model misspecification or nonpositivity is indicated. Thus selection bias may remain. It is often unrealistic to identify what is the cause of bias in high-dimensional data due to the difficulties of examining whether there are individuals who are not artificially censored for certain levels of common predictors. It is also usually not possible to identify the correctness of the specified functional form, and the no unmeasured confounders assumption cannot be tested. However, the existence of deterministic or nearly deterministic covariates can be assessed by examining the associated between common predictors and the endpoint of interest as well as the censoring mechanism. This is important because it shows the limitations of the IPCW method.</p> <p>In addition, the authors present a useful addition to the equations presented by Robins and Finkelstein (2000). They show that the survival function corrected for selection bias in the presence of informative censoring using the IPCW method can be estimated using the following equations, adapted from those presented by Robins and Finkelstein (2000) (assuming exchangeability and correct model specification):</p> $\hat{\lambda}(t_j) = \frac{\sum_{i \in D_j} \widehat{W}_i(t_j)}{\sum_{i \in R_j} \widehat{W}_i(t_j)} \quad [1]$ $\hat{S}(t) = \prod_{t_j \leq t} [1 - \hat{\lambda}(t_j)] \quad [2]$ <p>Where t_j is the time corresponding to the jth visit at which an event was observed to have occurred, R_j is the subset of the patient group for whom the minimum of their observed survival or censoring time is greater than or equal to t_j, and D_j is the subset of R_j who experience the event at t_j. $\widehat{W}_i(t_j)$ is the estimated weight for individual i at t_j, $\hat{\lambda}(t_j)$ is the estimated hazard at t_j, and $\hat{S}(t)$ is the estimated survival time at time t. It is clear that if crossover is random all weights will equal 1, and equations [1] and [2] will reduce to the standard Kaplan-Meier estimator.</p>
What are the key assumptions of the method?	See IPCW.
What are the theoretical advantages and disadvantages associated with the method?	See IPCW.
What are the potential biases associated with the method?	See IPCW.
Why might the method not be appropriate?	See IPCW.
How does the method compare to others identified (what are the similarities and differences of the method compared to others identified)?	See IPCW.
Does the method represent an extension to another method?	See IPCW.
Application	
Is there a worked example in the survival setting?	<p>Yes, there is simulated data to show the bias associated with the IPCW method in the presence of a small sample size, strong selection bias, unmeasured confounders, and model misspecification.</p> <p>For all examined scenarios, 500 simulations of sample size 50 or 500 were performed. Failure times were generated from a Weibull distribution (i.e., $S(t) \exp[-(t/\lambda)^\sigma]$) where λ and σ for the baseline survival function was 9.0 and 2.5, respectively. Failure times were generated as a function of time-fixed binary covariates z_1 and z_2 where the relative hazard of failure was specified to be 12.2 for both covariates. The prevalence of z_1 and z_2 was 50% and the proportion of failure times that were censored was 60%.</p>

	<p>Scenario (I) corresponded to a censoring mechanism that does not induce selection bias. The sample size was 500 and censoring times were generated from an exponential distribution independently of z_1 and z_2. The μ for the baseline survival function was 2.7.</p> <p>Scenario (II) corresponded to a censoring mechanism that induces selection bias. The sample size was 500 and censoring times were generated from an exponential distribution as a function of z_1 and z_2. The μ for the baseline survival function was 12.2 and the relative hazard of censoring as a function of z_1 and z_2 was 4.5.</p> <p>Scenario (III) corresponded to a censoring mechanism that induces selection bias in the context of small sample size. The sample size was 50 and censoring times were generated from an exponential distribution as a function of z_1 and z_2. The μ for the baseline survival function was 12.2 and the relative hazard of censoring as a function of z_1 and z_2 was 4.5.</p> <p>Scenario (IV) corresponded to a censoring mechanism that induces strong selection bias. The sample size was 500 and censoring times were generated from an exponential distribution as a function of z_1 and z_2. The μ for the baseline survival function was 33.1 and the relative hazard of censoring as a function of z_1 and z_2 was 12.2.</p> <p>Scenario (V) corresponded to a censoring mechanism that induces selection bias in the context of an unmeasured common predictor, z_2. The sample size was 500 and censoring times were generated from an exponential distribution as a function of z_1 and z_2. The μ for the baseline survival function was 12.2 and the relative hazard of censoring as a function of z_1 and z_2 was 4.5.</p> <p>Scenario (VI) corresponded to a censoring mechanism that induces selection bias in the context of a misspecified common predictor, z_2. The sample size was 500 and censoring times were generated from an exponential distribution as a function of z_1 and z_2. The μ for the baseline survival function was 12.2. The relative hazard of censoring as a function of z_1 and z_2 was 4.5.</p> <p>There is also a practical example of applying the IPCW method to a real-world dataset from an HIV study.</p>
Is the example relevant?	Yes.
What were the results (compared to other techniques, and compared to an intention to treat (ITT) and/or per protocol approach?	<p>The authors presented mean survival and mean squared error (MSE) for the standard KM and IPCW estimates for each of the above described simulation scenarios. Scenario (I) demonstrates that in the absence of selection bias due to censoring and a sufficiently large sample size the standard KM and IPCW estimators can be used to obtain unbiased estimates of survival with MSEs equal to the variance. However, as in scenario (II), in the presence of selection bias the standard KM estimator will likely be biased, while the IPCW estimator when necessary assumptions are met will yield an unbiased estimate of the survival function. When necessary assumptions are violated by small sample size (scenario (III)), strong selection bias (scenario (IV)), an unmeasured common predictor (scenario (V)), or model misspecification (scenario (VI)), the IPCW survival function may be biased with a substantial MSE. However like in scenario (VI), violation of necessary assumptions does not always result in biased estimates. A similar pattern was observed for the median survival for the standard KM and IPCW estimates for each of the examined simulation scenarios.</p> 

	 <p>In the real-world study, the IPCW method gave survival curves that were quite similar to the uncorrected survival curves, and thus the authors hypothesised that the method was unlikely to have corrected for the induced selection bias. The authors stated that this was likely to be due to small sample sizes ($n=52$ after 12 years, and 11 of 13 AIDS events after this time point were artificially censored), and strong induced selection bias associated with almost deterministic covariates.</p>
<p>Other Issues</p> <p>Are there any other relevant characteristics associated with the method?</p>	<p>The authors note that data pooling – that is, the use of data from other sources with similar populations to add to the available data – might solve some of the problems associated with the IPCW method, as this will increase the sample size. However, in the context of RCTs of novel cancer interventions, this may not be possible.</p> <p>It is notable that while the authors’ simulation study demonstrates the potential biases associated with the IPCW method, their simulated data are quite simple – for instance there are no time-dependent covariates – the included covariates, z_1 and z_2, were time-fixed and binary. The relative hazard of censoring as a function of z_1 and z_2 remained constant over time, and 60% of failure times were censored. With time-dependent covariates and a risk of censoring that changes over time the problems associated with applying the IPCW method may be expected to increase, especially if large proportions are censored. Hence while an n of 500 seemed reasonable in their simulations (provided that selection bias is not very strong and there are no unmeasured confounders), this may not be the case with higher proportions of censoring, more complex data generation, and if censoring proportions are very high (i.e. almost deterministic covariates).</p>

Appendix 5: Exclusion lists

Initial Search

Reference	Reason for Exclusion
<p>Cuzick J, Edwards R and Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. <i>Statistics in Medicine</i> 1997; 16:1017-1029</p>	<p>The authors present a ‘profile likelihood’ approach for adjusting for non-compliance and contamination, which respects the randomisation of the trial and is based upon identifying compliers and non-compliers in different treatment groups. However the model is developed for a situation in which endpoints are rare and where little power is lost by ignoring the time at which failures occur. The model is also primarily for a situation where the intervention is given only once, or over a very short time period. Thus the method is primarily for screening or prevention interventions. For cancer drugs, where endpoints are common, and the time to the endpoint is of key importance, and the intervention is given over time and switching (contamination) could occur at any time, the method is not suitable. The authors briefly consider circumstances where time-to-failure is important but state that this requires more technical details to develop and analyse and therefore they only give a brief description and they do not present a model for the situation in which contamination occurs during follow-up. Thus, the method is not suitable for dealing with the crossover problem as we have defined it, and thus this paper was excluded.</p>
<p>White IR, Walker S, Babiker AG and Darbyshire JH. Impact of treatment changes on the interpretation of the Concorde trial. <i>Aids</i> 1997;11;8:999-1006</p>	<p>In this paper the authors analyse data from an Aids trial in which immediate therapy was compared to delayed therapy. Many patients in the delayed therapy group received therapy before the planned time. The authors compare a simple Cox model with treatment group as a time-dependent covariate approach to the Robins and Tsiatis method for adjusting the analysis. The found that the simple approach was seriously affected by selection bias, because patients who changed treatment had a worse prognosis than those who did not change treatment. They attempted to control for this by correcting for CD4 (a key indicator of earlier treatment than scheduled) counts, but the analysis was still open to bias (even adjusting for CD4 counts the patients that changed treatment (received earlier active treatment) had poorer prognosis), and they noted that correcting using CD4 was flawed as changing CD4 counts were a consequence as well as a cause of treatment changes. The Robins and Tsiatis RPSFTM method was much preferred, although two minor disadvantages were noted: the method assumes a correct model and offers no way to check the model; and the method is computationally intensive. Because this paper is an application of a previously reviewed method, rather than a new method or an extension to an existing method, it is excluded.</p>
<p>Robins JM. Correcting for non-compliance in equivalence trials. <i>Statistics in Medicine</i> 1998; 17:269-302</p>	<p>In this paper the author compares several methods for adjusting estimates of treatment effectiveness in the presence of non-compliance. The methods considered include G-computation algorithm estimators, inverse probability of censoring weighted estimators, iterated conditional expectation estimators, G-estimation and likelihood-based estimation of rank-preserving structural models, structural nested distribution models, structural nested mean models, coarse structural nested models, non-nested marginal structural models and continuous-time nested models. The author presents an overview of each method, comparing their plausibility, robustness, strength of assumptions required, programming and computation burdens. However, the author assumes no treatment crossover and so the review of the methods is not relevant for the specific problem being addressed by this review. Thus this paper was excluded.</p>

<p>Korhonen PA, Laird NM and Palmgren J. Correcting for non-compliance in randomised trials: An application to the ATBC study. <i>Statistics in Medicine</i> 1999;18:2879-2897</p>	<p>In this paper the authors compare an ITT analysis, an as treated analysis, and a G-estimation method for estimating the treatment effect when participants drop out of the active treatment group. They undertake a simulation study and use real data. They use a RPSFTM type model, and class Robins and Tsiatis' (1991) approach a G-estimation approach (whereby the treatment effect is ascertained which would result in equal survival times in the control group and the treatment group if none were treated). This paper was excluded from the review because it only considers non-compliance in the form of drop-out from the treatment group – it does not consider switching from the control group onto the experimental treatment. Also it does not extend the RPSFTM.</p>
<p>Heller G. An adjustment for a post-randomisation variable in the comparison of two treatment for survival. <i>Statistics in Medicine</i> 2001;20:3475-3485</p>	<p>In this paper the author attempts to increase the power of a model comparing two treatments by accounting for unobserved baseline prognostic factors by employing a post-randomisation surrogate – thus it is not about adjusting for treatment crossover.</p>
<p>Duffy SW and Cuzick J. Correcting for non-compliance bias in case-control studies to evaluate cancer screening programmes. <i>Applied Statistics</i> 2002;51;2:235-243</p>	<p>This paper provides an example of a method using external data to control for non-compliance. However, the non-compliance considered is whether or not a patient is screened for cancer, and the method is developed for use on case-control studies. Therefore it is not relevant for the review.</p>
<p>Loeys T and Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomised trial with all-or-nothing compliance. <i>Biometrics</i> 2003;59:100-105</p>	<p>This paper presents an adjusted hazard ratio method for adjusting survival estimates for non-compliance. However, the compliance considered is all or nothing – i.e. if non-compliance occurs it is assumed to occur at time point zero – and the method is developed for a situation in which compliance is full in the control group, and partial in the experimental group. Although the method could be used to assess non-compliance (switching) in the control group, the method would still assume that switching occurred at time point zero. This does not fit in with the treatment crossover problem as specified in this thesis, whereby switching can occur at any time point, most likely upon disease progression, and whereby it is unlikely that switching will occur at time point zero.</p>
<p>Willan AR, Bingshu Chen E, Cook RJ and Lin DY. Incremental net benefit in randomized clinical trials with quality-adjusted survival. <i>Statistics in Medicine</i> 2003;22:353-362</p>	<p>This paper deals with methods for estimating the probability of surviving and duration of interest, mean quality-adjusted survival and costs in the presence of censoring. The paper does not address treatment crossover and so is excluded.</p>
<p>Sundstrom S, Brembes RM, Kaasa S, Aasebo U and Aamdal S. Second-line chemotherapy in recurrent small cell lung cancer: Results from a crossover schedule after primary treatment with cisplatin and etoposide (EP-regimen) or cyclophosphamide, epirubicin, and vincristin (CEV-regimen). <i>Lung Cancer</i> 2005;48:251-261</p>	<p>This paper analyses a crossover trial. The only attempt made at correcting the ITT analysis for crossover is to conduct an 'as treated' analysis which is not discussed in detail and is very likely to be prone to selection bias. Methods for dealing with crossover are not the subject of the paper and no novel methods are used or developed and thus this paper is excluded.</p>
<p>Baker SG, Fitzmaurice GM, Freedman LS and Kramer BS. Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. <i>Biostatistics</i> 2006;7;1:29-40</p>	<p>This paper concentrates on methods to adjust analyses in the presence of unobserved baseline covariates that are informative in that they are associated with both outcome and the probability that the outcome is missing or censored. The paper does not deal in any detail with crossover and the method developed is not developed for use in the presence of crossover. Very brief mention of using the method in the presence of crossover is made in the discussion, whereby groups would be compared as randomised, but data for patients who switch would be classed as missing. For the method discussed to be appropriate all informative covariates would have to be identified and there must be no other nonignorable missing data mechanism. The authors state that these are very strong assumptions. Because the method is not developed for use in a crossover context, and because if it were used in the context of crossover it would appear to be similar to the IPCW approach, this paper is excluded.</p>
<p>Hernandez AV, Eijkemans MJC and Steyerberg EW. Randomized controlled trials with time-to-event outcomes: How much does prespecified covariate adjustment increase power? <i>Annals of Epidemiology</i> 2006;16:41-48</p>	<p>In this paper the authors analyse different methods for adjusting for baseline covariates. The application is survival analysis, but no mention is made of treatment crossover and the method is not developed for dealing with crossover, thus the paper is excluded.</p>
<p>Tannen RL, Weiner MG and Xie D. Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. <i>Pharmacoepidemiology and Drug Safety</i> 2008;17:671-685</p>	<p>In this paper the authors present a method for adjusting for unmeasured confounders in the context of observational studies. The method is based upon the prior event rate ratio. It is not suitable for use in randomised controlled trials and does not deal with treatment crossover, and hence the paper is excluded.</p>
<p>Cain LE and Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of</p>	<p>This paper involves an application of Robins and Finkelstein's (2000) IPCW method. Therefore, as the method used in the paper does not extend the approach, the paper is excluded.</p>

randomized highly active antiretroviral therapy on incident AIDS or death. <i>Statistics in Medicine</i> 2009;28:1725-1738	
Holme I, Szarek M, Cater NB, Faergeman O, Kastelein JJP, Olsson AG <i>et al.</i> Adherence-adjusted efficacy with intensive versus standard statin therapy in patients with acute myocardial infarction in the IDEAL study. <i>European Journal of Cardiovascular Prevention and Rehabilitation</i> 2009;16:315-320	In this paper the authors re-estimate treatment effects to take into account different adherence levels of the two drugs included in the trial. They achieved this by setting 'high' and 'low' adherence levels and completing Cox regression analysis stratified by adherence status. The method is not developed to deal with treatment crossover, and the only indirect mention of crossover suggests that upon first cardiovascular event non-study treatments may be given in either treatment arm, and this is controlled for in one analysis by censoring data at the time of first cardiovascular event. The paper therefore does not develop a method relevant for inclusion in this review, and is excluded.
Stone A. The application of bespoke spending functions in group-sequential designs and the effect of delayed treatment switching in survival trials. <i>Pharmaceutical Statistics</i> 2010;9:151-161	In this paper the author discuss approaches for planning clinical trials when it is envisaged that treatment switching may occur. In the discussion the author very briefly considers methods for adjusting estimates of the treatment effect when switching occurs, but simply refers to other papers on the matter rather than offering a new approach (as this is not the focus of the paper). Thus this paper is excluded.

Exclusion List: Secondary Search

Reference	Reason for Exclusion
Peduzzi P, Detre K, Wittes J, Holford T. Intent-to-treat analysis and the problems of crossovers – An example from the veterans-administration coronary-bypass surgery study. <i>Journal of Thoracic and Cardiovascular Surgery</i> 1991;101;3:481-487.	This paper acknowledges the problems associated with an ITT analysis in the presence of crossover, but the methods suggested to account for it are naive. They include censoring crossovers when treatment changes, transferring crossovers from the original group to the new treatment group when treatment changes, excluding all crossovers from the analysis, and counting crossovers from the date of randomisation in the treatment ultimately received group. These are prone to selection bias. Hence, because this paper does not discuss methods that attempt to control for crossover in a way that avoids selection bias, this paper is excluded.
Goldman, AI. The cure model and time confounded risk in the analysis of survival and other time events. <i>Journal of Clinical Epidemiology</i> 1991;44;12:1327-1340	In this paper the authors compare the Kaplan-Meier estimator to the logrank test for the analysis and testing of cure model data. The paper does not deal with treatment crossover and therefore is excluded.
Sommer A and Zeger SL. On estimating efficacy from clinical trials. <i>Statistics in Medicine</i> 1991;10:45-52.	This paper appears to be one of the first attempts at developing a method for estimating 'efficacy' in clinical trials taking into account non-compliance. The authors develop a method that attempts to control for selection bias by comparing the compliers in a treatment group to an inferred control subgroup chosen to eliminate selection bias. The method involves observing outcomes in compliers and non-compliers in the treatment group, and assuming that the control group would include a group of patients with similar outcomes as the non-compliers in the treatment group. Then the true effect (in this paper, a relative risk) of the treatment can be calculated taking these groups into account. However, the method is not specifically for dealing with treatment crossover, or to deal with situations whereby compliance is not a discreet event and outcomes are 'time-to-event'. The authors state that for such circumstances a survival analysis analogue of the binary data methods discussed by them would be required, and that further research is required in this area. Hence because this paper does not develop a method that could be used for the treatment crossover problem as defined in this thesis it is excluded.
France LA, Lewis JA and Kay R. The analysis of failure time data in crossover studies. <i>Statistics in Medicine</i> 1991;10:1099-1113	The authors state that they develop a method for analysing failure time data in crossover studies. However, the context within which they provide a method is different from that being studied in this thesis. The authors analyse a crossover trial in which patients are treated for angina. They are given one treatment for a set amount of time, and are then given an exercise test, and then they are given the other treatment for a set amount of time again followed by an exercise test. The 'failure time' referred to is the time to failure in the exercise test – thus each patient has two failure times, one for each treatment. The authors compare the treatments using a Cox model with treatment and period covariates and use 'preference' scores based upon which treatment each patient did best on to compute a hazard ratio. They then use this HR to re-estimate Kaplan-Meier curves and then compare median survival. Thus this is not the type of crossover or analysis being studied in this thesis and this paper is excluded.
Robins J. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. <i>Biometrika</i> 1992;79;2:321-334	In this paper the author uses the 'strong version' of the accelerated failure time model to estimate the causal effect of a time varying treatment on time to an event in the presence of time-dependent confounding variables. This is a similar type of model as the RPSFTM, but this method is developed specifically for observational data and associated time dependent confounding variables. Thus it is not relevant for this review, with the same author having dealt specifically with methods for dealing with treatment crossover (non-compliance) in RCTs in a separate paper. Thus this paper is excluded.
Peduzzi P, Wittes J and Detre K. Analysis as-randomized and the problem of non-adherence: An example from the veterans affairs randomized trial of coronary artery bypass surgery. <i>Statistics in Medicine</i> 1993;12:1185-1195	This paper is very similar to the Peduzzi (1991) paper discussed above in this table. The authors acknowledge the problems associated with an ITT analysis in the presence of crossover, but the methods suggested to account for it are naive. They include censoring crossovers when treatment changes, transferring crossovers from the original group to the new treatment group when treatment changes, excluding all crossovers from the analysis, and counting crossovers from the date of randomisation in the treatment ultimately received group. These are prone to selection bias. Hence, because this paper does not discuss methods that attempt to control for crossover in a way that avoids selection bias, this paper is excluded.

Tudor G, Koch GG. Review of nonparametric methods for the analysis of crossover studies. <i>Statistical Methods in Medical Research</i> 1994;3;4:345-381	This paper reviews a variety of methods for estimating treatment effects in crossover trials. However, the context is not directly relevant for this review as we are interested in unplanned/non-randomised crossover. Hence, this paper is excluded.
Nieto FJ, Coresh J. Adjusting survival curves for confounders: A review and a new method. <i>American Journal of Epidemiology</i> 1996;143;10:1059-1068	In this paper the authors review methods for adjusting survival curves for covariates, typically for use in observational studies. They also develop their own method. However the paper is specifically about methods for comparing survival in treatment groups when patient populations differ, treatment crossover is not mentioned. Thus, the method is not relevant for this review and the paper is excluded.
McCall BP. The identifiability of the mixed proportional hazards model with time-varying coefficients. <i>Econometric Theory</i> 1996;12:733-738	In this paper the author establishes conditions for the nonparametric identifiability of the mixed proportional hazards model with time-varying coefficients. The paper is not about estimating survival in the presence of treatment crossover, and so the paper is excluded.
Lindsey JK, Jones B, Lewis JA. Analysis of cross-over trials for duration data. <i>Statistics in Medicine</i> 1996;15:527-535	In this paper the authors discuss methods for analysing crossover trials, rather than methods for adjusting for crossover. Primarily they consider methods for analysing crossover trials in which events are repeating, or when treatment is given for a period followed by a duration test (such as the exercise test analysed by France (1991), discussed above in this table. In these circumstances the outcomes are known for each patient with both treatments under consideration. This is not the case in the treatment crossover issue addressed in this thesis, whereby the problem is that survival in the absence of crossover is not known for crossover patients. Thus, this paper does not address the issue that is the focus of this review, and it is therefore excluded.
Schmooer C, Schumacher M. Effects of covariate omission and categorization when analysing randomized trials with the Cox model. <i>Statistics in Medicine</i> 1997;16:225-237	The authors demonstrate the effects of omitting or incorrectly categorising a relevant prognostic factor when estimating the effect of a treatment on survival based upon a randomised trial. The paper does not deal with treatment crossover and thus is not relevant for this review, and is excluded.
Goetghebeur E, Molenberghs G and Katz J. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. <i>Statistics in Medicine</i> 1998;17:341-355	Here the authors develop a method for estimating the treatment effect taking compliance into account, based on a missing data framework. However the authors assume that there is zero non-compliance in the control arm, and therefore assume that there is no treatment crossover. Thus they are dealing with situations in which there is a degree of non-compliance in the treatment arm. In this thesis we do not seek methods to adjust for this because such non-compliance could be due to factors such as toxicity. Therefore the methods developed in this paper are not relevant for the treatment crossover issue, and the paper is excluded from the review.
Baker SG. Analysis of survival data from a randomized trial with all-or-nothing compliance: Estimating the cost-effectiveness of a cancer screening program. <i>Journal of the American Statistical Association</i> 1998;93;443:929-934	In this paper the author develops a model for assessing treatment effects in the event of all-or-none compliance. He extends previous methods mainly by extending likelihood-based methodology for all-or-none compliance to the analysis of discrete-time survival data. However, the author states that his focus is only on all-or-none compliance, which is not the focus of this review. We are seeking methods that can adjust for compliance and non-compliance (treatment crossover) whereby non-compliance can occur at any point in time, as this is what is typically seen in cancer drug trials. The author of this paper focuses on a cancer screening trial, which places his research in a different context to that relevant for this review. Thus, this paper is excluded.
Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. <i>Biometrika</i> 1999;2:365-379	Similar to Baker's 1998 paper (discussed above in this table), this study develops a model for assessing treatment effects in the event of all-or-none compliance. The authors include a section in which they extend their model for use with survival outcomes. However, in developing their model the authors assume that patients in the control arm cannot receive the new intervention – non-compliance can only happen in the new treatment arm and thus crossover cannot occur. Thus the paper does not develop a method that is relevant for the problem studied in this thesis, and thus this paper is excluded.
Peduzzi P, Henderson W, Hartigan P, Lavori P. Analysis of randomized controlled trials. <i>Epidemiologic Reviews</i> 2002;24;1:26-38	This paper is a review article. Survival analysis is mentioned but treatment crossover is not. New methods are not developed. Hence the paper is excluded.
Loeys T and Goetghebeur E. Baseline information in structural failure time estimators for the effect of observed treatment compliance. <i>Statistics in Medicine</i> 2002;21:1173-1188	In this paper the authors note that under the RPSFTM method the structural parameter value which best achieves equality between treatment-free survival times is identified by comparing the plausible parameter values using weighted log-rank tests. This method protects the alpha level but gains no efficiency over the ITT analysis. They state that while the point estimator is further from the null, it has larger variance due to allowance for selective (where there is a relation between compliance and latent treatment-free survival) non-compliance. In this paper, the authors examine whether power can be safely recovered in the structural analysis by introducing baseline covariates or parametric assumptions on the baseline treatment-free survival. For this, the authors develop a procedure called Nuisance Estimation from Control (NEC). They find in simulation studies and an application to real data that their method leads to more precise (narrower confidence intervals) estimators of the structural effect. However, the authors develop their method assuming that there is non-compliance in the treatment group, but that patients in the control group cannot receive the new treatment – hence they assume no crossover. They state that their method is only applicable when there is no access to experimental treatment in the control arm, and hence the method is not relevant for this review of methods for dealing with treatment crossover. Hence the paper is excluded.
Hogan JW, Daniels MJ. A hierarchical modelling approach to analysing longitudinal data with drop-out and non-compliance, with application	In this paper the authors formulate a Bayesian hierarchical model that can be used to estimate causal effects in longitudinal clinical trials where patients spend differing amounts of time on assigned treatments, where a loss to follow-up occurs throughout the trial and where the non-compliance and drop-out processes are possibly dependent on both observed and unobserved repeated measures. However, treatment crossover is not considered and the paper is excluded.

to an equivalence trial in paediatric acquired immune deficiency syndrome. <i>Applied Statistics</i> 2002;51;1:1-21	
Dawson R, Lavori PW. Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard. <i>Statistics in Medicine</i> 2002;21:1641-1661	In this paper the authors estimate the effects of time-varying treatments on the discrete-time hazards using an inverse weighting method, in the context of observational data and antidepressant treatment for depression. The authors consider treatments that have a monotonic discontinuation pattern. Treatment crossover is not considered and hence the paper is excluded.
Diaz-Uriarte R. Incorrect analysis of crossover trials in animal behaviour research. <i>Animal Behaviour</i> 2002;63:815-822	This paper deals with methods for analysing trials that are designed to include crossover, rather than methods to adjust for the unplanned occurrence of crossover. Thus the paper is excluded from this review.
Lesaffre E, Kocmanova D, Lemos PA, Disco CMC, Serruys PW. A retrospective analysis of the effect of noncompliance on time to first major adverse cardiac event in LIPS. <i>Clinical Therapeutics</i> 2003;25;9:2431-2447	In this paper the authors attempt to correct for non-compliance (treatment discontinuation rather than crossover) using time-dependent covariates in a Cox model. This is open to selection bias and is an application of a naive method, and therefore this paper is excluded.
Yamaguchi T and Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. <i>Statistics in Medicine</i> 2004;23:2005-2022	This paper is excluded from the review of methodological papers as it represents an application of the methods developed in the other Yamaguchi and Ohashi (2004) paper. However, because this paper forms part II of the other paper, the results of the application are included in the Yamaguchi and Ohashi (2004) data extraction table included in Appendix 2.1.
Matsui S. Analysis of times to repeated events in two-arm randomized trials with noncompliance and dependent censoring. <i>Biometrics</i> 2004;60:965-976	In this paper the author develops a method using the RPSFTM framework to account for noncompliance and dependent censoring when the outcomes being measured are repeated events. This paper is excluded because repeated events are not relevant for the review presented here, as our focus is on time-to-disease progression and overall survival.
Lok J, Gill R, van der Vaart A, Robins J. Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. <i>Statistica Neerlandica</i> 2004;58;3:271-295	This paper demonstrates the need for structural nested failure time models for estimating the causal effect of a time-dependent treatment on time to an event of interest in the presence of time-dependent confounding variables. The focus is on varying the amount of a treatment received over time, rather than on treatment crossover. Also the method is for observational data, and assumes there is no censoring. Therefore the method is not applicable to the treatment crossover problem that is identified in this thesis. Hence the paper is excluded.
Barber JS, Murphy SA, Verbitsky N. Adjusting for time-varying confounding in survival analysis. <i>Sociological Methodology</i> 2004;34:163-192	This paper focuses on the application of the use of marginal structural models which use inverse probability of treatment weights to adjust for time-varying confounding in observational data survival analysis. The method discussed is the same as that developed in Hernan, Brumback and Robins (2001), which is included in the review. Hence this paper is excluded.
Hernan MA, Cole SR, Margolick R, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying covariates. <i>Pharmacoepidemiology and Drug Safety</i> 2005;14:477-491	In this paper the authors consider nested structural models and marginal structural models for adjusting survival estimates in the presence of time-dependent confounding caused by time-varying covariates which are affected by prior exposure. Papers which discuss these methods are already included in the review, so this paper is excluded.
Snapinn SM, Jiang Q, Iglewicz B. Illustrating the impact of a time-varying covariate with an extended Kaplan-Meier estimator. <i>The American Statistician</i> 2005;59;4:301-307	In this paper the authors develop a method for adjusting the Kaplan-Meier estimator to illustrate the association between a time-varying covariate (such as blood pressure over time) and the outcome. This is not relevant for adjusting survival estimates for treatment crossover, hence this paper is excluded.
Cuzick J, Sasieni P, Myles J, Tyrer J. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. <i>Journal of the Royal Statistical Society Series B-Statistical Methodology</i> 2007; 69:565-588	In this paper the author extends his previously developed method (Cuzick, Edwards, Segnan 1997). The previous method provided a way of adjusting estimates of the treatment effect in the presence of non-compliance and contamination (crossover), but did not allow for situations in which the time to event was important. In this paper the authors allow for time-to-event, but they assume that contamination occurs at the time of randomisation. This is not the case in trials of cancer drugs where crossover usually occurs upon disease progression. Hence the method developed in this paper is not relevant for the situation studied in this thesis, and the paper is excluded.

Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. <i>Clinical Trials</i> 2008;5:5-13	In this paper the authors seek to re-explain structural nested models such as the RPSFTM method in order to encourage their use in analysis of trials. The authors' focus is on non-compliance rather than crossover. The paper is excluded from this review as it does not extend the method, rather it reviews and re-states it.
Klungsoyr O, Sexton J, Sandanger I, Nygard JF. Sensitivity analysis for unmeasured confounding in a marginal structural Cox proportional hazards model. <i>Lifetime Data Analysis</i> 2009;15:278-294	In this paper the authors show how sensitivity to unmeasured confounding might be estimated in a marginal structural Cox proportional hazards model (which estimates parameters using an inverse probability weighted estimator) in the context of point exposure – where only one assessment of exposure is made. This context is not relevant for treatment crossover as defined in this thesis, and the paper is excluded.
Kerkhof M, Roobol MJ, Cuzick J, Sasieni P, Roemeling S, Schroder FH, Steyerberg EW. Effect of the correction for noncompliance and contamination on the estimated reduction of metastatic prostate cancer within a randomized screening trial (ERSPC section Rotterdam). <i>International Journal of Cancer</i> 2010;127:2639-2644	This paper applies the method developed by Cuzick <i>et al</i> (1997 and 2007, see above) to a metastatic prostate cancer screening RCT in which contamination (crossover) and non-compliance occurred. Because this paper is an application and does not extend the method, and because the original methods were excluded, this paper is excluded.
Korn EL, Freidlin B. Causal inference for definitive clinical end points in a randomized clinical trial with intervening nonrandomized treatments. <i>Journal of Clinical Oncology</i> 2010;28;24:3800-3802	This is an editorial in which the authors briefly discuss the problem of treatment crossover in cancer trials. They mention the RPSFTM method as a potential answer, as well as a method developed in the same issue of the journal by London <i>et al</i> (2010). The London <i>et al</i> (2010) method attempts to estimate the causal effect when non-randomised treatments are given in the intervening period between randomisation and the outcome of interest – death. In their example, stem-cell transplantation is given to patients with relapsed neuroblastoma depending upon their response to the two randomised chemotherapies. The authors note that the main problem with London's approach is that it is assumed that a patient's treatment arm and response are the only criteria that affect the decision as to whether the patient received transplantation. This is unrealistic. The authors note that a more common situation is where patients switch treatments upon disease progression, and it is with regard to this that they mention the RPSFTM method. However, again they note that the method requires three unverifiable assumptions – equal OS treatment effect no matter when the treatment is given; absolute OS benefit is never greater than the actual treatment time; all patients receive the same benefit from the treatment. The authors conclude that these analyses should only be exploratory and analysts should be aware of the strong unverifiable assumptions required. Because this editorial does not present a new method, it is excluded from this review.
Matsuyama Y. A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. <i>Statistics in Medicine</i> 2010;29:2107-2116	In this paper the authors compare ITT and per-protocol analyses to the RPSFTM method in the context of a simulation study of a non-inferiority trial in which treatment changes occurred. The RPSFTM method is tested rather than IPCW-type methods because they wished to avoid making assumptions about either observed or unobserved factors that influence a patient's compliance decision – thus they undertake a randomisation-based analysis. This paper is excluded because it does not extend any existing methods, or develop a new one.
Odondi L, McNamee R. Performance of statistical methods for analysing survival data in the presence of non-random compliance. <i>Statistics in Medicine</i> 2010;29:2994-3003	In this paper the authors conduct a simulation study to compare the RPSFTM method to methods developed by Loeys and Goetghebeur (2003) and White <i>et al</i> (unpublished) for analysing survival data in the presence of non-random compliance. However, although the methods compared allow for treatment switching in both directions, the authors only considered a case in which patients in the experimental treatment group switching onto the control – the experimental treatment was not available to the control group. The Loeys and Goetghebeur only allows for all or nothing compliance, and this method is not suitable for the treatment crossover problem considered in this thesis. However the RPSFTM method is included. The Odondi and McNamee paper is excluded because it does not extend existing methods or develop a new one, and also it does not consider the treatment crossover problem as identified in this thesis.
Simes J, Voysey M, O'Connell R, Glasziou P, Best JD, Scott R, Pardy C, Byth K, Sullivan DR, Ehnholm C, Keech A for the FIELD study investigators. A novel method to adjust efficacy estimates for uptake of other active treatments in long-term clinical trials. <i>PLoS One</i> 2010;5;1:e8580	In this paper the authors develop a method for adjusting estimates of the treatment effect in the presence of large imbalances in patients who commence active nonstudy medications. The reason for developing the method was the FIELD trial in the diabetes disease area, in which many patients started taking other drugs (such as statins) during the trial period. To account for this the authors applied efficacies of the nontrial drugs taken based on meta-analyses from external trials in a penalised Cox model. This paper is excluded because the method adjusts for non-trial treatments for which there is external data available. This is not a relevant context for the treatment crossover problem as identified in this thesis.

Exclusion List: Citation Search

Reference	Reason for Exclusion
Robins J. Correcting for non-compliance in randomized trials using structural nested mean models. <i>Communications in Statistics – Theory</i>	This paper introduces the method of structural nested mean models for adjusting estimates of treatment effects in the presence of non-compliance. The method could incorporate deviations from the treatment protocol, but is developed for continuous outcome measures, rather than for survival outcomes. Therefore the method described in the paper is not suitable for dealing with treatment crossover in the context of survival outcomes and so is excluded.

and Methods 1994;23;8:2379-2412	
Rochon J. Accounting for covariates observed post randomization for discrete and continuous repeated measures data. Journal of the Royal Statistical Society. Series B (Methodological) 1996;58;1:205-219	This paper presents a method for adjusting for covariates observed post-randomisation. However, the context is not survival analysis, and the method is not developed to deal with survival outcomes – thus the paper is excluded.
Goetghebeur EJT, Shapiro SH. Analysing non-compliance in clinical trials: Ethical imperative or mission impossible? Statistics in Medicine 1996; 15:2813-2826	The main focus of this paper is on discussing the importance of correcting for non-compliance. A method is discussed, but very much within the context of non-compliance in the active treatment arm, and not in the context of survival outcomes. Hence the method is not suitable for adjusting for treatment crossover in the context of survival, and the paper is excluded.
White IR and Pocock SJ. Statistical reporting of clinical trials with individual changes from allocated treatment. Statistics in Medicine 1996;15:249-262	In this paper the authors consider circumstances in which patients change treatments over time within an RCT. Changes considered include switches between randomised treatments, switches to combination treatments, switches to non-randomised treatments and switches to no treatment. The authors consider ITT analyses, on treatment analyses, and current treatment analyses, where treatment indicators were included in a time-dependent proportional hazards model. Although the authors tested adding additional parameters into their model, such as number of previous treatments, they acknowledged that doing so may indicate whether selection bias is present, but would not control for it. They do not offer methods for controlling for such selection bias. In addition their methods are developed primarily for situations in which several different types of treatment switch are possible, rather than treatment crossover as studied in this thesis. Hence, because the methods are fundamentally naive, this paper is excluded.
Goetghebeur E, Lapp K. The effect of treatment compliance in a placebo-controlled trial: Regression with unpaired data. Applied Statistics 1997;46;3:351-364	In this paper the authors use a structural nested mean model as developed by Robins (1994) (excluded in this table, above) to adjust treatment effects for compliance in a blood pressure trial. The outcome measure is a repeated measure rather than a survival outcome, and the authors assume that patients in the control arm cannot receive the new treatment – hence this paper is not relevant and is excluded.
White IR, Goetghebeur EJT. Clinical trials comparing two treatment policies: Which aspects of the treatment policies make a difference? Statistics in Medicine 1998;17:319-339	In this paper the authors use the RPSFTM method to identify the effect associated with several treatments, based upon a hypertension trial in which patients received a sequence of treatments following failure of their initial treatment. This included treatment crossover as well as additional treatments. The authors conducted two analyses, one using a univariate model and another based on a range of treatment-related covariates. They then used the RPSFTM method to assess the impact of each covariate. This could be regarded as an extension to the RPSFTM method, but it is in the context of testing the effects of several potential treatments, whereas our review is specific to treatment crossover. Hence the paper is an application of the RPSFTM method and the extension is not directly relevant for the treatment crossover problem as described in this thesis, thus the paper is not included in this review. Also, the multivariate use of the RPSFTM is included in other papers included in the review.
Albert JM. A threshold causal model for clinical trials with departures from intended treatment. Statistics in Medicine 1999;18:1615-1626	In this paper the author provides a method to estimate the threshold dose required beyond which no further treatment benefit is obtained. The context is not survival outcomes, and the method assumes no knowledge of compliance in the control group (as the method is specifically focussed on compliance in the treatment group). Therefore the method is not relevant and the paper is excluded.
Fischer-Lapp K, Goetghebeur E. Practical properties of some structural mean analyses of the effect of compliance in randomized trials. Controlled Clinical Trials 1999;20:531-546	In this paper the authors use a structural mean model to adjust for non-compliance and estimate the treatment effect. However the paper does not consider treatment crossover and concentrates specifically on compliance in the active treatment arm. Hence the paper is excluded.
Heitjan DF. Ignorability and bias in clinical trials. Statistics in Medicine 1999;18:2421-2434	In this paper the authors consider the problems associated with bias caused by issues such as non-ignorable treatment switching. They consider ITT analyses, as treated analyses, and adherers-only analyses using a Coarse-data model, and show the bias associated with these in the presence of non-ignorable confounding. The author states that if the parametric form of the non-ignorable coarsening mechanism is known, it can be adjusted for by estimating its parameter simultaneously with the parameters of the distribution of the outcome. However, the paper is not survival specific, and in the real-world example given it is assumed that crossover does not occur from the placebo group into the intervention group. Hence this paper is excluded.
White IR, Bamias C, Hardy P, Pocock S, Warner J. Randomized clinical trials with added rescue medication: some approaches to their analysis and interpretation. Statistics in Medicine 2001;20:2995-3008	In this paper the authors consider approaches for adjusting treatment effects in the presence of rescue medication being given to patients in a clinical trial. The authors specifically state that they are not dealing with treatment crossover, but some other treatment not randomised. The authors only develop methods in the context of a repeated outcome measure (activities of daily living score over time), rather than survival outcomes. Therefore the methods are not suitable for use in a survival context and the paper is excluded.
Korhonen P, Palmgren J. Effect modification in a randomized trial under non-ignorable non-compliance: an application to the alpha-tocopherol beta-carotene study. Applied Statistics 2002;51;1:115-133	In this paper the authors develop a method that allows estimation of the effects of treatment modifying factors on a survival end-point in the presence of non-compliance. The authors compare their method to the Robins and Tsiatis's RPSFTM method. However, their focus is on non-compliance in the experimental treatment arm, and they assume that the experimental treatment is not available to those in the control group. Hence, the method is not suitable for adjusting for treatment crossover, and the paper is excluded.
White IR, Carpenter J, Pocock S, Henderson RA. Adjusting treatment comparisons to account for	Here the authors suggest methods for estimating the treatment effect in the presence of various non-randomised treatments being given. The methods are developed in the context of an angina trial, in which several patients on medical management received a surgical intervention at some point, and the outcome of interest is a repeated outcome measure. The

non-randomized interventions: an example from an angina trial. <i>Statistics in Medicine</i> 2003;22:781-793	authors assume that any patient who received an unplanned surgical intervention would have had a poor outcome measure had they not received the intervention, and this forms the basis for estimating the treatment effect in the absence of such interventions. The authors state that for time-to-event outcomes, the method would involve assuming that the event would have occurred if the non-randomised treatment had not occurred. They state that this may be reasonable when considering the time taken to develop certain symptoms, but would not be reasonable when considering an outcome such as death. Hence, the method is not suitable for adjusting overall survival estimates in the presence of treatment crossover, and the paper is excluded.
Ten Have TR, Joffe M, Cary M. Causal logistic models for non-compliance under randomized treatment with univariate binary response. <i>Statistics in Medicine</i> 2003;22:1255-1283	In this paper the authors develop a method for estimating the marginal causal log-odds ratio for binary outcomes under treatment non-compliance. The method is potentially useful for adjusting for treatment crossover because patients in the control group are not assumed to not be able to receive the intervention – although in the case studies given such patients could not receive the intervention. However, the paper focuses specifically upon single binary endpoints, rather than time-to-event outcomes. Hence, the method is not suitable for survival analysis, and so is not included.
Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. <i>Epidemiology</i> 2004;15;5:615-625	This is primarily a review article, does not specifically mention treatment crossover, and does not develop new methods. Hence the paper is excluded.
Goetghebeur E, Vansteelandt S. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. <i>Statistical Methods in Medical Research</i> 2005;14:397-415	In this paper the authors review the literature on structural mean models for the analysis of exposures resulting from partial compliance in RCTs. They then discuss the potential problems associated with inaccurate measurement of how much of a treatment a patient has taken. In their consideration of mean models the authors note that they can be formed such that if the experimental treatment is available in the control group the treatment effect obtained may differ from that in the experimental group. This is worthy of consideration in crossover analyses. However, in the paper the focus is on repeated measure outcomes on a continuous scale, rather than survival outcomes. In fact the authors note that for right-censored survival data mean analysis is hard to justify and structural distribution models such as the RPSFTM offer an appropriate alternative. The authors note that these models make assumptions about the contrasts of distributions between observed and reference outcomes which are in general more demanding assumptions than those made in structural mean models, but that randomisation based inference shares with structural mean approaches the robustness property under the null hypothesis that randomisation protects. Because this paper does not consider survival outcomes in detail, because it discusses mean models which are not suitable for censored data, and because it is essentially a review article, it is excluded.
Loeys T, Goetghebeur E, Vandebosch A. Causal proportional hazards models and time-constant exposure in randomized clinical trials. <i>Lifetime Data Analysis</i> 2005;11:435-449	In this paper the authors extend their previously published causal proportional hazards model. Previously their model was limited to all-or-nothing compliance, whereas their new model allows for time-constant discrete and continuous exposure levels. However, their method assumes that the experimental treatment is not available in the control arm, and thus the method is not suitable for adjusting for treatment crossover. Hence this paper is excluded.
Vansteelandt S, Goetghebeur E. Sense and sensitivity when correcting for observed exposures in randomized clinical trials. <i>Statistics in Medicine</i> 2005;24:191-210	In this paper the authors show how generalised structural mean models can be used to correct estimates of treatment efficacy for measured non-compliance. However, the model developed involves a binary outcome measure, rather than a time-to-event outcome, and it is assumed that patients in the control group cannot receive the experimental treatment – hence the method is not suitable for dealing with treatment crossover and the paper is excluded from this review.
Walter SD, Guyatt G, Montori VM, Cook R, Prasad K. A new preference-based analysis for randomized trials can estimate treatment acceptability and effect in compliant patients. <i>Journal of Clinical Epidemiology</i> 2006;59:685-696	In this paper the authors develop a ‘preference-based analysis’ whereby patients are characterised by their preference for either of the treatments being compared in a study. The approach involves estimating the proportions of patients in various preference groups, and then estimating the outcome rates in the various groups, which provides information on the treatment effect based on compliance. This could be used to adjust estimates in the presence of treatment crossover. However the method presented in this paper focuses on a simple binary outcome rather than a time-to-event, assumes there is no missing data, and also assumes that if crossover occurs it occurs immediately. This does not reflect the treatment crossover problem defined in this thesis and therefore this paper is excluded.
Cai Z, Kuroki M, Sato T. Non-parametric bounds on treatment effects with non-compliance by covariate adjustment. <i>Statistics in Medicine</i> 2007;26:3188-3204	Here the authors derive non-parametric bounds on treatment effects by making use of observed covariate information, in the presence of non-compliance. Patients are stratified according to their covariates, and the bounds for the treatment effect are estimated for each stratum, allowing summary bounds to be developed for the treatment effect. However the method is developed for a simple binary outcome rather than for a time-to-event outcome, and the compliance appears to be all-or-nothing for each treatment arm. Hence it seems unlikely that this method is suitable for dealing with treatment crossover in metastatic cancer trials, and this paper is excluded.
Bellamy SL, Lin JY, Ten Have TR. An introduction to causal modelling in clinical trials. <i>Clinical Trials</i> 2007;4:58-73	This is a review article and does not develop any new methods, and hence it is excluded. The paper explores in detail structural mean modelling. The authors compare structural mean models to the principal stratification, instrumental variable causal modelling approach. Structural mean models are about efficacy – the effect of actually receiving treatment – whereas the principal stratification approach is about the effect of assigning a treatment in subgroups defined by compliance behaviour. Both can be used to estimate causal effects. The estimation of survival outcomes is not mentioned in relation to the principal stratification method. The context of the paper is a univariate continuous outcome measure, rather than survival.
Lui KJ, Chang KC. Five interval estimators for proportion ratio under a stratified randomized clinical trial with noncompliance. <i>Biometrical Journal</i> 2007;49;4:613-626	In this paper the authors consider five asymptotic interval estimators for the interval estimation of the the proportion ratio of probabilities of response between two treatments in an RCT that is subject to noncompliance. However the context is not time-to-event and the paper is excluded.
Lui KJ. Notes on test equality in stratified noncompliance randomized trials. <i>Drug</i>	In this paper the authors develop four asymptotic test procedures to assess the risk difference for two treatments in the context of an RCT in which noncompliance occurs. However, the context is not time-to-event and so the paper is excluded.

Information Journal 2007;41:607-618	
Chiba Y. Bounds on causal effects in randomized trials with noncompliance under monotonicity assumptions about covariates. <i>Statistics in Medicine</i> 2009;28:3249-3259	In this paper the authors consider reasonable bounds for estimates of causal effects in circumstances of treatment non-compliance, including treatment crossover. The authors note that reasonable bounds will not always be demonstrated by the ITT and the per-protocol estimates. They state that methods such as the RPSFTM will be unbiased under the null hypothesis, but may not be unbiased under a non-null hypothesis, which is the motivation for considering bounds around causal effects. However, the context within which the authors method lies is not time-to-event and it appears that a patient is defined either as a switcher or a non-switcher – crossover cannot occur over time. Hence it seems unlikely that this method will be suitable for dealing with treatment crossover.
Royston P, Parmar MKB, Altman D. Visualising length of survival in time-to-event studies: A complement to Kaplan-Meier plots. <i>Journal of the National Cancer Institute</i> 2008;100;2:92-97	In this paper the authors discuss the use of a log normal parametric model for imputing values for censored survival data. They use prognostic values for the unobserved survival time, but do not account for the fact that these may also influence the probability of being censored – hence censoring may also be dependent upon the prognostic variables and will therefore be informative. Hence this approach is not designed to take into account dependent censoring, which is very likely to be the case where patients are censored due to treatment crossover. Therefore this method is not suitable for the treatment crossover problem discussed here, and this paper is excluded.
Bond SJ, White IR. Estimating causal effects using prior information on nontrial treatments. <i>Clinical Trials</i> 2010;7:664-676	The authors state that they developed this method to deal with departures from randomised treatments – thus specifically not for the context of treatment crossover. However the method may still be of interest, since it may be reasonable to assume that patients in the control group receiving the experimental treatment after disease progression is akin to receiving a non-trial treatment. The authors develop a method using instrumental variables and prior information. However, the method was not developed for a survival context. Although the authors state that it could be extended to longitudinal data by replacing defined models with time-dependent versions, they have not yet done this. Censoring would also need to be taken into account. Therefore, though this method remains of interest, it is not yet in a form that is relevant and as such the paper is excluded.

Exclusion List: Reference Search

Reference	Reason for Exclusion
Robins J. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. <i>Mathematical Modelling</i> 1986;7:1393-1512	This paper introduces Robins’ approach to causal inference in the context of time varying exposures. Treatment crossover is not mentioned. Robins later built upon this to develop the RPSFTM method, which is specific to treatment crossover issues and RCTs, as well as observational SNMs which are included in the review. Hence this initial paper is excluded.
Robins J. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: <i>Health Service Research Methodology: A Focus on AIDS</i> . Eds: Sechrest L, Freeman H, Mulley A. Washington DC, US Public Health Service, National Centre for Health Services Research. 1989: pp 113-159	This paper introduces several methods for causal inference. The author states that the methods are suitable for estimating the effect of any potentially time-varying treatment or exposure on the time to some event of interest. In particular, the authors state that their methods are required in observational studies when there are extraneous risk factors for the event of interest that also predict subsequent exposure to treatment. Thus the context is not specific to treatment crossover in RCTs, although such a context is mentioned in the paper. However, one of the methods developed is the RPSFTM method, which Robins and Tsiatis later developed and suggested in the case of treatment crossover. The authors state that this method lacks biological plausibility due to its rank preserving nature, but that the RPSFTM is a sub-class of a more general class of models – the structural nested independence failure time models, which make no assumption about rank preservation. The authors also discuss a G-computation technique that can be used to estimate regime-specific survival under several conditions. This paper is not included in the review because the methods it develops are included in other papers which are included.
Robins J. The control of confounding by intermediate variables. <i>Statistics in Medicine</i> 1989;8:679-701	This paper is similar to Robins (1986) discussed above. The paper introduces Robins’ approach to causal inference in the context of time varying exposures, and specifically develops the extended standardized risk difference measure, which is an unbiased estimator of the overall effect of exposure in the presence of a covariate that is both a confounder and an intermediate variable. Treatment crossover is not mentioned. Robins later built upon this to develop the RPSFTM method, which is specific to treatment crossover issues and RCTs, and the observational SNMs that are included in the review. Hence this initial paper is excluded.
Efron B, Feldman D. Compliance as an explanatory variable in clinical trials. <i>Journal of the American Statistical Association</i> 1991;86;413:9-17	In this paper the authors seek to estimate the true dose-response curve using data from a trial in which compliance was imperfect. However treatment crossover is not considered, and this this paper is excluded.
Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In <i>AIDS Epidemiology: Methodological Issues</i> . Eds: Jewell NP, Dietz K, Farewell VT. Boston: Birkhauser, 1992: pp 297-331	Here the authors propose a method for adjusting for nonrandom noncompliance by estimating the the probability of becoming noncompliant at time t as a function of time-dependent prognostic factors prior to time t. A key part of the paper is demonstrating that when data are available for all time-dependent prognostic factors for mortality that independently predict censoring, then the dependence between the censoring and failure can be corrected for by replacing the Kaplan-Meier estimator, log-rank test and Cox partial likelihood estimator of the ratio of the treatment-arm-specific mortality rates by their inverse probability of censoring weighted versions (although IPCW is not specifically called IPCW in this paper). This method is developed further by Robins and Finkelstein (2000), which includes the context of treatment crossover. Hence this initial paper is excluded.
Robins J, Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. <i>Biometrika</i> 1992;79;2:311-319	In this paper the authors introduce a class of semiparametric accelerated failure time models that can be used in the presence of time-dependent covariates. Treatment crossover is not considered, and hence this paper is excluded.
Robins JM, Blevins D, Ritter G, Wulfsohn M. G-	Similar to Robins (1989) discussed above in this table, this paper discusses the use of structural nested failure time models and G-estimation to estimate the effect of an unrandomised

estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. <i>Epidemiology</i> 1992;3:319-336	treatment (prophylaxis) given as an additional treatment within an RCT comparing two separate treatment arms. This is done by estimating failure time if, contrary to fact, the patient had not received the additional treatment. The author states that the methods are necessary to control for bias in any epidemiologic study in which there exists a time-dependent risk factor for death (such as pneumocystis carinii pneumonia history) which influences subsequent exposure to the treatment under study (in this case prophylaxis), and which itself is influenced by past exposure to that treatment. This will be the case if there exists a time-dependent risk factor which is simultaneously a confounder and an intermediate variable. This paper is excluded because the relevant methods it discusses are covered by other papers that are included.
Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. <i>Biometrics</i> 1992;48:479-495	In this paper the authors present a method for estimating the causal effect of one or more treatment in the presence of confounding factors which covary with the treatments and are independent predictors of the outcome. The method is similar to a propensity score approach. The context is not survival or treatment crossover and so this paper is excluded.
Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. <i>Proceedings of the Biopharmaceutical Section, American Statistical Association</i> 1993:24-33. Alexandria, Virginia: American Statistical Association	This paper is similar to Robins and Rotnitzky (1992), included in this table above. The estimators suggested in that paper are slightly modified here. The authors demonstrate that when data are available for all time-dependent prognostic factors for mortality that independently predict censoring, then the dependence between the censoring and failure can be corrected for by replacing the Kaplan-Meier estimator, log-rank test and Cox partial likelihood estimator of the ratio of the treatment-arm-specific mortality rates by their inverse probability of censoring weighted versions (although IPCW is not specifically called IPCW in this paper). This method is developed further by Robins and Finkelstein (2000), which includes the context of treatment crossover. Hence this initial paper is excluded.
Robins JM. Analytic methods for estimating HIV-treatment and cofactor effects. In <i>Methodological Issues and AIDS Behavioural Research</i> , Eds: Ostrow DG, Kessler RC. Plenum Press, New York. 1993.	In this paper the author develops numerous methods for estimating the effect of HIV treatment taking into account cofactor impacts. The approaches are largely based upon G-estimation and structural nested models. The authors note that most of the methods that they develop are fundamentally 'observational' in that they require data on time-independent and time-dependent confounding factors – risk factors for the outcome of interest that also predict subsequent treatment with or exposure to the drug or cofactor under study. This paper is excluded because the relevant methods are covered by other papers that are included.
Finkelstein DM, Schoenfeld DA. Analysing survival in the presence of an auxiliary variable. <i>Statistics in Medicine</i> 1994;13:1747-1754	In this paper the authors use disease progression as an auxiliary variable (but state that any other variable that impacted overall survival could be used) to improve estimates of OS for patients that remain alive at then end of an AIDS trial (ie this is a method for extrapolation). The context of treatment crossover is not considered and so this paper is excluded. Other auxiliary variables methods that take account of informative/dependent censoring are included.
Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. <i>Statistics in Medicine</i> 1994;13:955-968	Similar to the paper by Finkelstein and Schoenfeld (1994) discussed above, the authors consider the use of auxiliary (surrogate) endpoints for improving estimates of OS in patients whose OS time is censored. A non-parametric approach is taken. Treatment crossover is not considered so this paper is excluded. Other auxiliary variables methods that take account of informative/dependent censoring are included.
Gray RJ. A kernel method for incorporating information on disease progression in the analysis of survival. <i>Biometrika</i> 1994;81;3:527-539	This paper considers the use of disease progression as an auxiliary variable to help in the estimation of OS for patients whose disease has progressed but who have not died. Treatment crossover is not considered so this paper is excluded. Other auxiliary variables methods that take account of informative/dependent censoring are included.
Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. <i>Journal of the American Statistical Association</i> 1996;91;434:444-455	In this paper the authors demonstrate how instrumental variables can be used to identify causal effects in the context of non-compliance. However the model developed is not for time-to-event, and treatment crossover is not specifically mentioned. Hence this paper is excluded.
Schenker N, Taylor JNG. Partially parametric techniques for multiple imputation. <i>Computational Statistics and Data Analysis</i> 1996;22:425-446	In this paper the authors consider parametric and partially parametric methods for conducting multiple imputation to estimate values for missing data. The example given is for time-to-event data, whereby right-censored estimates are re-estimated using multiple imputation. Treatment crossover is not mentioned, and this method has been developed by Hsu <i>et al</i> (2006) in a paper included in the review. Hence this paper is excluded.
Robins JM. Causal inference from complex longitudinal data, In "Latent variable modelling and applications to causality, Lecture notes in statistics (120)", Ed.: Berkane M, NY: Springer Verlag 1997:69-117	In this paper Robins summarises and unifies his previous 1986 and 1987 publications, in which he developed methods for estimating counterfactual causal inference for longitudinal studies with direct and indirect effects and time-varying treatments, confounders and concomitants (including structural nested models – there is some extension to these in the paper). As this paper is primarily a review, and relevant methods are covered by other papers included in the review, it is excluded.
Keiding N. Event history analysis and inference from observational epidemiology. <i>Statistics in Medicine</i> 1999;18:2353-2363	The authors state that "The theme of this paper is how the analysis of observational epidemiological studies may be enriched by more explicitly including the timing of the relevant events." However they state that time-dependent confounders complicate this analysis, and they consider the use of structural nested failure time models to deal with this. The RCT treatment crossover context is not considered, and the method used is not new, and hence this paper is excluded.
Keiding N, Filiberti M, Esbjerg S, Robins JM,	In this paper the author applies a structural nested failure time model using G-estimation to an observational leukaemia dataset. The context is not an RCT with treatment crossover,

Jacobsen N. The graft versus leukaemia effect after bone marrow transplantation: A cast study using structural nested failure time models. <i>Biometrics</i> 1999;55:23-28	rather it is an observational dataset with a time-dependent confounder, and the aim is not to estimate the effect of the treatment, but rather the effect of the confounder. Also, this paper is an application of an already identified method. Hence this paper is excluded.
Robins JM. Association, causation, and marginal structural models. <i>Synthese</i> 1999;121:151-179	In this paper the author describes the use of inverse probability of treatment weights to estimate the parameters of a marginal structural model for estimating the causal effects of time-varying treatments. Again, the emphasis is on observational studies and the effects of time-dependent treatments – not on treatment crossover. This method is covered by other papers included in this review, and hence this paper is excluded.
Korhonen P, Loeys T, Goetghebeur E, Palmgren J. Lifetime Data Analysis 2000;6:107-121	This paper is similar to that published by Loeys <i>et al</i> 2001, which is included in the review. The authors extend Robins and Tsiatis's 1991 method to take into account cluster randomisation. The method discussed in this paper is covered by Loeys <i>et al</i> 2001, and hence this paper is excluded.
Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. <i>Epidemiology</i> 2000;11;5:550-560	Like Robins 1999 papers, this paper introduces marginal structural models and inverse probability of treatment weights to estimate treatment effects in observational studies with exposures or treatments that vary over time. Thus the methods discussed are included in other papers included in the review, and this paper is excluded.
Murphy SA, van der Laan MJ, Robins JM. Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. <i>Journal of the American Statistical Association</i> 2001;96;456:1410-1423	In this paper the authors use a structural mean model (a model for the marginal mean of counterfactual responses) to estimate the effect of dynamic treatment regimes. A dynamic treatment regime is defined as a list of rules for how the level of treatment will be tailored through time given an individual's severity. However often treatment will be selected in an unplanned way. The authors consider a situation in which unplanned selection of the treatment level occurs (primarily in an observational setting), and use a marginal mean model to estimate a mean response to a particular dynamic treatment regime. Thus, the method attempts to deal with a situation in which compliance with a particular treatment plan, whereby dose is determined by severity, is not perfect in the intervention group. This is different from considering treatment crossover from a control group to an intervention group. Also, the model is developed in the context of measuring a mean effect, rather than a time-to-event. Hence, this paper is excluded.
Faucett CL, Schenker N, Taylor JMG. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. <i>Biometrics</i> 2002;58:37-47	In this paper the authors use CD4 counts as an auxiliary variable in an analysis of an AIDS trial. Censored data is classed as missing and survival times for the censored data are estimated using the auxiliary variable. Primarily this is an approach for extrapolation, but potentially it could be used in the presence of treatment crossover, whereby crossover patients are censored. The authors mention the possibility of using the method in the presence of dependent censoring. The method is built upon by Hsu <i>et al</i> (2006), which is included in the review. Hence this paper is excluded.
Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. <i>Biometrics</i> 2002;58;1:48-57	Here the authors develop a method for estimating the treatment effect of randomised sequences of treatments – i.e. where a patient is initially randomised between a treatment and a control, and then depending upon their response are then randomised to a second-line treatment. Treatment effects of the different treatments are combined, and this method does not represent a method for adjusting survival estimates in the presence of unplanned treatment crossover. Hence this paper is excluded.
Taylor JMG, Murray S, Hsu CH. Survival estimation and testing via multiple imputation. <i>Statistics and Probability Letters</i> 2002;58:221-232	In this paper the authors introduce a multiple imputation method for estimating survival for censored patients. The estimated survival time for a censored patient is a draw from the estimated distribution of event times amongst those at risk after the censoring time. No auxiliary variables are used. Clearly this is an unsuitable method for estimating survival for patients censored due to treatment crossover, as selection bias would be likely. Hence this paper is excluded. The multiple imputation method is built upon in a potentially more useful method developed by Hsu <i>et al</i> (2006), which is included in the review.
Murphy SA. Optimal dynamic treatment regimes. <i>Journal of the Royal Statistical Society. Series B.</i> 2003;65;2:331-366	In this paper the author considered a method to allow the dynamic treatment regime that would result in the maximal mean response to be identified. The focus is not on time-to-event outcomes, and the methodology is not developed in the context of treatment crossover. Hence this paper is excluded.
Brumback B, Greenland S, Redman M, Kiviat N, Diehr P. The intensity-score approach to adjusting for confounding. <i>Biometrics</i> 2003;59:274-285	Here the authors present an intensity score approach using a structural nested mean model and inverse probability of treatment weights to adjust for confounding in the context of an observational data set. The method is not specific to treatment crossover and the outcome considered is not time-to-event, hence this paper is excluded.
Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. <i>Biometrics</i> 2004;60:124-133	In this paper the authors derive more efficient estimators for the method presented by Lunceford <i>et al</i> (2002) (discussed above) for estimating combined treatment effects in two-stage randomisation trials. As in the Lunceford paper, the context is a situation in which a patient is initially randomised between a treatment and a control, and then depending upon their response they are randomised to a second-line treatment. Treatment effects of the different treatments are combined, and this method does not represent a method for adjusting survival estimates in the presence of unplanned treatment crossover. Hence this paper is excluded.

Appendix 6: Parametric models summarised

There are a wide range of parametric models available, and each have their own characteristics which make them suitable for different data sets. Exponential, Weibull, Gompertz, log-logistic, log normal and Generalised Gamma models may be classified as the “standard” parametric models. These models and their key characteristics are described briefly below. Further details on the properties of the individual parametric models that should be considered can be found in Collet (2003), [1] including diagrams of hazard, survivor and probability density functions which show the variety of shapes that the different models can take, depending upon their parameters. The hazard function is the event rate at time t conditional upon survival until time t . The survivor function is the probability that the survival time is greater than or equal to time t and is equivalent to $1 - F(t)$ where $F(t)$ is the probability density function, representing the probability that the survival time is less than t .

Exponential distribution

Hazard function: $h(t) = \lambda$ for $0 \leq t < \infty$ where λ is a positive constant and t is time.

Survivor function: $S(t) = \exp\left\{-\int_0^t \lambda du\right\} = e^{-\lambda t}$

The exponential distribution is the simplest parametric model as it incorporates a hazard function that is constant over time, and therefore it has only one parameter, λ . The exponential model is a proportional hazards model, which means that if two treatment groups are considered within the model, the hazard of the event for an individual in one group at any time point is proportional to the hazard of a similar individual in the other group – the treatment effect is measured as a hazard ratio. If the exponential distribution is to be used it is important to consider whether the hazard is likely to remain constant over an entire lifetime.

Weibull distribution

Hazard function: $h(t) = \lambda \gamma t^{\gamma-1}$ for $0 \leq t < \infty$ where λ is a positive value and is the scale parameter, and γ is a positive value and is the shape parameter.

Survivor function: $S(t) = \exp\left\{-\int_0^t \lambda \gamma u^{\gamma-1} du\right\} = \exp(-\lambda t^\gamma)$

The Weibull distribution can be parameterised either as a proportional hazards model (as shown in the survivor function above) or an accelerated failure time model. In an accelerated failure time model when two treatment groups are compared the treatment effect is in the form of an acceleration factor which acts multiplicatively on the time scale. Weibull models depend on two parameters – the shape parameter and the scale parameter. The Weibull distribution is more flexible than the exponential because the hazard function can either increase or decrease monotonically, but it cannot change direction. The exponential distribution is a special case of the Weibull, where $\gamma = 1$. Where $\gamma > 1$ the hazard function increases monotonically and where $\gamma < 1$ the hazard function decreases monotonically. When considering the applicability of a Weibull distribution the validity of monotonic hazards must be considered.

Gompertz distribution

Hazard function: $h(t) = \lambda e^{\theta t}$ for $0 \leq t < \infty$ where λ is a positive value and is the scale parameter, and θ is the shape parameter.

Survivor function: $S(t) = \exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t})\right\}$

Similar to the Weibull distribution the Gompertz has two parameters – a shape parameter and a scale parameter. Also similar to the Weibull distribution the hazard in the Gompertz distribution increases or decreases monotonically. Where $\theta = 0$ survival times have an exponential distribution, where $\theta > 0$ the hazard increases monotonically with time and where $\theta < 0$ the hazard decreases monotonically with time. The Gompertz distribution differs from the Weibull distribution because it has a log-hazard function which is linear with respect to time, whereas the Weibull distribution is linear with respect to the log of time. Also, the Gompertz model can only be parameterised as a proportional hazards model. When considering the applicability of a Gompertz distribution the validity of monotonic hazards must be considered.

Log-Logistic distribution

Hazard function: $h(t) = \frac{e^{\theta} \kappa t^{\kappa-1}}{1 + e^{\theta} t^{\kappa}}$ for $0 \leq t < \infty, \kappa > 0$

Survivor function: $S(t) = \{1 + e^{\theta} t^{\kappa}\}^{-1}$

The log-logistic distribution is an accelerated failure time model and has a hazard function which can be non-monotonic with respect to time. It has two parameters, θ and κ . If $\kappa \leq 1$ the hazard decreases monotonically with time, but if $\kappa > 1$ the hazard has a single mode whereby there is initially an increasing hazard, followed by a decreasing hazard. When considering the applicability of the log-logistic distribution the validity of non-monotonic hazards must be considered. Owing to their functional form, log-logistic models often result in long tails in the survivor function, and this must also be considered if they are to be used.

Log normal distribution

Hazard function: $h(t) = \frac{f(t)}{S(t)}$ for $0 \leq t < \infty$ where $f(t)$ is the probability density function of T .

Survivor function: $S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$ where Φ is the standard normal distribution function.

The log normal distribution is very similar to the log-logistic distribution, and has two parameters: μ and σ . The hazard increases initially to a maximum, before decreasing as t increases. The similarities between the logistic and normal distributions mean that the results of log-logistic models and log normal models are likely to be similar. As with log-logistic models, when considering the applicability of the log normal distribution the validity of non-monotonic hazards and potentially long tails in the survivor function must be considered.

Generalised Gamma

Hazard function: $h(t) = f(t)/S(t)$ where $f(t)$ is the probability density function of T .

Survivor function: $S(t) = 1 - \Gamma_{(\lambda t)^\theta}(\rho)$ where $\Gamma_{\lambda t}(\rho)$ is known as the incomplete gamma function.

The Generalised Gamma distribution is a flexible three-parameter model, with parameters λ , ρ and θ . It is a generalisation of the two parameter gamma distribution and it is useful because it includes the Weibull, exponential and log normal distributions as special cases, which means it can help distinguish between alternative parametric models. θ is the shape parameter of the distribution and when this equals 1 the generalised gamma distribution is equal to the

standard gamma distribution. When p equals 1 the distribution is the same as the Weibull distribution and as p becomes closer to infinity the distribution becomes more and more similar to the log normal distribution. Hence when a generalised gamma model is fitted the resulting parameter values can signify whether a Weibull, exponential, Gamma or log normal model may be suitable for the observed data.

Piecewise Models

Piecewise parametric models are more flexible than individual parametric models and provide a simple way for modelling a variable hazard function. They are generally referred to as piecewise constant models, as typically exponential models are fitted to different time periods, with each time period having a constant hazard rate.[2] Piecewise constant models are particularly useful for modelling datasets in which variable hazards are observed over time. Models other than the exponential also allow for non-constant hazards over time, but in the case of Weibull and Gompertz models the hazard must be monotonic, and in the case of log-logistic and log normal models the hazard is unimodal. Piecewise constant models do not restrict the hazard in this way. However, these models are less useful for the extrapolated portion of the survival curve, since in this portion hazards are not observed. Thus, as an alternative to the piecewise constant model consideration could be given to using a different parametric model (such as a Weibull, Gompertz, log-logistic, log normal or Generalised Gamma model) for the extrapolated portion of the survival curve, although an exponential should also be considered if it is deemed appropriate to extrapolate with a constant hazard rate. .

Other Models

Alongside the standard parametric models and piecewise models discussed above there are various other more weakly structured, flexible models available – such as Royston and Parmar’s spline-based models.[3] These have not been used in NICE Appraisals as yet, but are potentially very useful. They are flexible parametric survival models that resemble generalised linear models with link functions. In simple cases these models can simplify to Weibull, Log-logistic or log normal distributions – which demonstrates their flexibility and usefulness in discriminating between alternative parametric models. Jackson *et al* (2010) discuss and implement other flexible parametric distributions, such as the Generalised F – which has four parameters and which simplifies to the Generalised Gamma distribution when one of those parameters tends towards zero – as well as Bayesian semi-parametric models which allow an arbitrarily flexible baseline hazard, and which are extrapolated by making assumptions about the future hazard (ideally based upon additional data or expert judgement).[4] These more flexible methods have not been used in NICE Appraisals as yet, but Jackson *et al* provide a helpful case study of the application of these methods, and the determination of best fitting models.

References

1. Collett D. Modelling Survival Data in Medical Research, 2nd ed. Texts in Statistical Science. 2003. Boca Raton, Chapman & Hall/CRC CRC Press LLC.
2. Billingham LJ, Abrams KR and Jones DR. Methods for the analysis of quality-of-life and survival data in health technology assessment. Health Technology Assessment 1999;3;10.
3. Royston P and Parmar MKB. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 2002; 21:2175-2197.
4. Jackson CH, Sharples LD, Thompson SG. Survival models in health economic evaluations: Balancing fit and parsimony to improve prediction. The International Journal of Biostatistics 2010; 6;1;34.

Appendix 7: STATA ado file code for simulating initial survival times

Program written by Michael Crowther. Reproduced with permission.

```

program define survsimtrt
    syntax newvarname(min=1 max=2),          N(string)          ///          -Number of survival times to simulate-
                                          LAMBDAAs(numlist)  ///          -Scale parameters for Weibull's-
                                          GAMMAs(numlist)  ///          -Shape parameters for Weibull's-
                                          SCOVariates(string)  ///          -Baseline covariates, e.g. (sex 0.5 race -0.4)-
                                          CENTOL(real 0.0001)  ///          -Tolerance of Newton-Raphson iterations-
                                          /* Joint model - standard Weibull */  ///
                                          BETAS(numlist min=2 max=2)  ///          -Intercept and slope values-
                                          SD(real 1)          ///          -Standard deviations of random effect-
                                          ALPHA(real 0)        ///
                                          LCOVariates(string)  ///
                                          LINTERaction(string)  ///
                                          TDE(string)          ///
                                          NLS                  ///
                                          ///

    local newvarname `varlist'              // ignore
    local nvars : word count `varlist'      // ignore

    local nlambda : word count `lambdas'    // ignore
    local ngamma : word count `gammas'      // ignore
/*****
/* Baseline covariates */
/* Longitudinal covariates */
/* This bit creates the linear predictor for the baseline covariates in the cea model, so say lcovariates is (badprog 0.5 trt -0.5) then this will
create 2 variables to add into the linear predictor:
longvareffect1 = 0.5*badprog
longvareffect2 = -0.5*treat
So the local macros `longcov_linpred' will contain "longvareffect1 + longvareffect2" */
    if "`lcovariates'"!="" {
        tokenize `lcovariates'
        local nlongcovlist : word count `lcovariates'
        local nlongcovvars = `nlongcovlist'/2
        cap confirm integer number `nlongcovvars'
        if _rc>0 {
            di as error "Variable/number missing in covariates"
            exit 198
        }
        local ind = 1
        forvalues i=1/^nlongcovvars' {
            cap confirm var ``ind'
            if _rc {
                local errortxt "invalid lcovariates(... ``ind' ``='ind'+1' ...)"
                local error = 1
            }
            cap confirm num ``='ind'+1'
            if _rc {
                local errortxt "invalid lcovariates(... ``ind' ``='ind'+1' ...)"
                local error = 1
            }
            tempvar longvareffect`i'
            gen double `longvareffect`i' = ``ind'*``='ind'+1'
            local ind = `ind' + 2
        }
        if ``error'=="1" {
            di as error ``errortxt'
            exit 198
        }
        local longcov_linpred ``longvareffect1'
        if `nlongcovvars'>1 {
            forvalues k=2/^nlongcovvars' {
                local longcov_linpred ``longcov_linpred' + `longvareffect`k''
            }
        }
        local longcov_linpred "+ `longcov_linpred'"
    }
    if ""interaction""!="" {
        /* This bit creates the variable that interacts with log(time) in the longitudinal linear predictor and multiplies it by the
        parameter value alpha which is the coefficient for cea in the survival model: so if linter(trt -0.5), `extra' will contain alpha * trt *
        -0.5. Multiplying this by log(time) comes into the gen `nr_time' equation below */
        tokenize `interaction'
        tempvar trtime
        gen `trtime' = `1'*^2'
        local extra "+ `alpha'*`trtime'"
    }

```

```

if "`tde'"!="" {
    tokenize `tde'
    tempvar trttde
    gen `trttde' = `1'*2'
    local tdeeffect "+ `trttde'"
}
/* Survival covariates */
/* This bit does the same as longitudinal covariates above but for the baseline covariates in the survival linear predictor. Stores the
names of the generated variables in survcov_linpred*/
if ""scovariates""!="" {
    tokenize `scovariates'
    local nsurvcovlist : word count `scovariates'
    local nsurvcovvars = `nsurvcovlist'/2
    cap confirm integer number `nsurvcovvars'
    if _rc>0 {
        di as error "Variable/number missing in scovariates"
        exit 198
    }
    local ind = 1
    forvalues i=1/^nsurvcovvars' {
        cap confirm var ``ind"
        if _rc {
            local errortxt "invalid scovariates(... ``ind" ``= `ind'+1" ...)"
            local error = 1
        }
        cap confirm num ``= `ind'+1"
        if _rc {
            local errortxt "invalid scovariates(... ``ind" ``= `ind'+1" ...)"
            local error = 1
        }
        tempvar survvareffect`i'
        gen double `survvareffect`i" = ``ind"*``= `ind'+1"
        local ind = `ind' + 2
    }
    if ""error""="1" {
        di as error ""errortxt"
        exit 198
    }
    local survcov_linpred ""survvareffect1"
    if `nsurvcovvars'>1 {
        forvalues k=2/^nsurvcovvars' {
            local survcov_linpred ""survcov_linpred' + `survvareffect`k""
        }
        local survcov_linpred2 ""* exp(`survcov_linpred)""
    }
}
/*****
/* Preliminaries */
cap set obs `n' //n is the number of survival times to generate
tempvar lhs u
qui gen `u' = runiform()
qui gen double `lhs' = 1-`u' //This generates the random deviates from a U(0,1) distribution, representing the cumulative distribution
function.
/*****
tempvar nr_time
/* Joint model */
local lambdastart : word 1 of `lambdas' //Extracts value of lambda for the baseline weibull distribution
local gammastart : word 1 of `gammas' //Extracts value of gamma for the baseline weibull distribution
forvalues i=1/2 {
    local b`i' : word `i' of `betas' //b1 will contain the mean of the random intercept for the cea model, b2
will contain the value of the fixed slope
}
tempvar slope
gen intercept = normal(`b1',`sd') `longcov_linpred' //Generate the random intercept values for antigen
gen `slope' = `b2' //Generate slope variable
/*****
/* Calculate survival times */
tempvar test1 test2
gen double `test1' = -log(`lhs')
gen double `test2' = `test1'*(`gammastart'+ `alpha' * `b2' `extra' `tdeeffect')/(`lambdastart'*`gammastart' `survcov_linpred2'*exp(`alpha' *intercept))
gen double `nr_time' = (`test2') ^ (1/(`gammastart'+ `alpha'*`b2' `extra' `tdeeffect'))
/*****
/* Final variables */
qui gen double `newvarname' = `nr_time'

```

end

Appendix 8: STATA do file code for running simulation study

```

cd "C:\simulationsv5"
capture program drop simstudyv501
program define simstudyv501, rclass
version 10.1
syntax [, obs(int 500) bprog(real 0.5) lambdasim(real 0.0005) betain(real 20) betasl(real 15) alphasim(real 0.02)
bprogin(real 5) trtlghr(real -0.7) bprogsim(real 0.5) lintertr(real -4) tde(real 0.15) admin(real 1095) logitcea31(real
0.1) logitcea32(real 0.3) logitcea33(real 0.6) logcea21(real 2.2) logcea22(real 1.6) logcea23(real 2) logcea11(real
3.8) logcea12(real 2.4) logcea13(real 5) logxo2(real 0.7) logxo3(real 0.4) gammasim(real 0.9) xomult(real
0.50267)]
***Note, for each of the different scenarios the above syntax is altered***
clear
adopath ++ "C:\ado\survsimtrt"
/* Joint model */
clear
pr drop _all
set obs `obs'
gen trtrand = rbinomial(1,0.5)
gen bprog = rbinomial(1,`bprog')
survsimtrt timeOS, n(`obs') lambdasim(`lambdasim') gammasim(`gammasim') betasim(`betain' `betasl') sd(1)
alpha(`alphat') lcovariates(bprog `bprogin') scovariates(trtrand `trtlghr' bprog `bprogsim') lintertr(`lintertr')
tde(trtrand `tde')
***estimate average treatment effect with no crossover or censoring***
gen id = _n
gen dead=1

preserve
replace dead=timeOS<`admin'
replace timeOS=`admin' if timeOS>`admin'
stset timeOS, failure(dead) id(id)
***then find true treatment effect without bprog as just want overall tx effect(?)***
stcox trtrand
return scalar truecox_hr = exp(_b[trtrand])
return scalar truecox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar truecox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar truecox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
stci if trtrand==0, rmean
return scalar trueauc_con = r(rmean)
return scalar trueauc_con_SE = r(se)
return scalar trueauc_con_LB = r(lb)
return scalar trueauc_con_UB = r(ub)
stci if trtrand==1, rmean
return scalar trueauc_int = r(rmean)
return scalar trueauc_int_SE = r(se)
return scalar trueauc_int_LB = r(lb)
return scalar trueauc_int_UB = r(ub)
streg trtrand, dist(weibull) time
return scalar trueweib_oft_af = exp(_b[trtrand])
return scalar trueweib_oft_af_SE = exp(_b[trtrand])*_se[trtrand]
return scalar trueweib_oft_af_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar trueweib_oft_af_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
restore
stset timeOS, failure(dead) id(id)

```

the below generates times to disease progression. it is assumed that disease progression happens at some point between randomisation and death, with it most likely to happen at some point in the middle of this time period, and with it relatively less likely to happen either very close to randomisation or death. This reflects the uncertainty around the relationship between PFS and OS

because OS is determined by treatment group, prognosis and cea, PFS will also reflect this as it is a function of OS

```
gen timePFS = timeOS*rbeta(5,5)
```

need to also calculate when PFS is observed, as this is what will dictate treatment decisions, and will only occur at scheduled appointments. Hence need to split data here

```
stsplit timeOS2, every(21)
```

```
sort id
```

```
by id: gen PFSobsind=1 if timePFS<timeOS2
```

```
by id: gen timePFSobst=timeOS2 if PFSobsind==1
```

```
by id: egen timePFSobs=min(timePFSobst)
```

now have this, can collapse the dataset again.

```
collapse (max) trtrand bprog timeOS timePFS timePFSobs intercept, by(id)
```

below we estimate underlying cea levels at the time of disease progression, and at the two following consultations.

rnormal reflects a random error in the cea term

***note we only need to estimate actual cea values for the control group since it is only these that crossover.

Therefore we do not need to include a treatment term here***

```
gen cea1 = intercept + `betasl'*log(timePFSobs) + `bprogin'*bprog + mnormal(0,2)
```

```
gen cea2 = intercept + `betasl'*log(timePFSobs+21) + `bprogin'*bprog + mnormal(0,2)
```

```
gen cea3 = intercept + `betasl'*log(timePFSobs+42) + `bprogin'*bprog + mnormal(0,2)
```

the below cuts underlying cea into three groups based on 33.3% centiles. This allows us to allow the probability of treatment crossover to depend upon the category of cea levels at the point of disease progression and at the following two consultations

```
egen cea0grp = cut(intercept), group(3)
```

```
egen cea1grp = cut(cea1), group(3)
```

```
egen cea2grp = cut(cea2), group(3)
```

```
egen cea3grp = cut(cea3), group(3)
```

```
egen timePFSobsgrp = cut(timePFSobs), group(3)
```

the below dictates that crossover is most likely if proptime is high combined with a low cea score upon progression, reflecting that those who have performed well, and have the greatest expected capacity to benefit, are the ones that crossover

```
gen p1 = invlogit(logit(`logitcea31') + log(`logcea21')*(cea1grp==1) + log(`logcea11')*(cea1grp==0))
```

if timePFSobsgrp==0

```
replace p1 = invlogit(logit(`logitcea32') + log(`logcea22')*(cea1grp==1) +
```

```
log(`logcea12')*(cea1grp==0)) if timePFSobsgrp==1
```

```
replace p1 = invlogit(logit(`logitcea33') + log(`logcea23')*(cea1grp==1) +
```

```
log(`logcea13')*(cea1grp==0)) if timePFSobsgrp==2
```

```
gen xo1 = rbinomial(1,p1) if trtrand==0 & timeOS>timePFSobs
```

```
gen xotime = timePFSobs if xo1==1
```

the below dictates that crossover becomes more unlikely in the second and third consultations after progression

```
gen p2 = invlogit(logit(`logitcea31') + log(`logcea21')*(cea2grp==1) + log(`logcea11')*(cea2grp==0)
```

```
+log(`logxo2')) if timePFSobsgrp==0
```

```
replace p2 = invlogit(logit(`logitcea32') + log(`logcea22')*(cea2grp==1) +
```

```
log(`logcea12')*(cea2grp==0) +log(`logxo2')) if timePFSobsgrp==1
```

```
replace p2 = invlogit(logit(`logitcea33') + log(`logcea23')*(cea2grp==1) +
```

```
log(`logcea13')*(cea2grp==0) +log(`logxo2')) if timePFSobsgrp==2
```

```

***altered below so that switch2 can only occur if patient lives to see their second consultation after
progression***
    gen xo2 = rbinomial(1,p2) if trtrand==0 & xo1 == 0 & timeOS > timePFSobs+21
    replace xotime = timePFSobs +21 if xo2==1
    gen p3 = invlogit(logit(`logitcea31') + log(`logitcea21')*(cea3grp==1) + log(`logitcea11')*(cea3grp==0)
+log(`logxo3')) if timePFSobsgrp==0
    replace p3 = invlogit(logit(`logitcea32') + log(`logitcea22')*(cea3grp==1) +
log(`logitcea12')*(cea3grp==0) +log(`logxo3')) if timePFSobsgrp==1
    replace p3 = invlogit(logit(`logitcea33') + log(`logitcea23')*(cea3grp==1) +
log(`logitcea13')*(cea3grp==0) +log(`logxo3')) if timePFSobsgrp==2
***altered below so that switch2 can only occur if patient lives to see their second consultation after
progression***
    gen xo3 = rbinomial(1,p3) if trtrand==0 & xo1 == 0 & xo2==0 & timeOS > timePFSobs+42
    replace xotime = timePFSobs +42 if xo3==1
    gen xo= 1 if (xo1==1 | xo2==1 | xo3==1)
    gen xoOSgainobs = timeOS-xotime if xo==1
***we then work out what the cea level was at the point of treatment switch for each patient ***
    gen cea_xotime = intercept + `betas1'*log(xotime) + `bprogin'*bprog
***-0.5 is the log hazard ratio, 0.005 is alpha, and 0.5 is gamma***
    replace xoOSgainobs = xoOSgainobs*(exp(-(`trtlghr')/(`gammasim')))*`xomult'
    gen timeOS2 = cond(xo==1, xoOSgainobs + xotime,timeOS)
***we censor assuming a max follow-up time of 3 years***
    gen died=timeOS2<`admin'
    replace timeOS2=`admin' if timeOS2>`admin'
    replace xoOSgainobs=. if xotime>`admin'
    replace xo=. if xotime>`admin'
    replace xotime=. if xotime>`admin'
    replace timePFSobsgrp=0 if timePFSobsgrp==.
    stset timeOS2, failure(died) id(id)
***split the data every 3 weeks to create a panel***
    stsplit timeOS3, every(21)
    sort id
    by id: gen obsno=_n
    tsset id obsno
**generate an observed cea for each observation which reflect measured cea (which may differ to true underlying
cea due to random and individual error)***
    gen obscea = intercept + `betas1'*log(_t0) + `bprogin'*bprog - `lintert'*trtrand*log(_t0) + rnormal(0,2)
    replace died=0 if died==.
    sort id
    by id: replace obscea = intercept if _n==1
by id: gen finalobs = 0
    by id: replace finalobs = 1 if _n==_N
    gen cens=0
    replace cens=1 if finalobs==1 & _t==`admin' & died==0
    replace died=. if cens==1
***gen xo proportion***
    by id: egen xoind=max(xo)
    replace xoind=0 if xoind==.
    by id: egen xoprop=max(timeOS3)
    replace xoprop=xoprop/21
    replace xoprop=xoprop+1
    replace xoprop=xoind/xoprop
    summ xoprop if trtrand==0

```

```

    return scalar xo_number=r(mean)*r(N)
    summ cens
    return scalar cens_number=r(mean)*r(N)
***Naive - ITT***
***area under the curve***
stci if trtrand==0, rmean
return scalar itt_auc_con =r(rmean)
return scalar itt_auc_con_SE = r(se)
return scalar itt_auc_con_LB = r(lb)
return scalar itt_auc_con_UB = r(ub)
gen cov_itt_auc_con = 0
replace cov_itt_auc_con = 1 if return(trueauc_con) > return(itt_auc_con_LB) & return(trueauc_con) <
return(itt_auc_con_UB)
return scalar cov_itt_auc_con = cov_itt_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar itt_cox_hr = exp(_b[trtrand])
return scalar itt_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar itt_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar itt_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_itt_cox_hr = 0
replace cov_itt_cox_hr = 1 if return(truecox_hr) > return(itt_cox_hr_LB) & return(truecox_hr) <
return(itt_cox_hr_UB)
return scalar cov_itt_cox_hr = cov_itt_cox_hr[1]
***Weibull ITT***
streg trtrand, dist(weibull) time
return scalar itt_weib_aft_af = exp(_b[trtrand])
return scalar itt_weib_aft_af_SE = exp(_b[trtrand])*_se[trtrand]
return scalar itt_weib_aft_af_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar itt_weib_aft_af_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_itt_weib_aft_af = 0
replace cov_itt_weib_aft_af = 1 if return(itt_weib_aft_af_LB)<return(trueweib_aft_af) &
return(itt_weib_aft_af_UB)>return(trueweib_aft_af)
return scalar cov_itt_weib_aft_af = cov_itt_weib_aft_af[1]
***Naive - Exclude switchers.***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens xoind, by(id)
drop if xoind==1
stset timeOS2, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar ppxc_auc_con = r(rmean)
return scalar ppxc_auc_con_SE = r(se)
return scalar ppxc_auc_con_LB = r(lb)
return scalar ppxc_auc_con_UB = r(ub)
gen cov_ppxc_auc_con = 0
replace cov_ppxc_auc_con = 1 if return(trueauc_con) > return(ppxc_auc_con_LB) & return(trueauc_con) <
return(ppxc_auc_con_UB)
return scalar cov_ppxc_auc_con = cov_ppxc_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar ppxc_cox_hr = exp(_b[trtrand])
return scalar ppxc_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]

```

```

return scalar ppxc_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar ppxc_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_ppxc_cox_hr = 0
replace cov_ppxc_cox_hr = 1 if return(truecox_hr) > return(ppxc_cox_hr_LB) & return(truecox_hr) <
return(ppxc_cox_hr_UB)
return scalar cov_ppxc_cox_hr = cov_ppxc_cox_hr[1]
restore
***Naive – censor switchers***
gen xoti=0
replace xoti=1 if xo==1 & timeOS3>=xotime
gen pfsti=0
replace pfsti=1 if timeOS3>=timePFSobs
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens xoind, by(id)
replace died=0 if xoind==1
replace timeOS2=xotime if xoind==1
stset timeOS2, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar ppcens_auc_con = r(rmean)
return scalar ppcens_auc_con_SE = r(se)
return scalar ppcens_auc_con_LB = r(lb)
return scalar ppcens_auc_con_UB = r(ub)
gen cov_ppcens_auc_con = 0
replace cov_ppcens_auc_con = 1 if return(trueauc_con) > return(ppcens_auc_con_LB) & return(trueauc_con) <
return(ppcens_auc_con_UB)
return scalar cov_ppcens_auc_con = cov_ppcens_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar ppcens_cox_hr = exp(_b[trtrand])
return scalar ppcens_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar ppcens_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar ppcens_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_ppcens_cox_hr = 0
replace cov_ppcens_cox_hr = 1 if return(truecox_hr) > return(ppcens_cox_hr_LB) & return(truecox_hr) <
return(ppcens_cox_hr_UB)
return scalar cov_ppcens_cox_hr = cov_ppcens_cox_hr[1]
restore
***Naive: Standard Cox Model with Treatment as a Time-dependent Covariate, with and without other covariates
taken into account***
stset timeOS2, failure(died) id(id)
gen trtnew=0
replace trtnew=1 if trtrand==0 & xoti==1
replace trtnew=1 if trtrand==1
***with other time-dependent covariates included***
egen obsceagr = cut(obscea), group(6)
stcox trtnew bprog cea0grp pfsti
return scalar tdc_m_cox_hr = exp(_b[trtnew])
return scalar tdc_m_cox_hr_SE = exp(_b[trtnew])*_se[trtnew]
return scalar tdc_m_cox_hr_LB = exp((_b[trtnew])-(1.96*_se[trtnew]))
return scalar tdc_m_cox_hr_UB = exp((_b[trtnew])+(1.96*_se[trtnew]))
gen cov_tdc_m_cox_hr = 0

```

```

replace cov_tdc_m_cox_hr = 1 if return(truecox_hr) > return(tdc_m_cox_hr_LB) & return(truecox_hr) <
return(tdc_m_cox_hr_UB)
return scalar cov_tdc_m_cox_hr = cov_tdc_m_cox_hr[1]
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(tdc_m_cox_hr))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+suvfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdc_m_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(tdc_m_cox_hr_LB))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+suvfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdc_m_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve

```

```

collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(tdcm_cox_hr_UB))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcm_adj_auc_con_UB=r(mean)
gen cov_tdcm_adj_auc_con = 0
replace cov_tdcm_adj_auc_con = 1 if return(trueauc_con) > return(tdcm_adj_auc_con_LB) &
return(trueauc_con) < return(tdcm_adj_auc_con_UB)
return scalar cov_tdcm_adj_auc_con = cov_tdcm_adj_auc_con[1]
restore
***treatment as time dependent variable, with tdc and with weibull to get an AF***
streg trtnew bprog cea0grp pfsti, dist(weibull) time
return scalar tdcm_weib_af = exp(_b[trtnew])
return scalar tdcm_weib_af_SE = exp(_b[trtnew])*_se[trtnew]
return scalar tdcm_weib_af_LB = exp((_b[trtnew])-(1.96*_se[trtnew]))
return scalar tdcm_weib_af_UB = exp((_b[trtnew])+(1.96*_se[trtnew]))
gen cov_tdcm_weib_af = 0
replace cov_tdcm_weib_af = 1 if return(tdcm_weib_af_LB)<return(trueweib_aft_af) &
return(tdcm_weib_af_UB)>return(trueweib_aft_af)
return scalar cov_tdcm_weib_af = cov_tdcm_weib_af[1]
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(tdcm_weib_af) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(tdcm_weib_af)
gen auc=0
return scalar tdcm_adj_auc_con_UB=r(mean)
gen cov_tdcm_adj_auc_con = 0

```

```

gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcm_we_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(tdcm_weib_af_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(tdcm_weib_af_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcm_we_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(tdcm_weib_af_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(tdcm_weib_af_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcm_we_adj_auc_con_UB=r(mean)
gen cov_tdcm_adj_auc_con = 0

```

```

replace cov_tdc_m_we_adj_auc_con = 1 if return(trueauc_con) > return(tdc_m_we_adj_auc_con_LB) &
return(trueauc_con) < return(tdc_m_we_adj_auc_con_UB)
return scalar cov_tdc_m_we_adj_auc_con = cov_tdc_m_we_adj_auc_con[1]
restore
***without other time-dependent covariates included***
stcox trtnew
gen tdc_s_cox_hr = exp(_b[trtnew])
return scalar tdc_s_cox_hr = tdc_s_cox_hr[1]
return scalar tdc_s_cox_hr_SE = exp(_b[trtnew])*_se[trtnew]
return scalar tdc_s_cox_hr_LB = exp((_b[trtnew])-(1.96*_se[trtnew]))
return scalar tdc_s_cox_hr_UB = exp((_b[trtnew])+(1.96*_se[trtnew]))
gen cov_tdc_s_cox_hr = 0
replace cov_tdc_s_cox_hr = 1 if return(truecox_hr) > return(tdc_s_cox_hr_LB) & return(truecox_hr) <
return(tdc_s_cox_hr_UB)
return scalar cov_tdc_s_cox_hr = cov_tdc_s_cox_hr[1]
***AUC***
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(tdc_s_cox_hr))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+suvrfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdc_s_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1

```

```

gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(tdc_s_cox_hr_LB))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+suvrfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdc_s_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(tdc_s_cox_hr_UB))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+suvrfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdc_s_adj_auc_con_UB=r(mean)
gen cov_tdc_s_adj_auc_con = 0
replace cov_tdc_s_adj_auc_con = 1 if return(trueauc_con) > return(tdc_s_adj_auc_con_LB) & return(trueauc_con)
< return(tdc_s_adj_auc_con_UB)
return scalar cov_tdc_s_adj_auc_con = cov_tdc_s_adj_auc_con[1]
restore
***without other time-dependent covariates included - weibull***
streg trtnew, dist(weibull) time
gen tdc_s_weib_af = exp(_b[trtnew])
return scalar tdc_s_weib_af = tdc_s_weib_af[1]
return scalar tdc_s_weib_af_SE = exp(_b[trtnew])*_se[trtnew]
return scalar tdc_s_weib_af_LB = exp((_b[trtnew])-(1.96*_se[trtnew]))
return scalar tdc_s_weib_af_UB = exp((_b[trtnew])+(1.96*_se[trtnew]))
gen cov_tdc_s_weib_af = 0
replace cov_tdc_s_weib_af = 1 if return(tdc_s_weib_af_LB)<return(trueweib_af_af) &
return(tdc_s_weib_af_UB)>return(trueweib_af_af)
return scalar cov_tdc_s_weib_af = cov_tdc_s_weib_af[1]
***AUC***
***AUC mean***
preserve

```

```

collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(tdcs_weib_af) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(tdcs_weib_af)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcs_we_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(tdcs_weib_af_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(tdcs_weib_af_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcs_we_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(tdcs_weib_af_LB) if _n==_N

```

```

egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(tdcs_weib_af_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar tdcs_we_adj_auc_con_UB=r(mean)
gen cov_tdcs_we_adj_auc_con = 0
replace cov_tdcs_we_adj_auc_con = 1 if return(trueauc_con) > return(tdcs_we_adj_auc_con_LB) &
return(trueauc_con) < return(tdcs_we_adj_auc_con_UB)
return scalar cov_tdcs_we_adj_auc_con = cov_tdcs_we_adj_auc_con[1]
restore
***Naive: Standard Cox Model with Crossover as a Time-dependent Covariate, with and without other covariates
taken into account***
stset timeOS2, failure(died) id(id)
stcox trtrand xoti bprog cea0grp pfsti
return scalar exotdcm_cox_hr = exp(_b[trtrand])
return scalar exotdcm_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar exotdcm_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar exotdcm_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_exotdcm_cox_hr = 0
replace cov_exotdcm_cox_hr = 1 if return(truecox_hr) > return(exotdcm_cox_hr_LB) & return(truecox_hr) <
return(exotdcm_cox_hr_UB)
return scalar cov_exotdcm_cox_hr = cov_exotdcm_cox_hr[1]
return scalar xotdcm_cox_hr = exp(_b[xoti])
return scalar xotdcm_cox_hr_SE = exp(_b[xoti])*_se[xoti]
return scalar xotdcm_cox_hr_LB = exp((_b[xoti])-(1.96*_se[xoti]))
return scalar xotdcm_cox_hr_UB = exp((_b[xoti])+(1.96*_se[xoti]))
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surf[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(exotdcm_cox_hr))
gen survfcontrol=1
replace survfcontrol=surfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0

```

```

replace auc =(time-time[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcm_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surf[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(exotdcm_cox_hr_LB))
gen survfcontrol=1
replace survfcontrol=surfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcm_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf =1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surf[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(exotdcm_cox_hr_UB))
gen survfcontrol=1
replace survfcontrol=surfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcm_adj_auc_con_UB=r(mean)
gen cov_xotdcm_adj_auc_con = 0

```

```

replace cov_xotdcm_adj_auc_con = 1 if return(trueauc_con) > return(xotdcm_adj_auc_con_LB) &
return(trueauc_con) < return(xotdcm_adj_auc_con_UB)
return scalar cov_xotdcm_adj_auc_con = cov_xotdcm_adj_auc_con[1]
restore
***treatment as time dependent variable, with tdcx and with weibull to get an AF***
streg trtrand xoti bprog cea0grp pfsti, dist(weibull) time
return scalar exotdcm_weib_af = exp(_b[trtrand])
return scalar exotdcm_weib_af_SE = exp(_b[trtrand])*_se[trtrand]
return scalar exotdcm_weib_af_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar exotdcm_weib_af_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_exotdcm_weib_af = 0
replace cov_exotdcm_weib_af = 1 if return(exotdcm_weib_af_LB)<return(trueweib_af_af) &
return(exotdcm_weib_af_UB)>return(trueweib_af_af)
return scalar cov_exotdcm_weib_af = cov_exotdcm_weib_af[1]
return scalar xotdcm_weib_af = exp(_b[xoti])
return scalar xotdcm_weib_af_SE = exp(_b[xoti])*_se[xoti]
return scalar xotdcm_weib_af_LB = exp(_b[xoti])-(1.96*_se[xoti])
return scalar xotdcm_weib_af_UB = exp(_b[xoti])+(1.96*_se[xoti])
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(exotdcm_weib_af) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(exotdcm_weib_af)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcm_we_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(exotdcm_weib_af_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]

```

```

gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(exotdcm_weib_af_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcm_we_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(exotdcm_weib_af_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(exotdcm_weib_af_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcm_we_adj_auc_con_UB=r(mean)
gen cov_xotdcm_we_adj_auc_con = 0
replace cov_xotdcm_we_adj_auc_con = 1 if return(trueauc_con) > return(xotdcm_we_adj_auc_con_LB) &
return(trueauc_con) < return(xotdcm_we_adj_auc_con_UB)
return scalar cov_xotdcm_we_adj_auc_con = cov_xotdcm_we_adj_auc_con[1]
restore
***without other time-dependent covariates included***
stcox trtrand xoti
gen exotdcs_cox_hr = exp(_b[trtrand])
return scalar exotdcs_cox_hr = exotdcs_cox_hr[1]
return scalar exotdcs_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar exotdcs_cox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar exotdcs_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_exotdcs_cox_hr = 0
replace cov_exotdcs_cox_hr = 1 if return(truecox_hr) > return(exotdcs_cox_hr_LB) & return(truecox_hr) <
return(exotdcs_cox_hr_UB)
return scalar cov_exotdcs_cox_hr = cov_exotdcs_cox_hr[1]
gen xotdcs_cox_hr = exp(_b[xoti])
return scalar xotdcs_cox_hr = xotdcs_cox_hr[1]
return scalar xotdcs_cox_hr_SE = exp(_b[xoti])*_se[xoti]
return scalar xotdcs_cox_hr_LB = exp(_b[xoti])-(1.96*_se[xoti])

```

```

return scalar xotdcs_cox_hr_UB = exp(_b[xoti])+(1.96*_se[xoti])
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(exotdcs_cox_hr))
gen survfcontrol=1
replace survfcontrol=survcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcs_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(exotdcs_cox_hr_LB))
gen survfcontrol=1
replace survfcontrol=survcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcs_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.

```



```

stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surfv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(exotdcs_cox_hr_UB))
gen survfcontrol=1
replace survfcontrol=surfvcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+surfvcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcs_adj_auc_con_UB=r(mean)
gen cov_xotdcs_adj_auc_con = 0
replace cov_xotdcs_adj_auc_con = 1 if return(trueauc_con) > return(xotdcs_adj_auc_con_LB) &
return(trueauc_con) < return(xotdcs_adj_auc_con_UB)
return scalar cov_xotdcs_adj_auc_con = cov_xotdcs_adj_auc_con[1]
restore
***without other time-dependent covariates included - weibull***
streg trtrand xoti, dist(weibull) time
gen exotdcs_weib_af = exp(_b[trtrand])
return scalar exotdcs_weib_af = exotdcs_weib_af[1]
return scalar exotdcs_weib_af_SE = exp(_b[trtrand])*_se[trtrand]
return scalar exotdcs_weib_af_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar exotdcs_weib_af_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_exotdcs_weib_af = 0
replace cov_exotdcs_weib_af = 1 if return(exotdcs_weib_af_LB)<return(trueweib_aft_af) &
return(exotdcs_weib_af_UB)>return(trueweib_aft_af)
return scalar cov_exotdcs_weib_af = cov_exotdcs_weib_af[1]
gen xotdcs_weib_af = exp(_b[xoti])
return scalar xotdcs_weib_af = xotdcs_weib_af[1]
return scalar xotdcs_weib_af_SE = exp(_b[xoti])*_se[xoti]
return scalar xotdcs_weib_af_LB = exp(_b[xoti])-(1.96*_se[xoti])
return scalar xotdcs_weib_af_UB = exp(_b[xoti])+(1.96*_se[xoti])
***AUC***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(exotdcs_weib_af) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]

```

```

gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(exotdcs_weib_af)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+surfvcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcs_we_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(exotdcs_weib_af_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(exotdcs_weib_af_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+surfvcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar xotdcs_we_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(exotdcs_weib_af_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(exotdcs_weib_af_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+surfvcontrol[_n-1])/2)
egen auc1=total(auc)

```

```

summ auc1
return scalar xotdcs_we_adj_auc_con_UB=r(mean)
gen cov_xotdcs_we_adj_auc_con = 0
replace cov_xotdcs_we_adj_auc_con = 1 if return(trueauc_con) > return(xotdcs_we_adj_auc_con_LB) &
return(trueauc_con) < return(xotdcs_we_adj_auc_con_UB)
return scalar cov_xotdcs_we_adj_auc_con = cov_xotdcs_we_adj_auc_con[1]
restore
***IPCW***
preserve
by id: drop if timeOS3 >(xotime)
by id: replace finalobs = 0
by id: replace finalobs = 1 if _n==_N
gen infcensOS=0
replace infcensOS=1 if xoti==1 & trtrand==0
by id: replace died=. if (infcensOS==1 | cens==1)
gen pfsint = timePFSobsgrp*cea1grp
***Deriving IPCWs. Knots based on 5, 10, 25, 50, 75 and 95 percentiles in base case. but note it makes very
little diff***
spbase obsno, knots(2, 5, 11, 22, 41) gen(spline)
sort id obsno
capture xi: logistic infcensOS i.bprog i.timePFSobsgrp i.obsceagr i.cea0grp i.cea1grp i.pfsint obsno spline* if
trtrand==0 & timeOS3>=timePFSobs & timeOS3<=(timePFSobs+42)
***note, above dictates that we only want to apply weights after progression and during the 3 consultations after
disease prog, as we know xo can't occur after this***
***Predict is then used to estimate the probability of receiving crossover treatment for each subject-day included
in the regression:***
predict ptrtrec if e(sample)
**The above code estimates the probability of each individual receiving crossover treatment (and therefore being
informatively censored) each day. For the IPCW we need the probability of remaining uncensored, so we submit
the probabilities from 1:
replace ptrtrec=ptrtrec*infcensOS+(1-ptrtrec)*(1-infcensOS)
replace ptrtrec=1 if ptrtrec=.
**Now we estimate each individual's probability of their complete censoring history up to each day**
sort id obsno
by id: replace ptrtrec=ptrtrec*ptrtrec[_n-1] if _n!=1
rename ptrtrec censdenom
***The numerator of the IPCW is estimated in a similar way as above, with the only difference being that the
initial logistic regression only includes baseline covariates, and it is applied to all observations in control group.
Note baseline cea score isn't included as it will be highly correlated with bprog***
sort id obsno
capture xi: logistic infcensOS i.bprog i.cea0grp obsno spline* if trtrand==0
predict ptrtrec2 if e(sample)
replace ptrtrec2=ptrtrec2*infcensOS+(1-ptrtrec2)*(1-infcensOS)
replace ptrtrec2=1 if ptrtrec2=.
sort id obsno
by id: replace ptrtrec2 = ptrtrec2*ptrtrec2[_n-1] if _n!=1
rename ptrtrec2 censnum
***The stabilised weight is derived by dividing the numerator by the denominator:***
gen stabweightxo=censnum/censdenom
replace stabweightxo=1 if trtrand==1
***Under the IPCW approach, a time-dependent Cox proportional hazards model can then be estimated to
calculate the treatment effect, adjusting for baseline characteristics and using the time-varying stabilized
weights.***

```

```

***don't include the time to pfs indicator in this model, as it would bias estimates of the treatment effect***
capture xi: logistic died trtrand i.bprog i.cea0grp obsno spline*[pw=stabweightxo], cluster(id)
return scalar ipcw_cox_hr = exp(_b[trtrand])
return scalar ipcw_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar ipcw_cox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar ipcw_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_ipcw_cox_hr = 0
replace cov_ipcw_cox_hr = 1 if return(truecox_hr) > return(ipcw_cox_hr_LB) & return(truecox_hr) <
return(ipcw_cox_hr_UB)
return scalar cov_ipcw_cox_hr = cov_ipcw_cox_hr[1]
summ stabweightxo
return scalar ipcw_weight_min = r(min)
return scalar ipcw_weight_max = r(max)
***Lambda numerator for each time point for IPCW KM***
gen IPCWKMD21 = stabweightxo if (died==1 & timeOS2<=21 & trtrand==0)
gen IPCWKMD21 = stabweightxo if ((died==0|died==.|died==1) & timeOS2<=21 & trtrand==0)
forvalues min=21(21)1092 {
local max=`min'+21
gen IPCWKMD`max`= stabweightxo if (died==1 & (timeOS2>`min' & timeOS2<=`max') & trtrand==0)
egen IPCWKMD`min`=total(IPCWKM`min')
gen IPCWKMD`max`= stabweightxo if ((died==0|died==.|died==1) & (timeOS2>`min' & timeOS2<=`max') &
trtrand==0)
egen IPCWKMD`min`=total(IPCWKM`min')
}
keep IPCWKMD`min' IPCWKMD`max'
drop if _n>=54
gen time=(`min')*21
***lambda for IPCW KM***
gen IPCWLAMBDA=0
forvalues t=21(21)1092 {
replace IPCWLAMBDA=IPCWKMD`t'/IPCWKMD`min' if time==`t'
}
***gen survival probabilities for each time point***
gen IPCWKMS=1
replace IPCWKMS=(IPCWKMS[_n-1]-(IPCWKMS[_n-1]*IPCWLAMBDA)) if _n>=2
***AUC***
replace time= time-10.5 if _n>1
gen AUC=0
replace AUC=(time-time[_n-1])*IPCWKMS[_n-1]
egen IPCWAUC=total(AUC)
summ IPCWAUC
return scalar ipcw_adj_auc_conwkm = r(mean)
restore
sort id
***AUC***
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime=.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0

```

```

replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf=exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(ipcw_cox_hr))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc=(time-time[_n-1])*((survfcontrol+suvfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar ipcw_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf=exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(ipcw_cox_hr_LB))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc=(time-time[_n-1])*((survfcontrol+suvfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar ipcw_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf=exp((-exp(cons)*(time^exp(lnp)))) if _n>1

```

```

gen hazf=0
replace hazf=1-(survf/surv[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(ipcw_cox_hr_UB))
gen survfcontrol=1
replace survfcontrol=survfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc=(time-time[_n-1])*((survfcontrol+suvfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar ipcw_adj_auc_con_UB=r(mean)
gen cov_ipcw_adj_auc_con = 0
replace cov_ipcw_adj_auc_con = 1 if return(trueauc_con) > return(ipcw_adj_auc_con_LB) & return(trueauc_con)
< return(ipcw_adj_auc_con_UB)
return scalar cov_ipcw_adj_auc_con = cov_ipcw_adj_auc_con[1]
restore
***IPCW without knowing cea***
preserve
by id: drop if timeOS3 >(xotime)
by id: replace finalobs = 0
by id: replace finalobs = 1 if _n==_N
gen infcensOS=0
replace infcensOS=1 if xoti==1 & trtrand==0
by id: replace died=. if (infcensOS==1 | cens==1)
***Deriving IPCWs***
spbase obsno, knots(2, 5, 11, 22, 41) gen(spline)
sort id obsno
capture xi: logistic infcensOS i.bprog i.timePFSobsgrp obsno spline* if trtrand==0 & timeOS3>=timePFSobs &
timeOS3<=(timePFSobs+42)
***note, above dictates that we only want to apply weights after progression and during the 3 consultations after
disease prog, as we know xo can't occur after this***
***Predict is then used to estimate the probability of receiving crossover treatment for each subject-day included
in the regression:***
predict ptrtrec if e(sample)
**The above code estimates the probability of each individual receiving crossover treatment (and therefore being
informatively censored) each day. For the IPCW we need the probability of remaining uncensored, so we submit
the probabilities from 1:**
replace ptrtrec=ptrtrec*infcensOS+(1-ptrtrec)*(1-infcensOS)
replace ptrtrec=1 if ptrtrec==.
**Now we estimate each individual's probability of their complete censoring history up to each day**
sort id obsno
by id: replace ptrtrec=ptrtrec*ptrtrec[_n-1] if _n!=1
rename ptrtrec censdenom
***The numerator of the IPCW is estimated in a similar way as above, with the only difference being that the
initial logistic regression only includes baseline covariates, and it is applied to all observations in control group.
Note baseline cea score isn't included as it will be highly correlated with bprog***
sort id obsno
capture xi: logistic infcensOS i.bprog obsno spline* if trtrand==0
predict ptrtrec2 if e(sample)
replace ptrtrec2=ptrtrec2*infcensOS+(1-ptrtrec2)*(1-infcensOS)
replace ptrtrec2=1 if ptrtrec2==.
sort id obsno
by id: replace ptrtrec2 = ptrtrec2*ptrtrec2[_n-1] if _n!=1
rename ptrtrec2 censnum

```

```

***The stabilised weight is derived by dividing the numerator by the denominator:***
gen stabweightxo=censnum/censdenom
replace stabweightxo=1 if trtrand==1
***Under the IPCW approach, a time-dependent Cox proportional hazards model can then be estimated to
calculate the treatment effect, adjusting for baseline characteristics and using the time-varying stabilized
weights.***
***don't include the time to pfs indicator in this model, as it would bias estimates of the treatment effect***
capture xi: logistic died trtrand i.bprog obsno spline*[pw=stabweightxo], cluster(id)
return scalar ipcwnoce_a_cox_hr = exp(_b[trtrand])
return scalar ipcwnoce_a_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar ipcwnoce_a_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar ipcwnoce_a_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_ipcwnoce_a_cox_hr = 0
replace cov_ipcwnoce_a_cox_hr = 1 if return(truecox_hr) > return(ipcwnoce_a_cox_hr_LB) & return(truecox_hr) <
return(ipcwnoce_a_cox_hr_UB)
return scalar cov_ipcwnoce_a_cox_hr = cov_ipcwnoce_a_cox_hr[1]
summ stabweightxo
return scalar ipcwnoce_a_weight_min = r(min)
return scalar ipcwnoce_a_weight_max = r(max)
***Lambda numerator for each time point for IPCW KM***
gen IPCWKMD21 = stabweightxo if (died==1 & timeOS2<=21 & trtrand==0)
gen IPCWKMD21 = stabweightxo if ((died==0|died==.|died==1) & timeOS2<=21 & trtrand==0)
forvalues min=21(21)1092 {
  local max=`min'+21
  gen IPCWKMD`max`= stabweightxo if (died==1 & (timeOS2>`min' & timeOS2<=`max') & trtrand==0)
  egen IPCWKMDN`min`=total(IPCWKM`min')
  gen IPCWKMD`max`= stabweightxo if ((died==0|died==.|died==1) & (timeOS2>`min' & timeOS2<=`max') &
trtrand==0)
  egen IPCWKMDD`min`=total(IPCWKMD`min')
}
keep IPCWKMDN* IPCWKMDD*
drop if _n>=54
gen time=(_n-1)*21
***lambda for IPCW KM***
gen IPCWLAMBDA=0
forvalues t=21(21)1092 {
  replace IPCWLAMBDA=IPCWKMDN`t'/IPCWKMDD`t' if time==`t'
}
***gen survival probabilities for each time point***
gen IPCWKMS=1
replace IPCWKMS=(IPCWKMS[_n-1]-(IPCWKMS[_n-1]*IPCWLAMBDA)) if _n>=2
***AUC***
replace time= time-10.5 if _n>1
gen AUC=0
replace AUC=(time-time[_n-1])*IPCWKMS[_n-1]
egen IPCWAUC=total(AUC)
summ IPCWAUC
return scalar ipcwnoce_a_adj_auc_conwkm = r(mean)
restore
sort id
***AUC mean***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)

```

```

by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surf[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(ipcwnoce_a_cox_hr))
gen survfcontrol=1
replace survfcontrol=surfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+surfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar ipcwnoce_a_adj_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surf[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(ipcwnoce_a_cox_hr_LB))
gen survfcontrol=1
replace survfcontrol=surfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survfcontrol+surfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar ipcwnoce_a_adj_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0

```

```

replace time=_n/1.8255707 if _n>1
gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf=exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen hazf=0
replace hazf=1-(survf/surf[_n-1]) if _n>1
gen hazfcontrol=hazf*(1/return(ipcwnoce_a_cox_hr_UB))
gen survfcontrol=1
replace survfcontrol=surfcontrol[_n-1] * (1-hazfcontrol) if _n>1
gen auc=0
replace auc=(time-time[_n-1])*((survfcontrol+surfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar ipcwnoce_adj_auc_con_UB=r(mean)
gen cov_ipcwnoce_adj_auc_con = 0
replace cov_ipcwnoce_adj_auc_con = 1 if return(trueauc_con) > return(ipcwnoce_adj_auc_con_LB) &
return(trueauc_con) < return(ipcwnoce_adj_auc_con_UB)
return scalar cov_ipcwnoce_adj_auc_con = cov_ipcwnoce_adj_auc_con[1]
restore
***Standard Weibull with two treatment effects, single Cox – drop exp group and data prior to crossover, re-
calculate time and admin with time zero= to crossover time, then estimate treatment effect in those that remain.
Then restore dataset and apply treatment effect to xogain period for crossover patients.***
preserve
drop if trtrand==1
drop if timeOS3<timePFSobs
by id: replace obsno = _n
by id: egen minrisk=min(timeOS3)
by id: replace timeOS3=timeOS3-minrisk
by id: replace xotime=xotime-minrisk
by id: replace timeOS2=timeOS2-minrisk
stset timeOS2, failure(died) id(id)
streg trtnew, dist(weibull) time
return scalar weib2s_weib_af = exp(_b[trtnew])
return scalar weib2s_weib_af_SE = exp(_b[trtnew])*_se[trtnew]
return scalar weib2s_weib_af_LB = exp(_b[trtnew])-(1.96*_se[trtnew])
return scalar weib2s_weib_af_UB = exp(_b[trtnew])+(1.96*_se[trtnew])
restore
sort id
preserve
***AUC***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(weib2s_weib_af))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean

```

```

return scalar weib2s_adj_auc_con = r(rmean)
return scalar weib2s_adj_auc_con_SE = r(se)
return scalar weib2s_adj_auc_con_LB = r(lb)
return scalar weib2s_adj_auc_con_UB = r(ub)
gen cov_weib2s_adj_auc_con = 0
replace cov_weib2s_adj_auc_con = 1 if return(trueauc_con) > return(weib2s_adj_auc_con_LB) &
return(trueauc_con) < return(weib2s_adj_auc_con_UB)
return scalar cov_weib2s_adj_auc_con = cov_weib2s_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar weib2s_adj_cox_hr = exp(_b[trtrand])
return scalar weib2s_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar weib2s_adj_cox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar weib2s_adj_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_weib2s_adj_cox_hr = 0
replace cov_weib2s_adj_cox_hr = 1 if return(truecox_hr) > return(weib2s_adj_cox_hr_LB) & return(truecox_hr)
< return(weib2s_adj_cox_hr_UB)
return scalar cov_weib2s_adj_cox_hr = cov_weib2s_adj_cox_hr[1]
restore
***lower 95%CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(weib2s_weib_af_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar weib2s_adj_lowci_auc_con = r(rmean)
return scalar weib2s_adj_lowci_auc_con_SE = r(se)
return scalar weib2s_adj_lowci_auc_con_UB = r(ub)

***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar weib2s_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar weib2s_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar weib2s_adj_lowci_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
restore
***Upper 95%CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(weib2s_weib_af_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2

```

```

by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar weib2s_adj_upci_auc_con = r(rmean)
return scalar weib2s_adj_upci_auc_con_SE = r(se)
return scalar weib2s_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar weib2s_adj_upci_cox_hr = exp(_b[trtrand])
return scalar weib2s_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar weib2s_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***Standard Weibull with two treatment effects, multi Cox – drop exp group and data prior to crossover, re-
calculate time and admin with time zero= to crossover time, then estimate treatment effect in those that remain.
Then restore dataset and apply treatment effect to xogain period for crossover patients.***
preserve
drop if trtrand==1
drop if timeOS3<timePFSobs
by id: replace obsno = _n
by id: egen minrisk=min(timeOS3)
by id: replace timeOS3=timeOS3-minrisk
by id: replace xotime=xotime-minrisk
by id: replace timeOS2=timeOS2-minrisk
stset timeOS2, failure(died) id(id)
streg trtnew bprog timePFSobsgrp cea0grp cea1grp, dist(weibull) time
return scalar weib2m_weib_af = exp(_b[trtnew])
return scalar weib2m_weib_af_SE = exp(_b[trtnew])*_se[trtnew]
return scalar weib2m_weib_af_LB = exp((_b[trtnew])-(1.96*_se[trtnew]))
return scalar weib2m_weib_af_UB = exp((_b[trtnew])+(1.96*_se[trtnew]))
restore
sort id
preserve
***AUC***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(weib2m_weib_af))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar weib2m_adj_auc_con = r(rmean)
return scalar weib2m_adj_auc_con_SE = r(se)
return scalar weib2m_adj_auc_con_LB = r(lb)
return scalar weib2m_adj_auc_con_UB = r(ub)
gen cov_weib2m_adj_auc_con = 0
replace cov_weib2m_adj_auc_con = 1 if return(trueauc_con) > return(weib2m_adj_auc_con_LB) &
return(trueauc_con) < return(weib2m_adj_auc_con_UB)

```

```

return scalar cov_weib2m_adj_auc_con = cov_weib2m_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar weib2m_adj_cox_hr = exp(_b[trtrand])
return scalar weib2m_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar weib2m_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar weib2m_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_weib2m_adj_cox_hr = 0
replace cov_weib2m_adj_cox_hr = 1 if return(truecox_hr) > return(weib2m_adj_cox_hr_LB) &
return(truecox_hr) < return(weib2m_adj_cox_hr_UB)
return scalar cov_weib2m_adj_cox_hr = cov_weib2m_adj_cox_hr[1]
restore
***lower 95%CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(weib2m_weib_af_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar weib2m_adj_lowci_auc_con = r(rmean)
return scalar weib2m_adj_lowci_auc_con_SE = r(se)
return scalar weib2m_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar weib2m_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar weib2m_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar weib2m_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95%CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(weib2m_weib_af_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar weib2m_adj_upci_auc_con = r(rmean)
return scalar weib2m_adj_upci_auc_con_SE = r(se)
return scalar weib2m_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***

```

```

stcox trtrand
return scalar weib2m_adj_upci_cox_hr = exp(_b[trtrand])
return scalar weib2m_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar weib2m_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***observational SNM***
***2-stage approach***
preserve
drop if trtrand==1
***note, below code will drop patients who progressed and died before their first consultation. for this part,
where we are just estimating the treatment effect in crossover patients, this is ok as none of these patients will
have crossed over***
drop if timeOS3<timePFSobs
by id: replace obsno = _n
by id: egen minrisk=min(timeOS3)
by id: replace timeOS3=timeOS3-minrisk
by id: replace xotime=xotime-minrisk
by id: replace timeOS2=timeOS2-minrisk
gen admin = `admin'-minrisk
stset timeOS2, id(id) failure(died)
capture stgest trnew bprog timePFSobsgrp obsceagrps cea0grp cea1grp, visit(obsno) firstvis(1) lasttime(admin)
range(-2,2) round(1) saveres(gest1) replace
return scalar snm2 = exp(-r(psi))
return scalar snm2_UB = exp(-r(lcipsi))
return scalar snm2_LB = exp(-r(ucipsi))
gen cov_snm2 = 0
replace cov_snm2 = 1 if return(snm2_LB)<return(trueweib_aft_af) & return(snm2_UB)>return(trueweib_aft_af)
replace cov_snm2 = . if r(psi)==. | r(lcipsi)==. | r(ucipsi)==.
return scalar cov_snm2 = cov_snm2[1]
restore
sort id
preserve
***AUC***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(snm2))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar snm2_adj_auc_con = r(rmean)
return scalar snm2_adj_auc_con_SE = r(se)
return scalar snm2_adj_auc_con_LB = r(lb)
return scalar snm2_adj_auc_con_UB = r(ub)
gen cov_snm2_adj_auc_con = 0
replace cov_snm2_adj_auc_con = 1 if return(trueauc_con) > return(snm2_adj_auc_con_LB) &
return(trueauc_con) < return(snm2_adj_auc_con_UB)
return scalar cov_snm2_adj_auc_con = cov_snm2_adj_auc_con[1]

```

```

***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar snm2_adj_cox_hr = exp(_b[trtrand])
return scalar snm2_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar snm2_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar snm2_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_snm2_adj_cox_hr = 0
replace cov_snm2_adj_cox_hr = 1 if return(truecox_hr) > return(snm2_adj_cox_hr_LB) & return(truecox_hr) <
return(snm2_adj_cox_hr_UB)
return scalar cov_snm2_adj_cox_hr = cov_snm2_adj_cox_hr[1]
restore
***lower 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(snm2_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar snm2_adj_lowci_auc_con = r(rmean)
return scalar snm2_adj_lowci_auc_con_SE = r(se)
return scalar snm2_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar snm2_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar snm2_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar snm2_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(snm2_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar snm2_adj_upci_auc_con = r(rmean)
return scalar snm2_adj_upci_auc_con_SE = r(se)
return scalar snm2_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand

```

```

return scalar snm2_adj_upci_cox_hr = exp(_b[trtrand])
return scalar snm2_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar snm2_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***observational SNM with no bprog***
***2-stage approach***
preserve
drop if trtrand==1
***note, below code will drop patients who progressed and died before their first consultation. for this part,
where we are just estimating the treatment effect in crossover patients, this is ok as none of these patients will
have crossed over***
drop if timeOS3<timePFSobs
by id: replace obsno = _n
by id: egen minrisk=min(timeOS3)
by id: replace timeOS3=timeOS3-minrisk
by id: replace xotime=xotime-minrisk
by id: replace timeOS2=timeOS2-minrisk
gen admin = `admin'-minrisk
stset timeOS2, id(id) failure(died)
capture stgest trtnew timePFSobsgrp obsceagr cpa0grp cea1grp, visit(obsno) firstvis(1) lasttime(admin) range(-
2,2) round(1) saveres(gest1) replace
return scalar snm2nob = exp(-r(psi))
return scalar snm2nob_UB = exp(-r(lcipsi))
return scalar snm2nob_LB = exp(-r(ucipsi))
gen cov_snm2nob = 0
replace cov_snm2nob = 1 if return(snm2nob_LB)<return(trueweib_aft_af) &
return(snm2nob_UB)>return(trueweib_aft_af)
replace cov_snm2nob = . if r(psi)==. | r(lcipsi)==. | r(ucipsi)==.
return scalar cov_snm2nob = cov_snm2nob[1]
restore
sort id
preserve
***AUC***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(snm2nob))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar snm2nob_adj_auc_con = r(rmean)
return scalar snm2nob_adj_auc_con_SE = r(se)
return scalar snm2nob_adj_auc_con_LB = r(lb)
return scalar snm2nob_adj_auc_con_UB = r(ub)
gen cov_snm2nob_adj_auc_con = 0
replace cov_snm2nob_adj_auc_con = 1 if return(trueauc_con) > return(snm2nob_adj_auc_con_LB) &
return(trueauc_con) < return(snm2nob_adj_auc_con_UB)
return scalar cov_snm2nob_adj_auc_con = cov_snm2nob_adj_auc_con[1]

```

```

***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar snm2nob_adj_cox_hr = exp(_b[trtrand])
return scalar snm2nob_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar snm2nob_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar snm2nob_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_snm2nob_adj_cox_hr = 0
replace cov_snm2nob_adj_cox_hr = 1 if return(truecox_hr) > return(snm2nob_adj_cox_hr_LB) &
return(truecox_hr) < return(snm2nob_adj_cox_hr_UB)
return scalar cov_snm2nob_adj_cox_hr = cov_snm2nob_adj_cox_hr[1]
restore
***lower 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(snm2nob_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar snm2nob_adj_lowci_auc_con = r(rmean)
return scalar snm2nob_adj_lowci_auc_con_SE = r(se)
return scalar snm2nob_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar snm2nob_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar snm2nob_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar snm2nob_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(snm2nob_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar snm2nob_adj_upci_auc_con = r(rmean)
return scalar snm2nob_adj_upci_auc_con_SE = r(se)
return scalar snm2nob_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand

```



```

return scalar snm2nob_adj_upci_cox_hr = exp(_b[trtrand])
return scalar snm2nob_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar snm2nob_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***RPSFTM***
***note need to collapse data first***
gen admin=`admin'
preserve
collapse (max) trtrand xoti bprog admin timeOS2 timePFSobs timePFSobsgrp xotime died cens xoOSgainobs,
by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xo0(xotime xoti) test(logrank) endstudy(admin)
return scalar RPSFTM = exp(-r(psi))
return scalar RPSFTM_UB = exp(-r(psi_low))
return scalar RPSFTM_LB = exp(-r(psi_upp))
gen cov_RPSFTM = 0
replace cov_RPSFTM = 1 if return(RPSFTM_LB)<return(trueweib_aft_af) &
return(RPSFTM_UB)>return(trueweib_aft_af)
replace cov_RPSFTM = . if r(psi)==. | r(psi_low)==. | r(psi_upp)==.
return scalar cov_RPSFTM = cov_RPSFTM[1]
***AUC***
sort id
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(RPSFTM))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar RPSFTM_adj_auc_con = r(rmean)
return scalar RPSFTM_adj_auc_con_SE = r(se)
return scalar RPSFTM_adj_auc_con_LB = r(lb)
return scalar RPSFTM_adj_auc_con_UB = r(ub)
gen cov_RPSFTM_adj_auc_con = 0
replace cov_RPSFTM_adj_auc_con = 1 if return(trueauc_con) > return(RPSFTM_adj_auc_con_LB) &
return(trueauc_con) < return(RPSFTM_adj_auc_con_UB)
return scalar cov_RPSFTM_adj_auc_con = cov_RPSFTM_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar RPSFTM_adj_cox_hr = exp(_b[trtrand])
return scalar RPSFTM_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar RPSFTM_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar RPSFTM_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_RPSFTM_adj_cox_hr = 0

```

```

replace cov_RPSFTM_adj_cox_hr = 1 if return(truecox_hr) > return(RPSFTM_adj_cox_hr_LB) &
return(truecox_hr) < return(RPSFTM_adj_cox_hr_UB)
return scalar cov_RPSFTM_adj_cox_hr = cov_RPSFTM_adj_cox_hr[1]
restore
sort id
preserve
***Lower 95% CI***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(RPSFTM_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar RPSFTM_adj_lowci_auc_con = r(rmean)
return scalar RPSFTM_adj_lowci_auc_con_SE = r(se)
return scalar RPSFTM_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar RPSFTM_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar RPSFTM_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar RPSFTM_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(RPSFTM_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar RPSFTM_adj_upci_auc_con = r(rmean)
return scalar RPSFTM_adj_upci_auc_con_SE = r(se)
return scalar RPSFTM_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar RPSFTM_adj_upci_cox_hr = exp(_b[trtrand])
return scalar RPSFTM_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar RPSFTM_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore

```

```

***estimate from exp survivor function***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(RPSFTM) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(RPSFTM)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTM_sfunc_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(RPSFTM_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(RPSFTM_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTM_sfunc_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4

```

```

gen time = 0
replace time = (1095/1999)*return(RPSFTM_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(RPSFTM_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTM_sfunc_auc_con_UB=r(mean)
gen cov_RPSFTM_sfunc_auc_con = 0
replace cov_RPSFTM_sfunc_auc_con = 1 if return(trueauc_con) > return(RPSFTM_sfunc_auc_con_LB) &
return(trueauc_con) < return(RPSFTM_sfunc_auc_con_UB)
return scalar cov_RPSFTM_sfunc_auc_con = cov_RPSFTM_sfunc_auc_con[1]
restore
***RPSFTM extrapolate counterfactuals***
preserve
collapse (max) trtrand xoti bprog admin timeOS2 timePFSobs timePFSobsgrp xotime died cens xoOSgainobs,
by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xo0(xotime xoti) test(logrank) endstudy(admin) gen(counterf) hr nokmgraph
return scalar RPSFTM_extrap_cox_hr=r(HR_adj)
return scalar RPSFTM_extrap_cox_hr_LB=r(HR_adj_low)
return scalar RPSFTM_extrap_cox_hr_UB=r(HR_adj_upp)
gen cov_RPSFTM_extrap_cox_hr = 0
replace cov_RPSFTM_extrap_cox_hr = 1 if return(truecox_hr) > return(RPSFTM_extrap_cox_hr_LB) &
return(truecox_hr) < return(RPSFTM_extrap_cox_hr_UB)
return scalar cov_RPSFTM_extrap_cox_hr = cov_RPSFTM_extrap_cox_hr[1]

stset counterf, failure(dcounterf) id(id)
streg, dist(weibull) nohr
expand 4
gen time = 0
***note 1.8255707 means that when i have 2000 rows i go up to 1095 days***
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survf+survf[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTM_extrapauc3_weib_con=r(mean)
restore
***RPSFTM with covariates***
***note need to collapse data first***

```

```

preserve
collapse (max) trtrand xoti bprog cea0grp admin timeOS2 timePFSobs timePFSobsgrp xotime died cens
xoOSgainobs, by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xotime==0
capture strbee trtrand, xo0(xotime xoti) test(weibull) endstudy(admin) adjvars(bprog cea0grp)
return scalar RPSFTMwc = exp(-r(psi))
return scalar RPSFTMwc_UB = exp(-r(psi_low))
return scalar RPSFTMwc_LB = exp(-r(psi_upp))
gen cov_RPSFTMwc = 0
replace cov_RPSFTMwc = 1 if return(RPSFTMwc_LB)<return(trueweib_aft_af) &
return(RPSFTMwc_UB)>return(trueweib_aft_af)
replace cov_RPSFTMwc = . if r(psi)==. | r(psi_low)==. | r(psi_upp)==.
return scalar cov_RPSFTMwc = cov_RPSFTMwc[1]
***AUC***
sort id
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(RPSFTMwc))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar RPSFTMwc_adj_auc_con = r(rmean)
return scalar RPSFTMwc_adj_auc_con_SE = r(se)
return scalar RPSFTMwc_adj_auc_con_LB = r(lb)
return scalar RPSFTMwc_adj_auc_con_UB = r(ub)
gen cov_RPSFTMwc_adj_auc_con = 0
replace cov_RPSFTMwc_adj_auc_con = 1 if return(trueauc_con) > return(RPSFTMwc_adj_auc_con_LB) &
return(trueauc_con) < return(RPSFTMwc_adj_auc_con_UB)
return scalar cov_RPSFTMwc_adj_auc_con = cov_RPSFTMwc_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar RPSFTMwc_adj_cox_hr = exp(_b[trtrand])
return scalar RPSFTMwc_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar RPSFTMwc_adj_cox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar RPSFTMwc_adj_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_RPSFTMwc_adj_cox_hr = 0
replace cov_RPSFTMwc_adj_cox_hr = 1 if return(truecox_hr) > return(RPSFTMwc_adj_cox_hr_LB) &
return(truecox_hr) < return(RPSFTMwc_adj_cox_hr_UB)
return scalar cov_RPSFTMwc_adj_cox_hr = cov_RPSFTMwc_adj_cox_hr[1]
restore
sort id
preserve
***Lower 95% CI***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)

```

```

by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(RPSFTMwc_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar RPSFTMwc_adj_lowci_auc_con = r(rmean)
return scalar RPSFTMwc_adj_lowci_auc_con_SE = r(se)
return scalar RPSFTMwc_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar RPSFTMwc_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar RPSFTMwc_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar RPSFTMwc_adj_lowci_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(RPSFTMwc_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar RPSFTMwc_adj_upci_auc_con = r(rmean)
return scalar RPSFTMwc_adj_upci_auc_con_SE = r(se)
return scalar RPSFTMwc_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar RPSFTMwc_adj_upci_cox_hr = exp(_b[trtrand])
return scalar RPSFTMwc_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar RPSFTMwc_adj_upci_cox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
restore
***estimate from exp survivor function***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0

```

```

replace time = (1095/1999)*return(RPSFTMwc) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(RPSFTMwc)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTMwc_sfuc_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(RPSFTMwc_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(RPSFTMwc_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTMwc_sfuc_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(RPSFTMwc_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1

```

```

gen survfcontrol= survf
gen timecontrol=time/return(RPSFTMwc_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTMwc_sfuc_auc_con_UB=r(mean)
gen cov_RPSFTMwc_sfuc_auc_con = 0
replace cov_RPSFTMwc_sfuc_auc_con = 1 if return(trueauc_con) > return(RPSFTMwc_sfuc_auc_con_LB) &
return(trueauc_con) < return(RPSFTMwc_sfuc_auc_con_UB)
return scalar cov_RPSFTMwc_sfuc_auc_con = cov_RPSFTMwc_sfuc_auc_con[1]
restore
***RPSFTM extrapolate counterfactuals with covariates***
preserve
collapse (max) trtrand xoti bprog cea0grp admin timeOS2 timePFSobs timePFSobsgrp xotime died cens
xoOSgainobs, by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xo0(xotime xoti) test(weibull) endstudy(admin) adjvars(bprog cea0grp) gen(counterf) hr
nokmgraph
return scalar RPSFTMwcc_extrap_cox_hr=r(HR_adj)
return scalar RPSFTMwcc_extrap_cox_hr_LB=r(HR_adj_low)
return scalar RPSFTMwcc_extrap_cox_hr_UB=r(HR_adj_upp)
gen cov_RPSFTMwcc_extrap_cox_hr = 0
replace cov_RPSFTMwcc_extrap_cox_hr = 1 if return(truecox_hr) > return(RPSFTMwcc_extrap_cox_hr_LB) &
return(truecox_hr) < return(RPSFTMwcc_extrap_cox_hr_UB)
return scalar cov_RPSFTMwcc_extrap_cox_hr = cov_RPSFTMwcc_extrap_cox_hr[1]
stset counterf, failure(dcounterf) id(id)
streg, dist(weibull) nohr
expand 4
gen time = 0
replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survf+survf[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTMwcc_extrapauc3_weib_con=r(mean)
restore
***RPSFTM extrapolate counterfactuals with covariates included in initial strbee and following extrap***
preserve
collapse (max) trtrand xoti bprog cea0grp admin timeOS2 timePFSobs timePFSobsgrp xotime died cens
xoOSgainobs, by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xo0(xotime xoti) test(weibull) endstudy(admin) adjvars(bprog cea0grp) gen(counterf)
stset counterf, failure(dcounterf) id(id)
streg bprog cea0grp, dist(weibull) nohr
expand 4
gen time = 0

```

```

replace time=_n/1.8255707 if _n>1
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen bprog=_b[bprog]
gen cea0c=_b[cea0grp]
egen bprogmean2= mean(bprog)
egen cea0mean2= mean(cea0grp)
egen bprogmean = max(bprogmean2)
egen cea0mean = max(cea0mean2)
gen survf =1
replace survf= exp((-exp(cons+(bprogmean*bprog)+(cea0mean*cea0c))*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((survf+survf[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar RPSFTMwcc_extrapauc3_weib_con=r(mean)
restore
***IPE Algorithm - Weibull. Note HR option doesn't work in strbee with ipe. therefore transform af***
preserve
collapse (max) trtrand xoti bprog admin timeOS2 timePFSobs timePFSobsgrp xotime died cens xoOSgainobs,
by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, x0(xotime xoti) test(weibull) endstudy(admin) tol(3) maxiter(1000) ipe ipest(ss)
return scalar IPE = exp(-r(psi))
return scalar IPE_UB = exp(-r(psilow))
return scalar IPE_LB = exp(-r(psiupp))
gen cov_IPE = 0
replace cov_IPE = 1 if return(IPE_LB)<return(true weib_aft_af) & return(IPE_UB)>return(true weib_aft_af)
replace cov_IPE = . if r(psi)==. | r(psilow) ==. | r(psiupp)==.
return scalar cov_IPE = cov_IPE[1]
estimates replay ss
estimates restore ss
gen lnp=_b[ln_p]
gen shape=exp(lnp)
return scalar IPEhr = exp(-shape*_b[trtrand])
return scalar IPEhr_LB = exp(-shape*log(return(IPE_UB)))
return scalar IPEhr_UB = exp(-shape*log(return(IPE_LB)))
gen cov_IPEhr = 0
replace cov_IPEhr = 1 if return(IPEhr_LB)<return(truecox_hr) & return(IPEhr_UB)>return(truecox_hr)
replace cov_IPEhr = . if return(IPE)==. | return(IPE_UB) ==. | return(IPE_LB) ==.
return scalar cov_IPEhr = cov_IPEhr[1]
gen cons=_b[_cons]*-shape
drop _t0 _t_d _st xoOSgainobs cens died xotime timePFSobsgrp timePFSobs timeOS2 admin bprog xoti trtrand
id
drop if _n>1
expand 1096
gen time=_n-1
gen surv =1
replace surv= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc1=total(auc)

```

```

summ auc1
return scalar IPE_extrapauc3_weib_con=r(mean)
restore
***AUC***
sort id
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPE))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar IPE_adj_auc_con = r(rmean)
return scalar IPE_adj_auc_con_SE = r(se)
return scalar IPE_adj_auc_con_LB = r(lb)
return scalar IPE_adj_auc_con_UB = r(ub)
gen cov_IPE_adj_auc_con = 0
replace cov_IPE_adj_auc_con = 1 if return(trueauc_con) > return(IPE_adj_auc_con_LB) & return(trueauc_con) <
return(IPE_adj_auc_con_UB)
return scalar cov_IPE_adj_auc_con = cov_IPE_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPE_adj_cox_hr = exp(_b[trtrand])
return scalar IPE_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPE_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar IPE_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_IPE_adj_cox_hr = 0
replace cov_IPE_adj_cox_hr = 1 if return(truecox_hr) > return(IPE_adj_cox_hr_LB) & return(truecox_hr) <
return(IPE_adj_cox_hr_UB)
return scalar cov_IPE_adj_cox_hr = cov_IPE_adj_cox_hr[1]
restore
sort id
preserve
***Lower 95% CI***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPE_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean

```

```

return scalar IPE_adj_lowci_auc_con = r(rmean)
return scalar IPE_adj_lowci_auc_con_SE = r(se)
return scalar IPE_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPE_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar IPE_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPE_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPE_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPE_adj_upci_auc_con = r(rmean)
return scalar IPE_adj_upci_auc_con_SE = r(se)
return scalar IPE_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPE_adj_upci_cox_hr = exp(_b[trtrand])
return scalar IPE_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPE_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***estimate from exp survivor function***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPE) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPE)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)

```

```

summ auc1
return scalar IPE_sfunc_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPE_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPE_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPE_sfunc_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPE_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPE_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPE_sfunc_auc_con_UB=r(mean)
gen cov_IPE_sfunc_auc_con = 0
replace cov_IPE_sfunc_auc_con = 1 if return(trueauc_con) > return(IPE_sfunc_auc_con_LB) &
return(trueauc_con) < return(IPE_sfunc_auc_con_UB)
return scalar cov_IPE_sfunc_auc_con = cov_IPE_sfunc_auc_con[1]
restore

```

```

***IPE Algorithm - Exponential***
preserve
collapse (max) trtrand xoti bprog admin timeOS2 timePFSobs timePFSobsgrp xotime died cens xoOSgainobs,
by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xo0(xotime xoti) test(exponential) endstudy(admin) tol(3) maxiter(1000) ipe ipest(tt)
return scalar IPEexp = exp(-r(psi))
return scalar IPEexp_UB = exp(-r(psilow))
return scalar IPEexp_LB = exp(-r(psiupp))
gen cov_IPEexp = 0
replace cov_IPEexp = 1 if return(IPEexp_LB)<return(trueweib_aft_af) &
return(IPEexp_UB)>return(trueweib_aft_af)
replace cov_IPEexp = . if r(psi)==. | r(psilow) ==. | r(psiupp)==.
return scalar cov_IPEexp = cov_IPEexp[1]
estimates replay tt
estimates restore tt
gen cons=(-_b[_cons])
gen shape=1
return scalar IPEexphr = exp(-shape*_b[trtrand])
return scalar IPEexphr_LB = exp(-shape*log(return(IPEexp_UB)))
return scalar IPEexphr_UB = exp(-shape*log(return(IPEexp_LB)))
gen cov_IPEexphr = 0
replace cov_IPEexphr = 1 if return(IPEexphr_LB)<return(truecox_hr) &
return(IPEexphr_UB)>return(truecox_hr)
replace cov_IPEexphr = . if return(IPEexp)==. | return(IPEexp_UB) ==. | return(IPEexp_LB)==.
return scalar cov_IPEexphr = cov_IPEexphr[1]
drop _t0 _t_d _st xoOSgainobs cens died xotime timePFSobsgrp timePFSobs timeOS2 admin bprog xoti trtrand
id
drop if _n>1
expand 1096
gen time=_n-1
gen surv =1
replace surv= exp((-exp(cons))*time) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPE_extrapauc3_exp_con=r(mean)
***Lower CI***
estimates restore tt
replace cons=(-_b[_cons])+(1.96*_se[_cons])
replace surv= exp((-exp(cons))*time) if _n>1
replace auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc2=total(auc)
summ auc2
return scalar IPE_extrapauc3_exp_con_lowci=r(mean)
***Upper CI***
estimates restore tt
replace cons=(-_b[_cons])-(1.96*_se[_cons])
replace surv= exp((-exp(cons))*time) if _n>1
replace auc=0

```

```

replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc3=total(auc)
summ auc3
return scalar IPE_extrapauc3_exp_con_upci=r(mean)
restore
sort id
preserve
***AUC***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEexp))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEexp_adj_auc_con = r(rmean)
return scalar IPEexp_adj_auc_con_SE = r(se)
return scalar IPEexp_adj_auc_con_LB = r(lb)
return scalar IPEexp_adj_auc_con_UB = r(ub)
gen cov_IPEexp_adj_auc_con = 0
replace cov_IPEexp_adj_auc_con = 1 if return(trueauc_con) > return(IPEexp_adj_auc_con_LB) &
return(trueauc_con) < return(IPEexp_adj_auc_con_UB)
return scalar cov_IPEexp_adj_auc_con = cov_IPEexp_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEexp_adj_cox_hr = exp(_b[trtrand])
return scalar IPEexp_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEexp_adj_cox_hr_LB = exp(_b[trtrand])-(1.96*_se[trtrand])
return scalar IPEexp_adj_cox_hr_UB = exp(_b[trtrand])+(1.96*_se[trtrand])
gen cov_IPEexp_adj_cox_hr = 0
replace cov_IPEexp_adj_cox_hr = 1 if return(truecox_hr) > return(IPEexp_adj_cox_hr_LB) & return(truecox_hr)
< return(IPEexp_adj_cox_hr_UB)
return scalar cov_IPEexp_adj_cox_hr = cov_IPEexp_adj_cox_hr[1]
restore
sort id
preserve
***lower 95% CI***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEexp_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***

```

```

stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEexp_adj_lowci_auc_con = r(rmean)
return scalar IPEexp_adj_lowci_auc_con_SE = r(se)
return scalar IPEexp_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEexp_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar IPEexp_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEexp_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEexp_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEexp_adj_upci_auc_con = r(rmean)
return scalar IPEexp_adj_upci_auc_con_SE = r(se)
return scalar IPEexp_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEexp_adj_upci_cox_hr = exp(_b[trtrand])
return scalar IPEexp_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEexp_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***estimate from exp survivor function***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEexp) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEexp)

```

```

gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEexp_sfuc_auc_con=r(mean)
restore
***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEexp_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEexp_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEexp_sfuc_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEexp_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf = 1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEexp_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEexp_sfuc_auc_con_UB=r(mean)
gen cov_IPEexp_sfuc_auc_con = 0

```



```

replace cov_IPEexp_sfuc_auc_con = 1 if return(trueauc_con) > return(IPEexp_sfuc_auc_con_LB) &
return(trueauc_con) < return(IPEexp_sfuc_auc_con_UB)
return scalar cov_IPEexp_sfuc_auc_con = cov_IPEexp_sfuc_auc_con[1]
restore
***IPE Algorithm with covariates - Weibull***
sort id
preserve
collapse (max) trtrand xoti bprog cea0grp admin timeOS2 timePFSobs timePFSobsgrp xotime died cens
xoOSgainobs, by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xol(xotime xoti) test(weibull) endstudy(admin) tol(3) maxiter(1000) ipe ipest(ss)
adjvars(bprog cea0grp)
return scalar IPEwc = exp(-r(psi))
return scalar IPEwc_UB = exp(-r(psilow))
return scalar IPEwc_LB = exp(-r(psiupp))
gen cov_IPEwc = 0
replace cov_IPEwc = 1 if return(IPEwc_LB)<return(trueweib_aft_af) &
return(IPEwc_UB)>return(trueweib_aft_af)
replace cov_IPEwc = . if r(psi)==. | r(psilow) ==. | r(psiupp)==.
return scalar cov_IPEwc = cov_IPEwc[1]
estimates replay ss
estimates restore ss
gen lnp=_b[ln_p]
gen shape=exp(lnp)
return scalar IPEwchr = exp(-shape*_b[trtrand])
return scalar IPEwchr_LB = exp(-shape*log(return(IPEwc_UB)))
return scalar IPEwchr_UB = exp(-shape*log(return(IPEwc_LB)))
gen cov_IPEwchr = 0
replace cov_IPEwchr = 1 if return(IPEwchr_LB)<return(truecox_hr) & return(IPEwchr_UB)>return(truecox_hr)
replace cov_IPEwchr = . if return(IPEwc)==. | return(IPEwc_UB) ==. | return(IPEwc_LB)==.
return scalar cov_IPEwchr = cov_IPEwchr[1]
gen cons=_b[cons]*-shape
gen bprog=_b[bprog]*-shape
gen cea0c=_b[cea0grp]*-shape
egen bprogmean2= mean(bprog) if trtrand==0
egen cea0mean2= mean(cea0grp) if trtrand==0
egen bprogmean = max(bprogmean2)
egen cea0mean = max(cea0mean2)
drop _t0 _t_d _st xoOSgainobs cens died xotime timePFSobsgrp timePFSobs timeOS2 admin bprog xoti trtrand
id bprogmean2 cea0mean2
drop if _n>1
expand 1096
gen time=_n-1
gen surv =1
replace surv= exp((-exp(cons+(bprogmean*bprog)+(cea0mean*cea0c))*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*(surv+surv[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEwc_extrapauc3_weib_con=r(mean)
restore
***AUC***

```

```

sort id
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEwc))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
stci if trtrand==0, rmean
return scalar IPEwc_adj_auc_con = r(rmean)
return scalar IPEwc_adj_auc_con_SE = r(se)
return scalar IPEwc_adj_auc_con_LB = r(lb)
return scalar IPEwc_adj_auc_con_UB = r(ub)
gen cov_IPEwc_adj_auc_con = 0
replace cov_IPEwc_adj_auc_con = 1 if return(trueauc_con) > return(IPEwc_adj_auc_con_LB) &
return(trueauc_con) < return(IPEwc_adj_auc_con_UB)
return scalar cov_IPEwc_adj_auc_con = cov_IPEwc_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEwc_adj_cox_hr = exp(_b[trtrand])
return scalar IPEwc_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEwc_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar IPEwc_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_IPEwc_adj_cox_hr = 0
replace cov_IPEwc_adj_cox_hr = 1 if return(truecox_hr) > return(IPEwc_adj_cox_hr_LB) & return(truecox_hr) <
return(IPEwc_adj_cox_hr_UB)
return scalar cov_IPEwc_adj_cox_hr = cov_IPEwc_adj_cox_hr[1]
restore
sort id
preserve
***Lower 95% CI***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEwc_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEwc_adj_lowci_auc_con = r(rmean)
return scalar IPEwc_adj_lowci_auc_con_SE = r(se)
return scalar IPEwc_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***

```

```

stcox trtrand
return scalar IPEwc_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar IPEwc_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEwc_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEwc_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEwc_adj_upci_auc_con = r(rmean)
return scalar IPEwc_adj_upci_auc_con_SE = r(se)
return scalar IPEwc_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEwc_adj_upci_cox_hr = exp(_b[trtrand])
return scalar IPEwc_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEwc_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***estimate from exp survivor function***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEwc) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEwc)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEwc_sfunc_auc_con=r(mean)
restore
***lower CI***

```

```

preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEwc_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEwc_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEwc_sfunc_auc_con_LB=r(mean)
restore

***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEwc_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[_ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEwc_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEwc_sfunc_auc_con_UB=r(mean)
gen cov_IPEwc_sfunc_auc_con = 0
replace cov_IPEwc_sfunc_auc_con = 1 if return(trueauc_con) > return(IPEwc_sfunc_auc_con_LB) &
return(trueauc_con) < return(IPEwc_sfunc_auc_con_UB)
return scalar cov_IPEwc_sfunc_auc_con = cov_IPEwc_sfunc_auc_con[1]
restore
***IPE Algorithm with covariates - Exponential***
preserve

```

```

collapse (max) trtrand xoti bprog cea0grp admin timeOS2 timePFSobs timePFSobsgrp xotime died cens
xoOSgainobs, by(id)
stset timeOS2, failure(died) id(id)
replace xotime=0 if xoti==0
capture strbee trtrand, xo0(xotime xoti) test(exponential) endstudy(admin) tol(3) maxiter(1000) ipe ipest(tt)
adjvars(bprog cea0grp)
return scalar IPEexpwc = exp(-r(psi))
return scalar IPEexpwc_UB = exp(-r(psilow))
return scalar IPEexpwc_LB = exp(-r(psiupp))
gen cov_IPEexpwc = 0
replace cov_IPEexpwc = 1 if return(IPEexpwc_LB)<return(trueweib_aft_af) &
return(IPEexpwc_UB)>return(trueweib_aft_af)
replace cov_IPEexpwc = . if r(psi)==. | r(psilow) ==. | r(psiupp)==.
return scalar cov_IPEexpwc = cov_IPEexpwc[1]
estimates replay tt
estimates restore tt
gen cons=-(b[cons])
gen shape=1
return scalar IPEexpwchr = exp(-shape*_b[trtrand])
return scalar IPEexpwchr_LB = exp(-shape*log(return(IPEexpwc_UB)))
return scalar IPEexpwchr_UB = exp(-shape*log(return(IPEexpwc_LB)))
gen cov_IPEexpwchr = 0
replace cov_IPEexpwchr = 1 if return(IPEexpwchr_LB)<return(truecox_hr) &
return(IPEexpwchr_UB)>return(truecox_hr)
replace cov_IPEexpwchr = . if return(IPEexpwc)==. | return(IPEexpwc_UB) ==. | return(IPEexpwc_LB)==.
return scalar cov_IPEexpwchr = cov_IPEexpwchr[1]
gen bprog=-(_b[bprog])
gen cea0c=-(_b[cea0grp])
egen bprogmean2= mean(bprog) if trtrand==0
egen cea0mean2= mean(cea0grp) if trtrand==0
egen bprogmean = max(bprogmean2)
egen cea0mean = max(cea0mean2)
drop _t0 _t_d _st xoOSgainobs cens died xotime timePFSobsgrp timePFSobs timeOS2 admin bprog xoti trtrand
id
drop if _n>1
expand 1096
gen time=_n-1
gen surv =1
replace surv= exp((-exp(cons+(bprogmean*bprog)+(cea0mean*cea0c)))*time) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPE_extrapauc3_expwc_con=r(mean)
***Lower CI***
estimates restore tt
replace cons=-(b[cons])+(1.96*_se[cons])
replace bprog=-(_b[bprog])+(1.96*_se[bprog])
replace cea0c=-(_b[cea0grp])+(1.96*_se[cea0grp])
replace surv= exp((-exp(cons+(bprogmean*bprog)+(cea0mean*cea0c)))*time) if _n>1
replace auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc2=total(auc)

```

```

summ auc2
return scalar IPE_extrapauc3_expwc_con_lowci=r(mean)
***Upper CI***
estimates restore tt
replace cons=-(b[cons])-(1.96*_se[cons])
replace bprog=-(_b[bprog])-(1.96*_se[bprog])
replace cea0c=-(_b[cea0grp])-(1.96*_se[cea0grp])
replace surv= exp((-exp(cons+(bprogmean*bprog)+(cea0mean*cea0c)))*time) if _n>1
replace auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc3=total(auc)
summ auc3
return scalar IPE_extrapauc3_expwc_con_upci=r(mean)

restore
sort id
preserve
***AUC***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEexpwc))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEexpwc_adj_auc_con = r(rmean)
return scalar IPEexpwc_adj_auc_con_SE = r(se)
return scalar IPEexpwc_adj_auc_con_LB = r(lb)
return scalar IPEexpwc_adj_auc_con_UB = r(ub)
gen cov_IPEexpwc_adj_auc_con = 0
replace cov_IPEexpwc_adj_auc_con = 1 if return(trueauc_con) > return(IPEexpwc_adj_auc_con_LB) &
return(trueauc_con) < return(IPEexpwc_adj_auc_con_UB)
return scalar cov_IPEexpwc_adj_auc_con = cov_IPEexpwc_adj_auc_con[1]
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEexpwc_adj_cox_hr = exp(_b[trtrand])
return scalar IPEexpwc_adj_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEexpwc_adj_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
return scalar IPEexpwc_adj_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
gen cov_IPEexpwc_adj_cox_hr = 0
replace cov_IPEexpwc_adj_cox_hr = 1 if return(truecox_hr) > return(IPEexpwc_adj_cox_hr_LB) &
return(truecox_hr) < return(IPEexpwc_adj_cox_hr_UB)
return scalar cov_IPEexpwc_adj_cox_hr = cov_IPEexpwc_adj_cox_hr[1]
restore
sort id
preserve

```

```

***lower 95% CI***
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEexpwc_LB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEexpwc_adj_lowci_auc_con = r(rmean)
return scalar IPEexpwc_adj_lowci_auc_con_SE = r(se)
return scalar IPEexpwc_adj_lowci_auc_con_UB = r(ub)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEexpwc_adj_lowci_cox_hr = exp(_b[trtrand])
return scalar IPEexpwc_adj_lowci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEexpwc_adj_lowci_cox_hr_UB = exp((_b[trtrand])+(1.96*_se[trtrand]))
restore
***Upper 95% CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
by id: replace xoOSgainobs=0 if xoOSgainobs==.
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/return(IPEexpwc_UB))*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0

***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stci if trtrand==0, rmean
return scalar IPEexpwc_adj_upci_auc_con = r(rmean)
return scalar IPEexpwc_adj_upci_auc_con_SE = r(se)
return scalar IPEexpwc_adj_upci_auc_con_LB = r(lb)
***Cox Proportional Hazards Analysis***
stcox trtrand
return scalar IPEexpwc_adj_upci_cox_hr = exp(_b[trtrand])
return scalar IPEexpwc_adj_upci_cox_hr_SE = exp(_b[trtrand])*_se[trtrand]
return scalar IPEexpwc_adj_upci_cox_hr_LB = exp((_b[trtrand])-(1.96*_se[trtrand]))
restore
***estimate from exp survivor function***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)

```

```

streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEexpwc) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEexpwc)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEexpwc_sfunc_auc_con=r(mean)
restore

***lower CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEexpwc_UB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)
gen cons=_b[_cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEexpwc_UB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEexpwc_sfunc_auc_con_LB=r(mean)
restore
***upper CI***
preserve
collapse (max) trtrand bprog timeOS2 xoOSgainobs xotime died cens, by(id)
by id: replace xotime=0 if xotime==.
stset timeOS2, failure(died) id(id)
streg if trtrand==1, dist(weibull) nohr
expand 4
gen time = 0
replace time = (1095/1999)*return(IPEexpwc_LB) if _n==_N
egen timeunit=max(time)
replace time=time[_n-1]+timeunit if (_n>1)

```

```

gen cons=_b[cons]
gen lnp=_b[ln_p]
gen survf=1
replace survf= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen survfcontrol= survf
gen timecontrol=time/return(IPEexpwc_LB)
gen auc=0
replace auc =(timecontrol-timecontrol[_n-1])*((survfcontrol+survfcontrol[_n-1])/2)
egen auc1=total(auc)
summ auc1
return scalar IPEexpwc_sfunc_auc_con_UB=r(mean)
gen cov_IPEexpwc_sfunc_auc_con = 0
replace cov_IPEexpwc_sfunc_auc_con = 1 if return(trueauc_con) > return(IPEexpwc_sfunc_auc_con_LB) &
return(trueauc_con) < return(IPEexpwc_sfunc_auc_con_UB)
return scalar cov_IPEexpwc_sfunc_auc_con = cov_IPEexpwc_sfunc_auc_con[1]
restore
end
simulate truecox_hr = r(truecox_hr) ///
    truecox_hr_SE = r(truecox_hr_SE) ///
    truecox_hr_LB = r(truecox_hr_LB) ///
    truecox_hr_UB = r(truecox_hr_UB) ///
    trueauc_con = r(trueauc_con) ///
    trueauc_con_SE = r(trueauc_con_SE) ///
    trueauc_con_LB = r(trueauc_con_LB) ///
    trueauc_con_UB = r(trueauc_con_UB) ///
    trueauc_int = r(trueauc_int) ///
    trueauc_int_SE = r(trueauc_int_SE) ///
    trueauc_int_LB = r(trueauc_int_LB) ///
    trueauc_int_UB = r(trueauc_int_UB) ///
    trueweib_aft_af = r(trueweib_aft_af) ///
    trueweib_aft_af_SE = r(trueweib_aft_af_SE) ///
    trueweib_aft_af_LB = r(trueweib_aft_af_LB) ///
    trueweib_aft_af_UB = r(trueweib_aft_af_UB) ///
    xo_number = r(xo_number) ///
    cens_number = r(cens_number) ///
    itt_auc_con = r(itt_auc_con) ///
    itt_auc_con_SE = r(itt_auc_con_SE) ///
    itt_auc_con_LB = r(itt_auc_con_LB) ///
    itt_auc_con_UB = r(itt_auc_con_UB) ///
    cov_itt_auc_con = r(cov_itt_auc_con) ///
    itt_cox_hr = r(itt_cox_hr) ///
    itt_cox_hr_SE = r(itt_cox_hr_SE) ///
    itt_cox_hr_LB = r(itt_cox_hr_LB) ///
    itt_cox_hr_UB = r(itt_cox_hr_UB) ///
    cov_itt_cox_hr = r(cov_itt_cox_hr) ///
    itt_weib_aft_af = r(itt_weib_aft_af) ///
    itt_weib_aft_af_SE = r(itt_weib_aft_af_SE) ///
    itt_weib_aft_af_LB = r(itt_weib_aft_af_LB) ///
    itt_weib_aft_af_UB = r(itt_weib_aft_af_UB) ///
    cov_itt_weib_aft_af = r(cov_itt_weib_aft_af) ///
    ppxc_auc_con = r(ppxc_auc_con) ///
    ppxc_auc_con_SE = r(ppxc_auc_con_SE) ///
    ppxc_auc_con_LB = r(ppxc_auc_con_LB) ///

```

```

ppxc_auc_con_UB = r(ppxc_auc_con_UB) ///
cov_ppxc_auc_con = r(cov_ppxc_auc_con) ///
ppxc_cox_hr = r(ppxc_cox_hr) ///
ppxc_cox_hr_SE = r(ppxc_cox_hr_SE) ///
ppxc_cox_hr_LB = r(ppxc_cox_hr_LB) ///
ppxc_cox_hr_UB = r(ppxc_cox_hr_UB) ///
cov_ppxc_cox_hr = r(cov_ppxc_cox_hr) ///
ppcens_auc_con = r(ppcens_auc_con) ///
ppcens_auc_con_SE = r(ppcens_auc_con_SE) ///
ppcens_auc_con_LB = r(ppcens_auc_con_LB) ///
ppcens_auc_con_UB = r(ppcens_auc_con_UB) ///
cov_ppcens_auc_con = r(cov_ppcens_auc_con) ///
ppcens_cox_hr = r(ppcens_cox_hr) ///
ppcens_cox_hr_SE = r(ppcens_cox_hr_SE) ///
ppcens_cox_hr_LB = r(ppcens_cox_hr_LB) ///
ppcens_cox_hr_UB = r(ppcens_cox_hr_UB) ///
cov_ppcens_cox_hr = r(cov_ppcens_cox_hr) ///
tdcm_cox_hr = r(tdcm_cox_hr) ///
tdcm_cox_hr_SE = r(tdcm_cox_hr_SE) ///
tdcm_cox_hr_LB = r(tdcm_cox_hr_LB) ///
tdcm_cox_hr_UB = r(tdcm_cox_hr_UB) ///
cov_tdcm_cox_hr = r(cov_tdcm_cox_hr) ///
tdcm_adj_auc_con = r(tdcm_adj_auc_con) ///
tdcm_adj_auc_con_LB = r(tdcm_adj_auc_con_LB) ///
tdcm_adj_auc_con_UB = r(tdcm_adj_auc_con_UB) ///
cov_tdcm_adj_auc_con = r(cov_tdcm_adj_auc_con) ///
tdcm_weib_af = r(tdcm_weib_af) ///
tdcm_weib_af_SE = r(tdcm_weib_af_SE) ///
tdcm_weib_af_LB = r(tdcm_weib_af_LB) ///
tdcm_weib_af_UB = r(tdcm_weib_af_UB) ///
cov_tdcm_weib_af = r(cov_tdcm_weib_af) ///
tdcm_we_adj_auc_con = r(tdcm_we_adj_auc_con) ///
tdcm_we_adj_auc_con_LB = r(tdcm_we_adj_auc_con_LB) ///
tdcm_we_adj_auc_con_UB = r(tdcm_we_adj_auc_con_UB) ///
cov_tdcm_we_adj_auc_con = r(cov_tdcm_we_adj_auc_con) ///
tdcs_cox_hr = r(tdcs_cox_hr) ///
tdcs_cox_hr_SE = r(tdcs_cox_hr_SE) ///
tdcs_cox_hr_LB = r(tdcs_cox_hr_LB) ///
tdcs_cox_hr_UB = r(tdcs_cox_hr_UB) ///
cov_tdcs_cox_hr = r(cov_tdcs_cox_hr) ///
tdcs_adj_auc_con = r(tdcs_adj_auc_con) ///
tdcs_adj_auc_con_LB = r(tdcs_adj_auc_con_LB) ///
tdcs_adj_auc_con_UB = r(tdcs_adj_auc_con_UB) ///
cov_tdcs_adj_auc_con = r(cov_tdcs_adj_auc_con) ///
tdcs_weib_af = r(tdcs_weib_af) ///
tdcs_weib_af_SE = r(tdcs_weib_af_SE) ///
tdcs_weib_af_LB = r(tdcs_weib_af_LB) ///
tdcs_weib_af_UB = r(tdcs_weib_af_UB) ///
cov_tdcs_weib_af = r(cov_tdcs_weib_af) ///
tdcs_we_adj_auc_con = r(tdcs_we_adj_auc_con) ///
tdcs_we_adj_auc_con_LB = r(tdcs_we_adj_auc_con_LB) ///
tdcs_we_adj_auc_con_UB = r(tdcs_we_adj_auc_con_UB) ///
cov_tdcs_we_adj_auc_con = r(cov_tdcs_we_adj_auc_con) ///

```

```

exotdcm_cox_hr = r(exotdcm_cox_hr) ///
exotdcm_cox_hr_SE = r(exotdcm_cox_hr_SE) ///
exotdcm_cox_hr_LB = r(exotdcm_cox_hr_LB) ///
exotdcm_cox_hr_UB = r(exotdcm_cox_hr_UB) ///
cov_exotdcm_cox_hr = r(cov_exotdcm_cox_hr) ///
xotdcm_cox_hr = r(xotdcm_cox_hr) ///
xotdcm_cox_hr_SE = r(xotdcm_cox_hr_SE) ///
xotdcm_cox_hr_LB = r(xotdcm_cox_hr_LB) ///
xotdcm_cox_hr_UB = r(xotdcm_cox_hr_UB) ///
xotdcm_adj_auc_con = r(xotdcm_adj_auc_con) ///
xotdcm_adj_auc_con_LB = r(xotdcm_adj_auc_con_LB) ///
xotdcm_adj_auc_con_UB = r(xotdcm_adj_auc_con_UB) ///
cov_xotdcm_adj_auc_con = r(cov_xotdcm_adj_auc_con) ///
exotdcm_weib_af = r(exotdcm_weib_af) ///
exotdcm_weib_af_SE = r(exotdcm_weib_af_SE) ///
exotdcm_weib_af_LB = r(exotdcm_weib_af_LB) ///
exotdcm_weib_af_UB = r(exotdcm_weib_af_UB) ///
cov_exotdcm_weib_af = r(cov_exotdcm_weib_af) ///
xotdcm_weib_af = r(xotdcm_weib_af) ///
xotdcm_weib_af_SE = r(xotdcm_weib_af_SE) ///
xotdcm_weib_af_LB = r(xotdcm_weib_af_LB) ///
xotdcm_weib_af_UB = r(xotdcm_weib_af_UB) ///
xotdcm_we_adj_auc_con = r(xotdcm_we_adj_auc_con) ///
xotdcm_we_adj_auc_con_LB = r(xotdcm_we_adj_auc_con_LB) ///
xotdcm_we_adj_auc_con_UB = r(xotdcm_we_adj_auc_con_UB) ///
cov_xotdcm_we_adj_auc_con = r(cov_xotdcm_we_adj_auc_con) ///
exotdcs_cox_hr = r(exotdcs_cox_hr) ///
exotdcs_cox_hr_SE = r(exotdcs_cox_hr_SE) ///
exotdcs_cox_hr_LB = r(exotdcs_cox_hr_LB) ///
exotdcs_cox_hr_UB = r(exotdcs_cox_hr_UB) ///
cov_exotdcs_cox_hr = r(cov_exotdcs_cox_hr) ///
xotdcs_cox_hr = r(xotdcs_cox_hr) ///
xotdcs_cox_hr_SE = r(xotdcs_cox_hr_SE) ///
xotdcs_cox_hr_LB = r(xotdcs_cox_hr_LB) ///
xotdcs_cox_hr_UB = r(xotdcs_cox_hr_UB) ///
xotdcs_adj_auc_con = r(xotdcs_adj_auc_con) ///
xotdcs_adj_auc_con_LB = r(xotdcs_adj_auc_con_LB) ///
xotdcs_adj_auc_con_UB = r(xotdcs_adj_auc_con_UB) ///
cov_xotdcs_adj_auc_con = r(cov_xotdcs_adj_auc_con) ///
exotdcs_weib_af = r(exotdcs_weib_af) ///
exotdcs_weib_af_SE = r(exotdcs_weib_af_SE) ///
exotdcs_weib_af_LB = r(exotdcs_weib_af_LB) ///
exotdcs_weib_af_UB = r(exotdcs_weib_af_UB) ///
cov_exotdcs_weib_af = r(cov_exotdcs_weib_af) ///
xotdcs_weib_af = r(xotdcs_weib_af) ///
xotdcs_weib_af_SE = r(xotdcs_weib_af_SE) ///
xotdcs_weib_af_LB = r(xotdcs_weib_af_LB) ///
xotdcs_weib_af_UB = r(xotdcs_weib_af_UB) ///
xotdcs_we_adj_auc_con = r(xotdcs_we_adj_auc_con) ///
xotdcs_we_adj_auc_con_LB = r(xotdcs_we_adj_auc_con_LB) ///
xotdcs_we_adj_auc_con_UB = r(xotdcs_we_adj_auc_con_UB) ///
cov_xotdcs_we_adj_auc_con = r(cov_xotdcs_we_adj_auc_con) ///
ipcw_cox_hr = r(ipcw_cox_hr) ///

```

```

ipcw_cox_hr_SE = r(ipcw_cox_hr_SE) ///
ipcw_cox_hr_LB = r(ipcw_cox_hr_LB) ///
ipcw_cox_hr_UB = r(ipcw_cox_hr_UB) ///
cov_ipcw_cox_hr = r(cov_ipcw_cox_hr) ///
ipcw_weight_min = r(ipcw_weight_min) ///
ipcw_weight_max = r(ipcw_weight_max) ///
ipcw_adj_auc_conwkm = r(ipcw_adj_auc_conwkm) ///
ipcw_adj_auc_con = r(ipcw_adj_auc_con) ///
ipcw_adj_auc_con_LB = r(ipcw_adj_auc_con_LB) ///
ipcw_adj_auc_con_UB = r(ipcw_adj_auc_con_UB) ///
cov_ipcw_adj_auc_con = r(cov_ipcw_adj_auc_con) ///
ipcwnocea_cox_hr = r(ipcwnocea_cox_hr) ///
ipcwnocea_cox_hr_SE = r(ipcwnocea_cox_hr_SE) ///
ipcwnocea_cox_hr_LB = r(ipcwnocea_cox_hr_LB) ///
ipcwnocea_cox_hr_UB = r(ipcwnocea_cox_hr_UB) ///
cov_ipcwnocea_cox_hr = r(cov_ipcwnocea_cox_hr) ///
ipcwnocea_weight_min = r(ipcwnocea_weight_min) ///
ipcwnocea_weight_max = r(ipcwnocea_weight_max) ///
ipcwnocea_adj_auc_conwkm = r(ipcwnocea_adj_auc_conwkm) ///
ipcwnocea_adj_auc_con = r(ipcwnocea_adj_auc_con) ///
ipcwnocea_adj_auc_con_LB = r(ipcwnocea_adj_auc_con_LB) ///
ipcwnocea_adj_auc_con_UB = r(ipcwnocea_adj_auc_con_UB) ///
cov_ipcwnocea_adj_auc_con = r(cov_ipcwnocea_adj_auc_con) ///
msm_cox_hr = r(msm_cox_hr) ///
msm_cox_hr_SE = r(msm_cox_hr_SE) ///
msm_cox_hr_LB = r(msm_cox_hr_LB) ///
msm_cox_hr_UB = r(msm_cox_hr_UB) ///
cov_msm_cox_hr = r(cov_msm_cox_hr) ///
msm_weight_min = r(msm_weight_min) ///
msm_weight_max = r(msm_weight_max) ///
msmwkm_adj_auc_con = r(msmwkm_adj_auc_con) ///
msm_adj_auc_con = r(msm_adj_auc_con) ///
msm_adj_auc_con_LB = r(msm_adj_auc_con_LB) ///
msm_adj_auc_con_UB = r(msm_adj_auc_con_UB) ///
cov_msm_adj_auc_con = r(cov_msm_adj_auc_con) ///
msmnocea_cox_hr = r(msmnocea_cox_hr) ///
msmnocea_cox_hr_SE = r(msmnocea_cox_hr_SE) ///
msmnocea_cox_hr_LB = r(msmnocea_cox_hr_LB) ///
msmnocea_cox_hr_UB = r(msmnocea_cox_hr_UB) ///
cov_msmnocea_cox_hr = r(cov_msmnocea_cox_hr) ///
msmnocea_weight_min = r(msmnocea_weight_min) ///
msmnocea_weight_max = r(msmnocea_weight_max) ///
msmnoceaawkm_adj_auc_con = r(msmnoceaawkm_adj_auc_con) ///
msmnocea_adj_auc_con = r(msmnocea_adj_auc_con) ///
msmnocea_adj_auc_con_LB = r(msmnocea_adj_auc_con_LB) ///
msmnocea_adj_auc_con_UB = r(msmnocea_adj_auc_con_UB) ///
cov_msmnocea_adj_auc_con = r(cov_msmnocea_adj_auc_con) ///
weib2s_weib_af = r(weib2s_weib_af) ///
weib2s_weib_af_SE = r(weib2s_weib_af_SE) ///
weib2s_weib_af_LB = r(weib2s_weib_af_LB) ///
weib2s_weib_af_UB = r(weib2s_weib_af_UB) ///
weib2s_adj_auc_con = r(weib2s_adj_auc_con) ///
weib2s_adj_auc_con_SE = r(weib2s_adj_auc_con_SE) ///

```

```

weib2s_adj_auc_con_LB = r(weib2s_adj_auc_con_LB) ///
weib2s_adj_auc_con_UB = r(weib2s_adj_auc_con_UB) ///
cov_weib2s_adj_auc_con = r(cov_weib2s_adj_auc_con) ///
weib2s_adj_cox_hr = r(weib2s_adj_cox_hr) ///
weib2s_adj_cox_hr_SE = r(weib2s_adj_cox_hr_SE) ///
weib2s_adj_cox_hr_LB = r(weib2s_adj_cox_hr_LB) ///
weib2s_adj_cox_hr_UB = r(weib2s_adj_cox_hr_UB) ///
cov_weib2s_adj_cox_hr = r(cov_weib2s_adj_cox_hr) ///
weib2s_adj_lowci_auc_con = r(weib2s_adj_lowci_auc_con) ///
weib2s_adj_lowci_auc_con_SE = r(weib2s_adj_lowci_auc_con_SE) ///
weib2s_adj_lowci_auc_con_UB = r(weib2s_adj_lowci_auc_con_UB) ///
weib2s_adj_lowci_cox_hr = r(weib2s_adj_lowci_cox_hr) ///
weib2s_adj_lowci_cox_hr_SE = r(weib2s_adj_lowci_cox_hr_SE) ///
weib2s_adj_lowci_cox_hr_UB = r(weib2s_adj_lowci_cox_hr_UB) ///
weib2s_adj_upci_auc_con = r(weib2s_adj_upci_auc_con) ///
weib2s_adj_upci_auc_con_SE = r(weib2s_adj_upci_auc_con_SE) ///
weib2s_adj_upci_auc_con_LB = r(weib2s_adj_upci_auc_con_LB) ///
weib2s_adj_upci_cox_hr = r(weib2s_adj_upci_cox_hr) ///
weib2s_adj_upci_cox_hr_SE = r(weib2s_adj_upci_cox_hr_SE) ///
weib2s_adj_upci_cox_hr_LB = r(weib2s_adj_upci_cox_hr_LB) ///
weib2m_weib_af = r(weib2m_weib_af) ///
weib2m_weib_af_SE = r(weib2m_weib_af_SE) ///
weib2m_weib_af_LB = r(weib2m_weib_af_LB) ///
weib2m_weib_af_UB = r(weib2m_weib_af_UB) ///
weib2m_adj_auc_con = r(weib2m_adj_auc_con) ///
weib2m_adj_auc_con_SE = r(weib2m_adj_auc_con_SE) ///
weib2m_adj_auc_con_LB = r(weib2m_adj_auc_con_LB) ///
weib2m_adj_auc_con_UB = r(weib2m_adj_auc_con_UB) ///
cov_weib2m_adj_auc_con = r(cov_weib2m_adj_auc_con) ///
weib2m_adj_cox_hr = r(weib2m_adj_cox_hr) ///
weib2m_adj_cox_hr_SE = r(weib2m_adj_cox_hr_SE) ///
weib2m_adj_cox_hr_LB = r(weib2m_adj_cox_hr_LB) ///
weib2m_adj_cox_hr_UB = r(weib2m_adj_cox_hr_UB) ///
cov_weib2m_adj_cox_hr = r(cov_weib2m_adj_cox_hr) ///
weib2m_adj_lowci_auc_con = r(weib2m_adj_lowci_auc_con) ///
weib2m_adj_lowci_auc_con_SE = r(weib2m_adj_lowci_auc_con_SE) ///
weib2m_adj_lowci_auc_con_UB = r(weib2m_adj_lowci_auc_con_UB) ///
weib2m_adj_lowci_cox_hr = r(weib2m_adj_lowci_cox_hr) ///
weib2m_adj_lowci_cox_hr_SE = r(weib2m_adj_lowci_cox_hr_SE) ///
weib2m_adj_lowci_cox_hr_UB = r(weib2m_adj_lowci_cox_hr_UB) ///
weib2m_adj_upci_auc_con = r(weib2m_adj_upci_auc_con) ///
weib2m_adj_upci_auc_con_SE = r(weib2m_adj_upci_auc_con_SE) ///
weib2m_adj_upci_auc_con_LB = r(weib2m_adj_upci_auc_con_LB) ///
weib2m_adj_upci_cox_hr = r(weib2m_adj_upci_cox_hr) ///
weib2m_adj_upci_cox_hr_SE = r(weib2m_adj_upci_cox_hr_SE) ///
weib2m_adj_upci_cox_hr_LB = r(weib2m_adj_upci_cox_hr_LB) ///
snm2 = r(snm2) ///
snm2_LB = r(snm2_LB) ///
snm2_UB = r(snm2_UB) ///
cov_snm2 = r(cov_snm2) ///
snm2_adj_auc_con = r(snm2_adj_auc_con) ///
snm2_adj_auc_con_SE = r(snm2_adj_auc_con_SE) ///
snm2_adj_auc_con_LB = r(snm2_adj_auc_con_LB) ///

```

```

snm2_adj_auc_con_UB = r(snm2_adj_auc_con_UB) ///
cov_snm2_adj_auc_con = r(cov_snm2_adj_auc_con) ///
snm2_adj_cox_hr = r(snm2_adj_cox_hr) ///
snm2_adj_cox_hr_SE = r(snm2_adj_cox_hr_SE) ///
snm2_adj_cox_hr_LB = r(snm2_adj_cox_hr_LB) ///
snm2_adj_cox_hr_UB = r(snm2_adj_cox_hr_UB) ///
cov_snm2_adj_cox_hr = r(cov_snm2_adj_cox_hr) ///
snm2_adj_lowci_auc_con = r(snm2_adj_lowci_auc_con) ///
snm2_adj_lowci_auc_con_SE = r(snm2_adj_lowci_auc_con_SE) ///
snm2_adj_lowci_auc_con_UB = r(snm2_adj_lowci_auc_con_UB) ///
snm2_adj_lowci_cox_hr = r(snm2_adj_lowci_cox_hr) ///
snm2_adj_lowci_cox_hr_SE = r(snm2_adj_lowci_cox_hr_SE) ///
snm2_adj_lowci_cox_hr_UB = r(snm2_adj_lowci_cox_hr_UB) ///
snm2_adj_upci_auc_con = r(snm2_adj_upci_auc_con) ///
snm2_adj_upci_auc_con_SE = r(snm2_adj_upci_auc_con_SE) ///
snm2_adj_upci_auc_con_LB = r(snm2_adj_upci_auc_con_LB) ///
snm2_adj_upci_cox_hr = r(snm2_adj_upci_cox_hr) ///
snm2_adj_upci_cox_hr_SE = r(snm2_adj_upci_cox_hr_SE) ///
snm2_adj_upci_cox_hr_LB = r(snm2_adj_upci_cox_hr_LB) ///
snm2nob = r(snm2nob) ///
snm2nob_LB = r(snm2nob_LB) ///
snm2nob_UB = r(snm2nob_UB) ///
cov_snm2nob = r(cov_snm2nob) ///
snm2nob_adj_auc_con = r(snm2nob_adj_auc_con) ///
snm2nob_adj_auc_con_SE = r(snm2nob_adj_auc_con_SE) ///
snm2nob_adj_auc_con_LB = r(snm2nob_adj_auc_con_LB) ///
snm2nob_adj_auc_con_UB = r(snm2nob_adj_auc_con_UB) ///
cov_snm2nob_adj_auc_con = r(cov_snm2nob_adj_auc_con) ///
snm2nob_adj_cox_hr = r(snm2nob_adj_cox_hr) ///
snm2nob_adj_cox_hr_SE = r(snm2nob_adj_cox_hr_SE) ///
snm2nob_adj_cox_hr_LB = r(snm2nob_adj_cox_hr_LB) ///
snm2nob_adj_cox_hr_UB = r(snm2nob_adj_cox_hr_UB) ///
cov_snm2nob_adj_cox_hr = r(cov_snm2nob_adj_cox_hr) ///
snm2nob_adj_lowci_auc_con = r(snm2nob_adj_lowci_auc_con) ///
snm2nob_adj_lowci_auc_con_SE = r(snm2nob_adj_lowci_auc_con_SE) ///
snm2nob_adj_lowci_auc_con_UB = r(snm2nob_adj_lowci_auc_con_UB) ///
snm2nob_adj_lowci_cox_hr = r(snm2nob_adj_lowci_cox_hr) ///
snm2nob_adj_lowci_cox_hr_SE = r(snm2nob_adj_lowci_cox_hr_SE) ///
snm2nob_adj_lowci_cox_hr_UB = r(snm2nob_adj_lowci_cox_hr_UB) ///
snm2nob_adj_upci_auc_con = r(snm2nob_adj_upci_auc_con) ///
snm2nob_adj_upci_auc_con_SE = r(snm2nob_adj_upci_auc_con_SE) ///
snm2nob_adj_upci_auc_con_LB = r(snm2nob_adj_upci_auc_con_LB) ///
snm2nob_adj_upci_cox_hr = r(snm2nob_adj_upci_cox_hr) ///
snm2nob_adj_upci_cox_hr_SE = r(snm2nob_adj_upci_cox_hr_SE) ///
snm2nob_adj_upci_cox_hr_LB = r(snm2nob_adj_upci_cox_hr_LB) ///
RPSFTM = r(RPSFTM) ///
RPSFTM_LB = r(RPSFTM_LB) ///
RPSFTM_UB = r(RPSFTM_UB) ///
cov_RPSFTM = r(cov_RPSFTM) ///
RPSFTM_adj_auc_con = r(RPSFTM_adj_auc_con) ///
RPSFTM_adj_auc_con_SE = r(RPSFTM_adj_auc_con_SE) ///
RPSFTM_adj_auc_con_LB = r(RPSFTM_adj_auc_con_LB) ///
RPSFTM_adj_auc_con_UB = r(RPSFTM_adj_auc_con_UB) ///

```

```

cov_RPSFTM_adj_auc_con = r(cov_RPSFTM_adj_auc_con) ///
RPSFTM_adj_cox_hr = r(RPSFTM_adj_cox_hr) ///
RPSFTM_adj_cox_hr_SE = r(RPSFTM_adj_cox_hr_SE) ///
RPSFTM_adj_cox_hr_LB = r(RPSFTM_adj_cox_hr_LB) ///
RPSFTM_adj_cox_hr_UB = r(RPSFTM_adj_cox_hr_UB) ///
cov_RPSFTM_adj_cox_hr = r(cov_RPSFTM_adj_cox_hr) ///
RPSFTM_adj_lowci_auc_con = r(RPSFTM_adj_lowci_auc_con) ///
RPSFTM_adj_lowci_auc_con_SE = r(RPSFTM_adj_lowci_auc_con_SE) ///
RPSFTM_adj_lowci_auc_con_UB = r(RPSFTM_adj_lowci_auc_con_UB) ///
RPSFTM_adj_lowci_cox_hr = r(RPSFTM_adj_lowci_cox_hr) ///
RPSFTM_adj_lowci_cox_hr_SE = r(RPSFTM_adj_lowci_cox_hr_SE) ///
RPSFTM_adj_lowci_cox_hr_UB = r(RPSFTM_adj_lowci_cox_hr_UB) ///
RPSFTM_adj_upci_auc_con = r(RPSFTM_adj_upci_auc_con) ///
RPSFTM_adj_upci_auc_con_SE = r(RPSFTM_adj_upci_auc_con_SE) ///
RPSFTM_adj_upci_auc_con_LB = r(RPSFTM_adj_upci_auc_con_LB) ///
RPSFTM_adj_upci_cox_hr = r(RPSFTM_adj_upci_cox_hr) ///
RPSFTM_adj_upci_cox_hr_SE = r(RPSFTM_adj_upci_cox_hr_SE) ///
RPSFTM_adj_upci_cox_hr_LB = r(RPSFTM_adj_upci_cox_hr_LB) ///
RPSFTM_sfunc_auc_con = r(RPSFTM_sfunc_auc_con) ///
RPSFTM_sfunc_auc_con_LB = r(RPSFTM_sfunc_auc_con_LB) ///
RPSFTM_sfunc_auc_con_UB = r(RPSFTM_sfunc_auc_con_UB) ///
cov_RPSFTM_sfunc_auc_con = r(cov_RPSFTM_sfunc_auc_con) ///
RPSFTM_extrapauc3_weib_con = r(RPSFTM_extrapauc3_weib_con) ///
RPSFTM_extrap_cox_hr = r(RPSFTM_extrap_cox_hr) ///
RPSFTM_extrap_cox_hr_LB = r(RPSFTM_extrap_cox_hr_LB) ///
RPSFTM_extrap_cox_hr_UB = r(RPSFTM_extrap_cox_hr_UB) ///
cov_RPSFTM_extrap_cox_hr = r(cov_RPSFTM_extrap_cox_hr) ///
RPSFTMwc = r(RPSFTMwc) ///
RPSFTMwc_LB = r(RPSFTMwc_LB) ///
RPSFTMwc_UB = r(RPSFTMwc_UB) ///
cov_RPSFTMwc = r(cov_RPSFTMwc) ///
RPSFTMwc_adj_auc_con = r(RPSFTMwc_adj_auc_con) ///
RPSFTMwc_adj_auc_con_SE = r(RPSFTMwc_adj_auc_con_SE) ///
RPSFTMwc_adj_auc_con_LB = r(RPSFTMwc_adj_auc_con_LB) ///
RPSFTMwc_adj_auc_con_UB = r(RPSFTMwc_adj_auc_con_UB) ///
cov_RPSFTMwc_adj_auc_con = r(cov_RPSFTMwc_adj_auc_con) ///
RPSFTMwc_adj_cox_hr = r(RPSFTMwc_adj_cox_hr) ///
RPSFTMwc_adj_cox_hr_SE = r(RPSFTMwc_adj_cox_hr_SE) ///
RPSFTMwc_adj_cox_hr_LB = r(RPSFTMwc_adj_cox_hr_LB) ///
RPSFTMwc_adj_cox_hr_UB = r(RPSFTMwc_adj_cox_hr_UB) ///
cov_RPSFTMwc_adj_cox_hr = r(cov_RPSFTMwc_adj_cox_hr) ///
RPSFTMwc_adj_lowci_auc_con = r(RPSFTMwc_adj_lowci_auc_con) ///
RPSFTMwc_adj_lowci_auc_con_SE = r(RPSFTMwc_adj_lowci_auc_con_SE) ///
RPSFTMwc_adj_lowci_auc_con_UB = r(RPSFTMwc_adj_lowci_auc_con_UB) ///
RPSFTMwc_adj_lowci_cox_hr = r(RPSFTMwc_adj_lowci_cox_hr) ///
RPSFTMwc_adj_lowci_cox_hr_SE = r(RPSFTMwc_adj_lowci_cox_hr_SE) ///
RPSFTMwc_adj_lowci_cox_hr_UB = r(RPSFTMwc_adj_lowci_cox_hr_UB) ///
RPSFTMwc_adj_upci_auc_con = r(RPSFTMwc_adj_upci_auc_con) ///
RPSFTMwc_adj_upci_auc_con_SE = r(RPSFTMwc_adj_upci_auc_con_SE) ///
RPSFTMwc_adj_upci_auc_con_LB = r(RPSFTMwc_adj_upci_auc_con_LB) ///
RPSFTMwc_adj_upci_cox_hr = r(RPSFTMwc_adj_upci_cox_hr) ///
RPSFTMwc_adj_upci_cox_hr_SE = r(RPSFTMwc_adj_upci_cox_hr_SE) ///
RPSFTMwc_adj_upci_cox_hr_LB = r(RPSFTMwc_adj_upci_cox_hr_LB) ///

```

```

RPSFTMwc_sfunc_auc_con = r(RPSFTMwc_sfunc_auc_con) ///
RPSFTMwc_sfunc_auc_con_LB = r(RPSFTMwc_sfunc_auc_con_LB) ///
RPSFTMwc_sfunc_auc_con_UB = r(RPSFTMwc_sfunc_auc_con_UB) ///
cov_RPSFTMwc_sfunc_auc_con = r(cov_RPSFTMwc_sfunc_auc_con) ///
RPSFTMwc_extrapauc3_weib_con = r(RPSFTMwc_extrapauc3_weib_con) ///
RPSFTMwc_extrap_cox_hr = r(RPSFTMwc_extrap_cox_hr) ///
RPSFTMwc_extrap_cox_hr_LB = r(RPSFTMwc_extrap_cox_hr_LB) ///
RPSFTMwc_extrap_cox_hr_UB = r(RPSFTMwc_extrap_cox_hr_UB) ///
cov_RPSFTMwc_extrap_cox_hr = r(cov_RPSFTMwc_extrap_cox_hr) ///
RPSFTMwc_extrapauc3_weib_con = r(RPSFTMwc_extrapauc3_weib_con) ///
IPE = r(IPE) ///
IPE_LB = r(IPE_LB) ///
IPE_UB = r(IPE_UB) ///
cov_IPE = r(cov_IPE) ///
IPEhr = r(IPEhr) ///
IPEhr_LB = r(IPEhr_LB) ///
IPEhr_UB = r(IPEhr_UB) ///
cov_IPEhr = r(cov_IPEhr) ///
IPE_extrapauc3_weib_con = r(IPE_extrapauc3_weib_con) ///
IPE_adj_auc_con = r(IPE_adj_auc_con) ///
IPE_adj_auc_con_SE = r(IPE_adj_auc_con_SE) ///
IPE_adj_auc_con_LB = r(IPE_adj_auc_con_LB) ///
IPE_adj_auc_con_UB = r(IPE_adj_auc_con_UB) ///
cov_IPE_adj_auc_con = r(cov_IPE_adj_auc_con) ///
IPE_adj_cox_hr = r(IPE_adj_cox_hr) ///
IPE_adj_cox_hr_SE = r(IPE_adj_cox_hr_SE) ///
IPE_adj_cox_hr_LB = r(IPE_adj_cox_hr_LB) ///
IPE_adj_cox_hr_UB = r(IPE_adj_cox_hr_UB) ///
cov_IPE_adj_cox_hr = r(cov_IPE_adj_cox_hr) ///
IPE_adj_lowci_auc_con = r(IPE_adj_lowci_auc_con) ///
IPE_adj_lowci_auc_con_SE = r(IPE_adj_lowci_auc_con_SE) ///
IPE_adj_lowci_auc_con_UB = r(IPE_adj_lowci_auc_con_UB) ///
IPE_adj_lowci_cox_hr = r(IPE_adj_lowci_cox_hr) ///
IPE_adj_lowci_cox_hr_SE = r(IPE_adj_lowci_cox_hr_SE) ///
IPE_adj_lowci_cox_hr_UB = r(IPE_adj_lowci_cox_hr_UB) ///
IPE_adj_upci_auc_con = r(IPE_adj_upci_auc_con) ///
IPE_adj_upci_auc_con_SE = r(IPE_adj_upci_auc_con_SE) ///
IPE_adj_upci_auc_con_LB = r(IPE_adj_upci_auc_con_LB) ///
IPE_adj_upci_cox_hr = r(IPE_adj_upci_cox_hr) ///
IPE_adj_upci_cox_hr_SE = r(IPE_adj_upci_cox_hr_SE) ///
IPE_adj_upci_cox_hr_LB = r(IPE_adj_upci_cox_hr_LB) ///
IPE_sfunc_auc_con = r(IPE_sfunc_auc_con) ///
IPE_sfunc_auc_con_LB = r(IPE_sfunc_auc_con_LB) ///
IPE_sfunc_auc_con_UB = r(IPE_sfunc_auc_con_UB) ///
cov_IPE_sfunc_auc_con = r(cov_IPE_sfunc_auc_con) ///
IPEexp = r(IPEexp) ///
IPEexp_LB = r(IPEexp_LB) ///
IPEexp_UB = r(IPEexp_UB) ///
cov_IPEexp = r(cov_IPEexp) ///
IPEexphr = r(IPEexphr) ///
IPEexphr_LB = r(IPEexphr_LB) ///
IPEexphr_UB = r(IPEexphr_UB) ///
cov_IPEexphr = r(cov_IPEexphr) ///

```



```

IPE_extrapauc3_exp_con = r(IPE_extrapauc3_exp_con) ///
IPE_extrapauc3_exp_con_lowci = r(IPE_extrapauc3_exp_con_lowci) ///
IPE_extrapauc3_exp_con_upci = r(IPE_extrapauc3_exp_con_upci) ///
IPEexp_adj_auc_con = r(IPEexp_adj_auc_con) ///
IPEexp_adj_auc_con_SE = r(IPEexp_adj_auc_con_SE) ///
IPEexp_adj_auc_con_LB = r(IPEexp_adj_auc_con_LB) ///
IPEexp_adj_auc_con_UB = r(IPEexp_adj_auc_con_UB) ///
cov_IPEexp_adj_auc_con = r(cov_IPEexp_adj_auc_con) ///
IPEexp_adj_cox_hr = r(IPEexp_adj_cox_hr) ///
IPEexp_adj_cox_hr_SE = r(IPEexp_adj_cox_hr_SE) ///
IPEexp_adj_cox_hr_LB = r(IPEexp_adj_cox_hr_LB) ///
IPEexp_adj_cox_hr_UB = r(IPEexp_adj_cox_hr_UB) ///
cov_IPEexp_adj_cox_hr = r(cov_IPEexp_adj_cox_hr) ///
IPEexp_adj_lowci_auc_con = r(IPEexp_adj_lowci_auc_con) ///
IPEexp_adj_lowci_auc_con_SE = r(IPEexp_adj_lowci_auc_con_SE) ///
IPEexp_adj_lowci_auc_con_UB = r(IPEexp_adj_lowci_auc_con_UB) ///
IPEexp_adj_lowci_cox_hr = r(IPEexp_adj_lowci_cox_hr) ///
IPEexp_adj_lowci_cox_hr_SE = r(IPEexp_adj_lowci_cox_hr_SE) ///
IPEexp_adj_lowci_cox_hr_UB = r(IPEexp_adj_lowci_cox_hr_UB) ///
IPEexp_adj_upci_auc_con = r(IPEexp_adj_upci_auc_con) ///
IPEexp_adj_upci_auc_con_SE = r(IPEexp_adj_upci_auc_con_SE) ///
IPEexp_adj_upci_auc_con_LB = r(IPEexp_adj_upci_auc_con_LB) ///
IPEexp_adj_upci_cox_hr = r(IPEexp_adj_upci_cox_hr) ///
IPEexp_adj_upci_cox_hr_SE = r(IPEexp_adj_upci_cox_hr_SE) ///
IPEexp_adj_upci_cox_hr_LB = r(IPEexp_adj_upci_cox_hr_LB) ///
IPEexp_sfunc_auc_con = r(IPEexp_sfunc_auc_con) ///
IPEexp_sfunc_auc_con_LB = r(IPEexp_sfunc_auc_con_LB) ///
IPEexp_sfunc_auc_con_UB = r(IPEexp_sfunc_auc_con_UB) ///
cov_IPEexp_sfunc_auc_con = r(cov_IPEexp_sfunc_auc_con) ///
IPEwc = r(IPEwc) ///
IPEwc_LB = r(IPEwc_LB) ///
IPEwc_UB = r(IPEwc_UB) ///
cov_IPEwc = r(cov_IPEwc) ///
IPEwchr = r(IPEwchr) ///
IPEwchr_LB = r(IPEwchr_LB) ///
IPEwchr_UB = r(IPEwchr_UB) ///
cov_IPEwchr = r(cov_IPEwchr) ///
IPEwc_extrapauc3_weib_con = r(IPEwc_extrapauc3_weib_con) ///
IPEwc_adj_auc_con = r(IPEwc_adj_auc_con) ///
IPEwc_adj_auc_con_SE = r(IPEwc_adj_auc_con_SE) ///
IPEwc_adj_auc_con_LB = r(IPEwc_adj_auc_con_LB) ///
IPEwc_adj_auc_con_UB = r(IPEwc_adj_auc_con_UB) ///
cov_IPEwc_adj_auc_con = r(cov_IPEwc_adj_auc_con) ///
IPEwc_adj_cox_hr = r(IPEwc_adj_cox_hr) ///
IPEwc_adj_cox_hr_SE = r(IPEwc_adj_cox_hr_SE) ///
IPEwc_adj_cox_hr_LB = r(IPEwc_adj_cox_hr_LB) ///
IPEwc_adj_cox_hr_UB = r(IPEwc_adj_cox_hr_UB) ///
cov_IPEwc_adj_cox_hr = r(cov_IPEwc_adj_cox_hr) ///
IPEwc_adj_lowci_auc_con = r(IPEwc_adj_lowci_auc_con) ///
IPEwc_adj_lowci_auc_con_SE = r(IPEwc_adj_lowci_auc_con_SE) ///
IPEwc_adj_lowci_auc_con_UB = r(IPEwc_adj_lowci_auc_con_UB) ///
IPEwc_adj_lowci_cox_hr = r(IPEwc_adj_lowci_cox_hr) ///
IPEwc_adj_lowci_cox_hr_SE = r(IPEwc_adj_lowci_cox_hr_SE) ///

```

```

IPEwc_adj_lowci_cox_hr_UB = r(IPEwc_adj_lowci_cox_hr_UB) ///
IPEwc_adj_upci_auc_con = r(IPEwc_adj_upci_auc_con) ///
IPEwc_adj_upci_auc_con_SE = r(IPEwc_adj_upci_auc_con_SE) ///
IPEwc_adj_upci_auc_con_LB = r(IPEwc_adj_upci_auc_con_LB) ///
IPEwc_adj_upci_cox_hr = r(IPEwc_adj_upci_cox_hr) ///
IPEwc_adj_upci_cox_hr_SE = r(IPEwc_adj_upci_cox_hr_SE) ///
IPEwc_adj_upci_cox_hr_LB = r(IPEwc_adj_upci_cox_hr_LB) ///
IPEwc_sfunc_auc_con = r(IPEwc_sfunc_auc_con) ///
IPEwc_sfunc_auc_con_LB = r(IPEwc_sfunc_auc_con_LB) ///
IPEwc_sfunc_auc_con_UB = r(IPEwc_sfunc_auc_con_UB) ///
cov_IPEwc_sfunc_auc_con = r(cov_IPEwc_sfunc_auc_con) ///
IPEexpwc = r(IPEexpwc) ///
IPEexpwc_LB = r(IPEexpwc_LB) ///
IPEexpwc_UB = r(IPEexpwc_UB) ///
cov_IPEexpwc = r(cov_IPEexpwc) ///
IPEexpwchr = r(IPEexpwchr) ///
IPEexpwchr_LB = r(IPEexpwchr_LB) ///
IPEexpwchr_UB = r(IPEexpwchr_UB) ///
cov_IPEexpwchr = r(cov_IPEexpwchr) ///
IPE_extrapauc3_expwc_con = r(IPE_extrapauc3_expwc_con) ///
IPE_extrapauc3_expwc_con_lowci = r(IPE_extrapauc3_expwc_con_lowci) ///
IPE_extrapauc3_expwc_con_upci = r(IPE_extrapauc3_expwc_con_upci) ///
IPEexpwc_adj_auc_con = r(IPEexpwc_adj_auc_con) ///
IPEexpwc_adj_auc_con_SE = r(IPEexpwc_adj_auc_con_SE) ///
IPEexpwc_adj_auc_con_LB = r(IPEexpwc_adj_auc_con_LB) ///
IPEexpwc_adj_auc_con_UB = r(IPEexpwc_adj_auc_con_UB) ///
cov_IPEexpwc_adj_auc_con = r(cov_IPEexpwc_adj_auc_con) ///
IPEexpwc_adj_cox_hr = r(IPEexpwc_adj_cox_hr) ///
IPEexpwc_adj_cox_hr_SE = r(IPEexpwc_adj_cox_hr_SE) ///
IPEexpwc_adj_cox_hr_LB = r(IPEexpwc_adj_cox_hr_LB) ///
IPEexpwc_adj_cox_hr_UB = r(IPEexpwc_adj_cox_hr_UB) ///
cov_IPEexpwc_adj_cox_hr = r(cov_IPEexpwc_adj_cox_hr) ///
IPEexpwc_adj_lowci_auc_con = r(IPEexpwc_adj_lowci_auc_con) ///
IPEexpwc_adj_lowci_auc_con_SE = r(IPEexpwc_adj_lowci_auc_con_SE) ///
IPEexpwc_adj_lowci_auc_con_UB = r(IPEexpwc_adj_lowci_auc_con_UB) ///
IPEexpwc_adj_lowci_cox_hr = r(IPEexpwc_adj_lowci_cox_hr) ///
IPEexpwc_adj_lowci_cox_hr_SE = r(IPEexpwc_adj_lowci_cox_hr_SE) ///
IPEexpwc_adj_lowci_cox_hr_UB = r(IPEexpwc_adj_lowci_cox_hr_UB) ///
IPEexpwc_adj_upci_auc_con = r(IPEexpwc_adj_upci_auc_con) ///
IPEexpwc_adj_upci_auc_con_SE = r(IPEexpwc_adj_upci_auc_con_SE) ///
IPEexpwc_adj_upci_auc_con_LB = r(IPEexpwc_adj_upci_auc_con_LB) ///
IPEexpwc_adj_upci_cox_hr = r(IPEexpwc_adj_upci_cox_hr) ///
IPEexpwc_adj_upci_cox_hr_SE = r(IPEexpwc_adj_upci_cox_hr_SE) ///
IPEexpwc_adj_upci_cox_hr_LB = r(IPEexpwc_adj_upci_cox_hr_LB) ///
IPEexpwc_sfunc_auc_con = r(IPEexpwc_sfunc_auc_con) ///
IPEexpwc_sfunc_auc_con_LB = r(IPEexpwc_sfunc_auc_con_LB) ///
IPEexpwc_sfunc_auc_con_UB = r(IPEexpwc_sfunc_auc_con_UB) ///
cov_IPEexpwc_sfunc_auc_con = r(cov_IPEexpwc_sfunc_auc_con) ///
reps(1000) saving(simulationv501): simstudyv501, obs(500) bprog(0.5) lambdasim(0.0005) betain(20)
betas1(15) alphas(0.02) bprogin(5) trtlghr(-0.7) bprogs1(0.5) lintert(-4) tde(0.15) admin(1095) logitcea31(0.1)
logitcea32(0.3) logitcea33(0.6) logcea21(2.2) logcea22(1.6) logcea23(2) logcea11(3.8) logcea12(2.4) logcea13(5)
logxo2(0.7) logxo3(0.4) gammasim(0.9) xomult(0.50267)
***Note, for each of the different scenarios the above syntax is altered***

```


Appendix 10: Case study STATA do files

a) Parametric modelling analysis on data unadjusted for treatment crossover

```
***Extrap model diagnostics***

***para***
streg trtrand, dist(exponential)
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3000) outfile(exp)
estat ic
predict expOS, surv
sts graph, by(trtrand) risktable addplot((line expOS _t if trtrand==0) (line expOS _t if trtrand==1)) xlabel(#6)

**we**
streg trtrand, dist(weibull)
estat ic
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3000) outfile(weib)
predict weibOS, surv
sts graph, by(trtrand) risktable addplot((line weibOS _t if trtrand==0) (line weibOS _t if trtrand==1)) xlabel(#6)

**gomp**
streg trtrand, dist(gompertz)
estat ic
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3000) outfile(gomp)
predict gompOS, surv
sts graph, by(trtrand) risktable addplot((line gompOS _t if trtrand==0) (line gompOS _t if trtrand==1)) xlabel(#6)

**loglogistic**
streg trtrand, dist(loglogistic) time tr
estat ic
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3000) outfile(logl)
predict logIOS, surv
sts graph, by(trtrand) risktable addplot((line logIOS _t if trtrand==0) (line logIOS _t if trtrand==1)) xlabel(#6)

**lognormal**
streg trtrand, dist(lognormal) time tr
estat ic
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3000) outfile(logn)
predict lognOS, surv
sts graph, by(trtrand) risktable addplot((line lognOS _t if trtrand==0) (line lognOS _t if trtrand==1)) xlabel(#6)

**gamma**
streg trtrand, dist(gamma) time tr
estat ic
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3000) outfile(gamma)
predict gammaOS, surv
sts graph, by(trtrand) risktable addplot((line gammaOS _t if trtrand==0) (line gammaOS _t if trtrand==1))
xlabel(#6)

***plots from outfiles***
tway (line weibc _t) (line weibc _t) (line KMc timeKMc) (line KMc timeKMc)
tway (line expc _t) (line expc _t) (line KMc timeKMc) (line KMc timeKMc)
tway (line gompc _t) (line gompc _t) (line KMc timeKMc) (line KMc timeKMc)
```

```
tway (line loglc _t) (line loglc _t) (line KMc timeKMc) (line KMc timeKMc)
tway (line lognc _t) (line lognc _t) (line KMc timeKMc) (line KMc timeKMc)
tway (line gammac _t) (line gammac _t) (line KMc timeKMc) (line KMc timeKMc)
```

Models fit only to exp group

```
***para***
streg if trtrand==1, dist(exponential)
stcurv, survival range(0 3000) outfile(exp2)
estat ic
predict expOS, surv

**we**
streg if trtrand==1, dist(weibull)
estat ic
stcurv, survival range(0 3000) outfile(weib2)
predict weibOS, surv

**gomp**
streg if trtrand==1, dist(gompertz)
estat ic
stcurv, survival range(0 3000) outfile(gomp2)
predict gompOS, surv

**loglogistic**
streg if trtrand==1, dist(loglogistic) time tr
estat ic
stcurv, survival range(0 3000) outfile(logl2)
predict logIOS, surv

**lognormal**
streg if trtrand==1, dist(lognormal) time tr
estat ic
stcurv, survival range(0 3000) outfile(logn2)
predict lognOS, surv

**gamma**
streg if trtrand==1, dist(gamma) time tr
estat ic
stcurv, survival range(0 3000) outfile(gamma2)
predict gammaOS, surv

b) Naive and randomisation-based crossover adjustment methods

stset timeOS2, failure(died) id(id)
preserve
***ITT***
***cox*
stcox trtrand
***cox wc***
stcox trtrand liver metsite ecog
**gamma joint***
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(itt3)
**gamma joint wc***
```

```

streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(itt4)
**gamma ind***
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(itt5)
**gamma ind wc***
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(itt6)

restore
preserve

***Exclude switchers***
preserve
drop if xo==1
**cox*
stcox trtrand
**cox wc***
stcox trtrand liver metsite ecog
**gamma joint***
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(x3)
**gamma joint wc***
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(x4)
**gamma ind***
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(x5)
**gamma ind wc***
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(x6)

***Naive – censor switchers***
preserve
replace died=0 if xo==1
replace timeOS2=xotime if xo==1
stset timeOS2, failure(died) id(id)
**cox*
stcox trtrand
**cox wc***
stcox trtrand liver metsite ecog
**gamma joint***
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(cens3)

**gamma joint wc***
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(cens4)
**gamma ind***
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(cens5)
**gamma ind wc***
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(cens6)

```

*****Treatment as a time-dependent covariate*****

```

preserve
stset timeOS2, failure(died) id(id)

***split the data to one observation per day, so that all xo times can be captured accurately***
***generate a time-dependent indicator for xo***
stsplit timeOS3, every(1)
sort id
gen xoind=0
replace xoind=1 if timeOS3>= xotime
by id: gen finalobs=0
by id: replace finalobs=1 if _n==_N
by id: replace died=0 if _n!=_N
stset timeOS2, failure(died) id(id)
***Next create trtnew***
sort id
gen trtnew=0
replace trtnew=1 if trtrand==1 | xoind==1
**cox*
stcox trtnew
**cox wc***
stcox trtnew liver metsite ecog
**gamma joint***
streg trtnew, dist(gamma) tr
stcurv, survival at1(trtnew=0) at2(trtnew=1) range(0 3700) outfile(tdc3)
**gamma joint wc***
streg trtnew liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtnew=0) at2(trtnew=1) range(0 3700) outfile(tdc4)
**gamma ind***
streg if trtnew==0, dist(gamma)
stcurv, survival range(0 3700) outfile(tdc5)
streg if trtnew==1, dist(gamma)
stcurv, survival range(0 3700) outfile(tdc6)
**gamma ind wc***
streg liver metsite ecog if trtnew==0, dist(gamma)
stcurv, survival range(0 3700) outfile(tdc7)
streg liver metsite ecog if trtnew==1, dist(gamma)
stcurv, survival range(0 3700) outfile(tdc8)

```

*****XO as a time-dependent indicator*****

```

preserve
stset timeOS2, failure(died) id(id)
***split the data to one observation per day, so that all xo times can be captured accurately***
***generate a time-dependent indicator for xo***
stsplit timeOS3, every(1)
sort id
gen xoind=0
replace xoind=1 if timeOS3>= xotime
by id: gen finalobs=0
by id: replace finalobs=1 if _n==_N
by id: replace died=0 if _n!=_N
stset timeOS2, failure(died) id(id)
**cox*

```

```

stcox trtrand xoind
***cox wc***
stcox trtrand xoind liver metsite ecog
***gamma joint***
streg trtrand xoind, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(xo3)
***gamma joint wc***
streg trtrand xoind liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(xo4)
***gamma ind***
streg xoind if trtrand==0, dist(gamma)
stcurv, survival at1(xoind=0) range(0 3700) outfile(xo5)

***gamma ind wc***
streg xoind liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival at1(xoind=0) range(0 3700) outfile(xo7)

***RPSFTM - no covariates***

***note need to collapse data first***
preserve
stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
strbee trtrand, xo0(xotime xo) test(logrank) endstudy(admin)
local RPSFTM = exp(-r(psi))
local RPSFTM_UB = exp(-r(psi_low))
local RPSFTM_LB = exp(-r(psi_upp))
strbee trtrand, xo0(xotime xo) test(logrank) endstudy(admin) gen(counterf) hr
stset counterf, failure(dcounterf) id(id)
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) range(0 3700) outfile(rpsftm3)
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1^RPSFTM)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftm4)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftm5)
restore
sort id
preserve
***Lower 95% CI***

```

```

replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1^RPSFTM_LB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftm6)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftm7)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1^RPSFTM_UB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftm8)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftm9)
restore
sort id
preserve

**Note: another RPSFTM approach is to assume there is only a treatment effect while exp patients are receiving
the treatment**
***eg. Essentially assume that exp patients xo onto the control treatment upon disease prog, as they stop the exp
treatment***
***This follows the structural model more closely - that is  $U_i = T_{0i} + T_{1i}$ , where  $T_0$  is time not on treatment and
 $T_1$  is time on treatment***
***But we don't know treatment discontinuation time for xo patients, so assume that once they're exposed they
remain exposed***
***And likely to be effect of treatment post discontinuation - effect doesn't immediately disappear***
***Alternative is to assume that exp group patients are exposed throughout, as above***
stset timeOS2, failure(died) id(id)

```

```

replace xotime=0 if xo==0
gen xotime1= timePFS2
gen xoexp=0
replace xoexp=1 if trtrand==1 & progxoind==1
replace xotime1=0 if xoexp==0
strbee trtrand, xo0(xotime xo) xo1(xotime1 xoexp) test(logrank) endstudy(admin)
local RPSFTMot = exp(-r(psi))
local RPSFTMot_UB = exp(-r(psi_low))
local RPSFTMot_LB = exp(-r(psi_upp))
strbee trtrand, xo0(xotime xo) xo1(xotime1 xoexp) test(logrank) endstudy(admin) gen(counterf) hr
stset counterf, failure(dcounterf) id(id)
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) range(0 3700) outfile(rpsftmot3)
replace counterf= timeOS2 if trtrand==1
replace dcounterf=died if trtrand==1
stset counterf, failure(dcounterf) id(id)
stcox trtrand
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTMot)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftmot4)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftmot5)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTMot_LB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand

```

```

streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftmot6)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftmot7)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTMot_UB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftmot8)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftmot9)
restore
sort id
preserve

***RPSFTM - with covariates***

***note need to collapse data first***
stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
strbee trtrand, xo0(xotime xo) test(weibull) endstudy(admin) adjvars(liver metsite ecog)
local RPSFTM2 = exp(-r(psi))
local RPSFTM_UB2 = exp(-r(psi_low))
local RPSFTM_LB2 = exp(-r(psi_upp))
strbee trtrand, xo0(xotime xo) test(weibull) endstudy(admin) gen(counterf) adjvars(liver metsite ecog) hr
stset counterf, failure(dcounterf) id(id)
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) range(0 3700) outfile(rpsftm10)
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTM2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***

```



```

***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftm11)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftm12)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTM_LB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftm13)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftm14)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTM_UB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftm15)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftm16)
restore
sort id
preserve

```

```

***RPSFTM With covariates, on treatment version*****
stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
gen xotime1 = timePFS2
gen xoexp=0
replace xoexp=1 if trtrand==1 & progxoind==1
replace xotime1=0 if xoexp==0
strbee trtrand, xo0(xotime xo) xo1(xotime1 xoexp) test(weibull) endstudy(admin) adjvars(liver metsite ecog)
local RPSFTMot2 = exp(-r(psi))
local RPSFTMot_UB2 = exp(-r(psi_low))
local RPSFTMot_LB2 = exp(-r(psi_upp))
strbee trtrand, xo0(xotime xo) xo1(xotime1 xoexp) test(weibull) endstudy(admin) gen(counterf) adjvars(liver
metsite ecog) hr
stset counterf, failure(dcounterf) id(id)
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) range(0 3700) outfile(rpsftmot10)
replace counterf= timeOS2 if trtrand==1
replace dcounterf=died if trtrand==1
stset counterf, failure(dcounterf) id(id)
stcox trtrand liver metsite ecog
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTMot2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftmot11)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftmot12)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/RPSFTMot_LB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2

```

```

by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftmot13)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftmot14)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/^RPSFTMot_UB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(rpsftmot15)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(rpsftmot16)
restore
sort id
preserve

***IPE Algorithm - Weibull, no covariates***

stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
strbee trtrand, xo0(xotime xo) test(weibull) endstudy(admin) tol(3) maxiter(1000) ipe ipest(ss)
local IPE = exp(-r(psi))
local IPE_UB = exp(-r(psilow))
local IPE_LB = exp(-r(psiupp))
estimates replay ss
estimates restore ss
gen lnp=_b[ln_p]
gen shape=exp(lnp)
gen cons=_b[cons]*-shape
drop if _n>1
expand 3700
gen time=_n-1
gen surv =1
replace surv= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)

```

```

egen auc1=total(auc)
summ auc1
restore
sort id
preserve
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/^IPE)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipe2)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipe3)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/^IPE_LB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipe4)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipe5)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0

```

below allows for censoring which means we don't observe the true OS gain for some xo patients

```
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPE_UB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipe6)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipe7)
restore
sort id
preserve
```

IPE Algorithm On treatment - Weibull, no covariates**

```
stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
gen xotime1= timePFS2
gen xoexp=0
replace xoexp=1 if trtrand==1 & progxoid==1
replace xotime1=0 if xoexp==0
strbee trtrand, xo0(xotime xo) xo1(xotime1 xoexp) test(weibull) endstudy(admin) tol(3) maxiter(1000) ipe
gen(counter)
replace counterf= timeOS2 if trtrand==1
replace dcounterf=died if trtrand==1
stset counterf, failure(dcounterf) id(id)
stcox trtrand
ipest(ss)
local IPEot = exp(-r(psi))
local IPEot_UB = exp(-r(psilow))
local IPEot_LB = exp(-r(psiupp))
estimates replay ss
estimates restore ss
gen lnp=_b[ln_p]
gen shape=exp(lnp)
gen cons=_b[_cons]*-shape
drop if _n>1
expand 3700
gen time=_n-1
gen surv =1
replace surv= exp((-exp(cons)*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*(surv+surv[_n-1])/2)
egen auc1=total(auc)
summ auc1
restore
sort id
preserve
***AUC***
sort id
```

by id: replace xotime=0 if xotime==.

```
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPEot)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipeot2)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipeot3)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPEot_LB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipeot4)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipeot5)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPEot_UB)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
```

```

stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand
streg trtrand, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipeot6)
streg if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipeot7)
restore
sort id
preserve

***IPE Algorithm - Weibull, with covariates***
stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
strbee trtrand, xo0(xotime xo) test(weibull) endstudy(admin) tol(3) maxiter(1000) ipe ipest(ss) adjvars(liver
metsite ecog)
local IPE2 = exp(-r(psi))
local IPE_UB2 = exp(-r(psilow))
local IPE_LB2 = exp(-r(psiupp))
estimates replay ss
estimates restore ss
gen lnp=_b[ln_p]
gen shape=exp(lnp)
gen cons=_b[cons]*-shape
gen livercov=_b[liver]*-shape
gen metsitecov=_b[metsite]*-shape
gen ecogcov=_b[ecog]*-shape

egen liver2= mean(liver) if trtrand==0
egen metsite2= mean(metsite) if trtrand==0
egen ecog2= mean(ecog) if trtrand==0
egen livermean = max(liver2)
egen metsitemean = max(metsite2)
egen ecogmean = max(ecog2)
drop if _n>1
expand 3700
gen time=_n-1
gen surv =1
replace surv= exp((-
exp(cons+(livermean*livercov)+(metsitemean*metsitecov)+(ecogmean*ecogcov))*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc1=total(auc)
summ auc1
restore
sort id
preserve
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/IPE2)*xoOSgainobs if xotime>0

```

```

gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipe8)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipe9)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/IPE_LB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
****AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipe10)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipe11)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/IPE_UB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipe12)
streg liver metsite ecog if trtrand==0, dist(gamma)

```

```

stcurv, survival range(0 3700) outfile(ipe13)
restore
sort id
preserve

***IPE Algorithm on treatment - Weibull, with covariates***

stset timeOS2, failure(died) id(id)
replace xotime=0 if xo==0
gen xotime1= timePFS2
gen xoexp=0
replace xoexp=1 if trtrand==1 & progxoid==1
replace xotime1=0 if xoexp==0
strbee trtrand, xo0(xotime xo) xo1(xotime1 xoexp) test(weibull) endstudy(admin) tol(3) maxiter(1000) ipe
gen(counterf)
ipest(ss) adjvars(liver metsite ecog)
local IPEot2 = exp(-r(psi))
local IPEot_UB2 = exp(-r(psilow))
local IPEot_LB2 = exp(-r(psiupp))
replace counterf= timeOS2 if trtrand==1
replace dcounterf=died if trtrand==1
stset counterf, failure(dcounterf) id(id)
stcox trtrand liver metsite ecog
estimates replay ss
estimates restore ss
gen lnp=_b[ln_p]
gen shape=exp(lnp)
gen cons=_b[cons]*-shape
gen livercov=_b[liver]*-shape
gen metsitecov=_b[metsite]*-shape
gen ecogcov=_b[ecog]*-shape
egen liver2= mean(liver) if trtrand==0
egen metsite2= mean(metsite) if trtrand==0
egen ecog2= mean(ecog) if trtrand==0
egen livermean = max(liver2)
egen metsitemean = max(metsite2)
egen ecogmean = max(ecog2)
drop if _n>1
expand 3700
gen time=_n-1
gen surv =1
replace surv= exp((-
exp(cons+(livermean*livercov)+(metsitemean*metsitecov)+(ecogmean*ecogcov))*(time^exp(lnp)))) if _n>1
gen auc=0
replace auc =(time-time[_n-1])*((surv+surv[_n-1])/2)
egen auc1=total(auc)
summ auc1
restore
sort id
preserve
***AUC***
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***

```

```

by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPEot2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
*** Note, don't want to change whether an event was censored or not, so don't alter censoring and dead
indicators***
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipeot8)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipeot9)
restore
sort id
preserve
***Lower 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobsadj=0
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPEot_LB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0

***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
****AUC***
stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipeot10)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipeot11)
restore
sort id
preserve
***Upper 95% CI***
replace xotime=0 if xo==0
sort id
by id: replace xotime=0 if xotime==.
gen xoOSgainobs=0
***below allows for censoring which means we don't observe the true OS gain for some xo patients***
by id: replace xoOSgainobs=timeOS2-xotime
gen xoOSgainobsadj=0
by id: replace xoOSgainobsadj=(1/ IPEot_UB2)*xoOSgainobs if xotime>0
gen timeOS4 = timeOS2
by id: replace timeOS4 = (xotime+xoOSgainobsadj) if trtrand==0 & xotime>0
***do survival analysis on re-estimated survival times***
stset timeOS4, failure(died) id(id)
***AUC***

```

```

stcox trtrand liver metsite ecog
streg trtrand liver metsite ecog, dist(gamma) tr
stcurv, survival at1(trtrand=0) at2(trtrand=1) range(0 3700) outfile(ipeot12)
streg liver metsite ecog if trtrand==0, dist(gamma)
stcurv, survival range(0 3700) outfile(ipeot13)
restore
sort id
preserve
c) IPCW code

use "N:\HEDS\Health Economics\Nick Latimer\Fellowship\GSK data\GSK lap data, stata\final combined GSK
data inc ind ttp.dta", clear
preserve
***note, start off with a dataset that has one observation per patient.***
***Died is 1 or 0 (0 is censored); prog is the indicator for progression and is 1 or 0 (0 is censored); xotime gives
time of crossover or is '!'.***
***add 1 to timeOS2 for all patients so that when split the data have timeOS3 up until censored or event day***
replace timeOS2 = timeOS2+1
stset timeOS2, failure(died) id(id)
***make assumption that no-one who crossed over had progressed at time of crossover, therefore set time of
disease progression (timePFS2) equal to XO time+1, and progression indicator (prog) equal to 1 upon this prog
time***
***the rest here is just manipulating the data to look how I want it - 1 is the event, 0 is no event, . is missing.
Also a crossover indicator, xo, is defined***
replace xo=1 if xotime!=.
replace timePFS2=xotime+1 if xo==1
replace prog=. if prog==0
replace died=. if died==0
gen censOS=0
replace censOS=1 if died==.
***split the data to one observation per day, so that all xo times can be captured accurately***
***generate a time-dependent indicator for xo***
stsplit timeOS3, every(1)
sort id
by id: gen obsno=_n
gen xoind=0
replace xoind=1 if timeOS3>= xotime
***Next mark death as censored for xo patients. Drop data after crossover time. Update all censoring indicators
now that have panel dataset***
sort id
gen finalobs = 0
by id: replace finalobs = 1 if _n==_N
by id: replace died=0 if died==. & finalobs==0
by id: drop if timeOS3 >(xotime)
by id: replace finalobs = 1 if _n==_N
by id: replace died=. if died==0 & finalobs==1
by id: replace prog=0 if timeOS3<timePFS2
by id: replace prog=. if prog==0 & finalobs==1
by id: replace censOS=1 if died==.
gen infcensOS=0
replace infcensOS=1 if xoind==1 & trtrand==0
replace prog=. if id==717|id==1036|id==1091|id==1203|id==1244|id==1264|id==1423|id==1528|id==1529
****New approach - we want to treat crossover as random (just depended on time), assume that everyone had a
chance of crossover pre-progression, hence all at risk*****
gen atrisk=0

```

```

by id: replace atrisk=1 if (trtrand==0 & timeOS3<timePFS2)
***generate an indicator for all censoring of progressive disease - ie this censors patients upon cross-over, or upon
disease progression***
gen censallPD=0
by id: replace censallPD=1 if (infcensOS==1 | prog==1)
***Deriving IPCWs. Knots based on 5, 25, 50, 75 and 95 percentiles***
spbase obsno, knots(24, 124, 279, 497, 834) gen(spline)
sort id obsno
***first find hazard of crossover over time (up until disease progression) in all control group patients whose
disease had not progressed by April 3 2006, taking into account baseline covariates and time since randomisation
(obsno)***
xi: logistic infcensOS i.liver i.metsite i.ecog obsno spline* if trtrand==0 & timeOS3<timePFS2 & atrisk==1
***Predict is then used to estimate the probability of receiving crossover treatment for each subject-day included
in the regression:***
predict ptrtrec if e(sample)
**The above code estimates the probability of each individual being censored each day. For the IPCW we need
the probability of remaining uncensored, so we submit the probabilities from 1:**
replace ptrtrec=ptrtrec*infcensOS+(1-ptrtrec)*(1-infcensOS)
***the time-dependent probability of remaining uncensored is 1 in patients before becoming at risk of xo (pre
Apr3 2006), and after disease progression, and in the exp group***
replace ptrtrec=1 if (timeOS3>=timePFS2 | atrisk==0 | trtrand==1)
replace ptrtrec=1 if ptrtrec==.
**Now we estimate each individual's probability of their complete censoring history up to each day**
sort id obsno
by id: replace ptrtrec=ptrtrec*ptrtrec[_n-1] if _n!=1
rename ptrtrec censdenom
***Secondly we find the hazard for crossover (up until disease progression) in all of the control group and all time
intervals, again with baseline covariates and time since randomisation included in the regression***
sort id obsno
xi: logistic infcensOS i.liver i.metsite i.ecog obsno spline* if trtrand==0
predict ptrtrec2 if e(sample)
replace ptrtrec2=ptrtrec2*infcensOS+(1-ptrtrec2)*(1-infcensOS)
replace ptrtrec2=1 if trtrand==1
replace ptrtrec2=1 if ptrtrec2==.
sort id obsno
by id: replace ptrtrec2 = ptrtrec2*ptrtrec2[_n-1] if _n!=1
rename ptrtrec2 censnum
***The stabilised weight is derived by dividing the numerator by the denominator. This gives us the IPCW for
each patient, taking into account the probability of censoring over time in the xo patients and in the control group
as a whole:***
gen stabweightxo=censnum/censdenom
replace stabweightxo=1 if trtrand==1
***Under the IPCW approach, a time-dependent Cox proportional hazards model can then be estimated to
calculate the treatment effect, adjusting for baseline characteristics and using the time-varying stabilized
weights.***
xi: logistic died trtrand i.liver i.metsite i.ecog obsno spline*[pw=stabweightxo], cluster(id)
restore

```

d) Code for estimating the 'survivor function' survival curves for each method that incorporates this as an extrapolation option

```

***Survivor function code***
***Use this code to estimate 'survivor function' curves for the treatment effects estimated using the rand methods
do file***
***then use AUC on final survivor functions do file to get AUC estimates***

```

```

*****ITT*****
***ITT Cox***
gen hazfITTCoX=hazfexp*(1/0.8619932)
gen ITTCox=1
replace ITTCox=ITTCox[_n-1] * (1-hazfITTCoX) if _n>1
gen hazfITTCoXlb=hazfexp*(1/0.6927252)
gen ITTCoxlb=1
replace ITTCoxlb=ITTCoxlb[_n-1] * (1-hazfITTCoXlb) if _n>1
gen hazfITTCoXub=hazfexp*(1/1.072622)
gen ITTCoxub=1
replace ITTCoxub=ITTCoxub[_n-1] * (1-hazfITTCoXub) if _n>1
***ITT Cox WC***
gen hazfITTCoXwc=hazfexpwc*(1/0.8138756)
gen ITTCoxwc=1
replace ITTCoxwc=ITTCoxwc[_n-1] * (1-hazfITTCoXwc) if _n>1
gen hazfITTCoXwclb=hazfexpwc*(1/0.6519813)
gen ITTCoxwclb=1
replace ITTCoxwclb=ITTCoxwclb[_n-1] * (1-hazfITTCoXwclb) if _n>1
gen hazfITTCoXwcub=hazfexpwc*(1/1.01597)
gen ITTCoxwcub=1
replace ITTCoxwcub=ITTCoxwcub[_n-1] * (1-hazfITTCoXwcub) if _n>1
***GG joint***
gen ConGGITTjointlb= ExpGGITTjointlong
gen timeConGGITTjointlb=timeExpGGITTjointlong/1.340126
gen ConGGITTjointub= ExpGGITTjoint
gen timeConGGITTjointub=time/0.9556028
***GG joint wc***
gen ConGGITTwcjointlb= ExpGGITTwcjointlong
gen timeConGGITTwcjointlb=timeExpGGITTwcjointlong/1.357624
gen ConGGITTwcjointub= ExpGGITTwcjoint
gen timeConGGITTwcjointub=time/0.9898316

*****PPEXc*****
***Cox***
gen hazfxCox=hazfexp*(1/0.7721726)
gen xCox=1
replace xCox=xCox[_n-1] * (1-hazfxCox) if _n>1
gen hazfxCoxlb=hazfexp*(1/0.6156446)
gen xCoxlb=1
replace xCoxlb=xCoxlb[_n-1] * (1-hazfxCoxlb) if _n>1
gen hazfxCoxub=hazfexp*(1/0.9684981)
gen xCoxub=1
replace xCoxub=xCoxub[_n-1] * (1-hazfxCoxub) if _n>1
***Cox WC***
gen hazfxCoxwc=hazfexpwc*(1/0.7539506)
gen xCoxwc=1
replace xCoxwc=xCoxwc[_n-1] * (1-hazfxCoxwc) if _n>1
gen hazfxCoxwclb=hazfexpwc*(1/0.5998617)
gen xCoxwclb=1
replace xCoxwclb=xCoxwclb[_n-1] * (1-hazfxCoxwclb) if _n>1
gen hazfxCoxwcub=hazfexpwc*(1/0.947621)
gen xCoxwcub=1
replace xCoxwcub=xCoxwcub[_n-1] * (1-hazfxCoxwcub) if _n>1
***GG joint***

```

```

gen ConGGxjointlb= ExpGGxjointlong
gen timeConGGxjointlb=timeExpGGxjointlong/1.487681
gen ConGGxjointub= ExpGGxjointlong
gen timeConGGxjointub=timeExpGGxjointlong/1.033733
***GG joint wc***
gen ConGGxwcjointlb= ExpGGxwcjointlong
gen timeConGGxwcjointlb=timeExpGGxwcjointlong/1.485579
gen ConGGxwcjointub= ExpGGxwcjointlong
gen timeConGGxwcjointub=timeExpGGxwcjointlong/1.051388

```

```

*****Cens*****
***Cox***
gen hazfcensCox=hazfexp*(1/0.8192005)
gen censCox=1
replace censCox=censCox[_n-1] * (1-hazfcensCox) if _n>1
gen hazfcensCoxlb=hazfexp*(1/0.6529733)
gen censCoxlb=1
replace censCoxlb=censCoxlb[_n-1] * (1-hazfcensCoxlb) if _n>1
gen hazfcensCoxub=hazfexp*(1/1.027744)
gen censCoxub=1
replace censCoxub=censCoxub[_n-1] * (1-hazfcensCoxub) if _n>1
***Cox WC***
gen hazfcensCoxwc=hazfexpwc*(1/0.7886078)
gen censCoxwc=1
replace censCoxwc=censCoxwc[_n-1] * (1-hazfcensCoxwc) if _n>1
gen hazfcensCoxwclb=hazfexpwc*(1/0.6270101)
gen censCoxwclb=1
replace censCoxwclb=censCoxwclb[_n-1] * (1-hazfcensCoxwclb) if _n>1
gen hazfcensCoxwcub=hazfexpwc*(1/0.9918538)
gen censCoxwcub=1
replace censCoxwcub=censCoxwcub[_n-1] * (1-hazfcensCoxwcub) if _n>1
***GG joint***
gen ConGGcensjointlb= ExpGGcensjointlong
gen timeConGGcensjointlb=timeExpGGcensjointlong/1.393067
gen ConGGcensjointub= ExpGGcensjoint
gen timeConGGcensjointub=time/0.9841903
***GG joint wc***
gen ConGGcenswcjointlb= ExpGGcenswcjointlong
gen timeConGGcenswcjointlb=timeExpGGcenswcjointlong/1.408435
gen ConGGcenswcjointub= ExpGGcenswcjointlong
gen timeConGGcenswcjointub=timeExpGGcenswcjointlong/1.010483

*****TDCM*****
***Cox***
gen hazftdcMcox=hazfexp*(1/0.8093785)
gen tdcMcox=1
replace tdcMcox=tdcMcox[_n-1] * (1-hazftdcMcox) if _n>1
gen hazftdcMcoxlb=hazfexp*(1/0.6490149)
gen tdcMcoxlb=1
replace tdcMcoxlb=tdcMcoxlb[_n-1] * (1-hazftdcMcoxlb) if _n>1
gen hazftdcMcoxub=hazfexp*(1/1.009366)
gen tdcMcoxub=1
replace tdcMcoxub=tdcMcoxub[_n-1] * (1-hazftdcMcoxub) if _n>1
***Cox WC***
gen hazftdcMcoxwc=hazfexpwc*(1/0.7797533)

```

```

gen tdcMcoxwc=1
replace tdcMcoxwc=tdcMcoxwc[_n-1] * (1-hazftdcMcoxwc) if _n>1
gen hazftdcMcoxwclb=hazfexpwc*(1/0.6243915)
gen tdcMcoxwclb=1
replace tdcMcoxwclb=tdcMcoxwclb[_n-1] * (1-hazftdcMcoxwclb) if _n>1
gen hazftdcMcoxwcub=hazfexpwc*(1/0.9737723)
gen tdcMcoxwcub=1
replace tdcMcoxwcub=tdcMcoxwcub[_n-1] * (1-hazftdcMcoxwcub) if _n>1
***GG joint***
gen ConGGtdcmjointlb= ExpGGtdcmjointlong
gen timeConGGtdcmjointlb=timeExpGGtdcmjointlong/1.402805
gen ConGGtdcmjointub= ExpGGtdcmjoint
gen timeConGGtdcmjointub=time/0.9993053
***GG joint wc***
gen ConGGtdcmwcjointlb= ExpGGtdcmwcjointlong
gen timeConGGtdcmwcjointlb=timeExpGGtdcmwcjointlong/1.40892
gen ConGGtdcmwcjointub= ExpGGtdcmwcjointlong
gen timeConGGtdcmwcjointub=timeExpGGtdcmwcjointlong/1.024747

*****TDCM xo indicator*****
***Cox***
gen hazfxoCox=hazfexp*(1/0.8216395)
gen xoCox=1
replace xoCox=xoCox[_n-1] * (1-hazfxoCox) if _n>1
gen hazfxoCoxlb=hazfexp*(1/0.6549424)
gen xoCoxlb=1
replace xoCoxlb=xoCoxlb[_n-1] * (1-hazfxoCoxlb) if _n>1
gen hazfxoCoxub=hazfexp*(1/1.030765)
gen xoCoxub=1
replace xoCoxub=xoCoxub[_n-1] * (1-hazfxoCoxub) if _n>1
***Cox WC***
gen hazfxoCoxwc=hazfexpwc*(1/0.78278)
gen xoCoxwc=1
replace xoCoxwc=xoCoxwc[_n-1] * (1-hazfxoCoxwc) if _n>1
gen hazfxoCoxwclb=hazfexpwc*(1/0.6224807)
gen xoCoxwclb=1
replace xoCoxwclb=xoCoxwclb[_n-1] * (1-hazfxoCoxwclb) if _n>1
gen hazfxoCoxwcub=hazfexpwc*(1/0.9843589)
gen xoCoxwcub=1
replace xoCoxwcub=xoCoxwcub[_n-1] * (1-hazfxoCoxwcub) if _n>1
***GG joint***
gen ConGGxojointlb= ExpGGxojointlong
gen timeConGGxojointlb=timeExpGGxojointlong/1.394009
gen ConGGxojointub= ExpGGxojoint
gen timeConGGxojointub=time/0.9843857
***GG joint wc***
gen ConGGxowcjointlb= ExpGGxowcjointlong
gen timeConGGxowcjointlb=timeExpGGxowcjointlong/1.407628
gen ConGGxowcjointub= ExpGGxowcjointlong
gen timeConGGxowcjointub=timeExpGGxowcjointlong/1.015093

***RPSFTM*****
***survivor function***
gen rpsftmsurv= ExpGGrpsftmlong
gen timerpsftmsurv=timeExpGGrpsftmlong/1.17801014824349

```

```

gen rpsftmsurvlb= ExpGGrpsftmlonglb
gen timerpsftmsurvlb=timeExpGGrpsftmlonglb/1.45753055062422
gen rpsftmsurvub= ExpGG
gen timerpsftmsurvub=time/0.94865071876581

***RPSFTMwc*****
***survivor function***
gen rpsftmwcsurv= ExpGGrpsftmwclong
gen timerpsftmwcsurv=timeExpGGrpsftmwclong/1.2199322202138
gen rpsftmwcsurvlb= ExpGGrpsftmwclonglb
gen timerpsftmwcsurvlb=timeExpGGrpsftmwclonglb/1.45300627030695
gen rpsftmwcsurvub= ExpGGrpsftmwclong
gen timerpsftmwcsurvub=timeExpGGrpsftmwclong/0.992773338805073

***RPSFTM on treatment*****
***survivor function***
gen rpsftmotsurv= ExpGGrpsftmotlong
gen timerpsftmotsurv=timeExpGGrpsftmotlong/1.227750828
gen rpsftmotsurvlb= ExpGGrpsftmotublong
gen timerpsftmotsurvlb=timeExpGGrpsftmotublong/1.612919031
gen rpsftmotsurvub= ExpGGrpsftmotlong
gen timerpsftmotsurvub=timeExpGGrpsftmotlong/0.895934295

***RPSFTMwc on treatment*****
***survivor function***
gen rpsftmotwcsurv= ExpGGrpsftmotwclong
gen timerpsftmotwcsurv=timeExpGGrpsftmotwclong/1.276329908
gen rpsftmotwcsurvlb= ExpGGrpsftmotublong
gen timerpsftmotwcsurvlb=timeExpGGrpsftmotublong/1.606092002
gen rpsftmotwcsurvub= ExpGGrpsftmotlong
gen timerpsftmotwcsurvub=timeExpGGrpsftmotlong/0.984418468

***IPE*****
***survivor function***
gen ipesurv= ExpGGipelong
gen timeipesurv=timeExpGGipelong/1.194226863675
gen ipesurvlb= ExpGGipelonglb
gen timeipesurvlb=timeExpGGipelonglb/1.39477749394074
gen ipesurvub= ExpGGipelong
gen timeipesurvub=timeExpGGipelong/1.0225128744452

***IPEwc*****
***survivor function***
gen ipewcsurv= ExpGGipewclong
gen timeipewcsurv=timeExpGGipewclong/1.22187101157241
gen ipewcsurvlb= ExpGGipewclonglb
gen timeipewcsurvlb=timeExpGGipewclonglb/1.4124223371808
gen ipewcsurvub= ExpGGipelong
gen timeipewcsurvub=timeExpGGipelong/1.05702726367088

***IPE on treatment*****
***survivor function***
gen ipeotsurv= ExpGGipeotlong
gen timeipeotsurv=timeExpGGipeotlong/1.246802657
gen ipeotsurvlb= ExpGGipeotlongub

```



```

gen timeipeotsurvlb=timeExpGGipeotlongub/1.45335547
gen ipeotsurvub= ExpGGipeotlong
gen timeipeotsurvub=timeExpGGipeotlong/1.069605298

```

```

****IPEwc on treatment****

```

```

****survivor function****

```

```

gen ipeotwcurv= ExpGGrpsftmotwclong
gen timeipeotwcurv=timeExpGGrpsftmotwclong/1.275593678
gen ipeotwcurvlb= ExpGGipeotlongub
gen timeipeotwcurvlb=timeExpGGipeotlongub/1.47296533
gen ipeotwcurvub= ExpGGipeotlong
gen timeipeotwcurvub=timeExpGGipeotlong/1.104669063

```

```

****IPCW****

```

```

gen hazfipwCox=hazfexp*(1/0.7503294)
gen ipcwCox=1
replace ipcwCox=ipcwCox[_n-1] * (1-hazfipwCox) if _n>1
gen hazfipwCoxlb=hazfexp*(1/0.5792606)
gen ipcwCoxlb=1
replace ipcwCoxlb=ipcwCoxlb[_n-1] * (1-hazfipwCoxlb) if _n>1
gen hazfipwCoxub=hazfexp*(1/0.9719188)
gen ipcwCoxub=1
replace ipcwCoxub=ipcwCoxub[_n-1] * (1-hazfipwCoxub) if _n>1

```

e) Code for estimating mean survival (area under the curve) for each survival curve

```

****auc from final survivor functions****

```

```

****ITT****

```

```

gen Expauc=0
replace Expauc=(time-time[_n-1])*((ExpGG+ExpGG[_n-1])/2)
egen Expauc1=total(Expauc)
drop Expauc
gen Expaucwc=0
replace Expaucwc=(time-time[_n-1])*((ExpGGwc+ExpGGwc[_n-1])/2)
egen Expaucwc1=total(Expaucwc)
drop Expaucwc
gen ContWKMGGauc=0
replace ContWKMGGauc=(time-time[_n-1])*((ContGGIPCWwkm+ContGGIPCWwkm[_n-1])/2)
egen ContWKMGGauc1=total(ContWKMGGauc)
drop ContWKMGGauc
gen ITTCoxauc=0
replace ITTCoxauc=(time-time[_n-1])*((ITTCox+ITTCox[_n-1])/2)
egen ITTCoxauc1=total(ITTCoxauc)
drop ITTCoxauc
gen ITTCoxaucb=0
replace ITTCoxaucb=(time-time[_n-1])*((ITTCoxlb+ITTCoxlb[_n-1])/2)
egen ITTCoxaucb1=total(ITTCoxaucb)
drop ITTCoxaucb
gen ITTCoxaucub=0
replace ITTCoxaucub=(time-time[_n-1])*((ITTCoxub+ITTCoxub[_n-1])/2)
egen ITTCoxaucub1=total(ITTCoxaucub)
drop ITTCoxaucub
gen ITTCoxwcauc=0
replace ITTCoxwcauc=(time-time[_n-1])*((ITTCoxwc+ITTCoxwc[_n-1])/2)
egen ITTCoxwcauc1=total(ITTCoxwcauc)

```

```

drop ITTCoxwcauc
gen ITTCoxwcaucb=0
replace ITTCoxwcaucb=(time-time[_n-1])*((ITTCoxwclb+ITTCoxwclb[_n-1])/2)
egen ITTCoxwcaucb1=total(ITTCoxwcaucb)
drop ITTCoxwcaucb
gen ITTCoxwcaucub=0
replace ITTCoxwcaucub=(time-time[_n-1])*((ITTCoxwcub+ITTCoxwcub[_n-1])/2)
egen ITTCoxwcaucub1=total(ITTCoxwcaucub)
drop ITTCoxwcaucub
gen ContGGITTjointauc=0
replace ContGGITTjointauc=(time-time[_n-1])*((ContGGITTjoint+ContGGITTjoint[_n-1])/2)
egen ContGGITTjointauc1=total(ContGGITTjointauc)
drop ContGGITTjointauc
gen ContGGITTjointaucb=0
replace ContGGITTjointaucb=(time-time[_n-1])*((ContGGITTjointlb-timeConGGITTjointlb[_n-1])*((ContGGITTjointlb+ContGGITTjointlb[_n-1])/2))
egen ContGGITTjointaucb1=total(ContGGITTjointaucb)
drop ContGGITTjointaucb
gen ContGGITTjointaucub=0
replace ContGGITTjointaucub=(time-time[_n-1])*((ContGGITTjointub-timeConGGITTjointub[_n-1])*((ContGGITTjointub+ContGGITTjointub[_n-1])/2))
egen ContGGITTjointaucub1=total(ContGGITTjointaucub)
drop ContGGITTjointaucub
gen ExpGGITTjointauc=0
replace ExpGGITTjointauc=(time-time[_n-1])*((ExpGGITTjoint+ExpGGITTjoint[_n-1])/2)
egen ExpGGITTjointauc1=total(ExpGGITTjointauc)
drop ExpGGITTjointauc
gen ContGGITTwcjointauc=0
replace ContGGITTwcjointauc=(time-time[_n-1])*((ContGGITTwcjoint+ContGGITTwcjoint[_n-1])/2)
egen ContGGITTwcjointauc1=total(ContGGITTwcjointauc)
drop ContGGITTwcjointauc
gen ContGGITTwcjointaucb=0
replace ContGGITTwcjointaucb=(time-time[_n-1])*((ContGGITTwcjointlb-timeConGGITTwcjointlb[_n-1])*((ContGGITTwcjointlb+ContGGITTwcjointlb[_n-1])/2))
egen ContGGITTwcjointaucb1=total(ContGGITTwcjointaucb)
drop ContGGITTwcjointaucb
gen ContGGITTwcjointaucub=0
replace ContGGITTwcjointaucub=(time-time[_n-1])*((ContGGITTwcjointub-timeConGGITTwcjointub[_n-1])*((ContGGITTwcjointub+ContGGITTwcjointub[_n-1])/2))
egen ContGGITTwcjointaucub1=total(ContGGITTwcjointaucub)
drop ContGGITTwcjointaucub
gen ExpGGITTwcjointauc=0
replace ExpGGITTwcjointauc=(time-time[_n-1])*((ExpGGITTwcjoint+ExpGGITTwcjoint[_n-1])/2)
egen ExpGGITTwcjointauc1=total(ExpGGITTwcjointauc)
drop ExpGGITTwcjointauc
gen ContGGITTindauc=0
replace ContGGITTindauc=(time-time[_n-1])*((ContGGITTind+ContGGITTind[_n-1])/2)
egen ContGGITTindauc1=total(ContGGITTindauc)
drop ContGGITTindauc
gen ContGGITTwcindauc=0
replace ContGGITTwcindauc=(time-time[_n-1])*((ContGGITTwcind+ContGGITTwcind[_n-1])/2)
egen ContGGITTwcindauc1=total(ContGGITTwcindauc)
drop ContGGITTwcindauc

```

```

summ Expauc1 ContWKMGGauc1 ITTCoxauc1 ITTCoxauc1b1 ITTCoxaucub1 ITTCoxwcauc1
ITTCoxwcauc1b1 ITTCoxwcaucub1 ContGGITTjointauc1 ContGGITTjointauc1b1 ContGGITTjointaucub1
ExpGGITTjointauc1 ContGGITTwcjointauc1 ContGGITTwcjointauc1b1 ContGGITTwcjointaucub1
ExpGGITTwcjointauc1 ContGGITTindauc1 ContGGITTwcindauc1

```

```
***Exclude***
```

```

gen Expauc=0
replace Expauc =(time-time[_n-1])*((ExpGG+ExpGG[_n-1])/2)
egen Expauc1=total(Expauc)
drop Expauc
gen Expaucwc=0
replace Expaucwc =(time-time[_n-1])*((ExpGGwc+ExpGGwc[_n-1])/2)
egen Expaucwc1=total(Expaucwc)
drop Expaucwc
gen ContWKMGGauc=0
replace ContWKMGGauc =(time-time[_n-1])*((ContGGIPCWwkm+ContGGIPCWwkm[_n-1])/2)
egen ContWKMGGauc1=total(ContWKMGGauc)
drop ContWKMGGauc
gen xCoxauc=0
replace xCoxauc =(time-time[_n-1])*((xCox+xCox[_n-1])/2)
egen xCoxauc1=total(xCoxauc)
drop xCoxauc
gen xCoxauc1b=0
replace xCoxauc1b =(time-time[_n-1])*((xCox1b+xCox1b[_n-1])/2)
egen xCoxauc1b1=total(xCoxauc1b)
drop xCoxauc1b
gen xCoxaucub=0
replace xCoxaucub =(time-time[_n-1])*((xCoxub+xCoxub[_n-1])/2)
egen xCoxaucub1=total(xCoxaucub)
drop xCoxaucub
gen xCoxwcauc=0
replace xCoxwcauc =(time-time[_n-1])*((xCoxwc+xCoxwc[_n-1])/2)
egen xCoxwcauc1=total(xCoxwcauc)
drop xCoxwcauc
gen xCoxwcauc1b=0
replace xCoxwcauc1b =(time-time[_n-1])*((xCoxwclb+xCoxwclb[_n-1])/2)
egen xCoxwcauc1b1=total(xCoxwcauc1b)
drop xCoxwcauc1b
gen xCoxwcaucub=0
replace xCoxwcaucub =(time-time[_n-1])*((xCoxwcub+xCoxwcub[_n-1])/2)
egen xCoxwcaucub1=total(xCoxwcaucub)
drop xCoxwcaucub
gen ContGGxjointauc=0
replace ContGGxjointauc =(time-time[_n-1])*((ContGGxjoint+ContGGxjoint[_n-1])/2)
egen ContGGxjointauc1=total(ContGGxjointauc)
drop ContGGxjointauc
gen ContGGxjointauc1b=0
replace ContGGxjointauc1b =(timeConGGxjointb-timeConGGxjointb[_n-1])*((ConGGxjointb+
ConGGxjointb[_n-1])/2)
egen ContGGxjointauc1b1=total(ContGGxjointauc1b)
drop ContGGxjointauc1b
gen ContGGxjointaucub=0
replace ContGGxjointaucub =(timeConGGxjointub-timeConGGxjointub[_n-1])*((ConGGxjointub+
ConGGxjointub[_n-1])/2)
egen ContGGxjointaucub1=total(ContGGxjointaucub)

```

```

drop ContGGxjointaucub
gen ExpGGxjointauc=0
replace ExpGGxjointauc =(time-time[_n-1])*((ExpGGxjoint+ExpGGxjoint[_n-1])/2)
egen ExpGGxjointauc1=total(ExpGGxjointauc)
drop ExpGGxjointauc
gen ContGGxwcjointauc=0
replace ContGGxwcjointauc =(time-time[_n-1])*((ConGGxwcjoint+ConGGxwcjoint[_n-1])/2)
egen ContGGxwcjointauc1=total(ContGGxwcjointauc)
drop ContGGxwcjointauc
gen ContGGxwcjointauc1b=0
replace ContGGxwcjointauc1b =(timeConGGxwcjointb-timeConGGxwcjointb[_n-1])*((ConGGxwcjointb+
ConGGxwcjointb[_n-1])/2)
egen ContGGxwcjointauc1b1=total(ContGGxwcjointauc1b)
drop ContGGxwcjointauc1b
gen ContGGxwcjointaucub=0
replace ContGGxwcjointaucub =(timeConGGxwcjointub-timeConGGxwcjointub[_n-1])*((ConGGxwcjointub+
ConGGxwcjointub[_n-1])/2)
egen ContGGxwcjointaucub1=total(ContGGxwcjointaucub)
drop ContGGxwcjointaucub
gen ExpGGxwcjointauc=0
replace ExpGGxwcjointauc =(time-time[_n-1])*((ExpGGxwcjoint+ExpGGxwcjoint[_n-1])/2)
egen ExpGGxwcjointauc1=total(ExpGGxwcjointauc)
drop ExpGGxwcjointauc
gen ContGGxindauc=0
replace ContGGxindauc =(time-time[_n-1])*((ConGGxind+ConGGxind[_n-1])/2)
egen ContGGxindauc1=total(ContGGxindauc)
drop ContGGxindauc
gen ContGGxwcindauc=0
replace ContGGxwcindauc =(time-time[_n-1])*((ConGGxwcind+ConGGxwcind[_n-1])/2)
egen ContGGxwcindauc1=total(ContGGxwcindauc)
drop ContGGxwcindauc

```

```

summ Expauc1 ContWKMGGauc1 xCoxauc1 xCoxauc1b1 xCoxaucub1 xCoxwcauc1 xCoxwcauc1b1
xCoxwcaucub1 ContGGxjointauc1 ContGGxjointauc1b1 ContGGxjointaucub1 ExpGGxjointauc1
ContGGxwcjointauc1 ContGGxwcjointauc1b1 ContGGxwcjointaucub1 ExpGGxwcjointauc1 ContGGxindauc1
ContGGxwcindauc1

```

```
***cens***
```

```

gen Expauc=0
replace Expauc =(time-time[_n-1])*((ExpGG+ExpGG[_n-1])/2)
egen Expauc1=total(Expauc)
drop Expauc
gen Expaucwc=0
replace Expaucwc =(time-time[_n-1])*((ExpGGwc+ExpGGwc[_n-1])/2)
egen Expaucwc1=total(Expaucwc)
drop Expaucwc
gen ContWKMGGauc=0
replace ContWKMGGauc =(time-time[_n-1])*((ContGGIPCWwkm+ContGGIPCWwkm[_n-1])/2)
egen ContWKMGGauc1=total(ContWKMGGauc)
drop ContWKMGGauc
gen censCoxauc=0
replace censCoxauc =(time-time[_n-1])*((censCox+censCox[_n-1])/2)
egen censCoxauc1=total(censCoxauc)
drop censCoxauc
gen censCoxauc1b=0

```

```

replace censCoxauc1b=(time-time[_n-1])*((censCox1b+censCox1b[_n-1])/2)
egen censCoxauc1b1=total(censCoxauc1b)
drop censCoxauc1b
gen censCoxaucub=0
replace censCoxaucub=(time-time[_n-1])*((censCoxub+censCoxub[_n-1])/2)
egen censCoxaucub1=total(censCoxaucub)
drop censCoxaucub
gen censCoxwcauc=0
replace censCoxwcauc=(time-time[_n-1])*((censCoxwc+censCoxwc[_n-1])/2)
egen censCoxwcauc1=total(censCoxwcauc)
drop censCoxwcauc
gen censCoxwcauc1b=0
replace censCoxwcauc1b=(time-time[_n-1])*((censCoxwclb+censCoxwclb[_n-1])/2)
egen censCoxwcauc1b1=total(censCoxwcauc1b)
drop censCoxwcauc1b
gen censCoxwcaucub=0
replace censCoxwcaucub=(time-time[_n-1])*((censCoxwcub+censCoxwcub[_n-1])/2)
egen censCoxwcaucub1=total(censCoxwcaucub)
drop censCoxwcaucub
gen ContGGcensjointauc=0
replace ContGGcensjointauc=(time-time[_n-1])*((ContGGcensjoint+ContGGcensjoint[_n-1])/2)
egen ContGGcensjointauc1=total(ContGGcensjointauc)
drop ContGGcensjointauc
gen ContGGcensjointauc1b=0
replace ContGGcensjointauc1b=(time-ContGGcensjoint1b-timeContGGcensjoint1b[_n-1])*((ContGGcensjoint1b+ContGGcensjoint1b[_n-1])/2)
egen ContGGcensjointauc1b1=total(ContGGcensjointauc1b)
drop ContGGcensjointauc1b
gen ContGGcensjointaucub=0
replace ContGGcensjointaucub=(time-ContGGcensjointub-timeContGGcensjointub[_n-1])*((ContGGcensjointub+ContGGcensjointub[_n-1])/2)
egen ContGGcensjointaucub1=total(ContGGcensjointaucub)
drop ContGGcensjointaucub
gen ExpGGcensjointauc=0
replace ExpGGcensjointauc=(time-time[_n-1])*((ExpGGcensjoint+ExpGGcensjoint[_n-1])/2)
egen ExpGGcensjointauc1=total(ExpGGcensjointauc)
drop ExpGGcensjointauc
gen ContGGcenswcjointauc=0
replace ContGGcenswcjointauc=(time-time[_n-1])*((ContGGcenswcjoint+ContGGcenswcjoint[_n-1])/2)
egen ContGGcenswcjointauc1=total(ContGGcenswcjointauc)
drop ContGGcenswcjointauc
gen ContGGcenswcjointauc1b=0
replace ContGGcenswcjointauc1b=(time-ContGGcenswcjoint1b-timeContGGcenswcjoint1b[_n-1])*((ContGGcenswcjoint1b+ContGGcenswcjoint1b[_n-1])/2)
egen ContGGcenswcjointauc1b1=total(ContGGcenswcjointauc1b)
drop ContGGcenswcjointauc1b
gen ContGGcenswcjointaucub=0
replace ContGGcenswcjointaucub=(time-ContGGcenswcjointub-timeContGGcenswcjointub[_n-1])*((ContGGcenswcjointub+ContGGcenswcjointub[_n-1])/2)
egen ContGGcenswcjointaucub1=total(ContGGcenswcjointaucub)
drop ContGGcenswcjointaucub
gen ExpGGcenswcjointauc=0
replace ExpGGcenswcjointauc=(time-time[_n-1])*((ExpGGcenswcjoint+ExpGGcenswcjoint[_n-1])/2)
egen ExpGGcenswcjointauc1=total(ExpGGcenswcjointauc)
drop ExpGGcenswcjointauc

```

```

gen ContGGcensindauc=0
replace ContGGcensindauc=(time-time[_n-1])*((ContGGcensind+ContGGcensind[_n-1])/2)
egen ContGGcensindauc1=total(ContGGcensindauc)
drop ContGGcensindauc
gen ContGGcenswcindauc=0
replace ContGGcenswcindauc=(time-time[_n-1])*((ContGGcenswcind+ContGGcenswcind[_n-1])/2)
egen ContGGcenswcindauc1=total(ContGGcenswcindauc)
drop ContGGcenswcindauc

summ Expauc1 ContWKMGGauc1 censCoxauc1 censCoxauc1b1 censCoxaucub1 censCoxwcauc1
censCoxwcauc1b1 censCoxwcaucub1 ContGGcensjointauc1 ContGGcensjointauc1b1 ContGGcensjointaucub1
ExpGGcensjointauc1 ContGGcenswcjointauc1 ContGGcenswcjointauc1b1 ContGGcenswcjointaucub1
ExpGGcenswcjointauc1 ContGGcensindauc1 ContGGcenswcindauc1

***tdcm***
preserve
gen Expauc=0
replace Expauc=(time-time[_n-1])*((ExpGG+ExpGG[_n-1])/2)
egen Expauc1=total(Expauc)
drop Expauc
gen Expaucwc=0
replace Expaucwc=(time-time[_n-1])*((ExpGGwc+ExpGGwc[_n-1])/2)
egen Expaucwc1=total(Expaucwc)
drop Expaucwc
gen ContWKMGGauc=0
replace ContWKMGGauc=(time-time[_n-1])*((ContGGIPCWwkm+ContGGIPCWwkm[_n-1])/2)
egen ContWKMGGauc1=total(ContWKMGGauc)
drop ContWKMGGauc
gen tdcMcoxauc=0
replace tdcMcoxauc=(time-time[_n-1])*((tdcMcox+tdcMcox[_n-1])/2)
egen tdcMcoxauc1=total(tdcMcoxauc)
drop tdcMcoxauc
gen tdcMcoxauc1b=0
replace tdcMcoxauc1b=(time-time[_n-1])*((tdcMcox1b+tdcMcox1b[_n-1])/2)
egen tdcMcoxauc1b1=total(tdcMcoxauc1b)
drop tdcMcoxauc1b
gen tdcMcoxaucub=0
replace tdcMcoxaucub=(time-time[_n-1])*((tdcMcoxub+tdcMcoxub[_n-1])/2)
egen tdcMcoxaucub1=total(tdcMcoxaucub)
drop tdcMcoxaucub
gen tdcMcoxwcauc=0
replace tdcMcoxwcauc=(time-time[_n-1])*((tdcMcoxwc+tdcMcoxwc[_n-1])/2)
egen tdcMcoxwcauc1=total(tdcMcoxwcauc)
drop tdcMcoxwcauc
gen tdcMcoxwcauc1b=0
replace tdcMcoxwcauc1b=(time-time[_n-1])*((tdcMcoxwclb+tdcMcoxwclb[_n-1])/2)
egen tdcMcoxwcauc1b1=total(tdcMcoxwcauc1b)
drop tdcMcoxwcauc1b
gen tdcMcoxwcaucub=0
replace tdcMcoxwcaucub=(time-time[_n-1])*((tdcMcoxwcub+tdcMcoxwcub[_n-1])/2)
egen tdcMcoxwcaucub1=total(tdcMcoxwcaucub)
drop tdcMcoxwcaucub
gen ContGGtdcmjointauc=0
replace ContGGtdcmjointauc=(time-time[_n-1])*((ContGGtdcmjoint+ContGGtdcmjoint[_n-1])/2)
egen ContGGtdcmjointauc1=total(ContGGtdcmjointauc)

```

```

drop ContGGtdcmjointauc
gen ContGGtdcmjointauc1b=0
replace ContGGtdcmjointauc1b=(timeConGGtdcmjointlb-timeConGGtdcmjointlb[_n-1])*((ConGGtdcmjointlb+
ConGGtdcmjointlb[_n-1])/2)
egen ContGGtdcmjointauc1b1=total(ContGGtdcmjointauc1b)
drop ContGGtdcmjointauc1b
gen ContGGtdcmjointaucub=0
replace ContGGtdcmjointaucub=(timeConGGtdcmjointub-timeConGGtdcmjointub[_n-1])*((
ConGGtdcmjointub+ ConGGtdcmjointub[_n-1])/2)
egen ContGGtdcmjointaucub1=total(ContGGtdcmjointaucub)
drop ContGGtdcmjointaucub
gen ExpGGtdcmjointauc=0
replace ExpGGtdcmjointauc=(time-time[_n-1])*((ExpGGtdcmjoint + ExpGGtdcmjoint[_n-1])/2)
egen ExpGGtdcmjointauc1=total(ExpGGtdcmjointauc)
drop ExpGGtdcmjointauc
gen ContGGtdcmwcjointauc=0
replace ContGGtdcmwcjointauc=(time-time[_n-1])*((ConGGtdcmwcjoint+ ConGGtdcmwcjoint[_n-1])/2)
egen ContGGtdcmwcjointauc1=total(ContGGtdcmwcjointauc)
drop ContGGtdcmwcjointauc
gen ContGGtdcmwcjointauc1b=0
replace ContGGtdcmwcjointauc1b=(timeConGGtdcmwcjointlb-timeConGGtdcmwcjointlb[_n-
1])*((ConGGtdcmwcjointlb+ ConGGtdcmwcjointlb[_n-1])/2)
egen ContGGtdcmwcjointauc1b1=total(ContGGtdcmwcjointauc1b)
drop ContGGtdcmwcjointauc1b
gen ContGGtdcmwcjointaucub=0
replace ContGGtdcmwcjointaucub=(timeConGGtdcmwcjointub-timeConGGtdcmwcjointub[_n-
1])*((ConGGtdcmwcjointub+ ConGGtdcmwcjointub[_n-1])/2)
egen ContGGtdcmwcjointaucub1=total(ContGGtdcmwcjointaucub)
drop ContGGtdcmwcjointaucub
gen ExpGGtdcmwcjointauc=0
replace ExpGGtdcmwcjointauc=(time-time[_n-1])*((ExpGGtdcmwcjoint + ExpGGtdcmwcjoint[_n-1])/2)
egen ExpGGtdcmwcjointauc1=total(ExpGGtdcmwcjointauc)
drop ExpGGtdcmwcjointauc
gen ContGGtdcmindauc=0
replace ContGGtdcmindauc=(time-time[_n-1])*((ConGGtdcmind+ ConGGtdcmind[_n-1])/2)
egen ContGGtdcmindauc1=total(ContGGtdcmindauc)
drop ContGGtdcmindauc
gen ExpGGtdcmindauc=0
replace ExpGGtdcmindauc=(time-time[_n-1])*((ExpGGtdcmind+ ExpGGtdcmind[_n-1])/2)
egen ExpGGtdcmindauc1=total(ExpGGtdcmindauc)
drop ExpGGtdcmindauc
gen ContGGtdcmwcindauc=0
replace ContGGtdcmwcindauc=(time-time[_n-1])*((ConGGtdcmwcind+ ConGGtdcmwcind[_n-1])/2)
egen ContGGtdcmwcindauc1=total(ContGGtdcmwcindauc)
drop ContGGtdcmwcindauc
gen ExpGGtdcmwcindauc=0
replace ExpGGtdcmwcindauc=(time-time[_n-1])*((ExpGGtdcmwcind+ ExpGGtdcmwcind[_n-1])/2)
egen ExpGGtdcmwcindauc1=total(ExpGGtdcmwcindauc)
drop ExpGGtdcmwcindauc

summ Expauc1 ContWKMGGauc1 tdcxCoxauc1 tdcxCoxauc1b1 tdcxCoxaucub1 tdcxCoxwcauc1
tdcxCoxwcauc1b1 tdcxCoxwcaucub1 ContGGtdcmjointauc1 ContGGtdcmjointauc1b1 ContGGtdcmjointaucub1
ExpGGtdcmjointauc1 ContGGtdcmwcjointauc1 ContGGtdcmwcjointauc1b1 ContGGtdcmwcjointaucub1
ExpGGtdcmwcjointauc1 ContGGtdcmindauc1 ExpGGtdcmindauc1 ContGGtdcmwcindauc1
ExpGGtdcmwcindauc1

```

```

***tdcm with xo ind***
preserve
gen Expauc=0
replace Expauc=(time-time[_n-1])*((ExpGG+ExpGG[_n-1])/2)
egen Expauc1=total(Expauc)
drop Expauc
gen Expaucwc=0
replace Expaucwc=(time-time[_n-1])*((ExpGGwc+ExpGGwc[_n-1])/2)
egen Expaucwc1=total(Expaucwc)
drop Expaucwc
gen ContWKMGGauc=0
replace ContWKMGGauc=(time-time[_n-1])*((ContGGIPCWwkm+ContGGIPCWwkm[_n-1])/2)
egen ContWKMGGauc1=total(ContWKMGGauc)
drop ContWKMGGauc
gen xCoxauc=0
replace xCoxauc=(time-time[_n-1])*((xCox+xCox[_n-1])/2)
egen xCoxauc1=total(xCoxauc)
drop xCoxauc
gen xCoxauc1b=0
replace xCoxauc1b=(time-time[_n-1])*((xCox1b+xCox1b[_n-1])/2)
egen xCoxauc1b1=total(xCoxauc1b)
drop xCoxauc1b
gen xCoxaucub=0
replace xCoxaucub=(time-time[_n-1])*((xCoxub+xCoxub[_n-1])/2)
egen xCoxaucub1=total(xCoxaucub)
drop xCoxaucub
gen xCoxwcauc=0
replace xCoxwcauc=(time-time[_n-1])*((xCoxwc+xCoxwc[_n-1])/2)
egen xCoxwcauc1=total(xCoxwcauc)
drop xCoxwcauc
gen xCoxwcauc1b=0
replace xCoxwcauc1b=(time-time[_n-1])*((xCoxwclb+xCoxwclb[_n-1])/2)
egen xCoxwcauc1b1=total(xCoxwcauc1b)
drop xCoxwcauc1b
gen xCoxwcaucub=0
replace xCoxwcaucub=(time-time[_n-1])*((xCoxwcb+xCoxwcb[_n-1])/2)
egen xCoxwcaucub1=total(xCoxwcaucub)
drop xCoxwcaucub
gen ContGGxojointauc=0
replace ContGGxojointauc=(time-time[_n-1])*((ContGGxojoint+ ContGGxojoint[_n-1])/2)
egen ContGGxojointauc1=total(ContGGxojointauc)
drop ContGGxojointauc
gen ContGGxojointauc1b=0
replace ContGGxojointauc1b=(timeConGGxojointlb-timeConGGxojointlb[_n-1])*((ConGGxojointlb+
ConGGxojointlb[_n-1])/2)
egen ContGGxojointauc1b1=total(ContGGxojointauc1b)
drop ContGGxojointauc1b
gen ContGGxojointaucub=0
replace ContGGxojointaucub=(timeConGGxojointub-timeConGGxojointub[_n-1])*((ConGGxojointub+
ConGGxojointub[_n-1])/2)
egen ContGGxojointaucub1=total(ContGGxojointaucub)
drop ContGGxojointaucub
gen ExpGGxojointauc=0
replace ExpGGxojointauc=(time-time[_n-1])*((ExpGGxojoint + ExpGGxojoint[_n-1])/2)

```

```

egen ExpGGxojointauc1=total(ExpGGxojointauc)
drop ExpGGxojointauc
gen ContGGxowcjointauc=0
replace ContGGxowcjointauc =(time-time[_n-1])*((ConGGxowcjoint+ ConGGxowcjoint[_n-1])/2)
egen ContGGxowcjointauc1=total(ContGGxowcjointauc)
drop ContGGxowcjointauc
gen ContGGxowcjointauc1b=0
replace ContGGxowcjointauc1b =(timeConGGxowcjointlb-timeConGGxowcjointlb[_n-1])*((ConGGxowcjointlb+ ConGGxowcjointlb[_n-1])/2)
egen ContGGxowcjointauc1b1=total(ContGGxowcjointauc1b)
drop ContGGxowcjointauc1b
gen ContGGxowcjointaucub=0
replace ContGGxowcjointaucub =(timeConGGxowcjointub-timeConGGxowcjointub[_n-1])*((ConGGxowcjointub+ ConGGxowcjointub[_n-1])/2)
egen ContGGxowcjointaucub1=total(ContGGxowcjointaucub)
drop ContGGxowcjointaucub
gen ExpGGxowcjointauc=0
replace ExpGGxowcjointauc =(time-time[_n-1])*((ExpGGxowcjoint + ExpGGxowcjoint[_n-1])/2)
egen ExpGGxowcjointauc1=total(ExpGGxowcjointauc)
drop ExpGGxowcjointauc
gen ContGGxoindauc=0
replace ContGGxoindauc =(time-time[_n-1])*((ConGGxoind+ ConGGxoind[_n-1])/2)
egen ContGGxoindauc1=total(ContGGxoindauc)
drop ContGGxoindauc
gen ContGGxowcindauc=0
replace ContGGxowcindauc =(time-time[_n-1])*((ConGGxowcind+ ConGGxowcind[_n-1])/2)
egen ContGGxowcindauc1=total(ContGGxowcindauc)
drop ContGGxowcindauc

```

```

summ Expauc1 ContWKMGGauc1 xCoxauc1 xCoxauc1b1 xCoxaucub1 xCoxwcauc1 xCoxwcauc1b1
xCoxwcaucub1 ContGGxojointauc1 ContGGxojointauc1b1 ContGGxojointaucub1 ExpGGxojointauc1
ContGGxowcjointauc1 ContGGxowcjointauc1b1 ContGGxowcjointaucub1 ExpGGxowcjointauc1
ContGGxoindauc1 ContGGxowcindauc1

```

*****RPSFTM*****

```

gen rpsftmextauc=0
replace rpsftmextauc =(time-time[_n-1])*((rpsftmext+ rpsftmext[_n-1])/2)
egen rpsftmextauc1=total(rpsftmextauc)
drop rpsftmextauc
gen conrpsftmshrinkjointauc=0
replace conrpsftmshrinkjointauc =(time-time[_n-1])*((conrpsftmshrinkjoint+ conrpsftmshrinkjoint[_n-1])/2)
egen conrpsftmshrinkjointauc1=total(conrpsftmshrinkjointauc)
drop conrpsftmshrinkjointauc
gen exprpsftmshrinkjointauc=0
replace exprpsftmshrinkjointauc =(time-time[_n-1])*((exprpsftmshrinkjoint+ exprpsftmshrinkjoint[_n-1])/2)
egen exprpsftmshrinkjointauc1=total(exprpsftmshrinkjointauc)
drop exprpsftmshrinkjointauc
gen conrpsftmshrinkindauc=0
replace conrpsftmshrinkindauc =(time-time[_n-1])*((conrpsftmshrinkind+ conrpsftmshrinkind[_n-1])/2)
egen conrpsftmshrinkindauc1=total(conrpsftmshrinkindauc)
drop conrpsftmshrinkindauc
gen conrpsftmshrinkjointlbauc=0
replace conrpsftmshrinkjointlbauc =(time-time[_n-1])*((conrpsftmshrinkjointlb+ conrpsftmshrinkjointlb[_n-1])/2)
egen conrpsftmshrinkjointlbauc1=total(conrpsftmshrinkjointlbauc)

```

```

drop conrpsftmshrinkjointlbauc
gen exprpsftmshrinkjointlbauc=0
replace exprpsftmshrinkjointlbauc =(time-time[_n-1])*((exprpsftmshrinkjointlb+ exprpsftmshrinkjointlb[_n-1])/2)
egen exprpsftmshrinkjointlbauc1=total(exprpsftmshrinkjointlbauc)
drop exprpsftmshrinkjointlbauc
gen conrpsftmshrinkindlbauc=0
replace conrpsftmshrinkindlbauc =(time-time[_n-1])*((conrpsftmshrinkindlb+ conrpsftmshrinkindlb[_n-1])/2)
egen conrpsftmshrinkindlbauc1=total(conrpsftmshrinkindlbauc)
drop conrpsftmshrinkindlbauc
gen conrpsftmshrinkjointubauc=0
replace conrpsftmshrinkjointubauc =(time-time[_n-1])*((conrpsftmshrinkjointub+ conrpsftmshrinkjointub[_n-1])/2)
egen conrpsftmshrinkjointubauc1=total(conrpsftmshrinkjointubauc)
drop conrpsftmshrinkjointubauc
gen exprpsftmshrinkjointubauc=0
replace exprpsftmshrinkjointubauc =(time-time[_n-1])*((exprpsftmshrinkjointub+ exprpsftmshrinkjointub[_n-1])/2)
egen exprpsftmshrinkjointubauc1=total(exprpsftmshrinkjointubauc)
drop exprpsftmshrinkjointubauc
gen conrpsftmshrinkindubauc=0
replace conrpsftmshrinkindubauc =(time-time[_n-1])*((conrpsftmshrinkindub+ conrpsftmshrinkindub[_n-1])/2)
egen conrpsftmshrinkindubauc1=total(conrpsftmshrinkindubauc)
drop conrpsftmshrinkindubauc
gen rpsftmwextauc=0
replace rpsftmwextauc =(time-time[_n-1])*((rpsftmwext+ rpsftmwext[_n-1])/2)
egen rpsftmwextauc1=total(rpsftmwextauc)
drop rpsftmwextauc
gen conrpsftmweshrinkjointauc=0
replace conrpsftmweshrinkjointauc =(time-time[_n-1])*((conrpsftmweshrinkjoint+ conrpsftmweshrinkjoint[_n-1])/2)
egen conrpsftmweshrinkjointauc1=total(conrpsftmweshrinkjointauc)
drop conrpsftmweshrinkjointauc
gen exprpsftmweshrinkjointauc=0
replace exprpsftmweshrinkjointauc =(time-time[_n-1])*((exprpsftmweshrinkjoint+ exprpsftmweshrinkjoint[_n-1])/2)
egen exprpsftmweshrinkjointauc1=total(exprpsftmweshrinkjointauc)
drop conrpsftmweshrinkjointauc
gen conrpsftmweshrinkindauc=0
replace conrpsftmweshrinkindauc =(time-time[_n-1])*((conrpsftmweshrinkind+ conrpsftmweshrinkind[_n-1])/2)
egen conrpsftmweshrinkindauc1=total(conrpsftmweshrinkindauc)
drop conrpsftmweshrinkindauc
gen conrpsftmweshrinkjointlbauc=0
replace conrpsftmweshrinkjointlbauc =(time-time[_n-1])*((conrpsftmweshrinkjointlb+ conrpsftmweshrinkjointlb[_n-1])/2)
egen conrpsftmweshrinkjointlbauc1=total(conrpsftmweshrinkjointlbauc)
drop conrpsftmweshrinkjointlbauc
gen exprpsftmweshrinkjointlbauc=0
replace exprpsftmweshrinkjointlbauc =(time-time[_n-1])*((exprpsftmweshrinkjointlb+ exprpsftmweshrinkjointlb[_n-1])/2)
egen exprpsftmweshrinkjointlbauc1=total(exprpsftmweshrinkjointlbauc)
drop conrpsftmweshrinkjointlbauc
gen conrpsftmweshrinkindlbauc=0
replace conrpsftmweshrinkindlbauc =(time-time[_n-1])*((conrpsftmweshrinkindlb+ conrpsftmweshrinkindlb[_n-1])/2)

```

```

egen conrpsftmwshrinkindlbauc1=total(conrpsftmwshrinkindlbauc)
drop conrpsftmwshrinkindlbauc
gen conrpsftmwshrinkjointubauc=0
replace conrpsftmwshrinkjointubauc =(time-time[_n-1])*((conrpsftmwshrinkjointub+
conrpsftmwshrinkjointub[_n-1])/2)
egen conrpsftmwshrinkjointubauc1=total(conrpsftmwshrinkjointubauc)
drop conrpsftmwshrinkjointubauc
gen exprpsftmwshrinkjointubauc=0
replace exprpsftmwshrinkjointubauc =(time-time[_n-1])*((exprpsftmwshrinkjointub+
exprpsftmwshrinkjointub[_n-1])/2)
egen exprpsftmwshrinkjointubauc1=total(exprpsftmwshrinkjointubauc)
drop exprpsftmwshrinkjointubauc
gen conrpsftmwshrinkindubauc=0
replace conrpsftmwshrinkindubauc =(time-time[_n-1])*((conrpsftmwshrinkindub+
conrpsftmwshrinkindub[_n-1])/2)
egen conrpsftmwshrinkindubauc1=total(conrpsftmwshrinkindubauc)
drop conrpsftmwshrinkindubauc
gen rpsftmsurvauc=0
replace rpsftmsurvauc =(timerpsftmsurv-timerpsftmsurv[_n-1])*((rpsftmsurv+ rpsftmsurv[_n-1])/2)
egen rpsftmsurvauc1=total(rpsftmsurvauc)
drop rpsftmsurvauc
gen rpsftmsurvlbauc=0
replace rpsftmsurvlbauc =(timerpsftmsurvlb-timerpsftmsurvlb[_n-1])*((rpsftmsurvlb+ rpsftmsurvlb[_n-1])/2)
egen rpsftmsurvlbauc1=total(rpsftmsurvlbauc)
drop rpsftmsurvlbauc
gen rpsftmsurvubauc=0
replace rpsftmsurvubauc =(timerpsftmsurvub-timerpsftmsurvub[_n-1])*((rpsftmsurvub+ rpsftmsurvub[_n-1])/2)
egen rpsftmsurvubauc1=total(rpsftmsurvubauc)
drop rpsftmsurvubauc
gen rpsftmwcsurvauc=0
replace rpsftmwcsurvauc =(timerpsftmwcsurv-timerpsftmwcsurv[_n-1])*((rpsftmwcsurv+ rpsftmwcsurv[_n-1])/2)
egen rpsftmwcsurvauc1=total(rpsftmwcsurvauc)
drop rpsftmwcsurvauc
gen rpsftmwcsurvlbauc=0
replace rpsftmwcsurvlbauc =(timerpsftmwcsurvlb-timerpsftmwcsurvlb[_n-1])*((rpsftmwcsurvlb+
rpsftmwcsurvlb[_n-1])/2)
egen rpsftmwcsurvlbauc1=total(rpsftmwcsurvlbauc)
drop rpsftmwcsurvlbauc
gen rpsftmwcsurvubauc=0
replace rpsftmwcsurvubauc =(timerpsftmwcsurvub-timerpsftmwcsurvub[_n-1])*((rpsftmwcsurvub+
rpsftmwcsurvub[_n-1])/2)
egen rpsftmwcsurvubauc1=total(rpsftmwcsurvubauc)
drop rpsftmwcsurvubauc

```

```

summ conrpsftmshrinkjointauc1 conrpsftmshrinkjointlbauc1 conrpsftmshrinkjointubauc1
exprpsftmshrinkjointauc1 exprpsftmshrinkjointlbauc1 exprpsftmshrinkjointubauc1 conrpsftmshrinkindauc1
conrpsftmshrinkindlbauc1 conrpsftmshrinkindubauc1 conrpsftmwshrinkjointauc1 conrpsftmwshrinkjointlbauc1
conrpsftmwshrinkjointubauc1 exprpsftmwshrinkjointauc1 exprpsftmwshrinkjointlbauc1
exprpsftmwshrinkjointubauc1 conrpsftmwshrinkindauc1 conrpsftmwshrinkindlbauc1
conrpsftmwshrinkindubauc1 rpsftmsurvauc1 rpsftmsurvlbauc1 rpsftmsurvubauc1 rpsftmwcsurvauc1
rpsftmwcsurvlbauc1 rpsftmwcsurvubauc1 rpsftmextauc1 rpsftmwextauc1

```

*****RPSFTM on treatment*****

```
gen rpsftmotextauc=0
```

```

replace rpsftmotextauc =(time-time[_n-1])*((rpsftmotext+ rpsftmotext[_n-1])/2)
egen rpsftmotextauc1=total(rpsftmotextauc)
drop rpsftmotextauc
gen conrpsftmotshrinkjointauc=0
replace conrpsftmotshrinkjointauc =(time-time[_n-1])*((conrpsftmotshrinkjoint+ conrpsftmotshrinkjoint[_n-1])/2)
egen conrpsftmotshrinkjointauc1=total(conrpsftmotshrinkjointauc)
drop conrpsftmotshrinkjointauc
gen exprpsftmotshrinkjointauc=0
replace exprpsftmotshrinkjointauc =(time-time[_n-1])*((exprpsftmotshrinkjoint+ exprpsftmotshrinkjoint[_n-1])/2)
egen exprpsftmotshrinkjointauc1=total(exprpsftmotshrinkjointauc)
drop exprpsftmotshrinkjointauc
gen conrpsftmotshrinkindauc=0
replace conrpsftmotshrinkindauc =(time-time[_n-1])*((conrpsftmotshrinkind+ conrpsftmotshrinkind[_n-1])/2)
egen conrpsftmotshrinkindauc1=total(conrpsftmotshrinkindauc)
drop conrpsftmotshrinkindauc
gen conrpsftmotshrinkjointlbauc=0
replace conrpsftmotshrinkjointlbauc =(time-time[_n-1])*((conrpsftmotshrinkjointlb+
conrpsftmotshrinkjointlb[_n-1])/2)
egen conrpsftmotshrinkjointlbauc1=total(conrpsftmotshrinkjointlbauc)
drop conrpsftmotshrinkjointlbauc
gen exprpsftmotshrinkjointlbauc=0
replace exprpsftmotshrinkjointlbauc =(time-time[_n-1])*((exprpsftmotshrinkjointlb+
exprpsftmotshrinkjointlb[_n-1])/2)
egen exprpsftmotshrinkjointlbauc1=total(exprpsftmotshrinkjointlbauc)
drop exprpsftmotshrinkjointlbauc
gen conrpsftmotshrinkindlbauc=0
replace conrpsftmotshrinkindlbauc =(time-time[_n-1])*((conrpsftmotshrinkindlb+ conrpsftmotshrinkindlb[_n-1])/2)
egen conrpsftmotshrinkindlbauc1=total(conrpsftmotshrinkindlbauc)
drop conrpsftmotshrinkindlbauc
gen conrpsftmotshrinkjointubauc=0
replace conrpsftmotshrinkjointubauc =(time-time[_n-1])*((conrpsftmotshrinkjointub+
conrpsftmotshrinkjointub[_n-1])/2)
egen conrpsftmotshrinkjointubauc1=total(conrpsftmotshrinkjointubauc)
drop conrpsftmotshrinkjointubauc
gen exprpsftmotshrinkjointubauc=0
replace exprpsftmotshrinkjointubauc =(time-time[_n-1])*((exprpsftmotshrinkjointub+
exprpsftmotshrinkjointub[_n-1])/2)
egen exprpsftmotshrinkjointubauc1=total(exprpsftmotshrinkjointubauc)
drop exprpsftmotshrinkjointubauc
gen conrpsftmotshrinkindubauc=0
replace conrpsftmotshrinkindubauc =(time-time[_n-1])*((conrpsftmotshrinkindub+ conrpsftmotshrinkindub[_n-1])/2)
egen conrpsftmotshrinkindubauc1=total(conrpsftmotshrinkindubauc)
drop conrpsftmotshrinkindubauc
gen rpsftmotwextauc=0
replace rpsftmotwextauc =(time-time[_n-1])*((rpsftmotwext+ rpsftmotwext[_n-1])/2)
egen rpsftmotwextauc1=total(rpsftmotwextauc)
drop rpsftmotwextauc
gen conrpsftmotwshrinkjointauc=0
replace conrpsftmotwshrinkjointauc =(time-time[_n-1])*((conrpsftmotwshrinkjoint+
conrpsftmotwshrinkjoint[_n-1])/2)
egen conrpsftmotwshrinkjointauc1=total(conrpsftmotwshrinkjointauc)

```

```

drop conrpsftmotweshrinkjointauc
gen exprpsftmotweshrinkjointauc=0
replace exprpsftmotweshrinkjointauc =(time-time[_n-1])*((exprpsftmotweshrinkjoint+
exprpsftmotweshrinkjoint[_n-1])/2)
egen exprpsftmotweshrinkjointauc1=total(exprpsftmotweshrinkjointauc)
drop exprpsftmotweshrinkjointauc
gen conrpsftmotweshrinkindauc=0
replace conrpsftmotweshrinkindauc =(time-time[_n-1])*((conrpsftmotweshrinkind+ conrpsftmotweshrinkind[_n-
1])/2)
egen conrpsftmotweshrinkindauc1=total(conrpsftmotweshrinkindauc)
drop conrpsftmotweshrinkindauc
gen conrpsftmotweshrinkjointlbauc=0
replace conrpsftmotweshrinkjointlbauc =(time-time[_n-1])*((conrpsftmotweshrinkjointlb+
conrpsftmotweshrinkjointlb[_n-1])/2)
egen conrpsftmotweshrinkjointlbauc1=total(conrpsftmotweshrinkjointlbauc)
drop conrpsftmotweshrinkjointlbauc
gen exprpsftmotweshrinkjointlbauc=0
replace exprpsftmotweshrinkjointlbauc =(time-time[_n-1])*((exprpsftmotweshrinkjointlb+
exprpsftmotweshrinkjointlb[_n-1])/2)
egen exprpsftmotweshrinkjointlbauc1=total(exprpsftmotweshrinkjointlbauc)
drop exprpsftmotweshrinkjointlbauc
gen conrpsftmotweshrinkindlbauc=0
replace conrpsftmotweshrinkindlbauc =(time-time[_n-1])*((conrpsftmotweshrinkindlb+
conrpsftmotweshrinkindlb[_n-1])/2)
egen conrpsftmotweshrinkindlbauc1=total(conrpsftmotweshrinkindlbauc)
drop conrpsftmotweshrinkindlbauc
gen conrpsftmotweshrinkjointubauc=0
replace conrpsftmotweshrinkjointubauc =(time-time[_n-1])*((conrpsftmotweshrinkjointub+
conrpsftmotweshrinkjointub[_n-1])/2)
egen conrpsftmotweshrinkjointubauc1=total(conrpsftmotweshrinkjointubauc)
drop conrpsftmotweshrinkjointubauc
gen exprpsftmotweshrinkjointubauc=0
replace exprpsftmotweshrinkjointubauc =(time-time[_n-1])*((exprpsftmotweshrinkjointub+
exprpsftmotweshrinkjointub[_n-1])/2)
egen exprpsftmotweshrinkjointubauc1=total(exprpsftmotweshrinkjointubauc)
drop exprpsftmotweshrinkjointubauc
gen conrpsftmotweshrinkindubauc=0
replace conrpsftmotweshrinkindubauc =(time-time[_n-1])*((conrpsftmotweshrinkindub+
conrpsftmotweshrinkindub[_n-1])/2)
egen conrpsftmotweshrinkindubauc1=total(conrpsftmotweshrinkindubauc)
drop conrpsftmotweshrinkindubauc
gen rpsftmotsurvauc=0
replace rpsftmotsurvauc =(timerpsftmotsurv-timerpsftmotsurv[_n-1])*((rpsftmotsurv+ rpsftmotsurv[_n-1])/2)
egen rpsftmotsurvauc1=total(rpsftmotsurvauc)
drop rpsftmotsurvauc
gen rpsftmotsurvlbauc=0
replace rpsftmotsurvlbauc =(timerpsftmotsurvlb-timerpsftmotsurvlb[_n-1])*((rpsftmotsurvlb+ rpsftmotsurvlb[_n-
1])/2)
egen rpsftmotsurvlbauc1=total(rpsftmotsurvlbauc)
drop rpsftmotsurvlbauc
gen rpsftmotsurvubauc=0
replace rpsftmotsurvubauc =(timerpsftmotsurvub-timerpsftmotsurvub[_n-1])*((rpsftmotsurvub+
rpsftmotsurvub[_n-1])/2)
egen rpsftmotsurvubauc1=total(rpsftmotsurvubauc)
drop rpsftmotsurvubauc

```

```

gen rpsftmotwcsurvauc=0
replace rpsftmotwcsurvauc =(timerpsftmotwcsurv-timerpsftmotwcsurv[_n-1])*((rpsftmotwcsurv+
rpsftmotwcsurv[_n-1])/2)
egen rpsftmotwcsurvauc1=total(rpsftmotwcsurvauc)
drop rpsftmotwcsurvauc
gen rpsftmotwcsurvlbauc=0
replace rpsftmotwcsurvlbauc =(timerpsftmotwcsurvlb-timerpsftmotwcsurvlb[_n-1])*((rpsftmotwcsurvlb+
rpsftmotwcsurvlb[_n-1])/2)
egen rpsftmotwcsurvlbauc1=total(rpsftmotwcsurvlbauc)
drop rpsftmotwcsurvlbauc
gen rpsftmotwcsurvubauc=0
replace rpsftmotwcsurvubauc =(timerpsftmotwcsurvub-timerpsftmotwcsurvub[_n-1])*((rpsftmotwcsurvub+
rpsftmotwcsurvub[_n-1])/2)
egen rpsftmotwcsurvubauc1=total(rpsftmotwcsurvubauc)
drop rpsftmotwcsurvubauc
sum conrpsftmotsrinkjointauc1 conrpsftmotsrinkjointlbauc1 conrpsftmotsrinkjointubauc1
exprpsftmotsrinkjointauc1 exprpsftmotsrinkjointlbauc1 exprpsftmotsrinkjointubauc1
conrpsftmotsrinkindauc1 conrpsftmotsrinkindlbauc1 conrpsftmotsrinkindubauc1
conrpsftmotweshrinkjointauc1 conrpsftmotweshrinkjointlbauc1 conrpsftmotweshrinkjointubauc1
exprpsftmotweshrinkjointauc1 exprpsftmotweshrinkjointlbauc1 exprpsftmotweshrinkjointubauc1
conrpsftmotweshrinkindauc1 conrpsftmotweshrinkindlbauc1 conrpsftmotweshrinkindubauc1 rpsftmotsurvauc1
rpsftmotsurvlbauc1 rpsftmotsurvubauc1 rpsftmotwcsurvauc1 rpsftmotwcsurvlbauc1 rpsftmotwcsurvubauc1
rpsftmotextauc1 rpsftmotwextauc1

```

*****IPE*****

```

gen conipeshrinkjointauc=0
replace conipeshrinkjointauc =(time-time[_n-1])*((conipeshrinkjoint+ conipeshrinkjoint[_n-1])/2)
egen conipeshrinkjointauc1=total(conipeshrinkjointauc)
drop conipeshrinkjointauc
gen expipeshrinkjointauc=0
replace expipeshrinkjointauc =(time-time[_n-1])*((expipeshrinkjoint+ expipeshrinkjoint[_n-1])/2)
egen expipeshrinkjointauc1=total(expipeshrinkjointauc)
drop expipeshrinkjointauc
gen conipeshrinkindauc=0
replace conipeshrinkindauc =(time-time[_n-1])*((conipeshrinkind+ conipeshrinkind[_n-1])/2)
egen conipeshrinkindauc1=total(conipeshrinkindauc)
drop conipeshrinkindauc
gen conipeshrinkjointlbauc=0
replace conipeshrinkjointlbauc =(time-time[_n-1])*((conipeshrinkjointlb+ conipeshrinkjointlb[_n-1])/2)
egen conipeshrinkjointlbauc1=total(conipeshrinkjointlbauc)
drop conipeshrinkjointlbauc
gen expipeshrinkjointlbauc=0
replace expipeshrinkjointlbauc =(time-time[_n-1])*((expipeshrinkjointlb+ expipeshrinkjointlb[_n-1])/2)
egen expipeshrinkjointlbauc1=total(expipeshrinkjointlbauc)
drop expipeshrinkjointlbauc
gen conipeshrinkindlbauc=0
replace conipeshrinkindlbauc =(time-time[_n-1])*((conipeshrinkindlb+ conipeshrinkindlb[_n-1])/2)
egen conipeshrinkindlbauc1=total(conipeshrinkindlbauc)
drop conipeshrinkindlbauc
gen conipeshrinkjointubauc=0
replace conipeshrinkjointubauc =(time-time[_n-1])*((conipeshrinkjointub+ conipeshrinkjointub[_n-1])/2)
egen conipeshrinkjointubauc1=total(conipeshrinkjointubauc)
drop conipeshrinkjointubauc
gen expipeshrinkjointubauc=0
replace expipeshrinkjointubauc =(time-time[_n-1])*((expipeshrinkjointub+ expipeshrinkjointub[_n-1])/2)
egen expipeshrinkjointubauc1=total(expipeshrinkjointubauc)
drop expipeshrinkjointubauc

```

```

egen expipeshrinkjointubauc1=total(expipeshrinkjointubauc)
drop expipeshrinkjointubauc
gen conipeshrinkindubauc=0
replace conipeshrinkindubauc =(time-time[_n-1])*((conipeshrinkindub+ conipeshrinkindub[_n-1])/2)
egen conipeshrinkindubauc1=total(conipeshrinkindubauc)
drop conipeshrinkindubauc
gen conipewcshrinkjointauc=0
replace conipewcshrinkjointauc =(time-time[_n-1])*((conipewcshrinkjoint+ conipewcshrinkjoint[_n-1])/2)
egen conipewcshrinkjointauc1=total(conipewcshrinkjointauc)
drop conipewcshrinkjointauc
gen expipewcshrinkjointauc=0
replace expipewcshrinkjointauc =(time-time[_n-1])*((expipewcshrinkjoint+ expipewcshrinkjoint[_n-1])/2)
egen expipewcshrinkjointauc1=total(expipewcshrinkjointauc)
drop expipewcshrinkjointauc
gen conipewcshrinkindauc=0
replace conipewcshrinkindauc =(time-time[_n-1])*((conipewcshrinkind+ conipewcshrinkind[_n-1])/2)
egen conipewcshrinkindauc1=total(conipewcshrinkindauc)
drop conipewcshrinkindauc
gen conipewcshrinkjointlbauc=0
replace conipewcshrinkjointlbauc =(time-time[_n-1])*((conipewcshrinkjointlb+ conipewcshrinkjointlb[_n-1])/2)
egen conipewcshrinkjointlbauc1=total(conipewcshrinkjointlbauc)
drop conipewcshrinkjointlbauc
gen expipewcshrinkjointlbauc=0
replace expipewcshrinkjointlbauc =(time-time[_n-1])*((expipewcshrinkjointlb+ expipewcshrinkjointlb[_n-1])/2)
egen expipewcshrinkjointlbauc1=total(expipewcshrinkjointlbauc)
drop expipewcshrinkjointlbauc
gen conipewcshrinkindlbauc=0
replace conipewcshrinkindlbauc =(time-time[_n-1])*((conipewcshrinkindlb+ conipewcshrinkindlb[_n-1])/2)
egen conipewcshrinkindlbauc1=total(conipewcshrinkindlbauc)
drop conipewcshrinkindlbauc
gen conipewcshrinkjointubauc=0
replace conipewcshrinkjointubauc =(time-time[_n-1])*((conipewcshrinkjointub+ conipewcshrinkjointub[_n-1])/2)
egen conipewcshrinkjointubauc1=total(conipewcshrinkjointubauc)
drop conipewcshrinkjointubauc
gen expipewcshrinkjointubauc=0
replace expipewcshrinkjointubauc =(time-time[_n-1])*((expipewcshrinkjointub+ expipewcshrinkjointub[_n-1])/2)
egen expipewcshrinkjointubauc1=total(expipewcshrinkjointubauc)
drop expipewcshrinkjointubauc
gen conipewcshrinkindubauc=0
replace conipewcshrinkindubauc =(time-time[_n-1])*((conipewcshrinkindub+ conipewcshrinkindub[_n-1])/2)
egen conipewcshrinkindubauc1=total(conipewcshrinkindubauc)
drop conipewcshrinkindubauc
gen ipesurvauc=0
replace ipesurvauc =(timeipesurv-timeipesurv[_n-1])*((ipesurv+ ipesurv[_n-1])/2)
egen ipesurvauc1=total(ipesurvauc)
drop ipesurvauc
gen ipesurvlbauc=0
replace ipesurvlbauc =(timeipesurvlb-timeipesurvlb[_n-1])*((ipesurvlb+ ipesurvlb[_n-1])/2)
egen ipesurvlbauc1=total(ipesurvlbauc)
drop ipesurvlbauc
gen ipesurvubauc=0
replace ipesurvubauc =(timeipesurvub-timeipesurvub[_n-1])*((ipesurvub+ ipesurvub[_n-1])/2)
egen ipesurvubauc1=total(ipesurvubauc)

```

```

drop ipesurvubauc
gen ipewcsurvauc=0
replace ipewcsurvauc =(timeipewcsurv-timeipewcsurv[_n-1])*((ipewcsurv+ ipewcsurv[_n-1])/2)
egen ipewcsurvauc1=total(ipewcsurvauc)
drop ipewcsurvauc
gen ipewcsurvlbauc=0
replace ipewcsurvlbauc =(timeipewcsurvlb-timeipewcsurvlb[_n-1])*((ipewcsurvlb+ ipewcsurvlb[_n-1])/2)
egen ipewcsurvlbauc1=total(ipewcsurvlbauc)
drop ipewcsurvlbauc
gen ipewcsurvubauc=0
replace ipewcsurvubauc =(timeipewcsurvub-timeipewcsurvub[_n-1])*((ipewcsurvub+ ipewcsurvub[_n-1])/2)
egen ipewcsurvubauc1=total(ipewcsurvubauc)
drop ipewcsurvubauc

```

```

summ conipeshrinkjointauc1 conipeshrinkjointlbauc1 conipeshrinkjointubauc1 expipeshrinkjointauc1
expipeshrinkjointlbauc1 expipeshrinkjointubauc1 conipeshrinkindauc1 conipeshrinkindlbauc1
conipeshrinkindubauc1 conipewcshrinkjointauc1 conipewcshrinkjointlbauc1 conipewcshrinkjointubauc1
expipewcshrinkjointauc1 expipewcshrinkjointlbauc1 expipewcshrinkjointubauc1 conipewcshrinkindauc1
conipewcshrinkindlbauc1 conipewcshrinkindubauc1 ipesurvauc1 ipesurvlbauc1 ipesurvubauc1 ipewcsurvauc1
ipewcsurvlbauc1 ipewcsurvubauc1

```

```

****TPE on treatment****
gen conipeotshrinkjointauc=0
replace conipeotshrinkjointauc =(time-time[_n-1])*((conipeotshrinkjoint+ conipeotshrinkjoint[_n-1])/2)
egen conipeotshrinkjointauc1=total(conipeotshrinkjointauc)
drop conipeotshrinkjointauc
gen expipeotshrinkjointauc=0
replace expipeotshrinkjointauc =(time-time[_n-1])*((expipeotshrinkjoint+ expipeotshrinkjoint[_n-1])/2)
egen expipeotshrinkjointauc1=total(expipeotshrinkjointauc)
drop expipeotshrinkjointauc
gen conipeotshrinkindauc=0
replace conipeotshrinkindauc =(time-time[_n-1])*((conipeotshrinkind+ conipeotshrinkind[_n-1])/2)
egen conipeotshrinkindauc1=total(conipeotshrinkindauc)
drop conipeotshrinkindauc
gen conipeotshrinkjointlbauc=0
replace conipeotshrinkjointlbauc =(time-time[_n-1])*((conipeotshrinkjointlb+ conipeotshrinkjointlb[_n-1])/2)
egen conipeotshrinkjointlbauc1=total(conipeotshrinkjointlbauc)
drop conipeotshrinkjointlbauc
gen expipeotshrinkjointlbauc=0
replace expipeotshrinkjointlbauc =(time-time[_n-1])*((expipeotshrinkjointlb+ expipeotshrinkjointlb[_n-1])/2)
egen expipeotshrinkjointlbauc1=total(expipeotshrinkjointlbauc)
drop expipeotshrinkjointlbauc
gen conipeotshrinkindlbauc=0
replace conipeotshrinkindlbauc =(time-time[_n-1])*((conipeotshrinkindlb+ conipeotshrinkindlb[_n-1])/2)
egen conipeotshrinkindlbauc1=total(conipeotshrinkindlbauc)
drop conipeotshrinkindlbauc
gen conipeotshrinkjointubauc=0
replace conipeotshrinkjointubauc =(time-time[_n-1])*((conipeotshrinkjointub+ conipeotshrinkjointub[_n-1])/2)
egen conipeotshrinkjointubauc1=total(conipeotshrinkjointubauc)
drop conipeotshrinkjointubauc
gen expipeotshrinkjointubauc=0
replace expipeotshrinkjointubauc =(time-time[_n-1])*((expipeotshrinkjointub+ expipeotshrinkjointub[_n-1])/2)
egen expipeotshrinkjointubauc1=total(expipeotshrinkjointubauc)
drop expipeotshrinkjointubauc
gen conipeotshrinkindubauc=0

```



```

replace conipeotshrinkindubauc =(time-time[_n-1])*((conipeotshrinkindub+ conipeotshrinkindub[_n-1])/2)
egen conipeotshrinkindubauc1=total(conipeotshrinkindubauc)
drop conipeotshrinkindubauc
gen conipeotwshrinkjointauc=0
replace conipeotwshrinkjointauc =(time-time[_n-1])*((conipeotwshrinkjoint+ conipeotwshrinkjoint[_n-1])/2)
egen conipeotwshrinkjointauc1=total(conipeotwshrinkjointauc)
drop conipeotwshrinkjointauc
gen expipeotwshrinkjointauc=0
replace expipeotwshrinkjointauc =(time-time[_n-1])*((expipeotwshrinkjoint+ expipeotwshrinkjoint[_n-1])/2)
egen expipeotwshrinkjointauc1=total(expipeotwshrinkjointauc)
drop expipeotwshrinkjointauc

```

```

gen conipeotwshrinkindauc=0
replace conipeotwshrinkindauc =(time-time[_n-1])*((conipeotwshrinkind+ conipeotwshrinkind[_n-1])/2)
egen conipeotwshrinkindauc1=total(conipeotwshrinkindauc)
drop conipeotwshrinkindauc
gen conipeotwshrinkjointlbauc=0
replace conipeotwshrinkjointlbauc =(time-time[_n-1])*((conipeotwshrinkjointlb+ conipeotwshrinkjointlb[_n-1])/2)
egen conipeotwshrinkjointlbauc1=total(conipeotwshrinkjointlbauc)
drop conipeotwshrinkjointlbauc
gen expipeotwshrinkjointlbauc=0
replace expipeotwshrinkjointlbauc =(time-time[_n-1])*((expipeotwshrinkjointlb+ expipeotwshrinkjointlb[_n-1])/2)
egen expipeotwshrinkjointlbauc1=total(expipeotwshrinkjointlbauc)
drop expipeotwshrinkjointlbauc
gen conipeotwshrinkindlbauc=0
replace conipeotwshrinkindlbauc =(time-time[_n-1])*((conipeotwshrinkindlb+ conipeotwshrinkindlb[_n-1])/2)
egen conipeotwshrinkindlbauc1=total(conipeotwshrinkindlbauc)
drop conipeotwshrinkindlbauc
gen conipeotwshrinkjointubauc=0
replace conipeotwshrinkjointubauc =(time-time[_n-1])*((conipeotwshrinkjointub+ conipeotwshrinkjointub[_n-1])/2)
egen conipeotwshrinkjointubauc1=total(conipeotwshrinkjointubauc)
drop conipeotwshrinkjointubauc
gen expipeotwshrinkjointubauc=0
replace expipeotwshrinkjointubauc =(time-time[_n-1])*((expipeotwshrinkjointub+ expipeotwshrinkjointub[_n-1])/2)
egen expipeotwshrinkjointubauc1=total(expipeotwshrinkjointubauc)
drop expipeotwshrinkjointubauc
gen conipeotwshrinkindubauc=0
replace conipeotwshrinkindubauc =(time-time[_n-1])*((conipeotwshrinkindub+ conipeotwshrinkindub[_n-1])/2)
egen conipeotwshrinkindubauc1=total(conipeotwshrinkindubauc)
drop conipeotwshrinkindubauc
gen ipeotsurvauc=0
replace ipeotsurvauc =(timeipeotsurv-timeipeotsurv[_n-1])*((ipeotsurv+ ipeotsurv[_n-1])/2)
egen ipeotsurvauc1=total(ipeotsurvauc)
drop ipeotsurvauc
gen ipeotsurvlbauc=0
replace ipeotsurvlbauc =(timeipeotsurvlb-timeipeotsurvlb[_n-1])*((ipeotsurvlb+ ipeotsurvlb[_n-1])/2)
egen ipeotsurvlbauc1=total(ipeotsurvlbauc)
drop ipeotsurvlbauc
gen ipeotsurvbauc=0

```

```

replace ipeotsurvbauc =(timeipeotsurvub-timeipeotsurvub[_n-1])*((ipeotsurvub+ ipeotsurvub[_n-1])/2)
egen ipeotsurvbauc1=total(ipeotsurvbauc)
drop ipeotsurvbauc
gen ipeotwcurvauc=0
replace ipeotwcurvauc =(timeipeotwcurv-timeipeotwcurv[_n-1])*((ipeotwcurv+ ipeotwcurv[_n-1])/2)
egen ipeotwcurvauc1=total(ipeotwcurvauc)
drop ipeotwcurvauc
gen ipeotwcurvlbauc=0
replace ipeotwcurvlbauc =(timeipeotwcurvlb-timeipeotwcurvlb[_n-1])*((ipeotwcurvlb+ ipeotwcurvlb[_n-1])/2)
egen ipeotwcurvlbauc1=total(ipeotwcurvlbauc)
drop ipeotwcurvlbauc
gen ipeotwcurvubauc=0
replace ipeotwcurvubauc =(timeipeotwcurvub-timeipeotwcurvub[_n-1])*((ipeotwcurvub+ ipeotwcurvub[_n-1])/2)
egen ipeotwcurvubauc1=total(ipeotwcurvubauc)
drop ipeotwcurvubauc

```

```

summ conipeotshrinkjointauc1 conipeotshrinkjointlbauc1 conipeotshrinkjointubauc1 expipeotshrinkjointauc1
expipeotshrinkjointlbauc1 expipeotshrinkjointubauc1 conipeotshrinkindauc1 conipeotshrinkindlbauc1
conipeotshrinkindubauc1 conipeotshrinkjointauc1 conipeotshrinkjointlbauc1 conipeotshrinkjointubauc1
expipeotwshrinkjointauc1 expipeotwshrinkjointlbauc1 expipeotwshrinkjointubauc1 conipeotwshrinkindauc1
conipeotwshrinkindlbauc1 conipeotwshrinkindubauc1 ipeotsurvauc1 ipeotsurvlbauc1 ipeotsurvbauc1
ipeotwcurvauc1 ipeotwcurvlbauc1 ipeotwcurvubauc1

```

```

****IPCW****
gen ipcwCoxauc=0
replace ipcwCoxauc =(time-time[_n-1])*((ipcwCox+ipcwCox[_n-1])/2)
egen ipcwCoxauc1=total(ipcwCoxauc)
drop ipcwCoxauc
gen ipcwCoxauc1b=0
replace ipcwCoxauc1b =(time-time[_n-1])*((ipcwCox1b+ipcwCox1b[_n-1])/2)
egen ipcwCoxauc1b1=total(ipcwCoxauc1b)
drop ipcwCoxauc1b
gen ipcwCoxaucub=0
replace ipcwCoxaucub =(time-time[_n-1])*((ipcwCoxub+ipcwCoxub[_n-1])/2)
egen ipcwCoxaucub1=total(ipcwCoxaucub)
drop ipcwCoxaucub

```

f) Code for estimating IPCW weighted Kaplan-Meier

```

****WKM****
preserve
drop if trtrand==1
keep id died timeOS3 trtrand stabweightxo
***Lambda numerator for each time point for IPCW KM.***
gen IPCWKM1= stabweightxo if (died==1 & timeOS3<=1 & trtrand==0)
gen IPCWKMD1= stabweightxo if ((died==0|died==. |died==1) & timeOS3<=1 & trtrand==0)
forvalues min=1(1)689 {
  local max=`min'+1
  gen IPCWKM`max`= stabweightxo if (died==1 & (timeOS3>`min' & timeOS3<=`max') & trtrand==0)
  egen IPCWKMN`min`=total(IPCWKM`min')
  gen IPCWKMD`max`= stabweightxo if ((died==0|died==. |died==1) & (timeOS3>`min' & timeOS3<=`max') &
  trtrand==0)
  egen IPCWKMDD`min`=total(IPCWKMD`min')

```

```

drop IPCWKM`min` IPCWKMD`min`
}

gen IPCWKM704 = stabweightxo if (died==1 & timeOS3==704 & trtrand==0)
egen IPCWKMN704=total(IPCWKM704)
gen IPCWKMD704= stabweightxo if ((died==0|died==.|died==1) & timeOS3==704 & trtrand==0)
egen IPCWKMD704=total(IPCWKMD704)
drop IPCWKM704 IPCWKMD704
gen IPCWKM753 = stabweightxo if (died==1 & timeOS3==753 & trtrand==0)
egen IPCWKMN753=total(IPCWKM753)
gen IPCWKMD753= stabweightxo if ((died==0|died==.|died==1) & timeOS3==753 & trtrand==0)
egen IPCWKMD753=total(IPCWKMD753)
drop IPCWKM753 IPCWKMD753
gen IPCWKM754 = stabweightxo if (died==1 & timeOS3==754 & trtrand==0)
egen IPCWKMN754=total(IPCWKM754)
gen IPCWKMD754= stabweightxo if ((died==0|died==.|died==1) & timeOS3==754 & trtrand==0)
egen IPCWKMD754=total(IPCWKMD754)
drop IPCWKM754 IPCWKMD754
gen IPCWKM798 = stabweightxo if (died==1 & timeOS3==798 & trtrand==0)
egen IPCWKMN798=total(IPCWKM798)
gen IPCWKMD798= stabweightxo if ((died==0|died==.|died==1) & timeOS3==798 & trtrand==0)
egen IPCWKMD798=total(IPCWKMD798)
drop IPCWKM798 IPCWKMD798
gen IPCWKM800 = stabweightxo if (died==1 & timeOS3==800 & trtrand==0)
egen IPCWKMN800=total(IPCWKM800)
gen IPCWKMD800= stabweightxo if ((died==0|died==.|died==1) & timeOS3==800 & trtrand==0)
egen IPCWKMD800=total(IPCWKMD800)
drop IPCWKM800 IPCWKMD800
gen IPCWKM815 = stabweightxo if (died==1 & timeOS3==815 & trtrand==0)
egen IPCWKMN815=total(IPCWKM815)
gen IPCWKMD815= stabweightxo if ((died==0|died==.|died==1) & timeOS3==815 & trtrand==0)
egen IPCWKMD815=total(IPCWKMD815)
drop IPCWKM815 IPCWKMD815
gen IPCWKM829 = stabweightxo if (died==1 & timeOS3==829 & trtrand==0)
egen IPCWKMN829=total(IPCWKM829)
gen IPCWKMD829= stabweightxo if ((died==0|died==.|died==1) & timeOS3==829 & trtrand==0)
egen IPCWKMD829=total(IPCWKMD829)
drop IPCWKM829 IPCWKMD829
gen IPCWKM846 = stabweightxo if (died==1 & timeOS3==846 & trtrand==0)
egen IPCWKMN846=total(IPCWKM846)
gen IPCWKMD846= stabweightxo if ((died==0|died==.|died==1) & timeOS3==846 & trtrand==0)
egen IPCWKMD846=total(IPCWKMD846)
drop IPCWKM846 IPCWKMD846
gen IPCWKM847 = stabweightxo if (died==1 & timeOS3==847 & trtrand==0)
egen IPCWKMN847=total(IPCWKM847)
gen IPCWKMD847= stabweightxo if ((died==0|died==.|died==1) & timeOS3==847 & trtrand==0)
egen IPCWKMD847=total(IPCWKMD847)
drop IPCWKM847 IPCWKMD847
gen IPCWKM848 = stabweightxo if (died==1 & timeOS3==848 & trtrand==0)
egen IPCWKMN848=total(IPCWKM848)
gen IPCWKMD848= stabweightxo if ((died==0|died==.|died==1) & timeOS3==848 & trtrand==0)
egen IPCWKMD848=total(IPCWKMD848)
drop IPCWKM848 IPCWKMD848
gen IPCWKM857 = stabweightxo if (died==1 & timeOS3==857 & trtrand==0)
egen IPCWKMN857=total(IPCWKM857)

```

```

gen IPCWKMD857= stabweightxo if ((died==0|died==.|died==1) & timeOS3==857 & trtrand==0)
egen IPCWKMD857=total(IPCWKMD857)
drop IPCWKM857 IPCWKMD857
gen IPCWKM861 = stabweightxo if (died==1 & timeOS3==861 & trtrand==0)
egen IPCWKMN861=total(IPCWKM861)
gen IPCWKMD861= stabweightxo if ((died==0|died==.|died==1) & timeOS3==861 & trtrand==0)
egen IPCWKMD861=total(IPCWKMD861)
drop IPCWKM861 IPCWKMD861
gen IPCWKM862 = stabweightxo if (died==1 & timeOS3==862 & trtrand==0)
egen IPCWKMN862=total(IPCWKM862)
gen IPCWKMD862= stabweightxo if ((died==0|died==.|died==1) & timeOS3==862 & trtrand==0)
egen IPCWKMD862=total(IPCWKMD862)
drop IPCWKM862 IPCWKMD862
gen IPCWKM872 = stabweightxo if (died==1 & timeOS3==872 & trtrand==0)
egen IPCWKMN872=total(IPCWKM872)
gen IPCWKMD872= stabweightxo if ((died==0|died==.|died==1) & timeOS3==872 & trtrand==0)
egen IPCWKMD872=total(IPCWKMD872)
drop IPCWKM872 IPCWKMD872
gen IPCWKM879 = stabweightxo if (died==1 & timeOS3==879 & trtrand==0)
egen IPCWKMN879=total(IPCWKM879)
gen IPCWKMD879= stabweightxo if ((died==0|died==.|died==1) & timeOS3==879 & trtrand==0)
egen IPCWKMD879=total(IPCWKMD879)
drop IPCWKM879 IPCWKMD879
gen IPCWKM885 = stabweightxo if (died==1 & timeOS3==885 & trtrand==0)
egen IPCWKMN885=total(IPCWKM885)
gen IPCWKMD885= stabweightxo if ((died==0|died==.|died==1) & timeOS3==885 & trtrand==0)
egen IPCWKMD885=total(IPCWKMD885)
drop IPCWKM885 IPCWKMD885
gen IPCWKM896 = stabweightxo if (died==1 & timeOS3==896 & trtrand==0)
egen IPCWKMN896=total(IPCWKM896)
gen IPCWKMD896= stabweightxo if ((died==0|died==.|died==1) & timeOS3==896 & trtrand==0)
egen IPCWKMD896=total(IPCWKMD896)
drop IPCWKM896 IPCWKMD896
gen IPCWKM905 = stabweightxo if (died==1 & timeOS3==905 & trtrand==0)
egen IPCWKMN905=total(IPCWKM905)
gen IPCWKMD905= stabweightxo if ((died==0|died==.|died==1) & timeOS3==905 & trtrand==0)
egen IPCWKMD905=total(IPCWKMD905)
drop IPCWKM905 IPCWKMD905
gen IPCWKM907 = stabweightxo if (died==1 & timeOS3==907 & trtrand==0)
egen IPCWKMN907=total(IPCWKM907)
gen IPCWKMD907= stabweightxo if ((died==0|died==.|died==1) & timeOS3==907 & trtrand==0)
egen IPCWKMD907=total(IPCWKMD907)
drop IPCWKM907 IPCWKMD907
gen IPCWKM909 = stabweightxo if (died==1 & timeOS3==909 & trtrand==0)
egen IPCWKMN909=total(IPCWKM909)
gen IPCWKMD909= stabweightxo if ((died==0|died==.|died==1) & timeOS3==909 & trtrand==0)
egen IPCWKMD909=total(IPCWKMD909)
drop IPCWKM909 IPCWKMD909
gen IPCWKM931 = stabweightxo if (died==1 & timeOS3==931 & trtrand==0)
egen IPCWKMN931=total(IPCWKM931)
gen IPCWKMD931= stabweightxo if ((died==0|died==.|died==1) & timeOS3==931 & trtrand==0)
egen IPCWKMD931=total(IPCWKMD931)
drop IPCWKM931 IPCWKMD931
gen IPCWKM953 = stabweightxo if (died==1 & timeOS3==953 & trtrand==0)
egen IPCWKMN953=total(IPCWKM953)

```

```

gen IPCWKMD953= stabweightxo if ((died==0|died==.|died==1) & timeOS3==953 & trtrand==0)
egen IPCWKMD953=total(IPCWKMD953)
drop IPCWKMD953 IPCWKMD953
gen IPCWKMD1110 = stabweightxo if (died==1 & timeOS3==1110 & trtrand==0)
egen IPCWKMD1110=total(IPCWKMD1110)
gen IPCWKMD1110= stabweightxo if ((died==0|died==.|died==1) & timeOS3==1110 & trtrand==0)
egen IPCWKMD1110=total(IPCWKMD1110)
drop IPCWKMD1110 IPCWKMD1110
gen IPCWKMD1189 = stabweightxo if (died==1 & timeOS3==1189 & trtrand==0)
egen IPCWKMD1189=total(IPCWKMD1189)
gen IPCWKMD1189= stabweightxo if ((died==0|died==.|died==1) & timeOS3==1189 & trtrand==0)
egen IPCWKMD1189=total(IPCWKMD1189)
drop IPCWKMD1189 IPCWKMD1189
gen IPCWKMD1235 = stabweightxo if (died==1 & timeOS3==1235 & trtrand==0)
egen IPCWKMD1235=total(IPCWKMD1235)
gen IPCWKMD1235= stabweightxo if ((died==0|died==.|died==1) & timeOS3==1235 & trtrand==0)
egen IPCWKMD1235=total(IPCWKMD1235)
drop IPCWKMD1235 IPCWKMD1235
gen IPCWKMD1313 = stabweightxo if (died==1 & timeOS3==1313 & trtrand==0)
egen IPCWKMD1313=total(IPCWKMD1313)
gen IPCWKMD1313= stabweightxo if ((died==0|died==.|died==1) & timeOS3==1313 & trtrand==0)
egen IPCWKMD1313=total(IPCWKMD1313)
drop IPCWKMD1313 IPCWKMD1313
drop if _n>1314
gen time=( _n-1)*1

```

lambda for IPCW KM

```

gen IPCWLAMBDA=0
forvalues t=1(1)689 {
  replace IPCWLAMBDA=IPCWKMN`t'/IPCWKMDD`t' if time==`t'
}

```

gen survival probabilities for each time point

```

replace IPCWLAMBDA=IPCWKMN704/IPCWKMD704 if time==704
replace IPCWLAMBDA=IPCWKMN753/IPCWKMD753 if time==753
replace IPCWLAMBDA=IPCWKMN754/IPCWKMD754 if time==754
replace IPCWLAMBDA=IPCWKMN798/IPCWKMD798 if time==798
replace IPCWLAMBDA=IPCWKMN800/IPCWKMD800 if time==800
replace IPCWLAMBDA=IPCWKMN815/IPCWKMD815 if time==815
replace IPCWLAMBDA=IPCWKMN829/IPCWKMD829 if time==829
replace IPCWLAMBDA=IPCWKMN846/IPCWKMD846 if time==846
replace IPCWLAMBDA=IPCWKMN847/IPCWKMD847 if time==847
replace IPCWLAMBDA=IPCWKMN848/IPCWKMD848 if time==848
replace IPCWLAMBDA=IPCWKMN857/IPCWKMD857 if time==857
replace IPCWLAMBDA=IPCWKMN861/IPCWKMD861 if time==861
replace IPCWLAMBDA=IPCWKMN862/IPCWKMD862 if time==862
replace IPCWLAMBDA=IPCWKMN872/IPCWKMD872 if time==872
replace IPCWLAMBDA=IPCWKMN879/IPCWKMD879 if time==879
replace IPCWLAMBDA=IPCWKMN885/IPCWKMD885 if time==885
replace IPCWLAMBDA=IPCWKMN896/IPCWKMD896 if time==896
replace IPCWLAMBDA=IPCWKMN905/IPCWKMD905 if time==905
replace IPCWLAMBDA=IPCWKMN907/IPCWKMD907 if time==907
replace IPCWLAMBDA=IPCWKMN909/IPCWKMD909 if time==909
replace IPCWLAMBDA=IPCWKMN931/IPCWKMD931 if time==931
replace IPCWLAMBDA=IPCWKMN953/IPCWKMD953 if time==953
replace IPCWLAMBDA=IPCWKMN1110/IPCWKMD1110 if time==1110

```

```

replace IPCWLAMBDA=IPCWKMN1189/IPCWKMD1189 if time==1189
replace IPCWLAMBDA=IPCWKMN1235/IPCWKMD1235 if time==1235
replace IPCWLAMBDA=IPCWKMN1313/IPCWKMD1313 if time==1313
gen IPCWKMS=1
replace IPCWKMS=(IPCWKMS[_n-1]-(IPCWKMS[_n-1]*IPCWLAMBDA)) if _n>=2
***AUC***
gen AUC=0
replace AUC=(time-time[_n-1])*IPCWKMS[_n-1]
egen IPCWAUC=total(AUC)
summ IPCWAUC

```

g) Code for creating a dataset that approximates the IPCW weighted Kaplan-Meier

want to fit a Generalised Gamma to the WKM. However, WKM isn't an adjusted dataset, so don't have the patient data to do this. So have done: (1-survival prob(from wkm)) * 100000 over time to get approx patient numbers in an a hypothetical 100000 patient trial - have to round to whole numbers, but with such large trial this should be approx ok. Then fit a GG to this

```

set obs 100000
gen id = _n
gen OS=0
replace OS=8 if id>=1 & id<=504
replace OS=20 if id>504 & id<=1006
replace OS=21 if id>1006 & id<=1512
replace OS=23 if id>1512 & id<=2022
replace OS=26 if id>2022 & id<=3036
replace OS=29 if id>3036 & id<=3553
replace OS=43 if id>3553 & id<=4070
replace OS=49 if id>4070 & id<=4572
replace OS=58 if id>4572 & id<=5079
replace OS=60 if id>5079 & id<=6122
replace OS=64 if id>6122 & id<=7152
replace OS=71 if id>7152 & id<=7659
replace OS=78 if id>7659 & id<=8198
replace OS=87 if id>8198 & id<=8713
replace OS=89 if id>8713 & id<=9233
replace OS=97 if id>9233 & id<=9743
replace OS=98 if id>9743 & id<=10246
replace OS=102 if id>10246 & id<=10812
replace OS=111 if id>10812 & id<=11371
replace OS=114 if id>11371 & id<=11881
replace OS=119 if id>11881 & id<=12420
replace OS=130 if id>12420 & id<=13027
replace OS=137 if id>13027 & id<=13562
replace OS=140 if id>13562 & id<=14090
replace OS=146 if id>14090 & id<=14621
replace OS=147 if id>14621 & id<=15229
replace OS=157 if id>15229 & id<=15754
replace OS=158 if id>15754 & id<=16275
replace OS=166 if id>16275 & id<=17334
replace OS=168 if id>17334 & id<=17857
replace OS=179 if id>17857 & id<=18376
replace OS=184 if id>18376 & id<=18888
replace OS=185 if id>18888 & id<=19531
replace OS=189 if id>19531 & id<=20054

```

```

replace OS=192 if id>20054 & id<=20646
replace OS=204 if id>20646 & id<=21170
replace OS=211 if id>21170 & id<=21685
replace OS=216 if id>21685 & id<=22236
replace OS=220 if id>22236 & id<=22802
replace OS=226 if id>22802 & id<=23323
replace OS=234 if id>23323 & id<=23897
replace OS=243 if id>23897 & id<=24400
replace OS=244 if id>24400 & id<=24894
replace OS=245 if id>24894 & id<=25489
replace OS=253 if id>25489 & id<=26047
replace OS=256 if id>26047 & id<=26599
replace OS=260 if id>26599 & id<=27097
replace OS=261 if id>27097 & id<=27592
replace OS=262 if id>27592 & id<=28087
replace OS=263 if id>28087 & id<=28607
replace OS=266 if id>28607 & id<=29171
replace OS=269 if id>29171 & id<=29759
replace OS=282 if id>29759 & id<=30286
replace OS=290 if id>30286 & id<=30868
replace OS=300 if id>30868 & id<=31377
replace OS=303 if id>31377 & id<=31954
replace OS=304 if id>31954 & id<=33140
replace OS=305 if id>33140 & id<=33651
replace OS=308 if id>33651 & id<=34739
replace OS=310 if id>34739 & id<=35283
replace OS=314 if id>35283 & id<=35799
replace OS=316 if id>35799 & id<=36909
replace OS=317 if id>36909 & id<=37439
replace OS=335 if id>37439 & id<=37962
replace OS=344 if id>37962 & id<=38510
replace OS=354 if id>38510 & id<=39152
replace OS=355 if id>39152 & id<=39700
replace OS=359 if id>39700 & id<=40190
replace OS=373 if id>40190 & id<=40717
replace OS=374 if id>40717 & id<=41234
replace OS=377 if id>41234 & id<=41793
replace OS=380 if id>41793 & id<=42396
replace OS=381 if id>42396 & id<=42894
replace OS=386 if id>42894 & id<=43395
replace OS=392 if id>43395 & id<=43971
replace OS=395 if id>43971 & id<=44586
replace OS=396 if id>44586 & id<=47690
replace OS=413 if id>47690 & id<=48583
replace OS=432 if id>48583 & id<=49221
replace OS=435 if id>49221 & id<=49728
replace OS=438 if id>49728 & id<=50342
replace OS=453 if id>50342 & id<=50852
replace OS=461 if id>50852 & id<=51385
replace OS=466 if id>51385 & id<=52218
replace OS=468 if id>52218 & id<=52785
replace OS=477 if id>52785 & id<=53298
replace OS=480 if id>53298 & id<=53802
replace OS=497 if id>53802 & id<=54515
replace OS=500 if id>54515 & id<=55035

```

```

replace OS=507 if id>55035 & id<=60702
replace OS=516 if id>60702 & id<=61235
replace OS=517 if id>61235 & id<=62342
replace OS=518 if id>62342 & id<=62852
replace OS=522 if id>62852 & id<=63355
replace OS=525 if id>63355 & id<=64197
replace OS=528 if id>64197 & id<=64722
replace OS=536 if id>64722 & id<=65227
replace OS=538 if id>65227 & id<=65844
replace OS=542 if id>65844 & id<=66375
replace OS=549 if id>66375 & id<=66879
replace OS=558 if id>66879 & id<=67916
replace OS=559 if id>67916 & id<=68460
replace OS=566 if id>68460 & id<=69299
replace OS=571 if id>69299 & id<=69939
replace OS=590 if id>69939 & id<=70647
replace OS=591 if id>70647 & id<=71152
replace OS=610 if id>71152 & id<=71941
replace OS=631 if id>71941 & id<=72483
replace OS=634 if id>72483 & id<=73195
replace OS=664 if id>73195 & id<=73720
replace OS=672 if id>73720 & id<=74223
replace OS=675 if id>74223 & id<=74830
replace OS=689 if id>74830 & id<=75583
replace OS=704 if id>75583 & id<=76506
replace OS=753 if id>76506 & id<=77042
replace OS=754 if id>77042 & id<=77861
replace OS=798 if id>77861 & id<=78408
replace OS=800 if id>78408 & id<=78987
replace OS=815 if id>78987 & id<=79842
replace OS=829 if id>79842 & id<=80520
replace OS=846 if id>80520 & id<=81096
replace OS=847 if id>81096 & id<=81818
replace OS=848 if id>81818 & id<=82490
replace OS=861 if id>82490 & id<=83160
replace OS=862 if id>83160 & id<=83784
replace OS=879 if id>83784 & id<=84450
replace OS=885 if id>84450 & id<=85126
replace OS=896 if id>85126 & id<=85777
replace OS=907 if id>85777 & id<=86756
replace OS=909 if id>86756 & id<=87925
replace OS=931 if id>87925 & id<=88810
replace OS=953 if id>88810 & id<=89601
replace OS=1110 if id>89601 & id<=91133
replace OS=1189 if id>91133 & id<=93275
replace OS=1235 if id>93275 & id<=97163
replace OS=1313 if id>97163 & id<=100000

```

```

gen died=1
replace died=0 if id>97163 & id<=100000
stset OS, failure(died) id(id)
streg, dist(gamma) time tr
estat ic
stcurv, survival range(0 3700) outfile(WKMGG2v2)
predict gammaOS, sur

```