



An exploration of methods to incorporate external evidence into survival modelling in oncology

Ash Bullement

BSc, MSc

Thesis submitted for the degree of Doctor of Philosophy

Supervised by Prof. Nicholas Latimer, Prof. Matthew Stevenson, and Prof. Gianluca Baio

Word count: 94,363
(Excluding tables, figures, captions, footnotes, and summary boxes)

The University of Sheffield
Faculty of Health
School of Medicine and Population Health
Sheffield Centre for Health and Related Research (SCHARR)

Submitted: December 2025 (revised March 2026)

Abstract

This thesis explores the role of external evidence in survival analysis for health technology assessment (HTA), specifically when estimating overall survival (OS) over a lifetime horizon, for populations with cancer.

External evidence encompasses data and/or information from sources such as disease registries, observational studies, expert opinion, and general population statistics; which has the potential to enhance decision-making critical survival extrapolations. However, external evidence introduces challenges related to evidence compatibility, methodological rigour, and transparency.

A systematic review was undertaken to identify and categorise methods used in the literature to leverage external evidence. This review was complemented by an analysis of methods used in contemporary HTAs of cancer treatments. Together, these reviews provided a foundation for evaluating the feasibility and relevance of different approaches in real-world contexts.

To test the methods three case studies (squamous cell carcinoma of the head and neck, melanoma, and non-small-cell lung cancer) were used to test selected methods, which illustrated practical considerations and issues with implementation. The method testing resulted in a set of seven comparable methods which could then be explored further.

A neutral comparison study was conducted, evaluating the performance of the seven methods under scenarios with varying data-generating mechanisms. Characteristics that varied across scenarios included: sample size, study follow-up, hazard profiles, survival magnitude, and external evidence similarity (versus the pivotal trial). The findings revealed trade-offs across methods under different scenarios, demonstrating that no single approach was dominant. Instead, the choice of method must account for the decision context including the nature of the external evidence and the characteristics of the target population.

This thesis concludes with practical recommendations for selecting and applying methods to incorporate external evidence into survival analysis for HTA, as well as recommendations for future research.

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not previously been presented for an award at this, or any other, university.

Acknowledgements

- Prof. Nicholas Latimer, Prof. Matthew Stevenson, and Prof. Gianluca Baio for supervision, advice, and guidance throughout the development of the thesis.
- Dr. Anthony Hatswell for general advice and support, as well as guidance relating to simulation and neutral comparison studies (Chapter 6), and for the spline weight wedding present photographed in Chapter 2.
- Dr. Lesley Uttley and Dr. Ruth Wong for advice and guidance concerning the development of the search strategies for the systematic literature review (Chapter 3).
- Dr. Gemma Shields for assistance with the execution of the systematic literature review (Chapter 3).
- Prof. Mark Rutherford for explanations of, and advice related to, relative and net survival methods (Chapter 3).
- Dr. Christopher Jackson for assistance in verifying the outputs of the literature search (Chapter 3).
- Ms. Simone Critchlow for assistance with the review of Single Technology Appraisals (Chapter 4).
- Ms. Audrey Petitjean for collaboration with the update to the review of Single Technology Appraisals (Chapter 4).
- Mr. Mark Edmondson-Jones for advice relating to statistical programming and fitting of survival models (Chapter 5).
- Prof. Nicky Welton for advice and guidance relating to multi-parameter evidence synthesis, as well as simulation and neutral comparison studies (Chapters 5 and 6).

Publications and presentations

Publications

- Bullement A, Kearns B. Incorporating external trial data to improve survival extrapolations: a pilot study of the COU-AA-301 trial. *Health Serv Outcomes Res Method*. 2022. 22:317–331.¹
This study was undertaken early in the development of the thesis, but served as a foundation for the methods testing covered in Chapter 5. For this paper, my co-author Dr. Kearns developed the study design, and I performed the statistical analysis. Together, we co-wrote the manuscript.
- Vadgama S, Mann J, Bashir Z, Spooner C, Collins GP, Bullement A. Predicting Survival for Chimeric Antigen Receptor T-Cell Therapy: A Validation of Survival Models Using Follow-Up Data From ZUMA-1. *Value Health*. 2022 Jun;25(6):1010-1017.²
This study was undertaken as part of my role working in industry, collaborating with Sachin Vadgama and colleagues at Kite, a Gilead company, but covers some broad survival analysis methodology discussed throughout the thesis. For this paper, my contributions were inputting to the concept and design of the analysis, execution of the analysis and interpretation of data, drafting of the manuscript, and critical revision of the manuscript.
- Bullement A, Stevenson MD, Baio G, Shields GE, Latimer NR. A Systematic Review of Methods to Incorporate External Evidence into Trial-Based Survival Extrapolations for Health Technology Assessment. *Med Decis Making*. 2023 Jul;43(5):610-620.³
This study reports the findings of the SLR discussed in Chapter 3. For this paper, I was the primary reviewer and writer, with input from my supervisors (Prof. Latimer, Prof. Stevenson, and Prof. Baio), and assistance from Dr. Shields with screening.
- Bullement A, Edmondson-Jones M, Guyot P, Welton NJ, Baio G, Stevenson M, Latimer NR. MPES-R: Multi-Parameter Evidence Synthesis in R for Survival Extrapolation-A Tutorial. *Pharmacoeconomics*. 2024 Dec;42(12):1317-1327.⁴
This study provides a tutorial related to multi-parameter evidence synthesis methods, which was developed as part of the methods testing covered in Chapter 5. For this paper, my co-author Mr. Edmondson-Jones assisted with statistical programming. I developed the case studies and analyses presented in the tutorial. All co-authors inputted to the design and execution of the tutorial paper, which I drafted and then edited.
- Petitjean A, Shang H, Bullement A, Latimer NR. Use of External Data to Inform Overall Survival Extrapolation in NICE Technology Appraisals for Oncology Drugs. *J Med Econ*. 2025 Dec;28(1):803-813. doi: 10.1080/13696998.2025.2506968. Epub 2025 May 24.⁵
This study reports findings from a review of NICE appraisals covered in Chapter 4. For this paper, I collaborated with Audrey Petitjean and Huiyi Shang from Sanofi to update my original review to include later published appraisals.

Presentations

- Bullement A, Critchlow S, Stevenson MD, Baio G, Latimer N. A Review of External Evidence Used to Support Survival Extrapolation in Assessments of Cancer Drugs by the National Institute for Health and Care Excellence (NICE). Presented at the ISPOR 2022 Annual European Congress, Vienna, Austria. 6-9 November 2022.⁶
- McQueen RB, Bullement A, Heeg B. Incorporating External Evidence into Extrapolation: Making Sense of Seemingly Endless Options. Presented at the ISPOR 2023 Annual Congress, Boston, MA, USA. 7-10 May 2023.⁷
- McQueen RB, Bullement A, Heeg B, Lee D. Incorporating External Evidence into Extrapolations: A Debate on the Trade-Offs between Complexity and Simplicity. Presented at the ISPOR 2023 Annual European Congress, Copenhagen, Denmark. 12-15 November 2023.⁸
- Pedder H, Allignol A, Jen M, Bullement A. Survival at the Margins: Navigating HTA Challenges in Modern Oncology Trials. Presented at the ISPOR 2025 Annual European Congress, Glasgow, Scotland. 9-12 November 2025.⁹

Contents

1.	Introduction.....	1
1.1	Health technology assessment (HTA).....	1
1.2	Survival analysis	6
1.3	External evidence.....	18
1.4	Research questions	21
1.5	Thesis structure	24
2.	Foundations of producing lifetime survival estimates for HTA	26
2.1	Key terminology	27
2.2	Evidence sources.....	35
2.3	Survival modelling.....	40
3.	Systematic literature review (SLR)	58
3.1	Introduction	59
3.2	Methods	60
3.3	Results	74
3.4	Discussion.....	159
4.	Single Technology Appraisal (STA) review	167
4.1	Introduction	168
4.2	Methods	170
4.3	Results	172
4.4	Discussion.....	183
5.	Method testing and case studies	188
5.1	Introduction	189
5.2	Methods	191
5.3	Results	200
5.4	Discussion.....	224
6.	Neutral comparison study (NCS).....	234
6.1	Introduction	235
6.2	Methods	237
6.3	Results	264
6.4	Discussion.....	289
7.	Overview, discussion, recommendations, and conclusions.....	298
7.1	Overview	298
7.2	Discussion.....	299
7.3	Method-specific recommendations.....	313
7.4	General recommendations	317
7.5	Conclusions	324
	References.....	326
	Appendices	344

Tables

Table 1-1: Comparison of efficacy and effectiveness	20
Table 2-1: Mixture versus non-mixture cure models	49
Table 2-2: Illustration of survival analysis data format.....	50
Table 3-1: Inclusion and exclusion criteria	61
Table 3-2: Comparison of pearl-growing methods re-produced from Schlosser <i>et al.</i> , (2006)	63
Table 3-3: Pearls from the Jackson <i>et al.</i> , (2016) review	65
Table 3-4: Database checking for Jackson <i>et al.</i> , (2016) review pearls	66
Table 3-5: Ovid searching (first iteration)	68
Table 3-6: Web of Science searching (first iteration).....	69
Table 3-7: Search limits	69
Table 3-8: Data extraction table	72
Table 3-9: Expected hits for validation	74
Table 3-10: Methods identified in first iteration taken forward for further evaluation in later chapters	77
Table 3-11: Methods identified in first iteration not taken forward but included in the review	80
Table 3-12: Multiplicative versus additive relative survival models.....	85
Table 3-13: Ovid searching (second iteration, Medline and EconLit)	130
Table 3-14: Web of Science searching (second iteration).....	131
Table 3-15: Methods identified in second iteration taken forward for further evaluation in later chapters	133
Table 3-16: Methods identified in second iteration not taken forward but included in the review	134
Table 3-17: Citation tracking results.....	145
Table 3-18: Reference checking results.....	146
Table 3-19: ISPOR Scientific Presentations Database search strings	150
Table 3-20: Methods identified in conference proceedings search	151
Table 3-21: Methods identified in 2024 update taken forward for further evaluation in later chapters.....	156
Table 3-22: Methods identified in 2024 update not taken forward but included in the review	156
Table 3-23: Summary of retained methods for method testing	162
Table 3-24: Summary of discarded methods	163
Table 4-1: Inclusion and exclusion criteria	170
Table 4-2: Included appraisals by cancer type	173
Table 5-1: Summary of retained methods for method testing	189
Table 5-2: Method testing code availability	192
Table 5-3: Case studies for method testing.....	193
Table 5-4: Summary of method-specific requirements.....	196
Table 5-5: Summary of method-specific inputs (vanilla implementations).....	198
Table 5-6: Parameters for simulated data using in method testing	199
Table 5-7: Top-level findings from method investigations	226
Table 5-8: Summary of discarded methods following method testing	229
Table 6-1: Parameters for hazard profiles used in NCS.....	241
Table 6-2: Tabulated summary of scenarios in the neutral comparison study	245
Table 6-3: Truth lookup for scenarios in the neutral comparison study	250
Table 6-4: Summary of retained methods for neutral comparison study	251
Table 6-5: Implementation of retained methods for neutral comparison study.....	252
Table 6-6: Truth considerations for each method.....	256
Table 6-7: Computers used in the NCS	263
Table 6-8: Run time – model fitting (full run)	267
Table 6-9: Run time – model fitting (sample run)	267

Figures

Figure 1-1: Health technology assessment agencies in the UK	2
Figure 1-2: Kaplan-Meier estimator of OS from the FLAURA-1 trial.....	7
Figure 1-3: Re-created Kaplan-Meier estimator of overall survival from FLAURA-1	11
Figure 1-4: Illustration of a complex hazard function.....	15
Figure 1-5: Parametric models presented in NICE TA621	16
Figure 1-6: Thesis structure and research questions addressed.....	25
Figure 2-1: Example of impact of standardised mortality ratios on survival.....	34
Figure 2-2: Summary of clinical trial phases (in humans).....	37
Figure 2-3: Example Kaplan-Meier estimation	41
Figure 2-4: A lead spline weight	46
Figure 2-5: How lead spline weights are used in real-world applications	46
Figure 2-6: Example partitioned-survival analysis model structure	54
Figure 3-1: Summary of relevant themes for consideration within the search strategy	67
Figure 3-2: Schematic of comprehensive pearl growing search iterations	70
Figure 3-3: PRISMA diagram (after first iteration)	75
Figure 3-4: Illustration of additive and multiplicative hazards models.....	84
Figure 3-5: The graphical model, as presented by Demiris & Sharples, (2006)	88
Figure 3-6: Summary of Nelson <i>et al.</i> , (2008) approach	92
Figure 3-7: Predicted survival re-produced from Nelson <i>et al.</i> , (2008).....	93
Figure 3-8: Extension of MCM to include two Weibull models (Lambert <i>et al.</i> , 2010)	94
Figure 3-9: Summary of Hawe <i>et al.</i> , (2014) approach	98
Figure 3-10: Overview of extrapolations re-produced from Benaglia <i>et al.</i> , (2015)	101
Figure 3-11: Three-part curve fit re-produced from Larkin <i>et al.</i> , (2015)	104
Figure 3-12: 5-year overall survival curves re-produced from Barlev <i>et al.</i> , (2017).....	106
Figure 3-13: Extrapolations of overall survival re-produced from Fisher <i>et al.</i> , (2017)	108
Figure 3-14: Summary of Guyot <i>et al.</i> , (2017) approach.....	110
Figure 3-15: Illustration of principle behind rolling extrapolation approach.....	114
Figure 3-16: Rolling extrapolation algorithm flowchart	115
Figure 3-17: Rolling extrapolation example re-produced from Hwang <i>et al.</i> , (2017)	116
Figure 3-18: Illustration of Pennington <i>et al.</i> method with only summary data	118
Figure 3-19: Estimates of overall survival reproduced from Cope <i>et al.</i> , (2019).....	121
Figure 3-20: Illustration of Vickers method	128
Figure 3-21: PRISMA diagram (after second iteration)	132
Figure 3-22: Models with priors on shape parameter re-produced from Soikkeli <i>et al.</i> , (2019)	137
Figure 3-23: Summary of Felizzi <i>et al.</i> , (2021) mixture-cure modelling approach	140
Figure 3-24: Three-part curve fit re-produced from Aballéa <i>et al.</i> , (2007)	147
Figure 3-25: Survival extrapolation re-produced from Hisashige <i>et al.</i> , (2008)	148
Figure 3-26: Example survival extrapolation re-produced from Howard <i>et al.</i> , (2009).....	149
Figure 3-27: Graphical representation of the blended curve method	152
Figure 3-28: Summary of final set of methods	154
Figure 3-29: Summary of Jackson (2023) approach	158
Figure 3-30: Sankey diagram of methods identified	160
Figure 3-31: Methods summary following completion of SLR	165
Figure 4-1: Introducing external evidence into NICE STA process	168
Figure 4-2: Identification of appraisals included in synthesis	173
Figure 4-3: Breakdown of appraisals by use of external evidence	175
Figure 5-1: Overlay of SEER estimates from Guyot <i>et al.</i> , (2017) versus re-creation	195
Figure 5-2: Method testing – base-case results (SCCHN)	202
Figure 5-3: Method testing – base-case results (melanoma)	203
Figure 5-4: Method testing – base-case results (NSCLC).....	204
Figure 5-5: Melanoma results for ‘standard’ and ‘summary’ piecewise models	206
Figure 5-6: SCCHN results for standard piecewise and blended survival models.....	207
Figure 5-7: NSCLC results for Bayesian relative survival with long-term hazard adjustment.....	208
Figure 5-8: NSCLC results for expert opinion survival with standard deviation edits	209
Figure 5-9: SCCHN results for Bayesian versus frequentist relative survival.....	210
Figure 5-10: NSCLC results comparing Hawe and Soikkeli	211
Figure 5-11: SCCHN results comparing alternative relative survival regions	212
Figure 5-12: Melanoma results comparing alternative relative survival models	213
Figure 5-13: SCCHN results with varying strength of prior for common shape parameter(s).....	214
Figure 5-14: SCCHN results disabling hazard ratio and life table adjustments	216
Figure 5-15: SCCHN results changing number and location of knots	217
Figure 5-16: SCCHN results comparing both MPES approaches fitted to two arms	218
Figure 5-17: Method testing – simulated data results (scenario 1).....	220
Figure 5-18: Method testing – simulated data results (scenario 2).....	221
Figure 5-19: Method testing – simulated data results (scenario 3).....	222

Figure 5-20: Method testing – simulated data results (scenario 4).....	223
Figure 5-21: Methods summary following completion of method testing and case studies	232
Figure 6-1: Overlay of TSD 21 simulation study and Gray <i>et al.</i> , (2021) survival estimates.....	239
Figure 6-2: Gray <i>et al.</i> , (2021) hazard estimates.....	240
Figure 6-3: Hazard profiles used in the NCS (up to 10 years) for disease-specific mortality.....	241
Figure 6-4: Illustration of equal 40-year RMST but poor fit.....	248
Figure 6-5: Example nested loop plot by Latimer <i>et al.</i> , (2020).....	257
Figure 6-6: Description of script generation in relation to TSD 21	261
Figure 6-7: Example of two logic errors flagged for the blended survival method	265
Figure 6-8: NLP: Bias in mean survival, all scenarios, trial as truth.....	269
Figure 6-9: NLP: Bias in mean survival, all scenarios, external as truth.....	270
Figure 6-10: NLP: Mean survival, all scenarios, both trial and external as truth.....	274
Figure 6-11: NLP: Bias in mean survival, first eight scenarios, trial as truth.....	275
Figure 6-12: NLP: Bias in mean survival, first eight scenarios, external as truth.....	276
Figure 6-13: NLP: Absolute bias in mean survival, all scenarios, trial as truth	281
Figure 6-14: NLP: Percentage bias in mean survival, all scenarios, trial as truth.....	282
Figure 6-15: NLP: empSE in mean survival, all scenarios	283
Figure 6-16: NLP: Bias in RMST at 10 years, all scenarios, trial as truth.....	285
Figure 6-17: NLP: Bias in point estimate of survival at 10 years, all scenarios, trial as truth.....	286
Figure 6-18: NLP: Bias in mean survival (faint) versus RMST at 10 years (bold), all scenarios, trial as truth	287
Figure 6-19: Hazard profiles used in the NCS (up to 40 years).....	288
Figure 7-1: B-A-K-E: Recommendations for analysts (MPES methods)	316
Figure 7-2: Flowchart of recommendations for use of external evidence for survival in HTA.....	322

Abbreviations

ACE	Agency for Care Effectiveness	NIHR	National Institute for Health and Care Research
ACM	Appraisal committee meeting	NJR	National Joint Registry
ADEMP	Aim(s), Data generating mechanism(s), Estimand(s), Method(s), Performance measure(s)	NLP	Nested loop plot
AFT	Accelerated failure time	NMA	Network meta-analysis
AIC	Akaike's information criterion	NMCM	Non-mixture-cure model
AJCC	American Joint Committee on Cancer	NMH	Non-monotonic hazards
AUC	Area under the curve	ONS	Office for National Statistics
AWMSG	All Wales Medicines Strategy Group	OS	Overall survival
BIC	Bayesian information criterion	PAIC	Population-adjusted indirect comparison
BIRC	Blinded independent review committee	PartSA	Partitioned-survival analysis
BMJ	British Medical Journal	PBAC	Pharmaceutical Benefits Advisory Committee
BSC	Best supportive care	PD	Progressed disease
BUGS	Bayesian inference Using Gibbs Sampling	PF	Progression-free
CADTH	Canadian Agency for Drugs and Technologies in Health	PFS	Progression-free survival
CAR	Chimeric Antigen Receptor	PICOS	Population, Intervention, Comparison, Outcome(s), Study type
CCRP	Childhood Cancer Registry of Piedmont	PPS	Post-progression survival
CDA	Canada's Drug Agency	PrePS	Pre-progression survival
CDF	Cancer Drugs Fund	PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
CGP	Corrected group prognosis	PSA	Probabilistic sensitivity analysis
CI	Confidence interval	PSM	Propensity score matching
CMA	Cost-minimisation analysis	PSS	Personal Social Services
CPG	Comprehensive pearl growing	QALY(s)	Quality-adjusted life year(s)
CRAN	(The) Comprehensive R Archive Network	RCS	Restricted cubic spline
Cri	Credible interval	RCT	Randomised controlled trial
CUA	Cost-utility analysis	RFS	Recurrence- or relapse-free survival
DES	Discrete event simulation	RMST	Restricted mean survival time
DFS	Disease-free survival	RPSFT(M)	Rank-preserving structural failure time (model)
DGM	Data generating mechanism	RWD	Real-world data
DIC	Deviance information criterion	RWE	Real-world evidence
DSU	Decision Support Unit	SACT	Systemic anticancer therapy
EAG	External Assessment Group	SD	Standard deviation
EAR	External Assessment Report	SE	Standard error
ECOG	Eastern Cooperative Oncology Group	SEE	Structured Expert Elicitation
EFS	Event-free survival	SEER	Surveillance, Epidemiology, and End Results
EMA	European Medicines Agency	SHELF	Sheffield Elicitation Framework
ERG	Evidence Review Group	SLR	Systematic literature review
ESMO	European Society for Medical Oncology	SMC	Scottish Medicines Consortium
FAD	Final Appraisal Determination	SMR	Standardised mortality ratio
FDA	Food and Drug Administration	SoC	Standard of care
FP	Fractional polynomial	STA	Single technology appraisal
FP1	First-order fractional polynomial	STC	Simulated treatment comparison
FP2	Second-order fractional polynomial	STM	State-transition model
FP3	Third-order fractional polynomial	TA	Technology appraisal
FPNMA	Fractional polynomial network meta-analysis	TPG	Traditional pearl growing
FSEA	Fitted separately to each arm	TSD	Technical Support Document
GAD	Government Actuary's Department	TTD	Time to treatment discontinuation
GLM	Generalised linear model	TTE	Time-to-event
HES	Hospital Episode Statistics	TTP	Time to progression
HR	Hazard ratio	USRDS	United States Renal Data System
HMD	Human Mortality Database	WHO	World Health Organisation
HPC	High Performance Computing	WOS	Web of Science
HRQL	Health-related quality of life	WTP	Willingness to pay
HST	Highly specialised technology		
HTA	Health technology assessment		
ICER	Incremental cost-effectiveness ratio		
IPD	Individual patient data		
ISPOR	The Professional Society for Health Economics and Outcomes Research		
ITC	Indirect treatment comparison		
ITT	Intention to treat		
JCA	Joint Clinical Assessment		
JR	Jackson <i>et al.</i> , review		
KM	Kaplan-Meier		
LE	Life expectancy		
LY(s)	Life-year(s)		
LYG	Life-years gained		
MA	Marketing authorisation		
MAIC	Matching adjusted indirect comparison		
MCM	Mixture-cure model		
MCSE	Monte Carlo standard error		
MHRA	Medicines & Healthcare products Regulatory Agency		
ML-NMR	Multilevel network meta-regression		
MPES	Multi-parameter evidence synthesis		
MRC	Medical Research Council		
MSM	Multi-state model		
MTA	Multiple technology appraisal		
NED	No evidence of disease		
NCS	Neutral comparison study		
NHS	National Health Service		
NICE	The National Institute for Health and Care Excellence		
NIH	National Institutes of Health		

Note: Footnotes specific to specific diseases (e.g., acronyms for specific cancer types), treatments (e.g., acronyms for drug regimens), or countries are not included in the list of abbreviations for the sake of brevity, but are defined on first mention throughout the thesis.

1. Introduction

This thesis focuses on the use of survival analysis methods to estimate overall survival (OS) for health technology assessment (HTA) that take into consideration external evidence as well as data collected from a pivotal* clinical trial, focusing on a single treatment group context. 'External evidence' is defined in greater detail later in the thesis, but broadly encompasses data or information about OS collected outside of the pivotal clinical trial. Estimates of OS over a lifetime horizon (often termed as 'survival extrapolations') are often influential on cost effectiveness results – which then inform HTA decision making. Historically these extrapolations have been based on trial data alone.¹⁰ External evidence can help us provide more realistic extrapolations, subject to several important considerations about the applicability of the evidence to the decision problem, and the treatment groups being compared.

This thesis primarily discusses treatments for cancer and focuses on estimation of OS specifically in a single treatment group, as OS is generally one of the main outcomes of interest for HTAs of cancer treatments. However, the research presented in this thesis has applicability outside of cancer populations and addresses adjacent research topics where relevant (e.g., other survival-based outcomes). The focus on a single treatment group allows for separateion of absolute survival estimation, versus comparative effectiveness. Despite this, the findings could reasonably be considered to apply to multiple treatment groups, in order to generate comparisons for HTA. The main contribution of this thesis is to systematically identify and assess relevant methodological approaches, to draw recommendations for the use of external evidence to inform survival extrapolations for HTA.

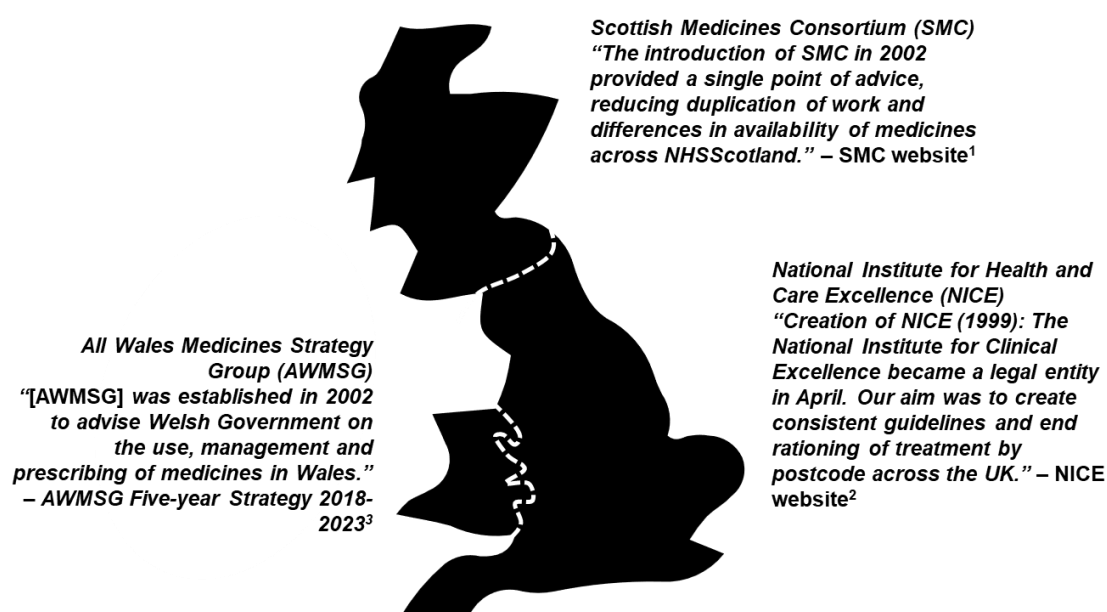
1.1 Health technology assessment (HTA)

HTA makes uses of clinical- and cost-effectiveness evidence regarding healthcare interventions to support policy decision-making. A precise definition of HTA is subject to some debate, though the World Health Organisation (WHO) defines HTA as a multidisciplinary process to systematically evaluate the properties, effects, and/or impacts of a given health technology (which may include social, economic, organisational, and ethical issues).¹¹ The Professional Society for Health Economics and Outcomes Research (ISPOR, previously known as the International Society for Pharmacoeconomics and Outcomes Research) also acknowledges that HTA uses explicitly-defined methods to determine the value of a given health technology at different points during its lifecycle.^{12, 13} These definitions highlight the use of a range of evidence sources, considering multiple perspectives, in order to allow for informed decision-making regarding healthcare resource allocation.

* A pivotal trial is a trial which is intended to serve as the primary basis from which regulatory approval is sought.

In line with the devolved nature of government in the UK, there are three established HTA bodies: The National Institute for Health and Care Excellence (NICE) in England, the Scottish Medicines Consortium (SMC), and the All Wales Medicines Strategy Group (AWMSG) (see Figure 1-1).[†] Based on a 2017 European Commission report, there are at least 56 European organisations that may be defined as HTA bodies, and many more internationally (such as Canada's Drug Agency [CDA, previously known as the Canadian Agency for Drugs and Technologies in Health, CADTH], the Pharmaceutical Benefits Advisory Committee [PBAC, Australia], the Agency for Care Effectiveness [ACE, Singapore], and the Institute for Clinical and Economic Review [United States]).¹⁴⁻¹⁸

Figure 1-1: Health technology assessment agencies in the UK



Key: AWMSG, All Wales Medicines Strategy Group; NHS, National Health Service; NHSScotland; National Health Service Scotland; NICE, National Institute for Health and Care Excellence; SMC, Scottish Medicines Consortium.

Note: NICE was renamed from 'National Institute for Clinical Excellence' to 'National Institute for Health and Care Excellence' in 2005.

Sources: ¹SMC website: <https://scottishmedicines.org.uk/about-us/>; ²NICE website: <https://www.nice.org.uk/about/who-we-are/history-of-nice>; ³AWMSG Five-year Strategy 2018–2023: <https://awttc.nhs.wales/files/awmsg/awmsg-medicines-strategy/awmsg-five-year-strategy-2018-2023/>. All websites accessed 27 April 2025.

Outside of country-specific HTA organisations, there are several influential organisations that influence HTA, and there are constantly changes in how HTAs are carried out internationally. From 12 January 2025, the European Union (EU) regulation on HTA came into force, which established a framework for conducting joint clinical assessment (JCA) of health technologies at the EU level.¹⁹ The outputs of JCAs (for relevant health technologies) are expected to be

[†] Northern Ireland does not have an equivalent HTA body. Instead, guidance issued by NICE is reviewed by the Department of Health in Northern Ireland, adapted where necessary (to ensure its applicability to Northern Ireland), and subsequently adopted. In addition, some guidance issued by NICE is adopted across the UK, including Scotland and Wales.

integral to HTAs conducted in EU member states, and while the regulation does not directly impact post-Brexit UK, some have speculated that JCA may indirectly influence UK HTA processes, and the influence of NICE internationally.^{20, 21}

A range of different types of health technology may be assessed by different HTA bodies, potentially using the same or distinct processes. In the case of NICE in England, pharmaceuticals are usually assessed via the Technology Appraisal (TA) and Highly Specialised Technology (HST) processes but can also be included as part of a clinical guideline. Interventional procedures, medical devices, digital, and diagnostic technologies each have separate processes for assessment by NICE (which in some cases could overlap), though in February 2025 a draft consultation document was published by NICE concerning the methods and processes that NICE intends to follow when evaluating interventional procedures and HealthTech products in the future.²² Unlike NICE, the SMC only issues guidance on the use of new medicines (and not, for example, medical devices).²³

For the purpose of this thesis, the focus is placed predominantly on pharmaceutical products undergoing HTA. Individual HTA bodies have variable influence (depending on their jurisdiction) but are nevertheless recognised to affect the use of interventions they issue guidance for (whether that be for, or against, their use). In England, decisions made by NICE for interventions assessed via the TA or HST appraisal processes are legally binding. These decisions commit the National Health Service (NHS) to making recommended interventions available to patients nationally within 3 months of the recommendation being made, though the use of specific interventions remains a decision made between the patient and the healthcare provider.^{24, 25}

While not an explicit requirement for HTA in general, contemporary HTA processes often involve a submission made by the marketing authorisation holder of the technology being assessed (for ease of readability, the marketing authorisation holder is henceforth termed ‘the company’). The company’s submission is then appraised by the HTA body; hence the onus is typically on the company itself to provide all necessary evidence to inform decision-making. The evidence required varies between HTA bodies, but as a minimum includes information regarding the clinical- and cost-effectiveness of the intervention versus the established standard of care (SoC) – the treatment(s) that would be theoretically displaced following a positive recommendation, where applicable (e.g., no treatment would be displaced if a new intervention is positioned after all other options have been exhausted).

The NICE Single Technology Appraisal (STA) process (and also the HST appraisal process) requires a submission from companies, whereas other HTA processes often rely on publicly available evidence. For HTAs where the company does not make a submission, the evidence

base upon which decisions may be based can be more variable and does not necessarily reflect the same context where external evidence could be used in a typical NICE STA. The use of external evidence in survival modelling, particularly for cancer treatments, has been identified as an area for further research (and in the context of a NICE STA specifically).^{26, 27}

HTAs of new interventions typically take place around the timing of regulatory approval, so as to minimise any potential delays in patient access to the new treatment. In October 2017, NICE stated its intention to publish guidance for all new drugs undergoing assessment within 90 days of the marketing authorisation being granted.²⁸ It is therefore unsurprising that a key source of evidence used to inform a company's submission is the registrational or 'pivotal' clinical trial(s) of the intervention. The trial(s) provide(s) information pertaining to the safety and efficacy of the intervention under investigation and, depending on the design and/or conduct of the trial(s), additional information such as outcomes for a control group(s), medical resource use, or health-related quality of life (HRQL) data may also be available.

While the trial(s) may provide an array of useful information for HTA decision-making, trial data are subject to limitations, especially in relation to the time at which HTA usually takes place. Clinical trials are seldom undertaken purely with the intention to inform cost-effectiveness analyses for reimbursement decision-making; rather, their focus is likely guided towards the requirements of regulatory bodies (such as the European Medicines Agency [EMA], the US Food and Drugs Administration [FDA] and the UK Medicines & Healthcare products Regulatory Agency [MHRA]). As a result, the endpoints used to assess a product's safety and efficacy may not align with those required for HTA. In particular, regulatory assessments may rely on surrogate endpoints, such as response rate, which may be insufficiently mature to support robust cost-effectiveness analysis at the time of HTA.

In addition to the potential misalignment of outcomes collected for regulatory and HTA submissions, clinical trials are designed to collect information over a specified follow-up period, which is rarely long enough to follow each patient from initiation of a given intervention until their death[‡]. Collecting information for all patients until death is both impractical (given the operational burden, both related to time and cost, of conducting clinical trials) and often unnecessary to establish the key differences in costs and effects required to inform decision-making – delaying decision-making may mean that cost-effective medicines are not provided to patients. As argued by Claxton (1999) with respect to the 'irrelevance of inference', there are clear implications associated with decisions to reimburse, or to not reimburse – both delaying or deferring a decision (e.g., waiting to collect further data) are also decisions, carrying important consequences.²⁹

[‡] After death, costs and benefits related to an intervention are typically assumed to cease for HTA purposes.

In NICE STAs, decisions related to cost effectiveness are made using the incremental cost-effectiveness ratio (ICER). The ICER represents the cost per quality-adjusted life year (QALY) gained, where the QALY represents the measure of effectiveness used in NICE appraisals. In brief, QALYs are estimated as the composite of the time spent alive and the utility weight representing HRQL, and one QALY is equivalent to one year of life (or one life-year [LY]) in perfect health. As part of the STA process, most new interventions are assessed based on the ICER being less than the specified willingness-to-pay (WTP) threshold, which is stated to be £20,000-30,000 per QALY gained, for most technologies assessed via the STA process.³⁰ Under certain circumstances (e.g., if a product is determined to meet the requirements for a disease severity modifier), a technology may be recommended with an ICER in excess of £30,000 (for example, QALYs could be assigned a weight of 1.7 if a disease severity modifier applies, equivalent to a threshold of up to £51,000 [i.e., £30,000 x 1.7]).³¹

Where costs incurred and/or health effects accrued (e.g., life-years [LYs]) after the end of follow-up in the trial(s) are important, it is necessary for HTA bodies to consider estimates of longer-term costs and/or effects in decision-making. On such occasions, the accurate estimation of OS may be critical to inform HTA decision-making. Only by explicitly estimating OS can the lifetime costs and effects associated with each treatment strategy (i.e., the intervention and its comparator[s]) be appropriately calculated and compared, and thus used to inform decision-making. More specifically, it is often the case that costs and effects extrapolated into the future have opposing effects on the estimated cost-effectiveness of a given intervention, though this relationship is context dependent. For example, additional survival for the new intervention may lead to improved health outcomes and thus (taken in isolation) may also lead to improved cost-effectiveness; however, extended survival is likely associated with additional medical management costs which could lead to worsened estimates of cost-effectiveness (again, taken in isolation).

HTAs for interventions for a range of health conditions usually require the estimation of OS (unless a simplified modelling approach is taken wherein OS is not directly estimated). This is almost always the case for HTAs of cancer medicines, as most new cancer treatments aim to extend survival either through a cure or by slowing the progression of the cancer. Consequently, in HTAs of new cancer treatments, it is often necessary to estimate lifetime survival for the new intervention over and above the established standard of care i.e., two estimates of survival are required.

As of 8th January 2025, NICE has published a total of 464 technology appraisals of cancer drugs (all being either STAs or multiple technology appraisals [MTAs], and none being HSTs). A previous review of OS estimation methods used in NICE appraisals of cancer treatments by Bell Gorrod *et al.* identified and assessed methodologies used in 58 appraisals, all of which

required the estimation of OS for decision-making.²⁷ Therefore, while the accurate estimation of OS may play an important role for the assessment of any life-extending intervention, the need to estimate OS is of particular interest within the context of cancer treatments, particularly those at advanced stages of disease, lead to substantially shorter life expectancy compared with other disease areas. Furthermore, published literature concerning OS estimation for HTA typically focus on cancer as a prominent application, given one of the intended benefits of many types of cancer treatment (though not all) is to extend survival, either through cure or delaying of disease progression.

1.2 Survival analysis

‘Survival analysis’ is used as a broad term to describe the analysis of data in the form of times from a definable entry point until the occurrence of a given event or events, that inform a corresponding endpoint.³² These data are often termed ‘time-to-event’ (TTE) data. In the case of estimating the endpoint of OS from trial data, the entry point typically constitutes the time of randomisation (which is usually shortly before when treatment is initiated), and the event of interest is death. Survival analysis methods aim to address specific features of TTE data that render traditional statistical methods (e.g., Student’s t-test to compare group means) inappropriate for consideration – namely, the potential for censoring.

In the case of OS, a censored observation means that a given patient may be known to have survived up until a given point in time, but after this time it is unknown when the patient died (given that death will eventually occur for all patients, regardless of whether this is related to the disease itself). For other TTE outcomes (such as time to disease progression), it may be the case that the event is never experienced (for example, if the patient was cured or if they died before disease progression). However, within the context of OS, it is clear that all patients will eventually experience the event of interest (death). The censoring of OS times in trials means that estimated survival curves (i.e., plotted survival proportions over time) will usually be incomplete. In other words, survival curves will comprise a combination of patients with recorded event times (deaths) and censor points. Survival curves are sometimes described as a survival function – $S(t)$, where t denotes a given timepoint (and $S(t)$ denotes the probability of survival beyond time t).

To illustrate the incompleteness of survival curves that may be seen within the context of cancer trials, it is helpful to consider the Kaplan-Meier (KM) estimator[§] of survival.³³ The KM estimator provides a means of inferring from TTE data what proportion of patients survive up

[§]KM estimators are sometimes described as KM *data*, *curves*, or *plots*. The output from a KM estimator can be plotted (which may resemble a curve for a large sample size), and x- and y-coordinates could be described as data for plotting. However, throughout this thesis the term ‘KM estimator’ is used, because the KM estimator is defined as a statistic used to estimate a survival (or survivor) function.

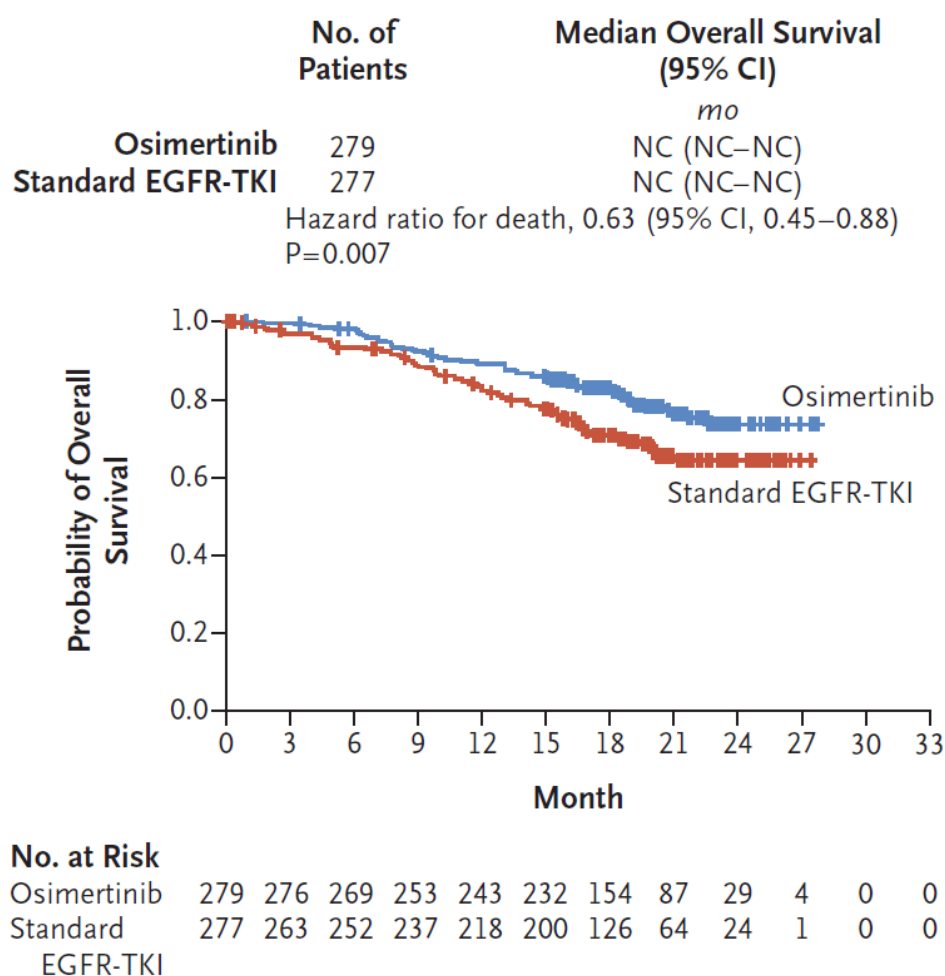
until a given point in time, and the resulting proportions may then be plotted in order to provide a visual estimate of the underlying survival curve.³⁴ The KM estimator represents a ‘stepped’ function, showing specific timepoints where at least one event has occurred (and so the proportion alive reduces), and periods where no events have occurred (and so the proportion alive remains the same). This is considered a non-parametric survival estimate – discussed further in Chapter 2.

An example of a KM estimator is used to illustrate some key features of survival data that are important for considering how survival estimates may be generated for HTA.

The FLAURA-1 trial (an illustrative example)

Figure 1-2 shows a plot of the KM estimator of OS from the FLAURA-1 trial of osimertinib (Tagrisso®, AstraZeneca) versus an SoC treatment (epidermal growth factor receptor tyrosine kinase inhibitor, EGFR-TKI) for patients with untreated EGFR-mutated advanced non-small-cell lung cancer (NSCLC). FLAURA-1 was the primary evidence source used to inform the company submission as part of NICE TA621 (later updated in NICE TA654).^{35, 36}

Figure 1-2: Kaplan-Meier estimator of OS from the FLAURA-1 trial



Key: CI, confidence interval; EGFR-TKI, epidermal growth factor receptor tyrosine kinase inhibitor (also referred to as SoC [standard of care] in NICE TA621); mo, month(s); NC, could not be calculated.

Note(s): This figure illustrates example Kapan-Meier estimates of overall survival from the FLAURA-1 trial (NCT02296125).

Source: Reproduced with permission from Soria J-C, Ohe Y, Vansteenkiste J *et al.* Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. *N. Engl. J. Med.* 378(2), 113–125 (2018), Copyright Massachusetts Medical Society. Figure 1, Panel D in Soria *et al.*³⁷

Initial observations

From Figure 1-2, there appears to be a survival advantage for patients receiving osimertinib versus EGFR-TKI – this is because the blue line (for osimertinib) sits above the red line (for standard EGFR-TKI), indicating a larger proportion of people still alive over time. However, the magnitude of this possible survival benefit is subject to substantial uncertainty given the large number of censored observations recorded after approximately 15 months on both arms. This censoring can be seen through the number of ‘tick’ marks, the drop in the number of patients still at risk, and the flattening of the KM estimates after 15 months). This is expected to be due to the nature of recruitment of patients in the FLAURA-1 study – patients were recruited between December 2014 and March 2016, and the KM estimate shown in Figure 1-2 was based on a data cut-off in June 2017.³⁷

Incomplete data and censoring

While it is important to be aware of the incompleteness of data, the reasons for censoring also warrant attention, as survival times recorded as part of a trial may be censored for a number of reasons. For example, reasons may include some patients dropping out of the study due to the tolerability intervention under investigation, or due to the end of the observation period of the trial. Both of these situations describe *right censoring*, which refers to the recorded survival time being less than the true, unknown survival time. However, depending on the type of right censoring the implications for survival analysis differ.

Censoring may be considered *informative* (i.e., the reason for censoring is related to the study itself and affects the estimation of outcomes). In the example above, if toxicity leads to a patient dropping out of the study their OS would be censored, but they would then be without active treatment, and thus likely have a relatively poorer prognosis versus if they had remained on treatment without toxicity. Biases that arise due to informative censoring may also act in opposing directions; for example, if an intervention leads to a ‘cure’ or opens up additional treatment options such as stem cell transplantation or surgical resection, patients may drop out of the study; or if an intervention is inferior then patients may drop out because they are too sick to attend clinic visits (and by consequence, may or may not discontinue treatment which could influence estimates of OS).³⁸ If informative censoring affects either or even both treatment arms, this will have implications both for the estimation of absolute outcomes for the affected arm(s), as well as estimates of relative effect.

Censoring related to the duration of follow-up is often described as *administrative censoring*, which (assuming the study is designed appropriately) can (usually) be considered as non-informative. The impact of administrative censoring on the estimation of OS can be seen in Figure 1-2 for the FLAURA-1 trial, where follow-up was limited after approximately 15 months. Administrative censoring *can* be addressed through further data collection if further collection is planned as part of the clinical trial. In the case of FLAURA-1, Figure 1-2 shows findings of an interim analysis with a data cut-off date of 12 June 2017 – these data were later updated in the final analysis.^{37, 39} However, some patients may still be alive at the end of the trial, with no further follow-up planned, and so there may remain a need to rely on extrapolation.

Median versus mean survival

Both arms in FLAURA-1 comprise of patients with a maximum follow-up of approximately 27 months, after which it is unknown what proportion of patients are expected to be alive. Importantly, it can be seen in Figure 1-2 for both arms that the median survival time cannot be estimated – the median survival estimate would correspond to the timepoint where the KM estimator is equal to 50%, and both of the KM estimators stop before reaching 50%. Median survival can be useful when describing the average survival for a given patient, as by definition, 50% of patients are expected to live less than the median survival time, and 50% of patients are expected to live longer than the median survival time. For HTA, *mean* survival is the primary measure of interest, as opposed to *median* survival. This is because HTA decisions rely on consideration of the whole disease population, and not just a ‘typical’ or ‘average’ patient.

Mean survival corresponds to the estimated LYs used to form the basis of the QALY gain which determines the ICER. However, even where QALYs and ICERs are not estimated for specific HTA bodies, consideration of the whole disease population is of paramount importance for decision making, as decisions are made for the whole eligible patient population, and not just an ‘average’ or ‘typical’ patient. Despite its importance, mean survival is not estimable from the FLAURA-1 data alone, and is highly unlikely to be estimable *without substantial additional data collection (i.e., much longer trial follow-up)*.

The hazard ratio

For the relative effect of osimertinib versus EGFR-TKI, the pivotal trial publication for FLAURA-1 presents an estimate in the form of a (single) hazard ratio (HR).³⁷ The hazard function evaluated at time t for the outcome of OS is defined as the instantaneous risk of death conditional upon an individual surviving up until time t . It follows therefore that the HR is defined as the ratio of the hazard of death in one group (e.g., those receiving the experimental intervention) versus another group (e.g., those receiving the control). The estimation of a

single summary HR is reliant upon the assumption that the ratio of hazards for each group is (at least approximately) constant over time, known more formally as the proportional hazards (PH) assumption.

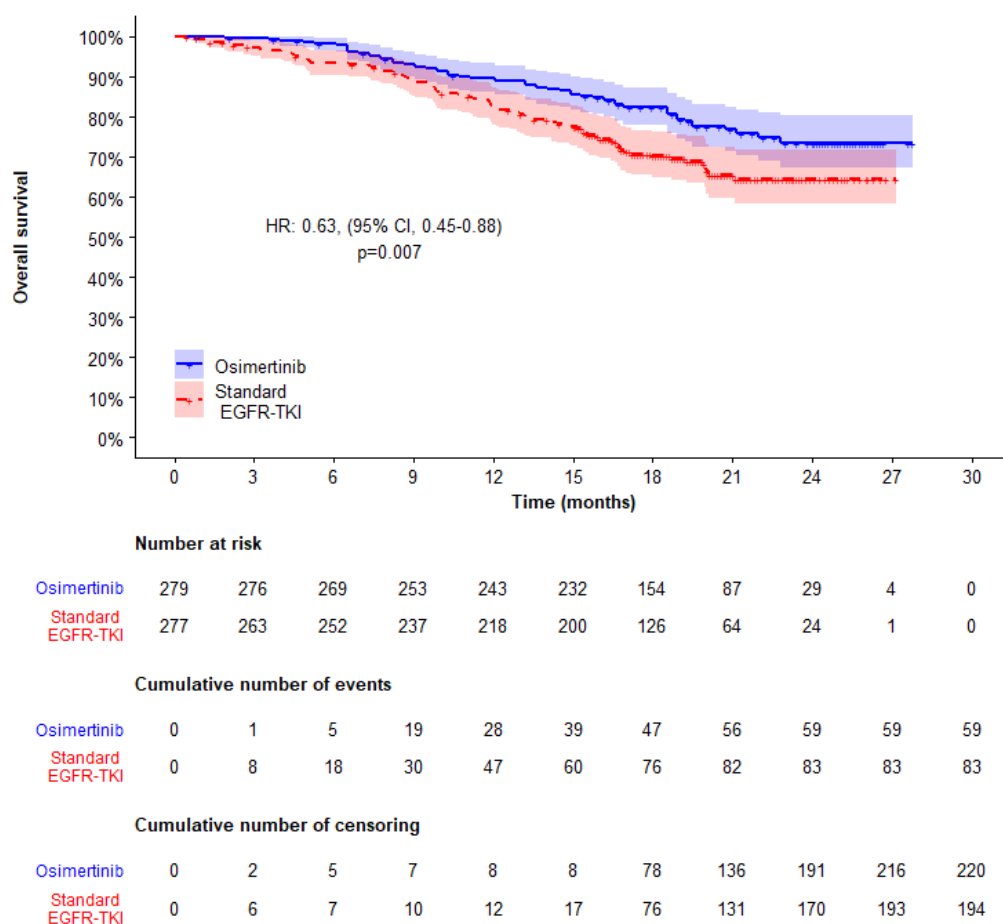
The potential issues with the use of HRs in the field of medical research are described by Miguel Hernán in his paper: "*The Hazard of Hazard Ratios*".⁴⁰ Hernan describes two main problems with the use of the HR. First, if the HR is not constant over time, then an estimated single summary HR may provide a misleading representation of the treatment effect. Second, if we consider period-specific HRs, these would have built-in selection bias because they would reflect differences in groups being compared contingent on earlier periods (e.g., to be eligible for an event in a second period, this means that susceptible patients must not have experienced the event in the first period). Despite these issues, HRs remain a commonly used means of measuring treatment effects in survival for cancer trials.

Numbers at risk and confidence intervals

The numbers at risk in Figure 1-2 also illustrate the increasing uncertainty in the KM estimator over time as patients either experience the event of interest or are censored. In addition, a confidence interval (CI) could be presented to demonstrate the uncertainty around the KM estimator. A survey conducted by Morris *et al.* suggested that both a number at risk table and a CI around the KM estimator are preferred means of visually presenting uncertainty. Without such features being presented alongside a KM estimator, it can be challenging to assess the inherent uncertainty in survival, and consequently the plausibility of different extrapolations.⁴¹

The 'true' CI limits around the FLAURA-1 KM estimates of OS are not known, as these were not provided in Figure 1-2. However, via a combination of digitisation software⁴² and an algorithm to re-create individual-patient data (IPD),⁴³ it is possible to re-evaluate the KM estimates of OS with reasonable certainty in this instance. The re-plotted KM estimates are shown in Figure 1-3 with estimated CI bounds, alongside the number of deaths and the number of censored observations. However, it should be noted that re-created KM estimates are imperfect, and without access to the raw data can only be estimated.

Figure 1-3: Re-created Kaplan-Meier estimator of overall survival from FLAURA-1



Key: CI, confidence interval; EGFR-TKI, epidermal growth factor receptor tyrosine kinase inhibitor (also referred to as SoC [standard of care] in NICE TA621); HR, hazard ratio.

Source: Figure re-created based on Figure 1-2.

As shown by Figure 1-2 and Figure 1-3, data from FLAURA-1 used to inform the company's submission in NICE TA621 were incomplete and subject to a high level of censoring. In September 2019, final results from the FLAURA-1 trial were announced, which included the following information regarding the endpoint of OS: *"Despite crossover** by one-fourth of the patients in the comparator arm, OS was prolonged with osimertinib versus the comparator. Median OS was 38.6 months (95.05% confidence interval [CI], 34.5-41.8) versus 31.8 months (95.05% CI, 26.6, 36.0), respectively (HR 0.799; 95.05% CI, 0.641-0.997; p = 0.0462)"* (European Society for Medical Oncology [ESMO] website).⁴⁴ These results highlight that while the HR met statistical significance, the CI limits overlap indicating uncertainty in the magnitude of benefit in terms of longer-term (in this case, median) survival.⁴⁴ Nevertheless, overlapping CIs for median survival ought not be conflated with evidence of no difference in survival –

** Here, crossover refers to patients in FLAURA-1 switching from the treatment they were assigned at randomisation (in this case, the SoC group), to the treatment received by the other group (osimertinib). Crossover is discussed further in Chapter 2.

median survival is one summary measure, and these CIs do not represent a formal test for a difference in median survival.

Survival analysis more broadly

It is important to note that the term ‘survival analysis’ may be used to describe any statistical method used to describe, examine, and/or interpret TTE data. These include the KM estimator (which aims to represent the survival function based on observed data), methods that focus on estimating relative effects between treatment arms in a clinical trial (such as a Cox PH model,⁴⁵ to obtain an estimate of the HR), and other methods which may be focused on estimating the ‘true’ underlying pattern of survival including the extrapolation of survival curves beyond the duration of trial follow up.

The commonality between all survival analysis methods is the ability to consider the dual nature of TTE data (i.e., the timing of the observation, and whether or not the event occurred). Despite the similarities, each specific method has its own prerequisite assumptions and limitations which may affect its ability to produce accurate results. That there are caveats associated with different analysis methods is not a unique issue to survival analysis, as such considerations apply across all statistical areas of study. For example, the choice of statistical model to describe bounded variables (e.g., between 0 and 1) may exclude certain distributions used to describe sampling variability from consideration. However, it is imperative that the nature of the data being analysed is acknowledged when undertaking survival analysis, such that appropriate methodologies may be considered.

When focussing on HTAs of cancer medicines, data collected as part of a trial are likely to be incomplete at the end of the observation period. In some situations, data may be considered sufficiently complete to permit HTA decision-making with little-to-no extrapolation of outcomes required. This may apply in cases where the intervention is associated with certain cost savings and a non-inferior safety and efficacy profile. However, in contemporary HTAs of cancer treatments this situation is seldom seen, especially acknowledging that new cancer drugs are generally recognised to have high acquisition costs (which are often justified by the manufacturers based on the costs of research and development).⁴⁶⁻⁴⁸ It may be the case that decision makers consider any additional costs are justified by the benefits seen during the trial follow-up period; but, even if this is true, downstream costs are likely to play a role in estimating cost-effectiveness results. Therefore, it is more often the case that OS needs to be estimated beyond follow up, in order to fully understand the magnitude of benefit provided.

For the purpose of producing estimates of lifetime survival from a clinical trial, parametric modelling methods are often used.¹⁰ Parametric methods make use of statistical models to describe the ‘true’ underlying sampling distribution of observed survival times. The estimation

of these parameters can then be used to produce an estimated survival curve, over a lifetime horizon, extending beyond the observed period of the clinical trial. The simplest example of a parametric model makes use of the exponential distribution, wherein the hazard of death is assumed to be constant over time. However, different parametric modelling approaches exist that may be considered, each of which impose different assumptions on the underlying hazard of the event of interest (e.g., death), resulting in different estimated survival curves^{††}.

A challenge often seen in HTAs of cancer treatments is when data available from the trial are based on limited numbers of events. For example, it will not be possible to estimate the mean or median via the KM if an insufficient number of events have been recorded at the time of analysis (the former of which is required to inform a cost-effectiveness analysis). As such, any extrapolations based on these data may be highly uncertain and could vary markedly depending on the properties of each survival model considered (which, based on limited data, may also be difficult to choose between).

Typically, as part of the NICE STA process, companies will submit a base-case analysis, ideally reflecting what it considers to be the most plausible estimate of cost-effectiveness. For STAs of cancer treatments, the base-case analysis usually includes the specification of a particular survival extrapolation to estimate survival, though this is subject to uncertainty. When faced with uncertainty, HTA decision makers consider the associated impact of various model settings and assumptions on cost-effectiveness results, including survival extrapolations (where relevant).

As part of a NICE STA, the following actions could follow based on the cost-effectiveness evidence presented:

- If either the base-case choice of a model, or an alternative, plausible survival extrapolation leads to an ICER above the range considered cost effective, this may lead to the intervention not being recommended based on the magnitude of uncertainty present.
 - Consequently, this may lead to sub-optimal decision-making, both from the perspective of the company and society (if the 'true' survival would lead to an ICER considered to be cost-effective).
- Conversely, should a decision maker be willing to recommend an intervention in spite of the uncertainty associated with survival extrapolation, this may commit the NHS to reimbursing an intervention that is not truly cost-effective (if, for example, the 'true' survival would lead to an ICER greater than the WTP threshold).

^{††} This is covered in detail in Chapter 2.

- This would be sub-optimal for society (based on the principle of opportunity cost), but not for the company (since it benefits from sales of its product).

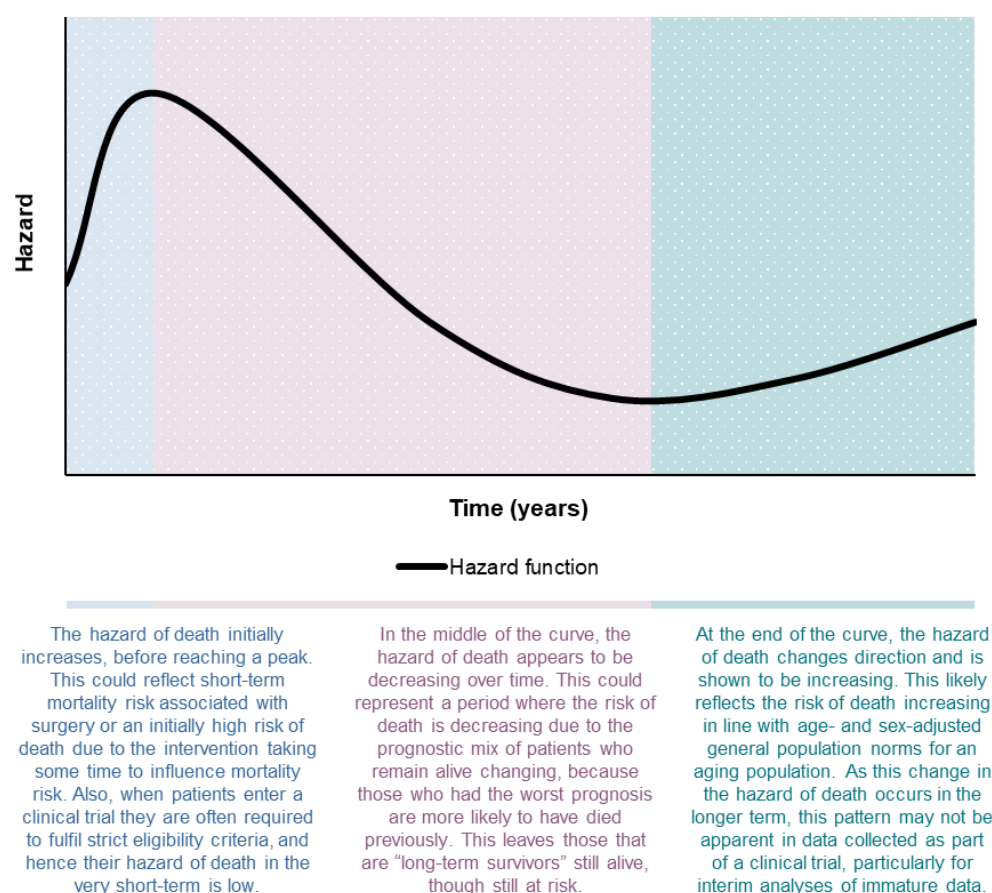
Due to uncertainty (including in survival analysis) some HTA bodies have introduced mechanisms by which a conditional recommendation may be made until further evidence is available to inform a final decision. In the case of NICE, the Cancer Drugs Fund (CDF) provides the NICE Appraisal Committee with the ability to recommend a potentially cost-effective oncology treatment while further evidence is collected.⁴⁹ The SMC allows for an interim decision to be made, subject to ongoing evaluation and future reassessment.⁵⁰ These mechanisms are designed to minimise the potential negative impacts associated with an ‘incorrect’ decision (i.e., the adoption of a treatment that is not cost effective, or *vice versa*).

Regardless of the specific HTA mechanisms that exist for each HTA body, a common requirement across many HTA bodies (as discussed previously) is for companies to submit a cost-effectiveness analysis, and for cancer medicines, this frequently involves estimating survival using some type of survival model. This is usually a parametric model (discussed further here, but in greater detail in Chapter 2). Although a range of alternative parametric models exist, the ability for models to produce realistic extrapolations of survival is contingent upon having sufficient data (or information) from which the model parameters may be robustly estimated. NICE Decision Support Unit (DSU) Technical Support Document (TSD) 14¹⁰ highlights six standard parametric survival models that are often used to inform company submissions^{‡‡}. However, only three of these (lognormal, log-logistic, and generalised gamma) are capable of reflecting *non-monotonic* hazards (i.e., a hazard function that includes at least one turning point, reflecting for example initially increasing then decreasing hazards).

To the point of changing hazards, nearly all HTAs involve the estimation of a survival curve over a lifetime horizon, either implicitly or explicitly, especially those concerning cancer interventions which predominantly affect the elderly – Cancer Research UK data suggest that approximately half of all cancer diagnoses in the period 2015-17 were in patients aged 70 or above.⁵¹ Two models that assume non-monotonic hazards (lognormal and log-logistic) are only capable of reflecting initially increasing hazards followed by decreasing hazards in the longer-term. Given that the hazard of death is expected to increase, *ceteris paribus*, as patients age, the likelihood of such models producing credible long-term estimates of the hazard of death for an elderly population is small. Figure 1-4 shows an example complex hazard function where a longer-term increase in the hazard is present.

^{‡‡} These are the exponential, generalised gamma, Gompertz, log-logistic, lognormal, and Weibull models. A more in-depth summary of survival analysis techniques is provided in Chapter 2 of the thesis.

Figure 1-4: Illustration of a complex hazard function



Source: Image based on a similar image presented in Figure 4 from TSD 21, by Rutherford *et al.*, (2021).⁵² A similar image is also presented in a methods primer by Latimer & Adler, (2022).⁵³

Three of the most commonly used models are capable of only considering either a constant hazard (exponential) or monotonically increasing or decreasing hazards (Gompertz and Weibull). Therefore, while long-term extrapolations may be seemingly credible, an initial reduction in the hazard of death associated with a given intervention may not be appropriately modelled by specifying a model that assumes constant or monotonic hazards. For example, a curative treatment may cause the short-term hazard of death to reduce markedly (compared with an untreated group), but in the longer-term the effects of aging will lead to an increase in the risk of death even if the curative treatment was successful. However, it could be argued that despite their simplicity, a model that assumes either a constant or monotonic hazard function could be sufficient for decision-making – particularly if the impact of choosing an alternative model on the ICER is small. For a Gompertz model specifically, a plateau can be seen in some cases where the effective hazard reduces to near zero, though this is unlikely to be plausible in the context of all-cause mortality.

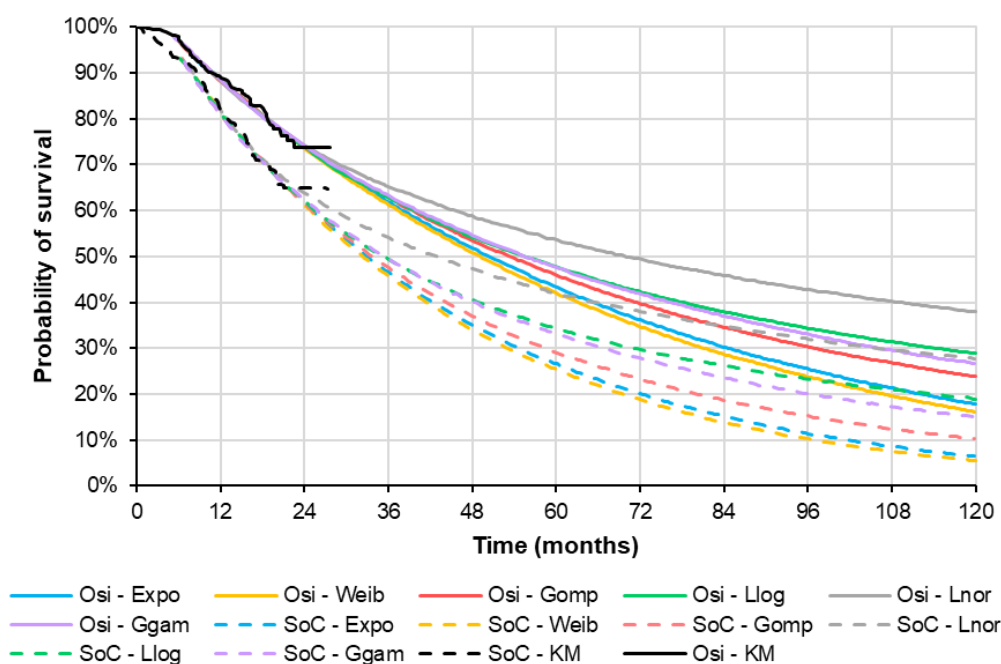
Consequently, the parametric models commonly used to estimate lifetime survival may or may not be sufficiently flexible to reflect the ‘true’ pattern of hazards. Moreover, it may be the case

that at the time of HTA, clinical trial data available to inform the estimation of a survival model may be especially limited in the longer-term (where the hazard of death may be expected to change). Therefore, even if a relatively flexible model is considered, it still may not be possible to reflect the ‘true’ pattern of hazards using the trial data alone.

The FLAURA-1 trial (revisited)

An additional consideration when considering a range of different parametric models is that each model can yield similar estimates of survival within the short-term (where the KM estimator may be produced from trial data), yet longer-term extrapolations (beyond the end of follow-up) can vary substantially. Based on the previous example of the FLAURA-1 trial used to inform NICE TA621, the extrapolations presented in the company’s submission for the outcome of OS are presented in Figure 1-5. It can be seen that depending on the functional form selected, the estimates of OS in the longer-term can vary greatly, with 10-year OS estimates varying between approximately 38% (lognormal model) and 17% (Weibull model) for the osimertinib arm.

Figure 1-5: Parametric models presented in NICE TA621



Key: Exp, exponential; Ggam, generalised gamma; Gomp, Gompertz; KM, Kaplan-Meier; Llog, log-logistic; Lnor, lognormal; Osi, osimertinib; SoC, standard of care; Weib, Weibull.

Source: Recreated from Figure 39 of NICE TA621 (osimertinib for untreated epidermal growth factor receptor mutation-positive non-small-cell lung cancer) company submission.³⁵ Image presented here with permission from AstraZeneca (the manufacturer of osimertinib). Copyright for the original image is retained by AstraZeneca.

Note: This image was re-created using WebPlotDigitizer Version 4.4, for the purpose of formatting.⁴² In the models shown above, a piecewise approach was used wherein the Kaplan-Meier estimator was used to inform estimates of overall survival up until 7.9 months, after which a fitted model was used to inform the remainder of the model time horizon. Piecewise models are discussed further in Chapter 2 of the thesis.

In the company's base-case analysis, the Weibull model was selected for both treatment arms. The model considered in the company's base-case analysis adopted a piecewise approach wherein the KM estimator was used to inform estimates of OS up until 7.9 months, after which a fitted model (in this case, a Weibull model) was used to inform the remainder of the model time horizon (fitted to data for patients that survived until at least 7.9 months). The base-case ICER was between £82,669 and £89,700 per QALY gained, depending on the choice of comparator (as multiple comparators were considered as part of this appraisal).³⁵ However, alternative models were also considered:

- Switching to an exponential model (though still adopting a piecewise approach) caused the ICER range to decrease slightly, to between £80,251 and £87,045 per QALY gained.³⁵
- Stepping away from a piecewise approach and considering a fully-parametric Weibull model without a cut-point caused the ICER range to increase markedly to between £93,388 and £124,833 per QALY gained.³⁵
- If a fully-parametric log-logistic approach was instead considered, the ICER range was between £102,422 to £111,395 per QALY gained.³⁵

These ICER ranges (specific to this particular appraisal) illustrate how influential different estimates of OS can be on the cost-effectiveness of the intervention, as different models could cause the ICER to decrease by around 3% (exponential, piecewise) or increase by as much as 39% (Weibull, non-piecewise), relative to the ICER associated with the model selected to inform the company's base-case analysis.³⁵ Though it may be noted that all of these ICERs are notably higher than the standard threshold of £20,000 to £30,000 per QALY gained, it is still important to estimate this accurately as the company may choose to submit a confidential discount as part of the NICE STA process.

It is also important to note that the ICER relies on the estimation of the *incremental* costs and QALYs – that is, the difference in costs and QALYs between the intervention and the comparator. Therefore, uncertainty present in both arms is important to take into consideration for decision-making. However, it is often the intervention group with the greatest amount of uncertainty for two main reasons: (i) it is usually a new treatment for which there is limited clinical knowledge of long-term outcomes, and (ii) it is usually associated with some kind of survival advantage, meaning that the KM estimate is less complete, versus the comparator (as can be seen in Figure 1-5 – the solid line for osimertinib sits above the dashed line for SoC).

Related to the other challenges highlighted above, survival extrapolation is based on the need to predict future outcomes. As such, there remains an unavoidable element of the unknown.

This is particularly important when considering HTAs of novel interventions with unique mechanistic properties, for which there may be limited information concerning the likely pattern of survival in the longer term.

1.3 External evidence

As described previously, there are a number of challenges that may arise when considering HTA based solely on data from the pivotal clinical trial(s), which in turn have implications for survival analysis (as well as other aspects of the decision problem under consideration). To address some of the limitations in a purely pivotal trial-based estimation of survival to inform HTA, other methods warrant consideration.

For simplicity, any evidence collected outside of the pivotal trial(s) of primary interest that may aid with survival estimation is considered ‘external evidence’ throughout the thesis. For now, the term ‘external evidence’ is intended to serve as an all-encompassing descriptor of both data and information obtained from outside of the pivotal trial(s) of primary interest. External evidence may take several forms, including (but not limited to):

- Data collected as part of another clinical trial (in the same or a similar indication).
- Observational data collected as part of a registry or other observational study.
- National statistics concerning general population mortality.
- Clinical expert opinion of survival for a given population receiving a given intervention.

The external evidence could theoretically be integrated within survival model estimation, used to aid in model selection, and/or validation of projections based on other evidence sources. Either approach (i.e., formal or informal [‘soft’] use of external evidence) may improve upon and/or reduce the uncertainty associated with purely trial-based extrapolations of survival. Moreover, external evidence could be sufficient to theoretically construct an estimate of survival without using the trial data at all – for example, by producing a synthetic control arm using external data, previously explored by Carrigan *et al.*, (2020).⁵⁴ Such an approach may avoid issues with clinical trial populations relating to their generalisability to a real-world clinical practice setting, with perhaps only the estimation of the treatment effect needing to be estimated from the clinical trial itself.

The use of external evidence to support and/or inform survival estimation has played a role in a number of previous HTAs, and in particular, NICE STAs. One example is TA319 (ipilimumab, melanoma),⁵⁵ in which longer-term registry data were used to inform the estimation of survival beyond 5 years where no direct evidence was available. Other examples of where external evidence has been used to inform OS estimates include the use of trial data from a related cohort combined with clinical opinion to guide extrapolations using immature data (TA517,⁵⁶ avelumab, Merkel cell carcinoma), and the use of background mortality rates adjusted by a

standardised mortality ratio (SMR) to account for ‘cured’ patients (TA559,⁵⁷ axicabtagene ciloleucel, B-cell lymphoma) – see Chapter 2 for more information on SMRs. External evidence has also been referenced to inform the expected pattern of hazards – for example, in TA658^{§§} (isatuximab, multiple myeloma [MM]),⁵⁸ the company cited data from a clinical study not considered in the scope of the STA to help inform the expected pattern of hazards (and consequently, inform model selection).

In the examples highlighted above, it may be inferred that external evidence may be applied in different forms to ultimately lead to the estimation of OS. For example, methods may focus on the estimation of the baseline survival curve itself (e.g., in the case of TA319,⁵⁵ where longer-term registry hazards were appended onto the end of the trial-based curve); or may instead focus on the estimation of relative survival effects (e.g., in the case of TA559,⁵⁷ where a ratio was applied to adjust longer-term survival projections). When considering these examples, it is important to acknowledge that while there is guidance available for the conduct of survival analysis for HTA in general^{10, 32, 52}, there is no established guidance regarding the use of external evidence in survival analysis specifically (please see Chapter 3 of the thesis for SLR findings illustrating this point further).

Beyond estimating an OS curve that addresses the limitations of clinical trial data, external evidence also presents a number of potential advantages over relying on trial data alone. For example, provision of external evidence may give decision makers more confidence when choosing the most appropriate model(s) to inform their decision-making. Such information, may essentially disqualify some models fitted to the trial data alone if considered implausible. Equally, the provision of external information may add justification to the selection of a given model(s). Furthermore, any clinical trial will consider a limited sample size, related to power calculations for the primary outcome (i.e., that the number of participants is sufficient to be reasonably sure that the result is not an artefact of the study involving too few participants). External evidence may allow for an increase in the number of data points available to inform estimates of survival (acknowledging implications related to randomisation and blinding; concepts which are covered later in the thesis – see Chapter 2).

Despite the potential advantages, inappropriate use of external evidence carries the potential to result in misleading or inaccurate projections of OS. An important consideration when faced with the decision to incorporate external evidence into survival estimation is to acknowledge that the trial-derived estimates of benefits of a new intervention are based on *efficacy*, which may or may not be considered a reasonably proxy for *effectiveness*. The differences between *efficacy* and *effectiveness* are described in Table 1-1, within the context of survival estimation.

^{§§} TA658 is currently under review (NICE [ID4067](#)).

Table 1-1: Comparison of efficacy and effectiveness

Item	Efficacy	Effectiveness
Description	The performance of a given health technology under 'ideal' and 'controlled' conditions, to assess if the health technology can produce an intended therapeutic result	The performance of a health technology in a 'real-world' setting, to measure the extent to which the health technology produces an intended therapeutic result
Examples	Registrational, Phase III, RCT	Observational cohort study
Advantages	• Able to control for various factors that may influence the estimation of outcomes and/or the effect(s) of a given health technology. • Usually possible to collect a large quantity of data, including information not routinely collected in routine practice (e.g., measures of patient-reported HRQL). • Able to consider a direct comparison of outcomes between two groups (e.g., patients given a new health technology versus an established standard of care).	• Assessments (e.g., scans for disease progression) and procedures (e.g., use of palliative radiotherapy) likely to align with routine clinical practice. • May be possible to make use of pre-existing data collection mechanisms (e.g., routinely collected data from patient charts). • Directly linked to the outcome measure of direct interest to an HTA agency (i.e., the performance of a given health technology in a 'real-world' setting).
Disadvantages	• Typically involves very specific inclusion/exclusion criteria to identify eligible patients. • Patient population may have limited applicability to a 'real-world' population, especially in the context of a global trial with patients recruited from many countries. • The design of an RCT may include various features which do not represent standard practice (e.g., ability for patients on the control arm to 'crossover' upon progression to receive the new health technology).	• Subject to various biases (e.g., selection and performance bias) as observational studies are often uncontrolled and no formal randomisation or blinding would be expected. • Usually not possible to undertake extensive 'real-world' studies until granting of the marketing authorisation for new health technologies, particularly related to longer-term outcomes. • May be difficult to identify large numbers of patients, particularly for rare diseases.

Key: HRQL, health-related quality of life; HTA, health technology assessment; RCT, randomised controlled trial.

It is perhaps important to consider that pragmatic trials are sometimes conducted – these are clinical trials which are designed to evaluate the *effectiveness* of an intervention, as it would be (or currently is) used in a real-world setting. Therefore, pragmatic trials may be viewed as a means of bridging what is often termed the 'efficacy-effectiveness gap'.⁵⁹ However, pragmatic trials are not typically conducted at a point in time to directly influence HTA processes.

The potential mismatch between the nature of evidence collected to inform regulatory approval and subsequently HTA (efficacy) versus the anticipated use of a given intervention in clinical practice (effectiveness) means that the combination of evidence sources requires careful consideration, as well as some determination of the applicability of either source to the decision problem of interest. Data from a randomised controlled trial (RCT) may be considered

suitable to estimate relative treatment effects in order to determine the efficacy of a given treatment, but these data are often then also assumed to be transportable across populations (including real-world populations considered as part of an HTA). The purpose of the RCT is ultimately not to estimate a baseline OS to populate an economic model, and there are a plethora of reasons why outcomes from an RCT might not be generalisable to a real-world population, some of which are noted in Table 1-1, and have been discussed elsewhere.^{60, 61}

There is also the need to acknowledge that extrapolations based on trial data in isolation may be sufficient for the purpose of HTA. For instance, it may be the case that the trial considers a relatively large sample of patients, was conducted recently, with a broadly generalisable population, and the data are mostly complete. The additional complexity of a modelling approach using external evidence in this scenario may be considered an unnecessary complication to achieve similar results, or even a 'red herring' which may lead to sub-optimal decision-making. For example, estimates incorporating external evidence may yield less optimistic projections due to the data considering a population of frailer (and less generalisable) patients, and thus result in an ICER higher than both the 'true' ICER and WTP threshold. Conversely, it may also be that the external evidence is too optimistic, and so the resultant ICER could be lower than the 'true' ICER.

1.4 Research questions

This research considers the following research questions (RQs):

- **Research question 1:** Which methods have been previously used to combine external evidence with trial data to make long-term projections of OS?
- **Research question 2:** How do these methods compare?
 - What assumptions and/or data are required?
 - How different are results associated with each method, and versus not using external evidence?
 - Which methods perform best?
- **Research question 3:** How should external evidence be incorporated into estimates of survival?
 - When should external evidence be used in general, and when should it not?
 - Which methods should be considered under specific circumstances?

The focus of my research is predominantly concerned with cancer interventions, for several distinct (yet related) reasons.

Firstly, as noted previously, cancer has been the most common subject of TAs considered by NICE.⁶² This may be (at least partly) because from 1 April 2016, all new cancer drugs and established cancer drugs with new licensed indications deemed to be significant are referred

automatically to NICE for appraisal.⁶³ The availability of a large number of published TA reports for a group of diseases with similar properties provides an evidence base from which to compare different analytical methods.

Secondly, related to the availability of data, NICE currently allows for cancer drugs to be recommended while funding is provided via the CDF provided a range of conditions are met. The CDF has undergone several changes in terms of how it operates since its original inception in 2011. For simplicity the CDF may be considered an alternative source of funding in England, separate from routine commissioning, and temporary in nature (in the sense that CDF funding is not provided indefinitely).^{49, 63} When appraising a cancer drug, the NICE Appraisal Committee has an additional option to consider: to recommend the drug to be made available via the CDF, subject to being plausibly cost-effective at the currently-proposed price though requiring further evidence. A data collection agreement is made to resolve major outstanding uncertainty in decision-making before the drug is later re-appraised.

The recommendation of a drug to be funded via the CDF is not necessarily related to uncertainty around the estimation of OS alone – further details on the role of managed access agreements as part of the CDF is discussed in Kang and Cairns (2023).⁶⁴ It is generally recognised that estimation of OS is a key driver of cost-effectiveness results, with associated uncertainty expected to be one of the main reasons why a CDF recommendation may be made. This further illustrates the relevance of cancer drugs within the context of using external evidence, as data collected (while funding is made available via the CDF) will likely inform the final decision reached by the NICE Appraisal Committee.

Lastly, outside of TAs conducted by NICE, a great deal of the published literature concerned with survival extrapolation for HTA focuses on cancer drugs, including NICE DSU guidance concerning survival extrapolation (TSD 14 and TSD 21).^{10, 52} More recently, TSD 26 was published concerning structured expert elicitation (SEE) for long-term survival outcomes, which includes a review of appraisals specifically of cancer interventions.⁶⁵ This is unsurprising given that cancer is a leading cause of death in the UK.⁶⁶ It is therefore expected that some methods identified may make use of external evidence sources that are specific to cancer (for example, data collected as part of established national registries such as the Systemic Anti-Cancer Therapy [SACT] dataset), for which equivalent sources may not exist for other disease areas. Given the number of assessments of cancer, should any disease specific methods have been developed, these are most likely to have been proposed in cancer, as opposed to other therapeutic areas.

It is important to note that while cancer is the main therapeutic area discussed throughout this thesis, it is anticipated that the findings of my research will have applications outside of

oncology. The ability to apply different methods in specific contexts then being subject to various caveats (such as data availability and the pattern of survival).

Another important restriction on the scope of this thesis is that any methods described or analysed as part of this research are intended to evaluate OS directly for a particular group of patients where the main source of data is from the pivotal clinical trial programme. Methods which evaluate OS-adjacent outcomes (e.g., recurrence-free survival) and/or relative effects (e.g., HRs) are considered beyond the scope of this research but may be described in brief and/or referred to where relevant. This includes methods where OS is estimated indirectly (such as via an assumed surrogacy relationship), or where OS is constructed via estimation of a treatment effect (i.e., indirect treatment comparison [ITC] methods, such as network meta-analysis [NMA] or population-adjusted indirect comparison [PAIC] – See Chapter 2, Section 2.3.4). This further restriction on the scope of the thesis was made to ensure that sufficiently similar methods can be applied, compared, and assessed for performance. This restriction is important given the primary aims of this research, and for practicality owing to the large volume of research undertaken in the field of survival analysis more generally. Later in the thesis, methods are compared to each other via a neutral comparison study (NCS), which requires that methods are aimed to estimating the same quantity (in this case, survival for a single cohort of patients). For further details of the NCS, please see Chapter 6.

The emphasis on OS is also closely related to the choice of a particular model structure frequently used to assess the cost effectiveness of life-extending cancer medicines – the partitioned-survival analysis (PartSA, discussed further in Chapter 2).^{67, 68} This model structure allows for the direct incorporation of OS estimates, rather than inferring OS through intermediate events or health state transitions (e.g., as would be the case in a Markov state-transition model). External evidence may play an important role in informing components of these alternative modelling approaches, for outcomes that are OS-like. For example, within a state-transition structure, external evidence may be helpful to characterise later clinical events such as post-surgery survival. These types of outcomes are fundamentally different to OS, and are likely to be highly context dependent and require specification of unique model structures, which would be challenging to compare within the context of this thesis.

Relatedly, the thesis focuses primarily on estimating survival for a single group of patients, as opposed to two or more groups of patients simultaneously. Many interventions undergoing HTA are evaluated within RCTs, which typically provide survival data for both the intervention and comparator arms. However, the comparator(s) used in the RCT may not align with the most relevant comparator for HTA decision-making, and treatment switching is often permitted (where patients initially randomised to the control arm later receive the intervention). Jointly modelling survival across multiple groups in such settings introduces additional

methodological challenges, including handling treatment switching, non-proportional hazards, subsequent therapies, and different mechanistic properties of treatments (and related expected differences in hazard trajectories). In HTA practice, it is also common to fit independent survival models to each trial arm, primarily to avoid assuming that proportional hazards hold.^{27, 69} Addressing these complexities was beyond the intended scope of the thesis, and so the primary focus was placed on single-group analyses. Despite this, Chapter 2 provides more information about clinical trial features, and Chapter 7 discusses key considerations and methodological approaches relevant to survival estimation in the context of RCTs.

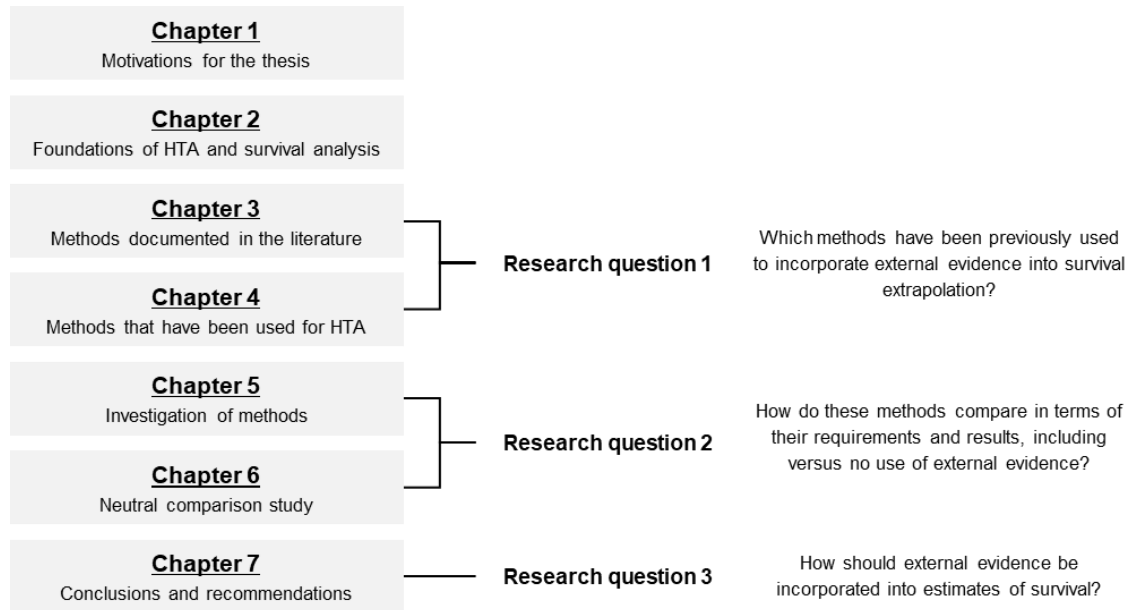
When considering the estimation of OS for a single treatment group, there is a rationale for expecting that the methods explored in this thesis may be more readily applicable to a comparator arm than the intervention arm. This reflects the general expectation that external evidence (particularly for longer-term OS) is more likely to be available for established treatments used in routine practice, rather than for novel interventions evaluated primarily (or perhaps only) in a clinical trial setting. Despite this, in some circumstances, estimates of long-term or conditional survival may be considered appropriate for both treatment groups, and there may be cases where longer-term data become available for the intervention arm specifically (e.g., after a treatment is made available via the CDF).

Addressing these research questions could usefully inform HTA decision-making. Gaining a better understanding of the available methods for survival extrapolation using external evidence, including their performance and appropriate contexts for use, would provide valuable insights for the HTA field. This, in turn, would support the selection and application of suitable methods, ultimately strengthening the evidence base that supports HTA decisions.

1.5 Thesis structure

A summary of the structure of this thesis is presented in Figure 1-6. The thesis is structured around the motivations behind the thesis, what research has been previously conducted, and the three key research questions described above.

Figure 1-6: Thesis structure and research questions addressed



Key: HTA, health technology assessment.

2. Foundations of producing lifetime survival estimates for HTA

Chapter synopsis

In this chapter, background information is provided concerning key terminology, different sources of evidence, and how survival modelling is typically carried out for HTA purposes. These concepts and terms are referred to extensively throughout the remainder of the thesis, including when describing the different methods that can be used to leverage external evidence for survival extrapolation, and are therefore important to define clearly.

The terms ‘survival’ and ‘hazard’ are used frequently throughout the course of the thesis, which are used to describe the mortality experience of individuals and groups of individuals being studied. While both of these terms are used in standard verbiage, explicit definitions of these terms with respect to statistical analysis are critical to understand the technical content of the thesis (including, for example, how different statistical methods translate inputs into survival outputs). In addition to these terms, language used to describe features of a clinical trial are described (for example, typical clinical trial endpoints, such as OS), along with terminology used to communicate and describe quantities that arise from a statistical analysis (estimates, estimators, and estimands).

Sources of evidence concerning the survival experience of a given population are described throughout the thesis. The term ‘evidence’ is used to refer to a body of facts or information which contribute to an overall understanding about whether or not something is true. This description covers both data (raw facts and figures, free of analysis or interpretation) and information (communication about how data have been understood or interpreted). The differences between ‘internal’ and ‘external’ sources of evidence are described, with reference to their role in informing the estimation of survival for HTA purposes, and different sources of evidence are also described. These include clinical trials, real-world and observational studies, disease registries, compassionate use/ expanded access programs, life tables, and clinical expert opinion.

Similarly in statistical analyses of survival for HTA, there are a range of established techniques and terms used. An introduction to the three different branches of survival analysis is provided (parametric versus semi-parametric versus non-parametric), the different types of survival models typically used in HTA, and a summary of how these models are typically produced (i.e., the ‘model fitting’ process). This includes importantly the most commonly seen parametric modelling approaches that are typically presented in company submissions as part of the NICE STA process.

2.1 Key terminology

This section covers key terminology used throughout the thesis. The terms discussed are well documented in several standard textbooks.^{32, 70, 71} However, key terms are summarised in this section for completeness.

2.1.1 Survival and hazards

The Britannica Dictionary defines ‘survival’ as “*the state or fact of continuing to live or exist especially in spite of difficult conditions*”.⁷² In this thesis, the term ‘survival’ is used to describe an individual (or group of individuals) that is (or are) still considered to be ‘at risk’ for a given event of interest to occur. In the case of OS, survival refers to any patient that is still alive at a given assessment time (or in other words, is still considered to be ‘at risk’ of death by virtue of being alive). However, the concept of survival can be extended to cover other TTE outcomes, described later in this chapter.

Given that TTE outcomes are expressly linked to a duration of time, it can be helpful to consider survival as a function of time, especially when survival is being discussed in the context of a population rather than an individual. The survival function evaluated at time t ($S(t)$) is defined as the probability of surviving beyond time t (from $t = 0$ [baseline]) (i.e., probability of surviving up until at least time t):

$$S(t) = \mathbb{P}(T > t)$$

In published literature, the term ‘survivor function’ is sometimes used in preference to ‘survival function’, as a means of acknowledging that $S(t)$ represents the outcomes associated with individuals at risk of the event(s) under consideration (i.e., it is about the survivors themselves).⁷³ However, ‘survival function’ is used throughout the thesis for consistency with the latest NICE DSU guidance concerning survival analysis: TSD 21 (where there are 107 mentions of ‘survival function’ versus 5 mentions of ‘survivor function’).⁵² The concept of survival will be referred to heavily throughout this thesis, though there are a number of related measures that are important to also consider which are defined and discussed in the subsections that follow.

One of the most important concepts linked to survival is the hazard of experiencing a given event of interest. As an example, let us consider the outcome of OS where there is one event of interest: death. If we are able to estimate $S(t)$ at two relevant time points (say, after 1 and 2 years of follow-up), we can then estimate a conditional survival probability as the probability of still being alive at 2 years *given that* a patient (or patients) was (or were) still alive at 1 year. If we reduce the difference between these times to increasingly shorter intervals, we can then estimate the hazard of death which is defined as the instantaneous failure rate, and like $S(t)$ we can consider the hazard of death as a function of time: $h(t)$ (or sometimes denoted $\lambda(t)$).

In some literature, $h(t)$ is referred to as the ‘force of mortality’, usually in studies involving the analysis of living organisms.⁷⁴ Owing to the fact that $h(t)$ refers to an *instantaneous* failure rate, it is defined according to the following limit formula:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}$$

Here, it is noted that the estimation of the hazard function, $h(t)$, is *conditional* upon survival until the timepoint under consideration, t . In other words, the hazard of death is estimated at a given time point only for those patients that are still considered to be ‘at risk’. This highlights an important consideration in the analysis of OS data – that individual patients can only experience the event of interest once. While seemingly obvious, this feature of TTE data has important implications for analysis.

The concept of $h(t)$ can be difficult to interpret because a ‘hazard’ is not observed, but instead hazards are *inferred* from inspection of observed survival data (while the same point could be said about survival at a population level [i.e., that it is not observed, *per se*], the concept of survival is arguably much simpler to understand because it refers to an event probability, rather than an instantaneous risk). Despite this, $h(t)$ provides a helpful way to describe the survival experience of a cohort over a given period of time.

There are a number of different presentations of hazard functions that can be used to further inspect survival data. A cumulative hazard function, typically denoted as $H(t)$ (or sometimes, $\Lambda(t)$), can be presented to illustrate the integral of the hazard function evaluated from time zero until time t (i.e., is the ‘accumulation’ of the hazard over time) which can help further identify patterns in survival data. Relatedly, the log-cumulative hazard function can be plotted against the log of time to assess whether for two groups hazards can be assumed to be proportional. Also, $h(t)$ describes the total hazard associated with a particular outcome, but in reality, this could be considered to comprise of a baseline (or expected) hazard experienced by a relatively ‘healthy’ group of people, combined with an excess hazard to capture the additional hazard of death associated with a particular condition (or conditions) under investigation. The concepts of ‘expected’ and ‘excess’ hazards are discussed further later in this chapter within the context of relative survival.

Returning to $h(t)$, in the most simplistic case, it may be that the hazard of death is approximately constant over time. A clear example of when this would be appropriate to assume is in the case of radioactive decay, first documented by Rutherford in 1900.⁷⁵ However, in the context of human survival, it is perhaps more likely that the hazard of death may change over time. As an illustrative example, if we consider a population of ‘healthy’ 45-year-old females, the hazard of death may consistently increase over time due to the effect of

aging on all-cause mortality. However, it may be the case that the hazard of death could continuously decrease over time (for example, in the case of a study where a curative intervention is provided). These hypothetical cases refer to a *monotonic* hazard function – that is, a hazard function which does not change direction, so that it is either consistently increasing or decreasing.

In reality however, hazard functions may be *non-monotonic* – i.e., there may be ‘turning points’ in $h(t)$ which contribute towards a more complex pattern of hazards over time. For example, the hazard of death may initially be increasing (e.g., immediately following surgery there may be a high risk of mortality due to complications), before reaching a peak and then decreasing in the longer term (e.g., reflecting a ‘cure-like’ plateau in survival for patients that survive the period immediately following surgery). However, in later years, the hazard is likely to increase again due to age-related mortality. Understanding what $h(t)$ is expected to be over a lifetime horizon is important to ensure estimates of survival are plausible, and therefore the hazard function is often a topic of debate in HTA decision making.

When producing estimates of survival using modelling techniques (discussed later in this chapter), it is often useful to consider the underlying hazard function, $h(t)$. Given that $h(t)$ represents the instantaneous risk of an event occurring, it is important for any representations of it to be ‘smooth’. Here, the term ‘smooth’ is used to describe a function which is *continuously differentiable*, meaning that there are no sudden breaks or gaps, or harsh jumps or drops. The presence of any discontinuities or abrupt changes in $h(t)$ can lead to implausible or unrealistic estimates of instantaneous risk, undermining the validity of the model. For example, if (hypothetically) there was a ‘jump’ in the hazard of death at a given time point (e.g., at 12 months), this suggests there is some external factor which means 12 months is some clinically relevant time point which triggers a difference in the underlying risk of death.

Nevertheless, in HTA, discontinuities or jumps in modelled estimates of $h(t)$ are sometimes reflected for practical purposes. An example of this is a model which caps the hazard of death at background mortality rates to ensure alignment with general population data – this can result in a jump in the hazard of death at annual time points, based on extracted estimates of hazards from published life tables. TSD 21 highlights that “*piecewise models may appear clinically unjustifiable and implausible, if sudden changes in hazards are modelled (i.e. with a discontinuity in the hazard function).*” (Rutherford *et al.*, 2020, p.35).⁵² Therefore, while discontinuities or jumps in $h(t)$ may be justified in specific contexts (primarily for practical reasons), they can introduce challenges for decision making in an HTA context – namely, the face validity of the modelled estimates.

2.1.2 Study endpoints and outcome measures

There are a number of different measures of survival-related outcomes that are typically collected in cancer studies, though these vary for a number of reasons, including the nature of the research question and the characteristics of the condition being studied. In clinical trials of cancer treatments, data for the outcome of OS are typically captured. The definition of OS is straightforward in most studies, where it typically represents the time from initiation of a given treatment until death or censoring. However, it may be the case that defining the start of treatment is less straightforward – for example, in the case of Chimeric Antigen Receptor (CAR) T-cell therapy, treatment could be defined as starting from infusion (when the product is administered to the patient), or it could instead be defined as starting from leukapheresis (when white blood cells are collected from the patient to create the CAR T-cell product).

Another commonly seen endpoint in clinical trials of cancer patients is progression-free survival (PFS), which refers to the time period over which patients are alive and without recorded disease progression. Progression itself is typically defined on a cancer-specific basis, but usually means that the cancer has in some way worsened, either in terms of tumour growth or spread of disease. Patients that remain ‘at risk’ for both disease progression and death are considered to be ‘progression-free’, and the first of these events to occur is then used to determine the PFS time. As PFS includes an additional event that can occur compared with OS (i.e., progression), PFS data are usually more complete compared with OS data, though in some studies measurement of progression may stop after a specific time period which can affect its interpretation.

In addition to PFS, there are a number of other measures that may be used where the concept of disease progression does not apply to a given group of patients. For example, in populations where disease progression is not well defined, event-free survival (EFS) may instead be used. For the outcome of EFS, a clinically relevant event is used in lieu of documented progression as a possible event of interest (for example, development of a secondary malignancy [a new cancer]). Alternatively, it may be the case that at baseline there is no evidence of disease from which to measure progression (e.g., populations that have undergone surgery with curative intent), and so a more relevant measure may be disease-free survival (DFS) – the period of time over which patients are alive with no evidence of disease recurrence.

Cause-specific hazards and relative survival

Returning to the outcome of OS, some studies may be interested in specific causes of death, which could be relevant to address specific research questions. In studies of population-based cancer registries, the outcomes of relative survival and net survival are sometimes used.⁷⁶

- *Relative survival*, $S_R(t)$, refers to the difference in survival experience between a population of patients with a given condition relative to a similarly defined ‘healthy’ population. More specifically, $S_R(t)$ is defined as a ratio between the observed survival of a group with a particular disease, $S_O(t)$, to that of the expected survival for the general population with comparable baseline characteristics, $S_P(t)$:

$$S_R(t) = \frac{S_O(t)}{S_P(t)}$$

By using relative survival, extrapolations of OS can be linked to the experience of an equivalent general population group, which can provide additional information on long-term outcomes.

- *Net survival* represents the probability of survival if all other causes of death were not applicable.

While similar, there are distinct reasons why one of these outcomes may be relevant to a particular research question over the other. Relative survival may help with understanding how the experience of a given group of patients relates to that of the general population (e.g., how survival of colorectal cancer patients in the UK compares with the general population), whereas net survival may help with understanding differences in survivorship for a given condition while accounting for differences in survival in the general population (e.g., how cancer-specific survival differs between the UK and the US).

Both relative and net survival can be estimated without specific cause of death information being recorded, however other measures could be considered if the cause of death was known. In this type of outcome (i.e., disease-specific survival), the measure of interest could be cancer-specific death where only deaths specifically due to cancer would be considered an event of interest, with deaths from all other causes being discarded in some way (either censored or removed from the analysis). However, cause of death information is known to be subject to limitations, including issues in determining cause and how it is recorded.^{77, 78}

Considering relative survival further, there are different ways of considering how the survival experience of the cohort of interest compares with the general population. For example, covariates of interest could be considered to act ‘multiplicatively’ or ‘additively’. In a *multiplicative* hazards model, the hazard function for a given individual i , $\lambda_i(t)$, may be represented as follows:

$$\lambda_i(t) = \lambda_0(t)e^{\beta^T x_i}$$

or alternatively:

$$\lambda_i(t|x) = \lambda_0(t)e^{\beta_1x_1+\beta_2x_2+\dots+\beta_nx_n}$$

Here, $\lambda_0(t)$ represents the underlying baseline hazard function (in the broader literature, the notation $h^*(t)$ is used for baseline hazard), and β represents a vector of n covariates with values x_i for a given patient, i . (Note: e is Euler's number – i.e., the base for natural logarithms, approximately 2.71828). In this model, covariates are assumed to have a multiplicative effect on the hazard (e.g., treatment assignment may double or halve the hazard of death), regardless of how $\lambda_0(t)$ changes over time, exactly how covariates are considered within a Cox PH model.^{*45}

An alternative to the multiplicative hazards model is the *additive* hazards model, represented as follows:

$$\lambda_i(t) = \lambda_0(t) + \beta^T x_i$$

or alternatively:

$$\lambda_i(t|x) = \lambda_0(t) + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n$$

In this model, at any point in time the overall hazard of death, $\lambda_i(t)$, comprises of a hazard function for the general population, $\lambda_0(t)$, plus a hazard function to capture the excess hazard of mortality, $\beta^T x_i$. The excess hazard may also be simplified in the literature to $\lambda_E(t)$. The additive hazards model (in general, not specific to relative survival analysis) was first proposed by Aalen.⁷⁹

If a relative survival modelling approach is used in the context of HTA, a survival model is fitted to patient-level survival data while incorporating external mortality rates, typically reflecting background mortality from national life tables stratified by age and sex. Background mortality rates are considered to be known, and so the model estimates the excess hazard associated with the disease under investigation. Therefore, compared to a model of OS, a relative survival model estimates the relative survival function, defined as the ratio of the observed survival in the patient cohort to the expected survival in a comparable general population, under the assumption of independence between the two components.

To produce an estimate of OS for use in a cost-effectiveness model, $S_O(t)$, the relative survival function, $S_R(t)$, needs to be combined with a general population survival function, $S_P(t)$, which can be achieved by rearranging the formula which defines relative survival:

* A short description of the Cox PH model is provided here for completeness. The Cox PH model is a frequently used survival analysis method for examining the association between survival time and predictor variable(s). It is semi-parametric, meaning that it does not assume a particular baseline hazard function. Instead, it assumes that the effect of covariates on the hazard is multiplicative and constant (with respect to time). The Cox PH model is not an extrapolation method – rather, it estimates hazard ratios (HRs) for covariates. For further details, refer to the citation included in the main body of the thesis.

$$S_O(t) = S_P(t)S_R(t)$$

This allows the extrapolated survival curve to reflect both background mortality and disease-specific excess mortality.

Other measures

For completeness, two other measures related to survival concerning prognosis given survivorship to a certain timepoint can be used: conditional survival and life expectancy (LE).

- **Conditional survival** is defined as the probability of surviving for a given period of time given that a patient has already survived up until a certain timepoint post baseline. For example, if at 1 and 2 years there are 80% and 60% of patients still alive, respectively; the conditional survival probability of living one additional year evaluated at 1 year would be 75% (i.e., 60% divided by 80%).
- **LE** is defined as the remaining estimated lifetime expected for a given individual, potentially conditional on surviving until a particular timepoint and/or having experienced specific events of interest. In some contexts, LE is defined as the estimated age at death. As an example, according to the Office for National Statistics (ONS), a 32-year-old male is expected to live until 85 years of age, meaning they would have a remaining LE of 53 years (as of 14 February 2025).⁸⁰

SMR

An SMR is sometimes referenced in HTAs with respect to estimating survival, particular in relation to the general population. An SMR is defined as the ratio between the number of *observed* deaths in a population compared to the number of *expected* deaths in a population. For example, if within a given time period, four deaths would be expected, but five were observed, the SMR would be calculated as $\frac{5}{4} = 1.25$.

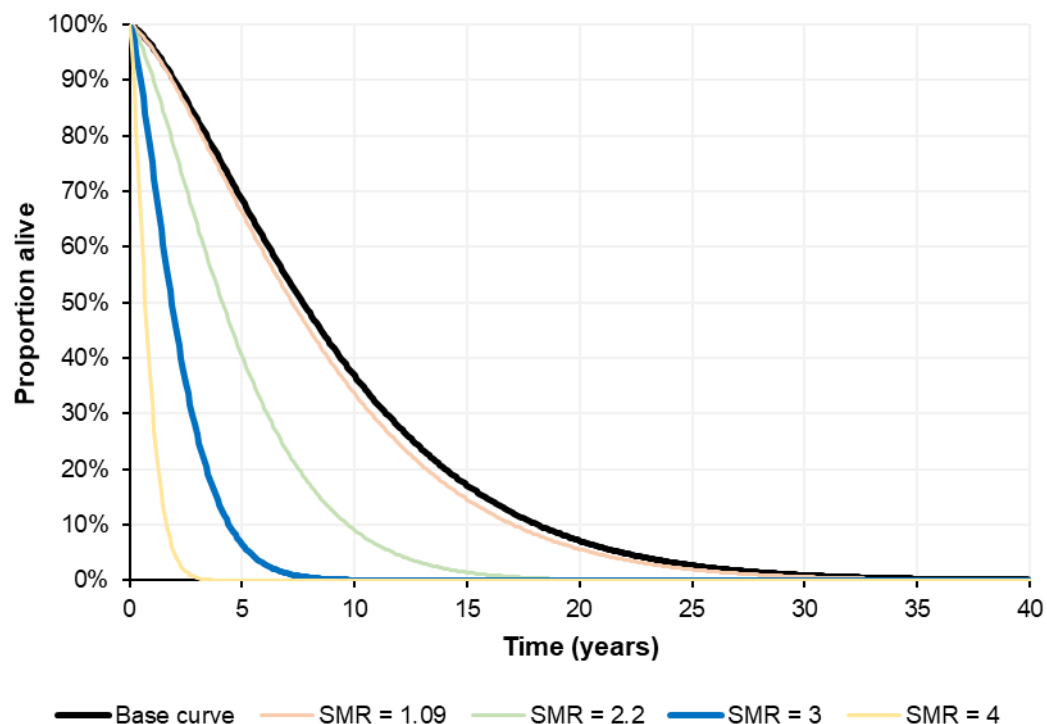
SMRs have been used in HTAs to adjust general population mortality estimates, accounting for residual elevated mortality risks in specific populations. This is usually in scenarios where patients, despite being considered 'cured' or 'in remission', continue to be experience higher mortality rates than the age- and sex-matched general population due to a combination of the impact of prior intensive treatments (which may have long-term effects, such as immunosuppression) or the presence of comorbidities. The SMR is typically applied as if it were an HR (though this is often not explicitly stated in real-world applications).

For example, in NICE TA893⁸¹ of brexucabtagene autoleucel for treating relapsed or refractory B-cell acute lymphoblastic leukaemia in people 26 years and over, the company initially

applied an SMR of 1.09 in its model to account for post-treatment mortality risk, based on data from a population with diffuse large B-cell lymphoma (DLBCL) published by Maurer *et al.*, (2014).⁸² The independent External Assessment Group (EAG) challenged this, and instead proposed an SMR of 4, based on a different study by Martin *et al.*, (2010). The company then proposed an alternative SMR of 2.2 (based on weighting populations by use of allogeneic stem cell transplant). Ultimately, the appraisal committee concluded that an SMR of 3 would be most appropriate to use. Further discussion on the NICE appraisal process, including the role of the EAG and the appraisal committee, is provided in Chapter 4.

The choice of SMR could have a large impact on estimates of survival, and consequently, cost-effectiveness results. Using a purely illustrative example, Figure 2-1 shows how estimates of mean survival (area-under-the-curve) are affected if an SMR is applied to a base curve with mean survival of 9.11 years, taking the range of values referenced in TA893. Depending on the SMR used, adjusted estimates of mean survival would range from 0.83 years (SMR = 4) to 8.57 years (SMR = 1.09).

Figure 2-1: Example of impact of standardised mortality ratios on survival



Key: SMR, standardised mortality ratio.

Note: The solid black line shows the base curve. In the context of TA893⁸¹, the blue line (SMR = 3) represents what the appraisal committee concluded was the most appropriate assumption for its decision making. The other curves illustrate other SMRs considered as part of this appraisal.

2.1.3 Estimates, estimators, and estimands

When considering survival-based measures and their role in HTA decision making, it is often necessary to be explicit about terminology to avoid confusion in interpretation. To assist, the following definitions relating to broader statistical analysis are often used:

- **Estimate:** An estimate is the value or outputs that are obtained from a statistical analysis using a given method and data set.
- **Estimator:** An estimator is the method that is used to obtain a given estimate.
- **Estimand:** An estimand is the quantity of interest to obtain from a statistical analysis, which the estimator is attempting to approximate.

In the context of survival analysis for HTA, focus is often placed on the sample truth (i.e., what is the long-term survival prospects of the trial cohort?). However, for decision-making, HTA decision makers are generally interested in the population truth (i.e., what are the survival outcomes for patients receiving a given intervention in practice, versus current care?). Estimates, estimators, and estimands feature throughout the thesis where necessary to distinguish between values, methods, and quantities.

2.2 Evidence sources

Throughout the thesis, a range of different sources of evidence concerning the estimation of survival for a given group of patients are described. Below, different types of evidence are described, including each of the different sources acknowledged throughout the thesis. For each source, the descriptions below are intended to briefly describe their main characteristics, what type of evidence, data, or information can usually be obtained from them (such as IPD from a trial, summary statistics for published sources, or KM estimates that could be digitised), and in what context(s) they are typically used in HTA (for example, estimation versus validation).

2.2.1 Data, information, and evidence

Researchers may learn about the survival experience of a given cohort of interest in a number of different ways, potentially using different methods of generation, with different implications for decision-makers. Data, information, and evidence are the three main terms used in HTA in this context, which are defined below. Throughout this thesis, the term 'evidence' is used in preference to 'data' or 'information' alone, to refer to all knowledge relevant to the research question (and not just data or information exclusively).

Data

Facts or figures about something, which are collected and reported without any attempt to explicitly interpret or describe them. Ultimately, data have to be processed in some way in

order for them to have meaning. For example, the set {1,2,3,4,5} represents data, but it is unclear without context what these data represent.

Information

Communication about how data have been understood or interpreted, which could include description, perspective, understanding, etc. Information provides context to data which has been collected (formally or informally), such that it can be communicated to an audience. For example, the set {1,2,3,4,5} may represent survival times for five patients measured in years, from which it could be inferred that median survival time was 3 years. In addition, as with most survival data, some observations may be censored, and so this added context is critical for the original data to be useful information.

Evidence

Body of facts or information which contribute to an overall understanding about whether or not something is true. Evidence encompasses both data and information, and represents the extent of knowledge gathered about something.

2.2.2 Internal and external evidence

As discussed in Chapter 1, this thesis focuses on estimating OS for HTA while taking into consideration external evidence, as well as data collected from a pivotal clinical trial (or trials). In this context, ‘external’ is used to describe any (relevant) evidence (i.e., data or information) collected outside of the main (pivotal) trial(s) that could help with estimating OS. By contrast, ‘internal’ refers to evidence collected as part of conducting the pivotal trial(s). For example, internal evidence could include data collected concerning patient survival times, comorbidities, demographics, etc. from the pivotal trial cohort(s). The sub-sections that follow provide examples of evidence sources that may be used in an HTA context.

Clinical trials

The National Institutes of Health (NIH) define a clinical trial as a *“research study in which one or more human subjects are prospectively assigned to one or more interventions ... to evaluate the effects of those interventions on health-related biomedical or [behavioural] outcomes”* (NIH, 2017).⁸³ For the purpose of this thesis, focus is placed on clinical trials conducted in populations with a particular type (or types) of cancer. There is some disagreement in the literature about the use of the term ‘trial’ to describe single-arm studies (i.e., studies without a control arm), but for the purpose of this thesis the NIH definition presented above is used.

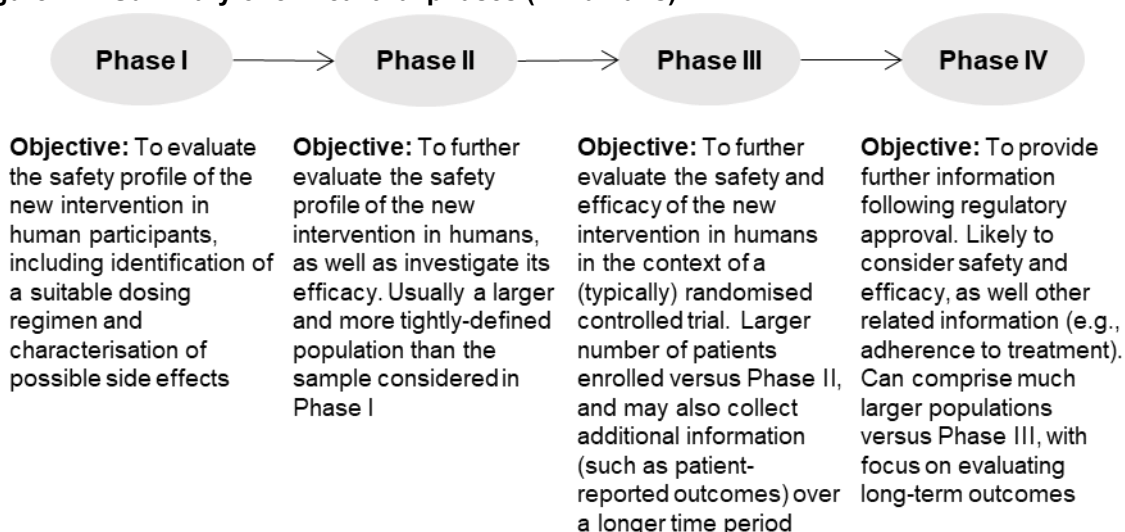
Clinical trials in cancer populations can be designed in many different ways depending on a range of considerations including, but by no means limited to:

- The overall objective(s) of the study (such as identifying an optimal dosing regimen, establishing whether or not a new intervention is superior to a given comparator, or considering the performance of a new intervention under specific circumstances).
- The rarity of the cancer(s) under study (and consequently the ability to recruit a sufficient number of patients).
- The risk-benefit ratio of exposure to an unlicensed intervention versus the current SoC (and how to balance this against the objective[s] of the study).
- The potential for confounding of results based on a range of factors (such as baseline patient characteristics and permitted procedures).
- The ethical implications of randomising patients to the current SoC (which may be particularly challenging for populations with no established active treatment options).

Generally speaking, at the time of HTA submissions, submitting companies will have conducted at least one clinical trial of the new intervention to establish its safety and efficacy. This is usually (but not always) in the context of an RCT in which the new intervention is typically compared to either the current SoC at the time the study was designed, an older treatment considered to be a suitable comparator, or placebo.

As part of the clinical development program for a new intervention, a number of studies are typically conducted, organised in different phases. Figure 2-2 summarises the different phases of clinical trials typically conducted throughout the development of a new intervention; however, in the context of HTA, submissions are typically based on data collected as part of Phase II or III clinical trials.

Figure 2-2: Summary of clinical trial phases (in humans)



In the context of HTA, the primary source of information available concerning the safety and efficacy of a new intervention typically comes from a clinical trial (or multiple trials) sponsored by the company (or conducted by the original developer of the new intervention prior to some form of commercialisation agreement). Consequently, at the time of an HTA submission, the company typically has access to the anonymised IPD from the aforementioned clinical trial(s) (often described as the pivotal trial[s] of the new intervention).

Furthermore, clinical trial data usually comprise the only data available at the time of HTA submission concerning outcomes for patients treated with the new intervention (given that at the time of submission, it is most often the case that the treatment is not available in routine clinical practice within the specific indication under assessment, and so access to treatment is most likely limited to the use of the intervention as part of a clinical trial or a compassionate use/ expanded access program [described further in a separate sub-section that follows]).

In the context of an RCT, there are two concepts of particular relevance to this thesis that differentiate it from other sources of evidence – randomisation and blinding:

- **Randomisation:** In a randomised trial, patients are assigned to each group based on a random allocation process, uninfluenced by patient or healthcare provider input.⁸⁴ This process helps ensure that differences between groups are due to chance rather than specific factors that were not accounted for (such as baseline demographics like age, disease-specific factors like stage or severity of disease, or the presence of comorbidities), thus reducing the risk of confounding.
- **Blinding:** In a single-blinded trial, study participants (i.e., patients) do not know which treatment group they are assigned to.⁸⁵ Double-blinded trials extend this lack of knowledge to the healthcare providers and other researchers involved in the trial, and triple-blinding extends this to include those involved with analysing the data collected from the trial. Blinding can reduce the risk of other forms of bias associated with awareness of assignment in a trial (for example, if a patient knew they were receiving the control drug, they may be more likely to drop out of the study).

It should be noted that evidence from other clinical trials (i.e., not the pivotal trial[s]) could be considered relevant external evidence for the purpose of informing HTA. In the most simplistic case, a clinical trial undertaken in a similar population for the control arm could serve as validation for extrapolations if longer-term follow-up data are available. As such, clinical trials may provide a combination of internal and external evidence relevant to this thesis.

Observational studies and disease registries

Outside a clinical trial setting, data may be collected through a number of other means. These include observational studies and disease registries (as well as any other type of database

wherein relevant survival data are collected). Observational studies are non-interventional studies (i.e., those where the investigators do not attempt to influence exposure to treatment or any other factors), which may provide external survival data (most likely for a control arm). Disease registries are databases set up to collect routinely data in a specific disease area (e.g., the Surveillance, Epidemiology, and End Results [SEER] Program in the United States).⁸⁶ As with observational studies, data from registries do not reflect a specific clinical trial design, but may involve some investigator involvement (i.e., may not strictly be deemed 'observational data').

Estimating treatment effects using these types of data sources presents several challenges. Unlike RCTs, observational study and disease registry data are subject to biases such as confounding, where differences in patient characteristics between treatment groups (such as the presence of comorbidities) may influence outcomes. Adjusting for differences through statistical methods (e.g., using propensity score-based methods), could mitigate some of these biases, but residual confounding may remain. In addition, these data sources often lack the detailed, standardised protocols developed for RCTs. Variability in inclusion criteria, data collection methods, follow-up schedules, and outcome definitions can make it difficult to compare results across sources or draw robust conclusions about treatment effects.

Compassionate use and expanded access programs

Some companies may enable patients to have early access to a drug undergoing development prior to it receiving a marketing authorisation through a compassionate use or expanded access programme. The FDA defines these programs as *“a potential pathway for a patient with a serious or immediately life-threatening disease or condition to gain access to an investigational medical product (drug, biologic, or medical device) for treatment outside of clinical trials when no comparable or satisfactory alternative therapy options are available”*.⁸⁷ Through providing early access to promising treatments, companies may collect data that could potentially be used to inform estimates of survival for HTA. However, data collected in this manner are understandably subjected to strict regulations and there are limitations around data collection (e.g., it is unlikely to be feasible or even permissible to collect the full set of data as would be considered standard in the context of a clinical trial).

Life tables (national population statistics)

A life table is a dataset (typically presented in tabular format) which provides estimates of survival (expressed as life expectancy, survival probabilities, or other related measures) separated by age (usually in yearly increments) and sex. These data can be used to infer what the survival experience of a comparably 'healthy' population could be (acknowledging that not all members of the general population are in 'perfect health').

Life tables from the UK are routinely published by ONS, and internationally can be accessed via the Human Mortality Database (HMD).^{88, 89} One of the key advantages of life tables is that these data are publicly available, and could be considered in a variety of contexts. However, by definition, these data reflect the survival experience of the general population, as opposed to the experience of a specific group of patients by disease area or treatment used.

Expert opinion

Another source of evidence for survival modelling comes in the form of expert opinion. In brief, expert opinion refers to any information about the survival experience of a given cohort of patients that is obtained or elicited from an individual (or group of individuals) considered to have relevant expertise in the treatment of similar patients in a clinical setting. In this sense, expert opinion refers to external *information*, as opposed to external *data*, since it refers to the perspective of clinical experts, rather than a presentation of raw facts and figures. Various methods exist to elicit the opinion of clinical experts, though most notably in the context of HTA submissions is the Sheffield Elicitation Framework (SHELF)⁹⁰, which has been used to inform previous survival analysis research.⁹¹ More recently, TSD 26 was published which lays out the principles for conducting SEE for long-term survival outcomes.⁶⁵ Although this guidance emphasises the role of SEE primarily in the validation and assessment of model-based extrapolations, it is still concerned with establishing clinically plausible survival projections and represents a mechanism through which expert opinion can be captured as part of the HTA decision-making process.

2.3 Survival modelling

In this thesis, focus is placed on estimating (using mathematical techniques) the survival experience of a given cohort of patients of interest. This overall process of estimation is termed 'survival modelling', where a mathematical model is used in the context of survival data.

Throughout this thesis, a survival model is defined as the culmination of any mathematical methods used to produce an estimate of survival for a given cohort. This does not necessarily mean that the survival model must be produced in one single step (e.g., via a regression model). Indeed, it is considered relatively commonplace for cost-effectiveness models to include *post-hoc* adjustments to survival extrapolations to account for background mortality (as evidenced through a review of cancer appraisals conducted by NICE, discussed further in Chapter 4). As such, 'survival model' could theoretically represent a combination of different statistical components and/or adjustments to ultimately yield estimates of OS for use in HTA; however, the term 'model' is often used to describe component statistical models, as well as the cost-effectiveness model used to inform an HTA submission. Therefore, it is important to note that the word 'model' is used to refer to several different (though related) topics in HTA.

The remainder of this sub-section describes different types of survival modelling methods, those commonly used to inform HTA submission specifically, and how these models are estimated.

2.3.1 Parametric, semi-parametric, and non-parametric

There are three broad types of survival models that are typically used in survival analysis, which are described in turn in the sub-sections that follow.

Non-parametric models

Non-parametric models do not assume a specific parametric form for $h(t)$ or $S(t)$. Instead, these types of model estimate these functions more flexibly, and can perhaps be thought of as being indexed by an infinite number of parameters, offering greater flexibility at the cost of less parsimony (i.e., simplicity in model specification). While they avoid strict assumptions like those made in parametric models (discussed in the next sub-section), they still make assumptions. The most common example of a non-parametric survival model is the KM estimator³³ (see Figure 1-2 for an example). In the KM estimator, $h(t)$ is assumed to change in a stepwise fashion over time. This is clear from its derivation, as the KM estimator evaluated at time t , $\hat{S}(t)$, is defined based on the following product:

$$\hat{S}(t) = \prod_{i: t_i \leq t} \left(1 - \frac{\text{events}_i}{\text{number at risk}_i}\right)$$

Here, events_i refers to the total number of events (e.g., deaths) that were recorded at time t_i , and number at risk_i refers to the total number of patients that were still considered to be at risk for the event of interest at time t_i . Taking an elementary example, if there are three patients with recorded (uncensored) survival times of 3, 6, and 9 months, respectively, the KM estimate of survival would be estimated as shown in Figure 2-3.

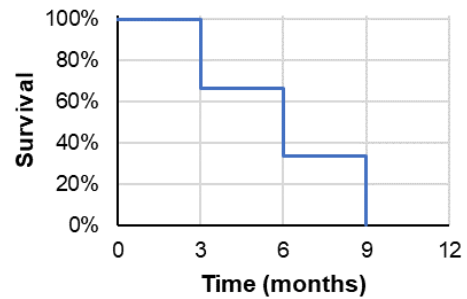
Figure 2-3: Example Kaplan-Meier estimation

$$t < 3 \text{ months} : \hat{S}(t) = 100\%$$

$$3 \leq t < 6 \text{ months} : \hat{S}(t) = 100\% * \left(1 - \frac{1}{3}\right) \approx 66.7\%$$

$$6 \leq t < 9 \text{ months} : \hat{S}(t) = 66.7\% * \left(1 - \frac{1}{2}\right) \approx 33.3\%$$

$$t \geq 9 \text{ months} : \hat{S}(t) = 33.3\% * \left(1 - \frac{1}{1}\right) = 0\%$$



Key: KM, Kaplan-Meier.

Semi-parametric

In a semi-parametric model, the form (or pattern) of the hazard function is not assumed to be represented by some parametric function, but some type of structural assumption(s) is (or are) made concerning the relationship between explanatory variables (such as treatment assignment) and the hazard. The most popular semi-parametric model cited in the literature is a Cox PH model⁴⁵.

The Cox PH model estimates the hazard of an event occurring at time t , i.e., $h(t)$, given a set of n explanatory variables (or covariates), defined as a $1 \times n$ vector: x . In other words, the Cox PH model estimates $h(t|x)$. The model assumes that the hazard function for an individual is the product of an unspecified baseline hazard function $h_0(t)$ and a function of n covariates, typically expressed as follows:

$$h(t|x) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}$$

Here, β represents a vector of coefficients, which can be interpreted as log HRs associated with the covariates. The key assumption made by the Cox PH model is that hazards are proportional between groups determined by explanatory variables, but the model makes no assumption about the underlying pattern of hazards.

Semi-parametric models, such as the Cox PH model, are typically used to describe differences in survival between groups. For example, using a Cox PH model an estimate of the HR for overall survival associated with a new drug versus standard of care could be estimated (see Figure 1-2 for the estimated HR in the FLAURA-1 trial). However, as described above, the baseline hazard function as part of a Cox PH model (i.e., $h_0(t)$) is unspecified – this means that without additional assumptions, it is not possible to obtain estimates of absolute survival from this type of model, which is necessary for HTA.

It should be noted that in HTA practice, the term ‘semi-parametric’ is sometimes used to describe a combination of a non-parametric and parametric model – e.g., a KM estimator evaluated until a given cut-point, followed by a parametric model. This generally is not a preferred descriptor, since ‘semi-parametric’ carries a specific meaning distinct from combinations of models. If using a KM estimator followed by a parametric model, this would be better described as a piecewise model (reflecting a model comprising two components), and is discussed further in a sub-section below.

Parametric

In a parametric survival model, the distribution of survival times is specified by assuming a particular functional form for the probability distribution of time to event. This begins by defining the survival time, T , as a random variable drawn from a known distribution – for example, a

Weibull distribution (commonly used in survival analysis), with scale parameter λ , and shape parameter k . This means that survival times are assumed to be sampled from the distribution $T \sim \text{Weibull}(\lambda, k)$. The assumed distribution (in this case, Weibull), means that there is a specific form for the survival function, $S(t)$, as well as related functions (such as $H(t)$).

These functions are all determined by a set of model parameters, and (where applicable) their relationship with explanatory variables is specified through a regression structure. In the case of a Weibull model, explanatory variables could be specified to impact the scale parameter λ , the shape parameter k , or both parameters. Because all aspects of the model are fully determined by this distributional assumption, parametric models can be used both to estimate survival within the observed data range and to extrapolate beyond it.

Parametric models have a number of advantages over semi-parametric and non-parametric models. Firstly, by making a parametric assumption, it is possible to extrapolate estimates of survival beyond the duration of available follow-up data. In addition, by assuming a parametric form, any undesired bumps/kinks, or flatlines/plateaus in a non-parametric model (such as the KM estimator) can be ‘smoothed’ to produce more realistic estimates of survival for a theoretically infinitely-sized patient population. Furthermore, through explicitly defining $h(t)$, it is possible to interrogate how the pattern of hazards change over time, and how treatment effects affect $h(t)$ (for example, does the treatment cause a ‘shift’ in $h(t)$, or does treatment only affect $h(t)$ in the short-term)

Each type of parametric model makes important assumptions as part of its specification, and therefore are associated with their own advantages and disadvantages. The goal of this chapter (and indeed the thesis as a whole) is not to investigate different parametric model specifications, and therefore a detailed account of all possible parametric model specifications is not provided here. However, a brief introduction to commonly used models in HTA is provided in the next section.

2.3.2 Models used in HTA

There are six standard parametric models that are commonly used to inform estimates of OS in HTA decision-making, which are described in detail within TSD 14 (first published in 2011, and later updated in 2013).¹⁰ These six ‘standard’ parametric models are:

- Exponential
- Generalised gamma
- Gompertz
- Log-logistic (or Fisk)
- Lognormal (or Log-Gaussian)
- Weibull

An additional model is sometimes also included as part of the suite of ‘standard’ parametric models used in HTA: the gamma model, which is a two-parameter special case of the

generalised gamma model. A helpful summary of the characteristics of these distributions is presented in a study by Kearns *et al.*, (2020), which is re-produced in Appendix A. There are other parametric models that could also be considered, but are less commonly seen in HTA submissions and published economic evaluations and hence are not discussed further at great length. These include:

- Generalised F; a four-parameter extension of the generalised gamma model, but requires integration to evaluate $S(t)$ at each timepoint, hence not used commonly in predominantly Excel-based economic models*.
- Rayleigh; a special case of the Weibull model with a fixed shape parameter, k , of 2, and thus seldom used as that the Weibull model offers a more flexible option.
- Extreme value; linked to Weibull, but is not bounded for survival times $t > 0$, so can produce unrealistic estimates of $S(t)$.

More recent HTAs have considered a variety of ‘advanced’ extrapolation methods. These methods are described in detail within TSD 21 (first published in 2020, and later updated in 2022).⁵² An introduction to these methods is provided in the sub-sections that follow, along with relevant literature containing additional information where available.

Piecewise models

Piecewise models consider different ‘sub-models’ between specific time points over the entire time horizon to reflect the expectation that trends differ between these time points. The full piecewise model therefore comprises the combination of each sub-model. The piecewise components could include non-parametric (e.g., KM estimates) and parametric (e.g., Weibull model) components. These models are sometimes referred to as being ‘semi-parametric’ piecewise models, when a combination of a KM estimate and a parametric tail is combined (i.e., a non-parametric KM + a parametric model = a semi-parametric piecewise model), though this terminology should not be conflated with the traditional definition of a semi-parametric model (see Chapter 2 for formal definitions).

Landmark models

Landmark models may be considered a specific type of piecewise model, in which a specific time point is selected where patients can be separated into different groups according to some clinical measure. This time point is called the ‘landmark’, and different survival extrapolations are used for the different groups for the remainder of the time horizon being considered (i.e., from the landmark time point onwards).

* Software typically used to develop cost-effectiveness models is discussed further in Chapter 2.

Landmark models require selection of a time point (e.g., 8 weeks), a clinical measure (e.g., response), and a means of separating groups of patients according to the chosen clinical measure (e.g., complete response versus all other response categories).

Mixture models

Mixture models refer to a group of models in which the overall model comprises of additive components that are weighted according to the proportion of individuals that belong within each 'mixture'. Some mixture models estimate the probability that each individual belongs to each group (by assigning each individual a value between 0 and 1 denoting their individual probability of being in one group), whereas other types of mixture models assume the population is made up of at least two distinct groups but without specifying individual probabilities of group membership.

Neither approach assigns individuals to a specific mixture based on known information (e.g., patients with no prior therapies assigned to one part of the model, and those with some prior therapy used assigned to another part of the model). The approach is a similar concept to a latent class analysis (i.e., an analysis in which group membership is not determined on the basis of observed groupings, but instead is inferred through other variables that are observed).

While not used often in HTA, a latent class analysis was used in a previous HTA for treatments in rheumatoid arthritis, in the context of an NMA to produce estimates of relative effects for different thresholds of response.⁹²

Spline-based and fractional polynomial models

Spline-based and fractional polynomial (FP) models represent a group of flexible parametric models.

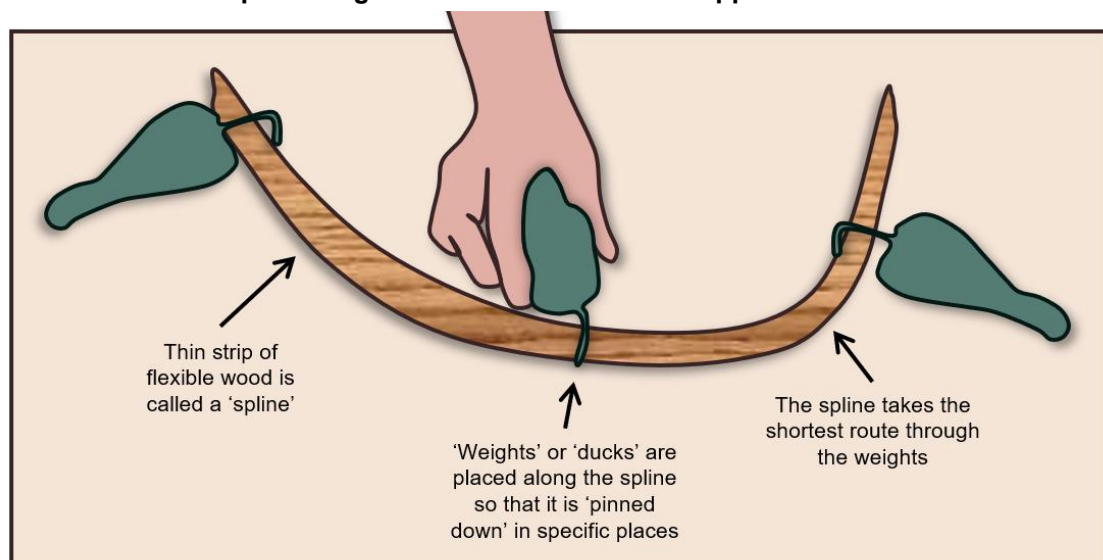
A spline function is a piecewise polynomial (an expression comprising of multiple terms), though traditionally a 'spline' referred to a long flexible strip of wood or rubber, used for drawing curves, held in place by lead weights (or 'knots' when described in the context of statistical analysis), and have been used in a variety of disciplines (e.g., boat building).⁹³ Figure 2-4 shows an example of the lead weights that have been used historically for drawing smooth curves, and Figure 2-5 shows a diagram of how these would be used.

Figure 2-4: A lead spline weight



Source: Original photograph taken by author.

Figure 2-5: How lead spline weights are used in real-world applications



Source: Original diagram produced by author.

The most common spline-based model used in survival analysis for HTA is the restricted cubic spline model, which are often termed 'Royston & Parmar' spline models, based on the authors' original paper concerning the use of these models in survival analysis in 2002.⁹⁴

- 'Restricted' relates to the fact that splines (particularly cubic splines) can be 'poorly behaved' at the two tails (before the first knot and after the last knot), and so in this type of spline model the model is constrained (or restricted) to be linear after the final knot.⁹⁵ The 'restricted' label therefore only refers to the model outside the boundary knots – inside the boundary knots, the model is cubic.

- ‘Cubic’ applies to the fact that in reality, splines that make use of polynomials of order 3 (e.g., $x^3 + 4x^2 - 5x + 7$) are predominantly used as an order of 3 is the simplest polynomial that allows an inflection point (i.e., a point where the second derivative changes direction or sign). Polynomials of lower degree, such quadratic, cannot model inflection points.

The standard application of a restricted cubic spline model is to represent the log-cumulative hazard, which can be considered an extension of a standard Weibull model. A standard Weibull model, with scale λ , and shape k , may be represented as follows:

$$\log H(t | x) = \log(\lambda) + k(\log t) + x^T \beta$$

The formula for the more flexible spline model is as follows:

$$\log H(t | x) = \gamma_0 + \gamma_1 z_1(\log t) + \gamma_2 z_2(\log t) + \dots + \gamma_n z_n(\log t) + x^T \beta$$

Comparing these formulae, it can be seen that the Weibull model is a special case of this spline model, where $\gamma_0 = \log(\lambda)$, $\gamma_1 = k$, $z_1 = 1$, and all other values of γ_i and z_i are set to 0.

It should also be noted that in addition to restricted-cubic spline models, there are other types of spline-based model that could be considered in the context of survival analysis, such as M-splines.⁹⁶

An FP is a polynomial which makes use of non-integer – or fractional – powers (e.g., $\frac{1}{2}$). FP models have received relatively less attention in the HTA field for producing survival models for a single treatment arm but have received substantially more attention in the context of generating ITCs.⁹⁷⁻⁹⁹ Spline-based and FP models are different in concept, yet both aim to increase flexibility in survival modelling compared to ‘standard’ parametric models (e.g., the Weibull model).

In FP models, the natural logarithm of the estimated hazard of the event of interest (e.g., death) is modelled using a polynomial function. FP models are fitted based on a selected order (usually first- or second-order models are considered), within a set of potential powers, often taking values from a restricted set originally proposed by Royston and Altman, (1994): $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$.¹⁰⁰ Of note, FP models are used more frequently in the context of estimating time-varying treatment effects within an evidence synthesis, as opposed to producing lifetime survival estimates for a single group, for example per the approach suggested in the frequently-cited work of Jansen (2011).⁹⁷

Flexible parametric models, including spline-based and FP models, are discussed further in TSD 21.⁵² However, for further details on restricted cubic spline-based models specifically, please refer to the original work of Royston and Parmar (2002).⁹⁴

Cure-based models

Cure-based models cover a range of modelling approaches wherein some patients are assumed to not experience the event of interest. Historically, cure models have been used in a variety of applications, including employee tenure, wild salmon monitoring, and credit scoring.¹⁰¹⁻¹⁰³ There are a broad range of different types of modelling methods that consider the concept of 'cure', but in HTA-related literature there are two common types of cure-based model that have been considered:

- **Mixture-cure models (MCMs):** A parametric MCM, in its most basic form, is defined as follows:

$$S(t) = \pi + (1 - \pi)S_U(t)$$

Here, π represents the proportion of patients that are 'cured' (i.e., do not experience the event of interest), which have a survival estimate equivalent to 100% forever, and $S_U(t)$ represents the survival function for the 'uncured' proportion of patients specifically. The overall survival function, $S(t)$, therefore represents a weighted average (i.e., a 'mixture') of these two groups.

- **Non-mixture cure models (NMCMs):** A parametric NMCM, in its most basic form, is defined as follows (using the same notation as per the MCM):

$$S(t) = \pi^{S_U(t)}$$

The fundamental difference between the MCM and the NMCM is that the MCM considers the population as a mixture of two groups (cured versus uncured), whereas the NMCM considers the survival of the population (as a whole) to eventually plateau at a specific value which is then revealed to be the cure fraction.

In brief, the NMCM provides a different means of modelling a cure fraction versus the more well-known MCM. Table 2-1 summarises the key differences between MCMs and NMCMs.

Table 2-1: Mixture versus non-mixture cure models

		Mixture	Non-mixture
Formula for...	... standard application	$S(t) = \pi + (1 - \pi)S_u(t)$	$S(t) = \pi^{F_Z(t)}$
	... relative survival application*	$S(t) = S^*(t)(\pi + (1 - \pi)S_u(t))$	$S(t) = S^*(t)(\pi^{F_Z(t)})$
Variables		π = proportion cured; $S^*(t)$ = expected survival function for the general population (in relative survival application)	
		$S_u(t)$ = survival function for the uncured proportion	$F_Z(t)$ = a cumulative distribution function, which is generally chosen to be $1 - S_Z(t)$, where $S_Z(t)$ is a choice of parametric (or non-parametric) survival model
Lay description		Patients considered a mixture of two groups: 'cured' patients who experience risk of death based on general population mortality only, and 'uncured' patients who experience risk of death based on general population (for the relative survival application) and disease-related mortality.	The disease-related survival curve eventually asymptotes, such that beyond a given time point the only risk of death is related to background mortality (for the relative survival application), or it is zero (if modelled via the standard application). However, before this time point, the risk of death is a blend of both general population and disease-related mortality.

Note: *In a relative survival application, all patients are subjected to the risk of death in the general population. In the standard application, this is not captured, and so the model will plateau at the proportion cured (π).

A book by Peng & Yu provides a comprehensive overview of the different cure modelling methods, their applications, and how they can be implemented in statistical analysis software packages.¹⁰⁴ In addition, a published tutorial on cure models by Latimer and Rutherford was published in 2024, which considers key issues that analysts should be aware of when considering the use of cure models in an HTA context.⁶⁹ It is also important to acknowledge that in the context of human survival, it is critical for sensible predictions for cure models to account for external data for background mortality rates, else 'cured' people would be predicted to live forever. For this reason, it is perhaps most suitable for these types of models to be fitted in a relative survival framework, as opposed to an all-cause framework, where the analyst would need to determine how best to account for background mortality.

A note on the popularity of parametric models for HTA

Parametric survival models are widely used in HTA primarily because they offer a practical and transparent means of estimating long-term survival outcomes from limited trial data. The ability to produce smooth survival curves and to extrapolate beyond the period of observed follow-up, makes them particularly attractive for HTA submissions that rely upon cost-effectiveness analysis, given that they can be easily incorporated into a cost-effectiveness model. Nevertheless, their popularity does not mean they are without limitations. Parametric models rely on assumptions about the shape of the hazard function, which may not reflect the

true underlying data-generating process. Therefore, while parametric models are imperfect, it could be argued that they are some of the best tools available for producing survival estimates for use in HTA.

2.3.3 Model fitting and incorporation in HTA

Data requirements

Data for survival analysis (assuming a single source of IPD) typically resemble the example format presented in Table 2-2. Ultimately, the minimum data required to inform a standard parametric survival model are a series of survival times (measured in a common unit, usually either days, months, or years), with paired event (or censor) flags to determine whether or not the patient was censored at this time. In most data sets, a large volume of other data is also included, such as baseline demographics (e.g., age, sex), clinical variables (e.g., disease state, performance status), and analysis set flags (for example, a flag for if the patient was intended to be treated, a flag for if the patient received the intervention thus is included in a safety analysis set).

Table 2-2: Illustration of survival analysis data format

Patient ID (USUBJID)	Outcome measure (PARAM)	Treatment arm (ARM)	Survival time (AVAL)	Event flag (EVENT)
<i>A unique subject identifier. Subject identification usually comprises a string of digits, which may include some reference to the study and/or site ID</i>	<i>Different outcome measures may be defined within the same data set. Here, a simple description of the outcome is provided, and usually defines the time scale</i>	<i>The data file may include patients that received different treatments. This column provides an identifier for treatment arm</i>	<i>Based on the outcome description, this column provides the specific value associated with each patient</i>	<i>Whether the event occurred at this time ('event'), or not ('censor'). Event coded as 1 or 0 (1 = Yes, 0 = No). Could be presented as 1 – EVENT (i.e., censors)</i>
1001-000-001	OS (days)	Placebo	100	1
1001-000-002	OS (days)	Active	150	0
1001-000-003	OS (days)	Placebo	80	0
1001-000-004	OS (days)	Active	110	1
1001-000-005	OS (days)	Active	115	1
1001-000-006	OS (days)	Placebo	95	1

Key: OS, overall survival.

The example shown in Table 2-2 refers to 'true' IPD – that is, these data refer to individually recorded patient survival times, along with all other information related to each record. In this thesis, it is assumed that IPD are available for at least the main source of data of interest (i.e., the clinical trial for the intervention under investigation). This level of data granularity may not always be available, especially when data are owned by a pharmaceutical company or other commercial organisation. Consequently, it may be the case that only summary-level data are available, which may be presented in several ways:

- Summary measures presented in prose (e.g., “Median survival was 24.3 months”).
- Raw data points (e.g., “of the original 100 patients, 50 survived until at least 5 years”).

- KM estimates in picture format (e.g., see Figure 1-2).

Data in these formats cannot be applied directly within a traditional parametric survival analysis. However, it may be possible to ‘calibrate’ model parameters such that the estimates produced correspond to the known data, though this is challenging when considering models with multiple parameters (e.g., any number of Weibull models will yield a median survival estimate that is the same). If KM estimates are available, pseudo-IPD can be re-created using a back-calculation algorithm, such as the approach of Guyot *et al.*, (2012).⁴³

Software

Assuming data are available to inform a survival model, it is necessary to identify suitable statistical analysis software to perform survival analysis in. In the HTA field, survival modelling is usually performed in one of three software packages/languages: *R*, SAS, or Stata.¹⁰⁵⁻¹⁰⁷

- ***R*** is an open-source, freely available statistical programming language that has become increasingly popular in HTA, particularly for survival modelling. Its appeal lies in its flexibility, and user-led development of specialised packages; such as *flexsurv*, which is widely used in NICE appraisals for fitting a broad range of parametric survival models.¹⁰⁸ In addition, *R* is considered relatively simple to learn, with lots of freely-available online resources and tutorials, and there is an ‘*R* for HTA’ academic consortium dedicated to the use of *R* in cost-effectiveness analysis.¹⁰⁹ However, unlike commercial software such as SAS, *R* lacks formal validation and quality assurance processes. This can raise concerns about reproducibility and long-term stability.
- **SAS** is widely considered the gold standard for analysing clinical trial data submitted for regulatory assessment (e.g., in submissions to the FDA).¹¹⁰ Its longstanding use in the pharmaceutical industry, along with its robust suite of validated procedures for statistical analysis, makes it the preferred choice of software for many clinical trial statisticians. Regulatory agencies often require or strongly prefer datasets and analysis outputs in formats that are compatible with SAS, and the reproducibility and quality assurance of SAS are viewed as key advantages compared with other software packages (such as *R*). Despite this, SAS is relatively high cost (compared to *R* which is free), it is less straightforward to learn versus other software packages, and may be considered to have less flexibility in terms of new methodologies.
- **Stata** is another widely used statistical software package, and is particularly popular among econometricians and applied health researchers. It offers a user-friendly interface that combines point-and-click functionality with the option to write code. Stata is commercial software, so it requires a paid licence, though it is more affordable than SAS and widely available through institutional subscriptions. It includes a range of tools

for survival analysis, and its well-documented syntax (including published guidance in the Stata journal) and reproducible output make it suitable for many HTA applications. However, compared to *R*, it is less flexible and has a smaller number of specialised packages for advanced survival modelling, particularly in the context of HTA. While reliable and well-supported, Stata may be limited for analysts working in HTA.

Other software packages are also available (e.g., python) but are not as popular in HTA. SAS, Stata, and *R* may be used to fit standard parametric survival models, which are done in a standard frequentist inference framework. However, some types of survival analysis require specialist Bayesian software – specific cases of Bayesian methods are described in Chapter 3. Bayesian survival modelling is typically undertaken in one of the following software packages: BUGS (Bayesian inference Using Gibbs Sampling), JAGS (Just Another Gibbs Sampler), or Stan.¹¹¹⁻¹¹³ The statistical software *R* is then frequently used as a user interface and for post-processing of results (e.g., visualisations).¹⁰⁵

The case for R

For the research presented throughout this thesis, I sought to use one piece of software for statistical analysis, primarily for ease of comparisons and to avoid potential issues in comparing results from different software packages. As noted above, *R* is widely adopted in HTA, and is a free, open-source statistical programming language. It has an active user community and there are large number of packages specifically developed for health economic modelling and survival analysis. *R* also provides a convenient interface for implementing Bayesian methods, for example through packages such as ‘rstan’ and ‘survHE’.^{114, 115} Given my existing experience with using *R*, its ability to leverage a large number of packages, and its extensive use in HTA, it was the natural choice for the research presented in this thesis.

Model selection

Plausibly, an analyst could produce a very large number of alternative survival model fits. However, decision-makers need to understand which model (or models) are the most suitable to inform decision making. Therefore, carefully selecting and justifying a given model (or set of models) is essential. TSD 14 provides a widely followed framework for this process.¹⁰ TSD 14 recommends fitting a range of parametric models and assessing their performance using both statistical and clinical considerations. While the more recent TSD 21 provides information on more advanced survival modelling techniques, the principles outlined in TSD 14 remain relevant to determining a suitable survival model for decision making.

Statistical measures of goodness of fit, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), provide a means of comparing how well different models

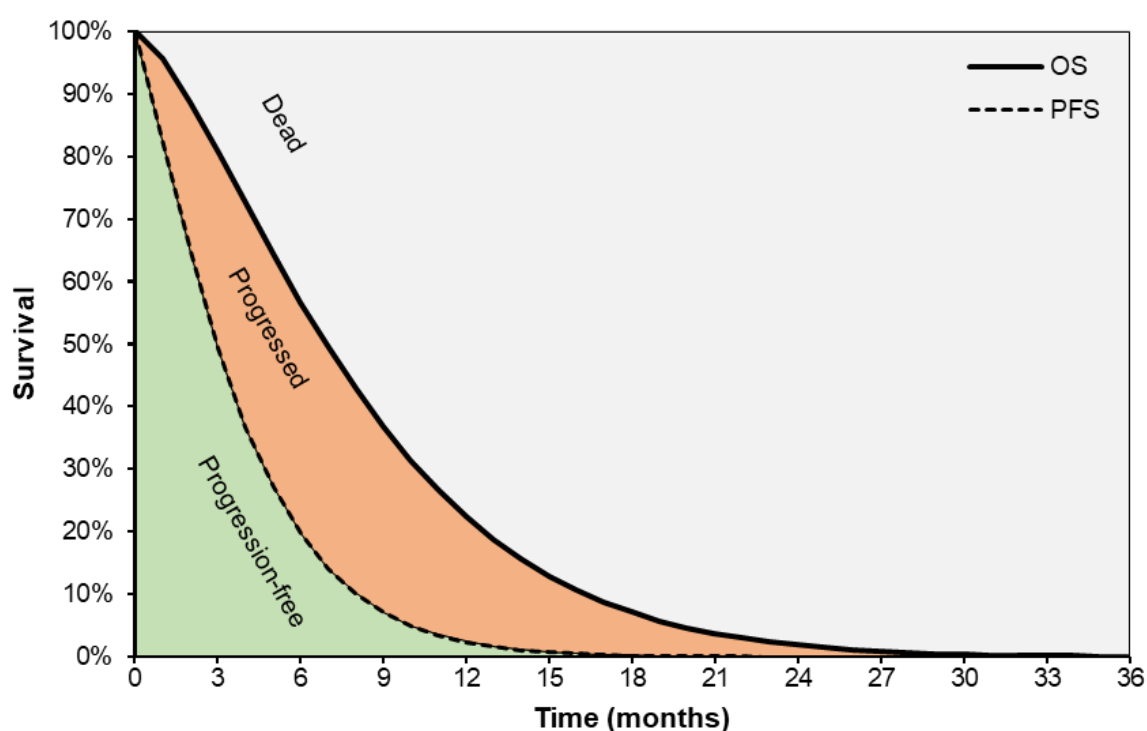
describe the observed data. Visual inspection of fitted survival models against the KM estimates is also important to ensure that a given model accurately represents the estimated survival curve over the period of clinical trial follow-up. However, a combination of statistical and visual goodness of fit is not sufficient when choosing a survival model that will be used for decision making, given that extrapolation will also be required beyond the duration of follow-up available for the clinical trial.

In addition to statistical and visual goodness of fit, it is therefore also important to assess the plausibility of estimated hazards, particularly in relation to the disease biology and the mechanism of action of the treatment. For example, potentially curative treatments (e.g., a stem cell transplant) may be expected to exhibit hazard functions that plateau, which certain types of survival models (e.g., a mixture-cure model) may be better able to capture. Clinical expert input can also help with this process, as they can help judge whether modelled long-term projections are realistic given what is known about the progression of disease and how outcomes might compare given how it is currently managed in real-world practice. Despite this, there will always be an element of the unknown involved in determining the plausibility of long-term survival estimates, and therefore clinical experts may find it challenging to choose between alternative survival models. Therefore, it is good practice to conduct sensitivity analyses using alternative plausible survival models.

Model implementation

Following model fitting and selection, the next step is to implement the model within a cost-effectiveness analysis. In the context of cancer therapies, the majority of cost-effectiveness models developed to inform HTA submissions have adopted a PartSA structure.^{67, 116} A schematic of a typical PartSA model is presented in Figure 2-6, which illustrates how this type of model works.

Figure 2-6: Example partitioned-survival analysis model structure



Key: OS, overall survival; PFS, progression-free survival.

In brief, an estimate of OS is used to determine patients who are alive or dead over the course of the model time horizon, which is then split (or 'partitioned') into patients with and without progressed disease through an estimate of PFS. Differences in costs and health effects for progressed and non-progressed patients are then used to determine differences in overall costs and QALYs. Variations of this type of model have been used, including those that introduce a time-to-treatment discontinuation (TTD) curve to reflect the duration of treatment.

The main benefits of a PartSA model lie in its simplicity and transparency. It is straightforward to explain to clinicians how the model has been populated and how the data (typically OS and PFS) have been analysed. This aligns well with the co-primary endpoints commonly reported in clinical trials of cancer medicines. In addition, PartSA models can often be populated using published survival data, and can incorporate indirect comparisons using NMAs of HRs for OS and PFS. Despite these advantages, the PartSA approach has important limitations. It relies heavily on extrapolation, which may introduce uncertainty when trial follow-up is short. The post-progression health state is also often modelled in a simplistic way, which can be problematic in diseases where multiple progression events occur but are not captured within the follow-up period of the trial. For further discussion of PartSA models, please see TSD 19 and Woods *et al.*, (2020).^{67, 68}

Other types of model structures have also been used to inform HTA submissions of cancer therapies, including cohort-level state-transition models and discrete event simulation (DES) models (see Chapter 4 for a review of cancer appraisals by NICE). However, in these types of models it is usually the case that survival is not based on a directly estimated OS model – instead, mortality is linked to the time spent in a particular health state and/or following a given type of event, which may be specific to a certain type of disease. As such, these types of models are not discussed further in the context of this thesis since focus is placed on the direct estimation of OS, without determining specific components which would be disease area specific and may rely upon a formal consideration of surrogacy relationships (e.g., how time spent without disease progression impacts the risk of death after progression).

2.3.4 ITCs in HTA

ITCs appear regularly in HTAs, particularly in the context of cancer medicines, and where head-to-head evidence for differences in OS between the intervention and at least one comparator is unavailable. ITC methods aim to estimate relative treatment effects, usually in the form of HRs. Given their widespread use and link to estimating OS, this sub-section introduces the most common ITC methods used in HTA, and explains how they relate to survival analysis.

Bucher comparison

A Bucher comparison (named after the original publication by Bucher *et al.*, [1997]¹¹⁷) is a simple ITC method used when two treatments (A and B) have each been compared with a common comparator (C) in different studies. The HR for A versus C is estimated by dividing the HR for A versus B by the HR for C versus B (though on a log scale, this is a subtraction). This method assumes that the populations in both trials are similar and that the relative effects are consistent across studies.

NMA

NMA combines both direct and indirect evidence across a network of clinical trials to estimate relative treatment effects, such as HRs for OS. It assumes that the included studies are sufficiently similar and that treatment effects are *exchangeable* across comparisons. This means (subject to evidence being sufficiently comparable), comparisons for multiple treatments can be made simultaneously, including treatments that have not been compared in a head-to-head study. Guidance for conducting NMA is provided in TSD 2.¹¹⁸

PAIC

A PAIC refers to a method in which IPD that are available for one or more data sources (usually a clinical trial) are used to adjust for between-study differences in the distribution of variables

that influence a particular outcome or outcomes.¹¹⁹ PAIC methods attempt to reduce bias compared with unadjusted comparisons that may arise from cross-trial differences in the characteristics of the individual study populations when a direct comparison is not available. PAIC methods are particularly relevant in HTA as many decisions rely on the ability to compare treatments that are not compared directly, and where there may be important differences across trials.

One of the most commonly-used PAIC methods in HTA is matching-adjusted indirect comparison (MAIC).¹²⁰ MAIC is used when IPD are available for one treatment of interest (e.g., A), but only published aggregate data are available for the comparator (e.g., C), first proposed by Signorovitch *et al.*, (2012).¹²¹ In MAIC, the IPD for A are reweighted so that the baseline characteristics match those reported in the comparator trial for C. A weighted treatment effect (e.g., HR for OS) can then be estimated. MAICs can be conducted in the anchored setting (where there is still a connected network of evidence), or the unanchored setting (where there is no connected network of evidence). MAICs rely on the assumption that all treatment effect modifiers have been accounted for (in both the anchored and unanchored setting), as well as all variables of prognostic importance (in the unanchored setting).

Other than MAIC, there are several other PAIC methods that have gained prominence in HTA practice. One of these is simulated treatment comparison (STC).^{122, 123} In STC, a regression model is fitted to the available IPD (e.g., for A versus B, in study *AB*) to estimate the relationship between baseline covariates (including treatment group) and outcomes. The fitted model is then used to predict outcomes for a target population defined by the published aggregate characteristics of the comparator trial (e.g., for study *BC*). This generates an adjusted estimate of the treatment effect in the comparator population (i.e., for A versus B, in study *BC*). Unlike MAIC, which reweights individual patients to match aggregate characteristics, STC relies on explicit outcome modelling to quantify the impact of specific covariates on the outcome (e.g., OS), and therefore depends on correct specification of the regression model and inclusion of all relevant effect modifiers. STC can also be performed in the unanchored setting, but relies on additional assumptions (like MAIC).

Another more recent approach is multilevel network meta-regression (ML-NMR). ML-NMR extends the standard NMA model by integrating IPD and aggregate data within a single hierarchical modelling framework.¹²⁴ Compared with MAIC and STC, ML-NMR can (in principle) use all available evidence simultaneously and avoid discarding information, while maintaining consistency with the underlying network structure. It can also be used to generate estimates of relative effects within a specified target population, which need not be identical to one of the study populations in the network. However, ML-NMR requires a more complex model specification (versus the standard NMA model), and in practice, reliable estimation

requires there to be both sufficient overlap in covariate distributions and adequate data to support modelling of treatment effect modifiers.

General guidance for conducting PAIC is provided in TSD 18, including reference to other methods (such as network meta-regression).¹¹⁹

Relevance to the thesis

While ITCs are widely used in HTA to estimate relative treatment effects, they are not the focus of this thesis. The methods discussed here are acknowledged to provide context, but the primary aim of the thesis is to explore approaches for estimating survival for a single group of patients, particularly in situations where external evidence may be used.

2.3.5 Conclusion

In this chapter, key terminology, sources of evidence, and fundamental concepts in survival modelling for HTA have been introduced and defined. This provides a foundation for the methods described in later chapters, which is of critical importance when discussing specific statistical approaches and the integration of external evidence. The distinction between survival and hazard functions, as well as other key statistical terms, has been outlined to support a precise understanding of survival analysis techniques.

Additionally, this chapter has explored the various sources of evidence that inform survival estimation, distinguishing between internal and external evidence, as well as the terms ‘data’ and ‘information’. The principles of survival analysis have been introduced, covering parametric, semi-parametric, and non-parametric approaches, along with the key steps involved in model fitting. This groundwork enables a structured and informed discussion of survival modelling throughout the remainder of the thesis.

3. Systematic literature review (SLR)

Chapter synopsis

In this chapter, the findings from a systematic literature review (SLR) of published methods studies concerning the use of external evidence to inform survival extrapolation are presented. The SLR was conducted to identify and synthesise all relevant published methods studies concerning the incorporation of external evidence (in some form) into estimates of survival based on data from a clinical trial. Methods were not specifically restricted to those applied in health technology assessment (HTA), though it was expected that most methods would be related.

Formal database searches were carried out using a search strategy developed using a ‘comprehensive pearl growing’ technique, supplemented by hand searching and citation tracking. Two searching iterations were initially undertaken, followed by an update in 2024. The 2024 update confirmed that future studies that are published are likely to be identified via the same search strategies. Relevant studies were summarised and classified according to how the external evidence influenced estimates of survival (i.e., how the external evidence was used as part of the model fitting process).

In total, 19 unique methods to incorporate external evidence into survival extrapolations were identified, covering four key themes: (i) methods that adopted a piecewise approach to incorporate external evidence, (ii) methods that involved specification of a Bayesian analysis using informative priors, (iii) methods that considered the role of general population data to adjust survival outcomes, and (iv) methods that included some other approach to leverage external evidence. The category of ‘other’ approaches included the analyst making an assumption (either based on their own judgement or based on input from a clinical expert), or introducing some other link to evidence from an external source (such as how data from one source [possibly for a ‘similar’ population] may relate to another source).

Through undertaking the SLR, it was noted that many methods exist, though relatively few appear to have received attention in contemporary HTA despite the papers themselves referring to HTA-specific examples (discussed further in Chapter 4). In addition, newly proposed methods introduced in papers did not provide comparisons to existing methodology, which introduces challenges with respect to identifying and justifying use of specific methodology for HTA purposes. In conclusion, a comprehensive set of methods used to incorporate external evidence into survival estimates was obtained systematically via this review. The systematic identification of methods from the SLR serves as a foundation for the remainder of the thesis.

The initial review was published in Medical Decision Making in 2023.³

3.1 Introduction

The first stated aim of this thesis is to identify which methods have been previously used to combine external evidence with trial data to make long-term projections of OS. To address this research question, it is necessary to conduct a review of the literature to find relevant methods that have been used in the past, though some methods may have only been introduced as part of an HTA. This chapter describes a review of the published literature, whereas Chapter 4 describes a review of published NICE STAs; with both reviews aiming to identify which methods have been used in this area.

Previously, Jackson *et al.* conducted a review of methods used when extrapolating survival from randomised trials using external data, published in 2017 (henceforth termed JR, for brevity).²⁶ The JR authors highlighted a number of methods that have been previously documented in the published literature, alongside the key underlying assumptions made and circumstances under which they have been applied. Also discussed were the implications of some of the models chosen and fitted by study authors (which provide some insights into the reason[s] why methods were used by study authors), as well as how the uncertainty in estimates may be represented. The JR concludes with recommendations for future research and considerations beyond the scope of their review, including issues related to generalisability of trial populations and methods that use non-data external evidence sources (such as judgements obtained from clinical experts).

The JR serves as a helpful description of methods that have been previously proposed to incorporate external data into survival extrapolations. However, the review was not systematic, and was concerned only with studies that made use of external *data* (i.e., methods that use other information sources, such as clinical expert opinion, were not in the scope of the JR). Consequently, while the JR is certainly a helpful summary of some potentially relevant methods, it has a number of limitations with respect to the scope of the thesis.

To address these limitations of the JR within the context of this thesis, and identifying studies published following its acceptance for publication, a *de novo* SLR was conducted. The SLR aimed to identify methods for estimating survival that incorporate an element of external evidence within the estimation of the individual survival models[†]. More specifically, the SLR aimed to identify *primary* studies of methods that include the use of external evidence for the estimation of a survival-based outcome (i.e., studies which present a ‘novel’ method, and not either an application of, or minor modification to, an established approach). The systematic

[†] External information may be considered within a validation context (e.g., to choose one modelling approach over another or to comment on the plausibility of the models’ projections), though as this does not impact the estimation of the survival model itself, such approaches are not considered relevant in this chapter. The role of external information outside the model fitting process will be discussed later in the thesis.

identification of relevant methods studies allows for an in-depth exploration of how different methods compare, which will be presented later in the thesis.

While some of these methods are relatively well-known, to date there has been no published SLR that has aimed to identify all potentially relevant methods that have been proposed and documented within the literature. By undertaking an SLR, a reproducible synthesis of the published literature can be obtained (and potentially updated periodically). The SLR can then be used to understand which methods are available for consideration when faced with undertaking survival extrapolation using external evidence, and in which circumstances the different methods have been applied.

The scope of the methods literature search is focused predominantly on the identification of original methods studies (i.e., those that propose either a new method, or an extension/adaptation to an existing method, to estimate survival using an element of external evidence). It is expected that some studies may have been published after a given original study, where the approach has perhaps been considered in other situations, or simply applied as part of a broader piece of research (for example, to inform an aspect of an economic model). Applications of previously developed methods (henceforth termed secondary methods studies) are beyond the scope of the SLR, but may be discussed later in the thesis when investigating specific methods in more detail, and are of particular relevance for the appraisals review (see Chapter 4).

Within the context of the SLR, it is important to note that some methods may appear relevant to the exact scope of the thesis, but this may not be apparent upon an initial inspection of the title and abstract, or perhaps even based on an initial read of the full paper. For example, methods that can be used to estimate survival for a single group of patients are the primary focus of the thesis, but some previous methods may have been developed to simultaneously estimate survival for two (or more) groups of patients. In the SLR, the goal was to identify all potentially relevant methods, noting that some of these methods may not be relevant throughout the subsequent chapters of the thesis. To this end, some methods are not taken forward to later parts of the thesis, and where applicable, rationale is provided for why a specific method was not taken forward.

3.2 Methods

3.2.1 Searching methods

Inclusion and exclusion criteria

A summary of the inclusion and exclusion criteria is provided in Table 3-1. The PICOS framework (as recommended for use in the Cochrane handbook¹²⁵: **P**opulation, **I**ntervention, **C**omparison, **O**utcome[s], **S**tudy type) is not directly applicable to this SLR due to the objective

being to identify methods literature, and so instead the inclusion and exclusion criteria were focused predominantly on the design of studies.

Table 3-1: Inclusion and exclusion criteria

Inclusion	Exclusion
Primary* methodological study or review of methodological studies; including reference to use of external evidence for the estimation of survival-based outcome(s)	All other study types, including (but not limited to): - Economic evaluations (unless they are presented as an illustrative case study within the context of a methodological study) - Clinical studies (including analyses of registry data, claims databases and other data sources) - Commentaries Grey literature sources will <u>not</u> be excluded if identified, though the search strategy is focused predominantly on the identification of peer-reviewed journal publications Studies that consider applications of previously developed methods (i.e., secondary studies**) Methods that do not incorporate external evidence within the estimation of the survival model itself, but instead informally refer to external evidence for validation or contextual purposes

Note: *Within the context of this review, a primary study is assumed to be any study which presents a 'novel' method concerning the use of external evidence in survival extrapolation. Some methods may be (minor) alterations of pre-existing methods and/or adaptations of methods used in other disciplines. For this review, these are deemed primary studies. Reviews of methodological studies were also included for two reasons: (1) they may contain a novel method developed by the review authors as a result of the review, and (2) such reviews may aid with the identification of other primary methodological studies. **Aligned with the description of a primary methodological study, secondary studies refer to those which directly apply methods already presented in the literature. For example, if a study presents an analysis described as applying the method of Smith *et al.*, this constitutes a secondary study. However, if the authors describe adapting a previous method, this is sufficient for it to be determined a primary study.

Economic evaluations were considered out of the scope of this SLR given that the approach to estimate survival may not be reflected within the title or abstract, and so (in principle) if these studies were included, it would be necessary to review the full text of all published economic evaluations to ascertain the approach taken to estimate survival. However, published HTA reports (including company submissions and associated critique from assessors) are instead reflected within a distinct review of published NICE appraisal documentation (see Chapter 4).

Search approach

SLRs of methods-based studies may be considered more challenging than SLRs of clinical studies for several reasons. These reasons include differences in terminology used to describe studies, a lack of validated search filters, and potential difficulties in trading off recall (proportion of relevant articles identified by a search versus all relevant articles) and precision (proportion of relevant articles identified by a search versus all identified articles). There are, however, a number of previously conducted SLRs which aimed to identify published methods-based literature.¹²⁶⁻¹²⁹ To assist with the design of the SLR, a brief Google search was

undertaken to consider points raised by authors of previous methods studies, covering a broad range of research disciplines (not all related to medical research).

Johnson *et al.*, (2009) reported the findings of an SLR concerning methods to elicit beliefs for the specification of Bayesian priors.¹³⁰ The authors of this study did not specify how appropriate search terms were determined, but a database search strategy was presented. From this, it may be inferred that search terms were employed based on the expectation that all relevant published studies would use relatively consistent terminology (e.g., ‘Bayes’ or ‘Bayesian’ when describing Bayesian methodology), and revolve around three central themes (i.e., ‘priors’, ‘obtaining opinion’, and ‘Bayesian methods’). Henrys *et al.*, (2014) conducted an SLR of statistical methods to monitor the potential impacts of products used in agriculture on the environment.¹³¹ However, unlike the study by Johnson *et al.*, (2009) where search terms were relatively straightforward, Henrys *et al.*, (2014) noted that in order to capture all relevant studies, the search terms needed would “*effectively cover the whole range of ecology and environmental studies*” (Henrys *et al.*, 2014, p.12).¹³¹

More specifically related to HTA, previous studies have undertaken methods-based reviews and encountered similar challenges when developing search strategies. Latimer, (2012) published on the role of crossover adjustment methods in the context of economic evaluation included a review of methods.¹³² As part of describing the approach taken, he states that “*systematically reviewing methodologies presents significantly different problems to systematically reviewing clinical interventions*” (Latimer, 2012, p.59).¹³² Ultimately, Latimer adopted a pearl-growing technique to identify relevant studies. Alshreef *et al.*, (2019) conducted a systematic review of methodological papers within the context of patient non-adherence when estimating treatment effectiveness and conducting HTA.¹²⁹ The authors of this study describe a “comprehensive pearl growing” (CPG) technique and 2-stage iterative searching approach to identify relevant papers, for which a review of information-seeking strategies by Schlosser *et al.* is cited.¹³³ Schlosser *et al.* note that there are different types of pearl-growing search techniques, including traditional pearl growing (TPG) as well as CPG.¹³³ A comparison of TPG and CPG techniques presented by Schlosser *et al.* is re-produced in Table 3-2 for context.

Table 3-2: Comparison of pearl-growing methods re-produced from Schlosser *et al.*, (2006)

Features	Traditional Pearl Growing	Comprehensive Pearl Growing
Purpose	To aid a practitioner's search for evidence to engage in evidence-based practice	To aid a researcher's search for studies to be included in a systematic review
Information source needed	One or a few relevant article/s – the 'pearl/s' (a very small sample from the population of relevant studies)	Compilation of relevant articles – lots of 'pearls' (a large sample from the population of relevant studies)
Functions	<i>Suggests potentially relevant:</i> • Databases • Keywords • Quality filters Suggests indexing problems	<i>Systematically identifies relevant:</i> • Databases • Keywords • Quality filters Reveals indexing problems and informs their remediation
Representativeness	May not be representative	Highly representative
Effort	Fairly time efficient	Time consuming

Source: Re-produced from Schlosser *et al.*, (2006): 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. *Int J Lang Commun Disord*. 2006 Sep-Oct;41(5):567-82. Table presented here with permission from John Wiley and Sons. Copyright for this table is retained by John Wiley and Sons.

Information reported in the above studies, as well as general SLR best practice guidance (detailed in the Cochrane handbook¹²⁵), were considered within the design of the overall searching approach. In addition, advice was sought from an experienced systematic reviewer (Dr. Lesley Uttley) and an experienced information specialist (Dr. Ruth Wong) to identify any potential issues with the searching approach.

A pearl-growing searching approach was adopted for the SLR, which has been used in a number of previous reviews especially within the medical field.^{129, 134-136} In addition, the JR provides a set of pearls from which other relevant studies may be found. Furthermore, by specifying a pearl-growing searching approach with more than one iteration, the most appropriate search terms can be established as the review is developed – this is especially useful within the context of a review where potentially-relevant search terms are not expected to be universally adopted in the published literature and may only be revealed through inspection of identified studies (i.e., pearls). However, as acknowledged by Schlosser *et al.*, there are different pearl-growing approaches that may be considered.¹³³

Given that the JR provides a set of pearls to inform a pearl-growing searching approach, and owing to it adopting a systematic searching approach, a CPG search was considered a better option versus a TPG search.[‡] A CPG searching approach requires the following steps to be taken, as outlined by Schlosser *et al.*¹³³:

[‡] A TPG search would involve searching based on a small number of relevant papers, as opposed to a CPG search where a set of papers obtained from a previous review can be used as an information source to base searching upon.

1. A collection of studies from a relevant review or bibliography is required (i.e., the JR).
2. Based on the collection of studies, relevant database(s) should then be determined.
 - The JR does not describe any specific database(s) other than Web of Science (WOS), so it is necessary to establish in which database(s) the pearls are indexed in order to determine which database(s) should be searched to find other pearls.
3. Using the studies and the corresponding database(s), it should then be established how studies are indexed in terms of keywords and/or quality filters.
 - Given that the JR is concerned with methods papers (for which quality filters are not anticipated to be available or consistent across studies), this part of the process is informed through the determination of keywords.
4. Using the identified keywords and/or quality filters, a building block query should then be developed in order to identify other relevant articles in the database(s).
 - This process involves using papers from step 1 to determine a search strategy.
5. Finally, the end of the search should be determined when articles retrieved provide diminishing relevance.
 - There is no gold-standard definition of diminishing relevance, and so the decision of when to stop searching requires careful consideration.

A more in-depth explanation of how the searching strategy was developed is described later in this sub-section.

Identification of initial pearls

The JR cited 14 distinct published studies that reported survival analysis methods involving external evidence. For completeness, an additional 24 HTA reports were also identified in the JR and were referenced by the authors, although published economic evaluations are out of the scope of this SLR as noted previously.

While a total of 14 published studies are noted in the JR, a comprehensive list of these studies is not provided within the article itself. The corresponding author (Dr. Christopher Jackson) was contacted to request a list of the studies that may be suitable to consider pearls, and the list of pearls was kindly provided via email correspondence. Of the 14 studies reported within the JR, three were published economic evaluations, leaving 11 studies which were original research articles with a focus on the development and/or application of novel methods for estimating survival incorporating external data. Of note, Jackson *et al.* did not undertake a formal systematic review, and so formal inclusion/exclusion criteria and search strategies were not provided.

The remaining 11 studies were identified as ‘pearls’ to develop the search strategy to identify additional pearls. The three excluded studies from the JR were considered after the formal database searching stage for the purpose of forwards and backwards citation tracing (discussed further later in this chapter). Table 3-3 contains the set of the 11 ‘pearls’ identified via the JR.²⁶

Table 3-3: Pearls from the Jackson *et al.*, (2016) review

#	Reference
1	Andersson TM-LL, Dickman PW, Eloranta S, <i>et al.</i> Estimating the loss in expectation of life due to cancer using flexible parametric survival models. <i>Stat Med.</i> 2013 Dec 30;32(30):5286–300.
2	Benaglia T, Jackson CH, Sharples LD. Survival extrapolation in the presence of cause specific hazards. <i>Stat Med.</i> 2014;34(5):796–811.
3	Chu PC, Wang JD, Hwang JS, Chang YY. Estimation of life expectancy and the expected years of life lost in patients with major cancers: Extrapolation of survival curves under high- censored rates. <i>Value Heal.</i> 2008;11(7):1102–9.
4	Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. <i>Stat Med.</i> 2006;25(11):1960–75.
5	Fang, C. T., Chang, Y. Y., Hsu, H. M., <i>et al.</i> Life expectancy of patients with newly- diagnosed HIV infection in the era of highly active antiretroviral therapy. <i>QJM An Int J Med.</i> 2007;100(2):97–105.
6	Guyot P, Welton NJ, Beasley M, Ades AE. Extrapolation of Trial-Based Survival Curves Using External Information. <i>Value Heal. Elsevier;</i> 2014 Nov 11;17(7):A326.
7	Hwang J, Wang J. Monte Carlo estimation of extrapolation of quality-adjusted survival for follow-up studies. <i>Stat Med.</i> 1999;18(13):1627–40.
8	Maetani S, Nakajima T, Nishikawa T. Parametric Mean Survival Time Analysis in Gastric Cancer Patients. <i>Med Decis Mak.</i> 2004;24(2):131–41.
9	Messori A, Trippoli S. A new method for expressing survival and life expectancy in lifetime cost-effectiveness studies that evaluate cancer patients (review). <i>Oncol Rep.</i> 1999;6(5):1135– 76.
10	Nelson CL, Sun JL, Tsiatis AA, Mark DB. Empirical estimation of life expectancy from large clinical trials: Use of left-truncated, right-censored survival analysis methodology. <i>Stat Med. Wiley Online Library;</i> 2008;27(26):5525–55.
11	Viscomi, S., Pastore, G., Dama, E., <i>et al.</i> Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukemia. <i>Ann Oncol.</i> 2006;17(1):167–71.

Development of initial search strategy

The following databases were searched:

- Medical databases:
 - Embase (via Ovid)
 - MEDLINE and MEDLINE In-Process (via Ovid)
- Other databases:
 - EconLit (via Ovid)
 - WOS

These databases were selected based on where the pearls were indexed, summarised in Table 3-4. All of the pearls were indexed in Embase, MEDLINE, and WOS (noting that MEDLINE is indexed within the WOS platform). Given that all the studies identified by the JR were indexed in both Embase and MEDLINE (even though formal database searching was not considered as part of this study), it was considered unlikely that the inclusion of further databases would identify additional hits. While none of the 11 pearls were indexed in EconLit,

the additional workload to include EconLit was relatively small (simply an extra 'tick box' when searching via Ovid), and so EconLit was included for completeness.

Table 3-4: Database checking for Jackson *et al.*, (2016) review pearls

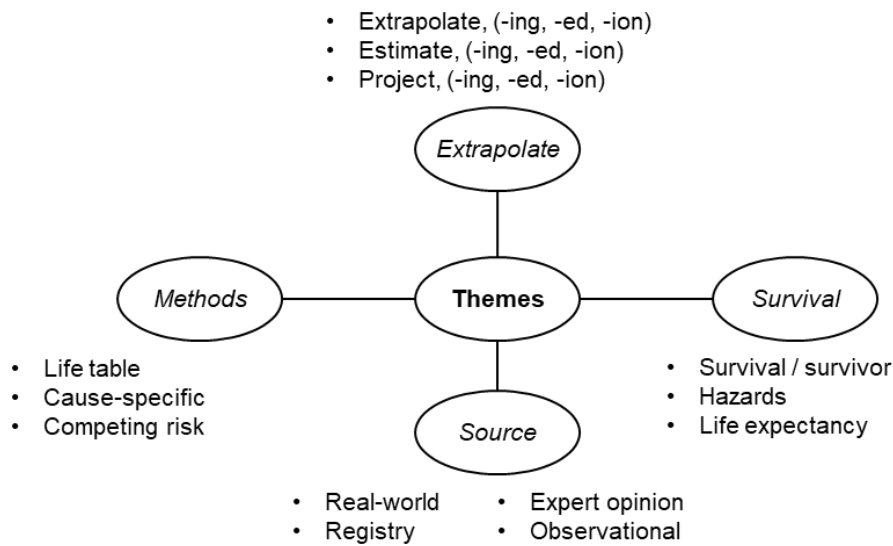
#*	Embase	MEDLINE	EconLit	Web of Science [†]
1	✓	✓	✗	✓
2	✓	✓	✗	✓
3	✓	✓	✗	✓
4	✓	✓	✗	✓
5	✓	✓	✗	✓
6	✓	✓	✗	✓
7	✓	✓	✗	✓
8	✓	✓	✗	✓
9	✓	✓	✗	✓
10	✓	✓	✗	✓
11	✓	✓	✗	✓

Note: *The study number corresponds to Table 3-3. [†]Web of Science includes MEDLINE. However, for later efficiency, the Web of Science search is edited to remove MEDLINE records to avoid superfluous duplication of search results.

Based on the 11 identified pearls, the titles and abstracts were run through a text analyser to identify common terms used across the pearls.¹³⁷ Relatively few common phrases were identified, which may be expected given that survival extrapolation could be described in a number of different ways (e.g., 'extrapolation of survival', 'extrapolated survival', 'survival extrapolations', etc.). Unsurprisingly, the most common word cited (after 'the' and 'of') was 'survival', appearing a total of 82 times. 'External' only appeared a total of 11 times, meaning that different authors did not always explicitly refer to the external evidence source using the word 'external'. 'Extrapolation' appeared 22 times, whereas 'extrapolate' appeared 4 times, and 'extrapolating' appeared twice.

Four broad themes (or 'conceptual strings') were identified as potentially relevant to include based on the text analyser, which are summarised in Figure 3-1.

Figure 3-1: Summary of relevant themes for consideration within the search strategy



The combination of the themes is directly related to the objective of this SLR (that is, to identify studies wherein survival has been extrapolated via methods using external evidence). The inclusion of terms related to specific methodology was considered necessary to ensure that well established approaches (e.g., those considering a life-table approach) were identified, which may otherwise not be described using lay terminology.

The overall intention of the search strategy was to ensure that identified studies:

- Reference extrapolation (or equivalent terminology).
- Discuss survival, either as an outcome (such as OS), statistical method (such as survival analysis), or other equivalent terminology.
- Note an external evidence source (e.g., real-world data) and/or approach to make use of external evidence to inform survival estimation.

Use of terms such as ‘survival’ and ‘extrapolation’ separately has the potential to yield a very large number of hits, most of which are unlikely to be relevant to this SLR. This is because it is common for survival to be discussed as a broad concept within a range of clinical studies, and extrapolation may be related to a multitude of other outcomes unrelated to survival. For this reason, search terms were designed (where possible) to ensure discussion concerning extrapolation was specifically related to survival (and not another, unrelated measure). Therefore, when developing the search strategy, different possible combinations of these terms (and others) were considered.

For the purpose of the initial search strategy, the goal was to successfully identify all of the pearls from the JR, while also being able to identify a number of expected additional hits

(explained further later in this chapter). Due to the iterative nature of the CPG approach, any terms deemed to be missing from the initial search strategy were noted throughout the searching process for consideration in subsequent searches.

The final search strategy used for the first iteration of CPG is presented in Table 3-5 and Table 3-6. Two variations of the strategy were developed, based on the need to run searches for three databases (Embase, MEDLINE, and EconLit) via Ovid, and one database (WOS, excluding MEDLINE records) separately. The WOS search was run excluding the MEDLINE database as a more granular search of the MEDLINE database was possible through Ovid (allowing, for example, the inclusion of MEDLINE In-Process hits), and given the expected degree of overlap between Embase and MEDLINE hits. This resulted in a more refined WOS search. The search strategies for each platform are presented in Table 3-5 (Ovid) and Table 3-6 (WOS).

Table 3-5: Ovid searching (first iteration)

#	Search term(s)
1	((extrap* or estimat* or project* or predict* or expect* or express*) adj3 survival).tw.
2	(parametric adj2 survival).tw.
3	(extrap* adj2 (trial* or mortality)).tw.
4	((estim* or express*) adj2 life-expectancy).tw.
5	(RWE or RWD).tw.
6	((real-life or real-world) adj3 (data or evidence)).tw.
7	(external adj (information or data)).tw.
8	background mortality.tw.
9	competing risk*.tw.
10	(excess adj (mortality or hazard*)).tw.
11	life-table*.tw.
12	(expert* adj (information or opinion)).tw.
13	(observational adj2 data*).tw.
14	((disease or cause) adj (related or specific) adj (survival* or hazard*)).tw.
15	(registry* or registries).tw.
16	(reference adj2 population).tw.
17	or/1-4
18	or/5-16
19	17 and 18
20	letter.pt.
21	historical article.pt.
22	editorial.pt.
23	conference abstract.pt.
24	or/20-24
25	19 not 24
26	exp animals/ not humans.sh.
27	25 not 26
28	limit 27 to english language

Table 3-6: Web of Science searching (first iteration)

#	Search term(s)
1*	(TI= ((extrap* OR estimat* OR project* OR predict* OR expect* OR express*) near/3 survival) OR AB= ((extrap* OR estimat* OR project* OR predict* OR expect* OR express*) near/3 survival))
2*	(TI = (parametric near/2 survival) OR AB= (parametric near/2 survival))
3*	(TI = (extrap* near/2 (trial* or mortality)) OR AB= (extrap* near/2 (trial* or mortality)))
4*	(TI = ((estim* or express*) near/3 (life-expectancy)) OR AB = ((estim* or express*) near/3 (life-expectancy)))
5*	(TI = (RWE or RWD) OR AB= (RWE or RWD))
6*	(TI = ((real-life or real-world) near/3 (data or evidence)) OR AB= ((real-life or real-world) near/3 (data or evidence)))
7*	((TI = (external near/1 (information or data))) OR (AB = (external near/1 (information or data))))
8*	(TI = background-mortality OR AB = background-mortality)
9*	(TI = competing risk* OR AB = competing risk*)
10*	(TI = (excess near/1 (mortality or hazard*)) OR (AB = (excess near/1 (mortality or hazard*)))
11*	(TI = life-table*) OR (AB = life-table*)
12*	(TI = (expert* near/1 (information or opinion))) OR (AB = (expert* near/1 (information or opinion)))
13*	((TI = (observational near/2 data*)) OR (AB = (observational near/2 data*)))
14*	(TI = ((disease or cause) near/1 (related or specific) near/1 (survival* or hazard*)) OR (AB= ((disease or cause) near/1 (related or specific) near/1 (survival* or hazard*)))
15*	(TI = (registry* OR registries)) OR (AB = (registry* OR registries))
16*	(TI = (reference near/2 population)) OR (AB = (reference near/2 population))
17*	#4 OR #3 OR #2 OR #1
18*	#16 OR #15 OR #14 OR #13 OR #12 OR #11 OR #10 OR #9 OR #8 OR #7 OR #6 OR #5
19	#18 AND #17 Refined by: [excluding] databases: (MEDLINE) AND databases: (WOS) <i>Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English</i>
20†	ANIMAL STUDIES FILTER (see Appendix A)
21	(#19 not #20) Refined by: document types: (ARTICLE OR EARLY ACCESS OR BOOK OR REVIEW) <i>Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English</i>

Note: *Each of these lines also include the following additional information: 'Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English'.

†Owing to its length, the filter applied to remove animal studies from the search is presented separately in Appendix A for simplicity.

Search limits and saturation

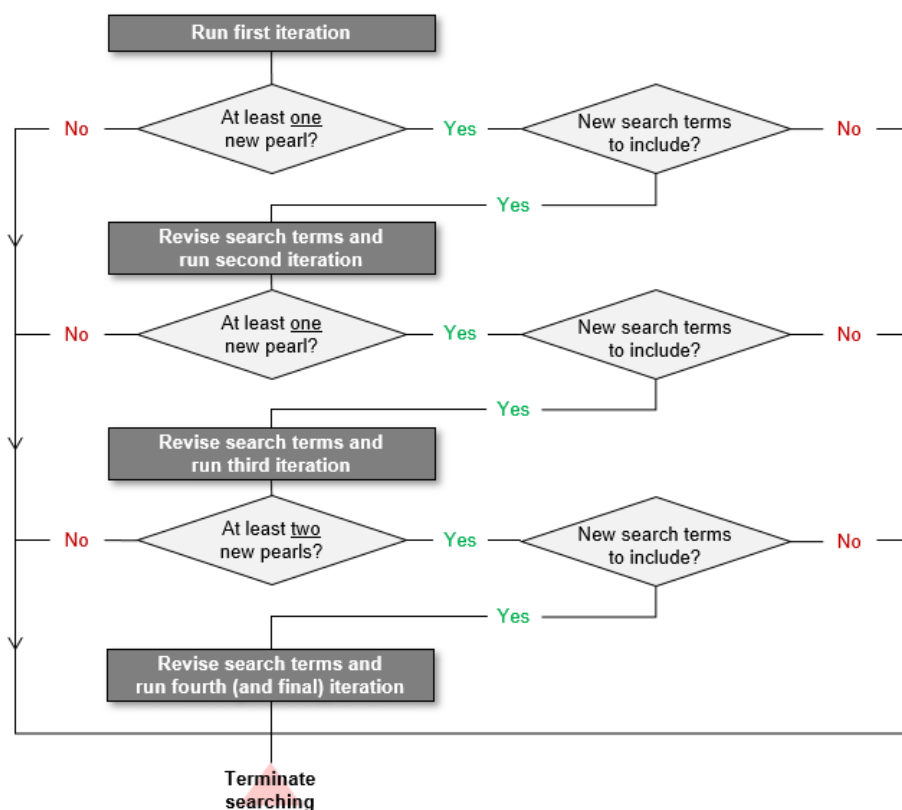
A summary of the search limits (and why these limits were imposed) is provided in Table 3-7.

Table 3-7: Search limits

Category	Restriction(s)	Rationale
Date of publication	1946 to the present	Online database limits (MEDLINE stops at 1946). This was also considered reasonable given that the Jackson <i>et al.</i> , (2016) review ²⁶ is anticipated to have identified the majority of relevant studies published prior to 2016.
Language	English language	Non-English language studies were excluded as a practical limitation based on lack of translation resources available. The majority of relevant studies are anticipated to have been written in English language.
Other limits	Accessible studies	Studies that could not be accessed via the University of Sheffield online portal were requested via ResearchGate/ author contact/request via the University of Sheffield library. If still not retrievable, studies were excluded based on inability to access.

Based on the initial strategy (presented in Table 3-5 and Table 3-6), an initial search was run, and new pearls were identified. This initial search may be considered the first iteration of the CPG searching approach. Petticrew and Roberts describe the need to determine a cut-off point for searching, as otherwise the review “*would never end*”; yet also acknowledge the lack of a definitive “*stopping rule*” for SLRs (Petticrew and Roberts, 2006, p.100).¹³⁸ As part of this study, the authors cite a previous study by Chilcott *et al.*, wherein TPG was stopped after the third iteration for pragmatic reasons.¹³⁹ For the purpose of this SLR, the process shown in Figure 3-2 was followed to determine when saturation was likely reached, stopping theoretically after a maximum of four iterations to mirror the approach of Chilcott *et al.* but with one additional search if it was deemed necessary (where another search would need to be based on at least two new pearls, indicating the need for further searching).

Figure 3-2: Schematic of comprehensive pearl growing search iterations



Using the approach described in Figure 3-2, it was expected that the majority of relevant studies that would be possible to identify through formal database searching would be identified. The results of each search were exported as a .RIS file, and de-duplicated using the built-in functionality of EndNote. A two-stage screening process was conducted, involving primary and secondary screening.

3.2.2 Screening and eligibility

Primary screening was conducted based on the study titles and abstract, independently performed by two reviewers (myself and Dr. Shields). Secondary screening was conducted on the full papers of included studies and was also independently performed by two reviewers. Any disagreements with primary or secondary screening were resolved by a third party, with expertise in the subject of survival extrapolation (Prof. Nicholas Latimer).

After identifying the final list of pearls obtained through formal database searching, a series of additional checks were performed for completeness. Firstly, citation searches for each pearl were conducted to identify any relevant papers that referenced a given pearl (or pearls) that were not found through formal searching (i.e., forwards citation tracing). In addition, a review of the reference lists for each pearl was also performed to establish if any older studies were missed through formal database searching (i.e., backwards citation tracing). Included within the set of papers for citation tracing were also the three papers not considered pearls from the JR (for completeness).

In addition to the formal database searching, materials presented at previous ISPOR meetings were also considered, through keyword searching via the ISPOR Presentations Database.¹⁴⁰ Emergent research in the context of HTA methodology has previously been presented at ISPOR meetings, including one of the pearls identified in the JR by Guyot *et al.*¹⁴¹

This search was considered to identify any emergent methods that are either not yet published (and therefore of interest to be aware of, though technically out of the scope of the SLR) or reference a method that may otherwise be difficult to identify through formal database searching (and therefore provide an alternative means of identifying such methods to avoid issues in findings such studies through formal database searching).

Finally, a complete set of pearls was provided to Prof. Nicholas Latimer, an expert in survival extrapolation for HTA (and co-author of the JR), to check for any missing, known studies.

3.2.3 Data extraction

A data extraction table was developed based on the features of a number of previous studies aiming to identify and/or evaluate methods within the HTA field.

Bell Gorrod *et al.* performed a review of survival analysis methods used in NICE TAs of cancer treatments.²⁷ The data extraction table developed for this study comprised of three section headings, which were chosen to align with the key recommendations from TSD 14 (which was used as a basis for comparing the different methods).¹⁰ In this study, the majority of the extraction grid comprised of yes/no questions (for example, “*was an assessment made of the proportional hazards assumption? [yes/no]*”, Bell Gorrod *et al.*, 2019, supplementary appendix

table A1). However, such questions are not as relevant within the context of this SLR, as the purpose of the review is to understand specific features of methods used, including how they work and whether they have been used previously. Consequently, the majority of the fields in the data extraction table for the SLR undertaken comprises of free-text fields (in order to ensure all relevant information is recorded).

As part of his PhD thesis, Latimer (2012) presented a data extraction *framework*, based around “*factors*” and “*considerations*”.¹³² Four factors were presented: “*origin*”, “*theoretical suitability*”, “*application*”, and “*other issues*”. Each factor was associated with considerations, which took the form of questions to be answered (e.g., within the “*theoretical suitability*” factor, one consideration was “*what are the theoretical advantages and disadvantages associated with the method?*” Latimer, 2012, p.60). This style of data extraction table is particularly well suited to this SLR, as it allows for the incorporation of all relevant considerations within the context of the review objective. The final data extraction table presented in Table 3-8 was developed.

The final data extraction table presented in Table 3-8 was developed.

Table 3-8: Data extraction table

Factor	Considerations
Summary information	• Name of method(s) used [free text] • Key reference(s) [free text] • Description of method, including the model fitting process [free text] • If the method is an extension to a previous approach, and if so, what the approach is [free text]
Requirements	• Data requirements to use the method(s) [free text] • Key assumptions made in survival estimation [free text] • Statistical software package(s), including associated functions/commands used in analysis [free text]
Application	• Context(s) in which the method(s) have been applied (e.g., simulated or ‘real’ data; used to inform HTA/ epidemiological studies) [free text] • Approach(es) used to validate and/or compare projections (versus other methods that use external evidence, or ‘standard’ approaches that do not make use of external evidence) [free text] • Code used to fit models (if reported) – yes/no with link to access code (e.g., supplementary material) [free text]
Commentary	• Advantages over other approaches (as described by authors) [free text] • Limitations highlighted by the authors [free text] • Any other comments noted (e.g., similarity between method identified versus another known method) [free text]

Key: HTA, health technology assessment.

Data extraction was performed by one reviewer, with a random sample (25%) repeated by a second reviewer (Dr. Shields), to ensure similarity in extracted information. No formal quality assessment or synthesis of the identified studies was planned. This is because the purpose of the SLR was to identify all relevant published information concerning the incorporation of

external evidence into estimates of survival from a clinical trial, regardless of the quality of the studies themselves. Any limitations in the reporting of specific studies were factored into the narrative review of the SLR findings (and are included in the data extraction table). An assessment of specific methods is nevertheless important to address the research questions for the thesis, and is therefore considered later in Chapters 5 and 6.

3.2.4 Limitations

Several limitations are acknowledged within the context of this review. Firstly, novel methods that are used within the context of other research (e.g., a *de novo* economic evaluation) were not identified by this SLR, given that as previously mentioned, in order to comprehensively identify such methods, the SLR would need to include all previously published economic evaluations. While this is a limitation, it is noted that a separate review of published NICE assessments was conducted to identify methods that have been used in contemporary HTA (see Chapter 4).

There is no standard terminology that exists to describe external evidence, and authors use different terminology based on the specific source of information used. Some evidence sources (e.g., a disease-specific registry) may be referenced by certain studies instead of 'registry', and so may not have been identified by the search strategy if insufficient detail is otherwise provided concerning the estimation of survival within the title or abstract. However, it is likely the case that disease-specific registries would be referred to in methods papers as a case study, and thus described accordingly as a registry within the abstract.

The effectiveness of specific methods was not explicitly evaluated as part of this review, as it is expected that the performance of each method may be affected greatly by the context to which it has been applied within. Nevertheless, the strengths and weaknesses of each approach (as stated by the study authors) were extracted and summarised in a narrative synthesis, and a further comparison of the strengths and limitations of the methods themselves is presented later in this chapter, as part of a narrative review of the methods identified.

The CPG searching approach was designed such that the pearls reported in the JR will be identified by the first iteration of the formal database searching. However, in addition to the pearls from the JR, there are several studies that were published after 2016 that were also expected to be identified by the search (which were identified via a brief Google search and/or were known of prior to executing the search), summarised in Table 3-9. In addition, an experienced systematic reviewer (Dr. Lesley Uttley) and an experienced information specialist (Dr. Ruth Wong) each kindly provided advice concerning the draft SLR protocol. The protocol was also registered as a record on PROSPERO (ID CRD42020202034).^{142, 143}

Table 3-9: Expected hits for validation

#	Reference
1	Vickers A. An Evaluation of Survival Curve Extrapolation Techniques Using Long-Term Observational Cancer Data. <i>Med Decis Making</i> . 2019 Nov;39(8):926-938.
2	Cope S, Ayers D, Zhang J, Batt K, Jansen JP. Integrating Expert Opinion With Clinical Trial Data to Extrapolate Long-Term Survival: A Case Study of CAR-T Therapy for Children and Young Adults With Relapsed or Refractory Acute Lymphoblastic Leukemia. <i>BMC Med Res Methodol</i> . 2019 Sep 2;19(1):182.
3	Pennington M, Grieve R, van der Meulen J, Hawkins N. Value of External Data in the Extrapolation of Survival Data: A Study Using the NJR Data Set. <i>Value Health</i> . 2018 Jul;21(7):822-829.

3.2.5 Synthesis

The types of methods to be identified via the SLR were expected to have been presented in a non-standard manner. Some studies may have opted to present plots of extrapolated survival estimates over the modelled time horizon, whereas others may have instead opted for tabulated summaries of mean survival estimates or simply equations.

For each method, a short explanation was prepared, supported where appropriate with a diagram. If the original study publication included a plot of extrapolated survival estimates, this was included without alteration (pending permission from the publisher). Next, if permission to present the original image(s) was denied, or no suitable plot was presented in the original study publication[§], the next best alternative was considered to be the preparation of an example using information provided in the study publication. Finally, if the proposed method could not be simply reflected using a plot of extrapolated survival estimates over a modelled time horizon, the process of model fitting was instead presented as a flowchart, explaining the key stages.

Ultimately, it was deemed useful for understanding each method to present the overall approach as a diagram, and so where feasible a diagram was either extracted or produced to illustrate each method.

3.3 Results

3.3.1 First iteration

Number of identified studies

The search strategies presented in Table 3-5 (Ovid) and Table 3-6 (WOS) were run on 1 October 2020. Searches were run to the latest date available as of the search date for each database. The date limits for each database are provided below:

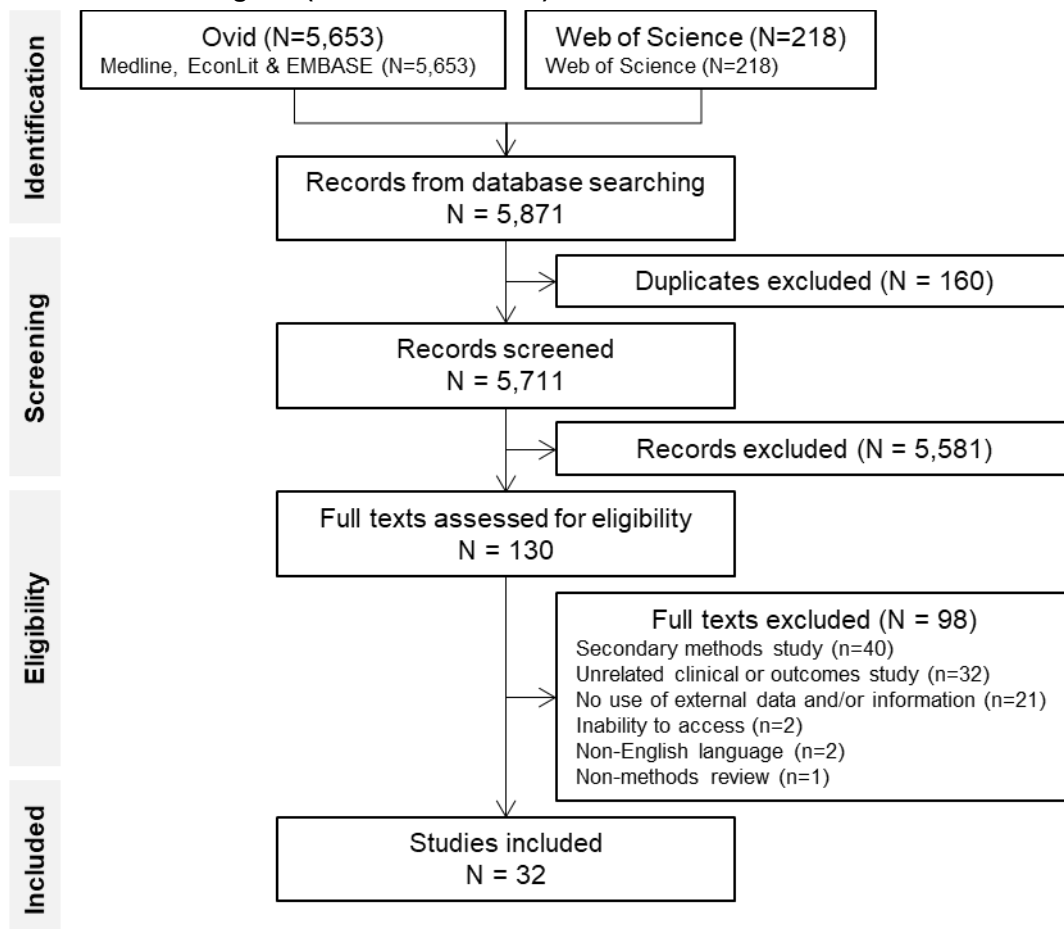
- Ovid
 - Econlit <1886 to 24 September 2020>,

[§] In this context, a plot was considered 'suitable' if it was of reasonable quality (e.g., plotted survival curves were easily distinguishable if plotted on the same axes), and it showed estimates of OS extrapolated over a relevant time horizon.

- Embase <1974 to 30 September 2020>,
- Ovid MEDLINE® and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions® <1946 to 29 September 2020>
- WOS
 - WOS (All databases, later filtered) <1864 to 1 October 2020>

For the first iteration, a total of 5,653 records from the Ovid search were identified, with an additional 218 records from the WOS search, giving a grand total of 5,871 records prior to de-duplication. After de-duplication (using the built-in functionality of EndNote), 160 duplicate records were removed leaving a total of 5,711 records for primary screening. After primary screening based on titles and abstracts, 5,581 records were excluded, leaving a total of 130 records eligible for secondary screening. After full-text screening, 98 records were excluded, which resulted in a total of 32 records suitable for data extraction.^{91, 144-174} Note that the 32 records include the original pearls as identified via the JR. The process of identifying suitable studies is presented in Figure 3-3, as a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram.¹⁷⁵

Figure 3-3: PRISMA diagram (after first iteration)



The original number of records (n=5,711 after de-duplication) for this search is relatively high compared to the number of records considered eligible for secondary screening (n=130), meaning that only ~2.3% of hits were assessed based on the full texts. While this proportion may appear low at face value, this was anticipated given the need to search based on a broad range of key terms to avoid missing potentially relevant hits if a narrower search strategy was employed – in other words, preference was placed on recall rather than precision.

Tabulation of identified methods

Based on the information extracted from each of the identified method studies, three main themes were identified:

- Piecewise methods: These methods involved use of external evidence to inform extrapolations only after a specified point in time.
- Methods based on the use of informative prior(s): Several methods included specification of informative priors within a Bayesian framework as a means of using external evidence.
- General population adjustment methods: These methods incorporated information relating to general population survival as the source of external evidence.

In addition, some methods used other approaches either alone or with one (or more) of the categories above. These methods were grouped under a theme called ‘other methods’ (though the same method could be grouped under multiple themes). Further explanation of these methods is provided when discussing each method in turn (see ‘Presentation of individual methods’), and full data extraction tables are presented in Appendix C.

Individual methods were considered on a case-by-case basis to establish whether or not they should be taken forward for further evaluation in later chapters of this thesis. The reasons for inclusion or exclusion for further consideration are provided in the ‘Presentation of individual methods’ sub-section, and for brevity a condensed summary of the methods taken forward and those not taken forward are provided in the following sub-sections.

Methods taken forward for further evaluation in later chapters

A tabulated summary of the methods that are considered further is provided in Table 3-10. In total, 12 distinct methods were taken forward, reported in 17 studies. Relevant groupings were also highlighted for each method summarised in Table 3-10. Further details are each method are presented in the next sub-section.

Table 3-10: Methods identified in first iteration taken forward for further evaluation in later chapters

Reference	Method(s)	Theme(s)	Rationale for inclusion
Lambert <i>et al.</i> , (2005) ¹⁵⁶ , Lambert <i>et al.</i> , (2007) ¹⁵⁸ , Lambert <i>et al.</i> , (2010) ¹⁶⁰ , Andersson <i>et al.</i> , (2013) ¹⁶³ , Jakobsen <i>et al.</i> , (2019) ¹⁶⁵	Relative survival models, considering any of the following approaches to measure disease-specific survival: • Standard parametric models • Mixture-/ non-mixture-cure models • Spline-based models • Fractional polynomial models • Mixture of two Weibull models	• General population	Each of these five studies build upon a 'standard' relative survival model (see Chapter 2) by considering extensions for fractional polynomials (Lambert <i>et al.</i> , [2005]), cure-based models (Lambert <i>et al.</i> , [2007]), mixtures of two Weibull models (Lambert <i>et al.</i> , [2010]), spline-based models (Andersson <i>et al.</i> , [2013]), or a combination of both (Jakobsen <i>et al.</i> , [2019]). These extensions to a relative survival model, as well as a standard parametric relative survival model, will be taken forward as applications of relative survival methodology using established survival extrapolation techniques (i.e., approaches listed in the left-hand column that have been used extensively in HTA).
Demiris & Sharples, (2006) ¹⁵¹	Bayesian relative survival model	• Informative prior • General population	This method is a Bayesian relative survival model, which makes use of data from multiple sources (illustrated via a worked example). The method is included as an extension of a 'standard' relative survival model, making use of multiple evidence sources.
Nelson <i>et al.</i> , (2008) ¹⁶⁹	Age-based piecewise model	• Piecewise	This approach provides a novel means of estimating a model which relies on patient age to predict long-term survival, as opposed to basing extrapolations on observed survival times. While this approach is somewhat linked with general piecewise approaches, it provides a unique means of estimating survival through basing extrapolations on a non-time measure beyond a given cut-point.
Hawe <i>et al.</i> , (2014) ¹⁷¹	Non-Bayesian specification of shape parameter from external data	• Other	The authors of this study make use of a specific parameter from a model fitted to an external data source, which represents a non-Bayesian means of making use of 'prior' information. This approach is included as a pragmatic, simplified approach to making use of external evidence.
Benaglia <i>et al.</i> , (2015) ¹⁵²	Poly-hazard model	• Informative prior • General population	The poly-hazard model has similarities with a relative survival approach, where multiple hazards (i.e., disease-specific and non-disease-specific) are combined to represent the overall hazard of death. However, in this approach, the model itself includes parameters for the contributing hazards, rather than making a <i>post-hoc</i> adjustment to a model fitted to all-cause mortality. Therefore, this approach is included as an alternative means of considering multiple hazards that make up an overall hazard.

Reference	Method(s)	Theme(s)	Rationale for inclusion
Larkin <i>et al.</i> , (2015) ¹⁴⁶ , Barlev <i>et al.</i> , (2017) ¹⁴⁷ , Fisher <i>et al.</i> , (2017) ¹⁴⁸	'Standard' piecewise model	Piecewise	A 'standard' piecewise model was identified in several studies with some minor variations, but ultimately the core mechanics of this approach was consistent. Though there are different variations of the overarching method, each study assumes conditional survival after a given cut-point (close to end of follow-up in study of interest) can be inferred from an external source. This approach represents a commonly used method to incorporate external evidence into extrapolations of overall survival.
Guyot <i>et al.</i> , (2017) ¹⁵³	Bayesian multi-parameter evidence synthesis	- Piecewise - Informative prior - General population - Other	The approach taken by the authors is one of the only approaches identified that attempts to take into consideration multiple sources of information via a multi-parameter evidence synthesis. It is included on the basis of it incorporating multiple sources of external evidence as part of one cohesive analysis.
Hwang <i>et al.</i> , (2017) ¹⁷²	Rolling extrapolation	General population	This approach builds upon the authors previous 'Monte Carlo extrapolation' method (see Table 3-11). While complex and linked to a piecewise/ relative survival approach, it represents an alternative means of making use of data for a reference population to extrapolate survival, and so is included for completeness.
Pennington <i>et al.</i> , (2018) ¹⁷³	'Summary data only' piecewise model	Piecewise	If only very limited external evidence is available (for example, no Kaplan-Meier estimates of overall survival, but reported point estimates of overall survival), many of the methods identified could not be considered. However, this method allows for consideration of a single point estimate (i.e., a summary measure of survival) to inform long-term survival estimates, and so is included for consideration under such circumstances.
	Parametric model with covariate(s) based on external data, with or without HR adjustment	Other	In this approach, patient-level data from multiple sources are combined through fitting parametric models with covariates. This method is included in addition to the other approach proposed by Pennington <i>et al.</i> as a novel approach that could be considered under certain circumstances.
Cope <i>et al.</i> , (2019) ⁹¹	Expert opinion Bayesian model	- Piecewise - Informative prior	This method was the only approach identified in which the only external evidence used is expert opinion. The method itself could be described as a piecewise approach, but owing to the Bayesian method of analysis, is considered a standalone method in its own right (and is discussed in further detail later in this chapter).
Vickers (2019) ¹⁷⁴	'Modified' piecewise model	- Piecewise - Other	Only one 'method' of the six approaches explored by Vickers in this study can be considered truly novel in the context of this review, which is a specific type of piecewise modelling approach with several adjustments related to treatment effects and long-term outcomes. This approach is included separately to the 'standard' piecewise model for completeness (and is discussed in further detail later in this chapter).

Methods not taken forward

A tabulated summary of the methods that are not considered further is provided in Table 3-11. Method studies that were discarded are presented separately in Appendix D (with reasons provided for why each method was discarded), to keep the main text focused on the methods most relevant to the thesis as a whole. Unlike the methods taken forward, relevant groupings were not determined since not all of these methods could be grouped. In addition, while the methods could have each been excluded on the basis that they did not meet the specific inclusion/exclusion criteria of the search, further details are provided for completeness.

Table 3-11: Methods identified in first iteration not taken forward but included in the review

Reference	Method(s)	Rationale for exclusion
Turner <i>et al.</i> , (1982) ¹⁶⁷	Bounded survival	The bounded survival model aims to address a limitation of standard parametric models, which is that they can (without adjustment) project a non-zero proportion of patients to still be alive beyond a realistic life expectancy. However, I was unable to reproduce the example provided by the authors through simple substitution. In addition, the concept of an upper limit on survival is routinely considered within contemporary cost-effectiveness analyses making use of a simpler 'cap' on the maximum plausible survival time. This is because nearly all models are not capable of predicting exactly 100% of patients to have died within a 'normal' life expectancy time period because of their mathematical properties (e.g., an exponential model can never hit zero because a fixed proportion is removed per unit time).
Hwang & Wang, (1999) ¹⁶⁸	Monte Carlo extrapolation	The authors of this study later developed the 'rolling extrapolation' method which addresses some limitations of the Monte Carlo extrapolation method. More specifically, the Monte Carlo method imposes the assumption of a constant (proportional) excess hazard, and the regression model fitted to estimate the ratio of OS between the index and reference populations must be linear. The Monte Carlo method was therefore excluded from consideration in the remainder of this thesis on the basis of the rolling extrapolation method effectively superseding this approach.
Messori & Trippoli, (1999) ¹⁴⁴	Piecewise extrapolation assuming cure	This method relies on the assumption that patients are effectively 'cured' if they survive longer than a given time point. This assumption is unlikely to hold in other situations and may not truly hold within the context considered by the authors of this study. Elements of the approach (e.g., making use of general population statistics) feature in other piecewise approaches that are considered further. However, this approach was excluded in favour of the other approaches as this considers a special case of another more flexible approach.
Heinävaara & Hakulinen, (2002) ¹⁵⁴	Parametric mixture model for multiple cancers	The focus of this study was placed on comparing estimates of cancer-specific survival based on patients with multiple cancers. The method does not represent a novel extension to relative survival analysis as a framework to inform estimates of overall survival in general. Rather, it provides an extension to consider relative survival of patients with multiple cancers, allowing for a comparison across groups.
Maetani <i>et al.</i> , (2004) ¹⁵⁵	Competing risks Boag model	This study presents an application of the Boag model (a lognormal mixture-cure model) to estimate disease-related survival and combines this with a background mortality curve. This represents a relatively simple application of a cure model in a relative survival framework. In addition, the assumption of cure is highly context-dependent, and is unlikely to hold across all possible applications. The key elements of this method (i.e., the use of a cure model with a relative survival framework) are considered further, but the application of this specific method (i.e., a Boag lognormal mixture-cure model to estimate disease-related survival) is not.
Stare <i>et al.</i> , (2005) ¹⁵⁷	Outcome measure for individualised relative survival	The outcome measure proposed by Stare <i>et al.</i> gives an alternative means of exploring relative survival, with a focus placed at the individual level. Therefore, as the focus of the thesis is on estimating survival for a cohort of patients for decision-making, this particular method does not meet the criteria to be defined as a novel method to extrapolate estimates of OS and is therefore not considered further.

Reference	Method(s)	Rationale for exclusion
Viscomi <i>et al.</i> , (2006) ¹⁴⁵	Piecewise calculation of mean life expectancy	The authors of this study were interested in the ability to estimate mean life expectancy, rather than a survival curve. Therefore, the proposed method does not provide information concerning the proportion of patients expected to be alive at specific time points. As such, the method is not considered further owing to its inability to provide information concerning the full distribution of expected survival for a cohort.
Pohar Perme <i>et al.</i> , (2009) ¹⁵⁹	Expectation-maximisation algorithm for fitting an additive hazards model	The approach presented in this study is not a unique method to extrapolate overall survival, as this method offers an alternative means of obtaining the parameters of an additive hazards model (a pre-existing method).
Andersson <i>et al.</i> , (2011) ¹⁶¹	Flexible parametric relative survival models with a cure fraction	The approach used in this study is superseded by a later study by Jakobsen <i>et al.</i> , (2019). ¹⁶⁵ Jakobsen <i>et al.</i> found that the method of Andersson <i>et al.</i> appeared to generally yield more biased estimates when compared to other (similar) approaches. Therefore, the other methods explored by Jakobsen <i>et al.</i> were assumed to supersede the approach proposed by Andersson <i>et al.</i>
Ishak <i>et al.</i> , (2011) ¹⁷⁰	Crossover adjustment using external data and predictive equations	The authors propose the use of external evidence to address the issue of crossover. The method proposed therefore has a very specific application and would not normally be considered when considering survival extrapolation in general. In addition, the approach used by the authors considers overall survival to comprise of sub-components, sub-divided by the time of disease progression (i.e., survival time until progression and after progression). The estimation of OS as a combination of interim clinical events is beyond the scope of the review, as such approaches necessitate the determination of a clinically relevant event.
Hinchcliffe <i>et al.</i> , (2012) ¹⁶²	Adjustment for deaths in general population within estimation of relative survival	The authors of this study consider adjustments to relative survival analysis to account for there being a hypothetically large proportion of deaths due to the cause under consideration in the general population (i.e., to avoid double counting). However, the approach itself is not a novel means of incorporating external evidence into estimates of overall survival and so is excluded from further consideration. In addition, the majority of HTAs of cancer treatments consider a relatively small population of patients (accounting for specific treatment lines and disease stage), and so adjusting for a large proportion of deaths observed in the general population is unlikely to be consequential for decision making.
Seppä <i>et al.</i> , (2014) ¹⁶⁴	Cure fraction model with random effects to explain regional differences	This method is based on reflecting potential differences by patient subgroups (in this case, according to regional variation in relative survival). While findings are described as potentially important from a public health perspective, it is extremely unlikely that regional variation in relative survival would be possible to consider within the context of estimating OS for HTA (owing to the sample size of studies typically available to inform decision making), and so this method is not taken further.

Reference	Method(s)	Rationale for exclusion
Thornton Snider <i>et al.</i> , (2017, 2018) ^{149, 150}	Method to consider the option value of new therapies	Application of the method requires several assumptions to be made, which are linked both to the expected difference in treatment effects for given therapies but are also linked to the expected uptake of new treatments. The method has limited applicability outside of the context considered by the authors, and the latter assumptions related to market uptake are beyond the scope of extrapolation exercises typically required for contemporary HTA.
Rubio <i>et al.</i> , (2019) ¹⁶⁶	Correction for misspecification of background mortality rates in relative survival analysis	The authors of this study focus on adjusting background mortality rates to account for differences in the population under consideration and available life table information to inform the estimation of background mortality. However, the approach itself is not a novel means of incorporating external evidence into estimates of OS and so is excluded from further consideration.

Presentation of individual methods

In the sub-sections that follow, each of the identified method studies that were taken forward for the remainder of the thesis are discussed.

Method studies are presented in chronological order, with studies published in the same year presented in alphabetical order based on the surname of the lead author. Individual references are provided in the first sentence following the sub-heading for ease of cross-referencing and formatting.

Within the presentation of each method study, key assumptions are presented, alongside strengths and limitations, as well as ultimately the rationale for taking the method forward for further consideration in later chapters of this thesis. The assumptions presented when describing each method are specific to each method, and so general assumptions that may apply across several (or even all) methods are not included – for example, the assumption of non-informative censoring is relevant to all the methods considered involving censored survival data. Common assumptions underpinning survival analysis methods are described in Chapter 2.

Lambert *et al.*, (2005)

Grouping(s): General population.

Note: This is one of five methods that describe different applications of a relative survival approach (the others being Lambert *et al.*, [2007]¹⁵⁸, Lambert *et al.*, [2010]¹⁶⁰, Andersson *et al.*, [2013]¹⁶³, and Jakobsen *et al.*, [2019]¹⁶⁵). Please refer to Chapter 2 for a description of how a relative survival approach can be used in general when producing estimates of OS.

Lambert *et al.* presented two extensions of generalised linear models (GLMs) to estimate relative survival as an alternative to assuming that covariate effects act multiplicatively on the excess hazard (in a PH framework).¹⁵⁶ The two extensions considered by Lambert *et al.* are:

- An *additive* covariate model, which provides estimates of the absolute difference in the excess mortality rate (instead of a *multiplicative* covariate model).
- The use of FP models within the estimation of relative survival to express the baseline excess mortality rate and time-dependent effects, considering both an additive and multiplicative covariate model.

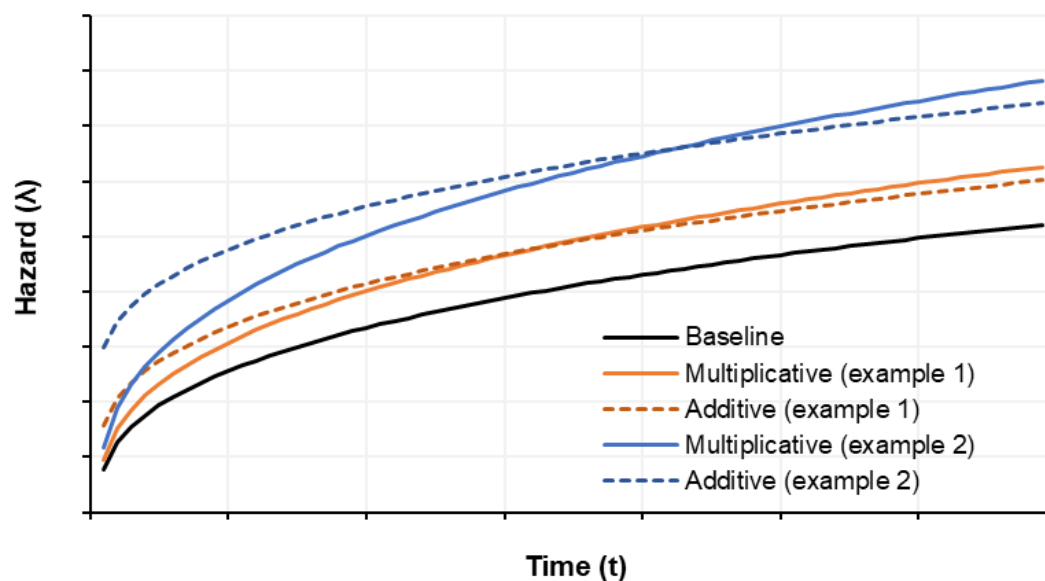
These two extensions were then combined to allow for a comparison of additive versus multiplicative FP models. Lambert *et al.* acknowledged that an additive covariate model had been used in a number of previous studies, though not always within the context of a relative survival model, and not described in such a way that it could be fitted using a GLM.

The hazard function for a given individual i , $\lambda_i(t)$, may be represented using a PH model (or a *multiplicative hazards model*), which was defined earlier in Chapter 2. An alternative to the PH (*multiplicative hazards*) model is the *additive hazards model* (also defined in Chapter 2). Both models refer to a static change to the hazard of death, but affect the resultant hazard function, $\lambda(t)$, differently, which is illustrated using two examples in Figure 3-4.

- For the *additive hazards model*, the difference in the hazard function is constant over time (i.e., the additive hazards model could be viewed as a movement of the baseline curve up the y axis).
- For the *multiplicative hazards model*, the ratio of the curves is constant over time, which in Figure 3-4 leads to the multiplicative and baseline curves diverging in the longer-term.

The two examples in Figure 3-4 show that seemingly similar results can be achieved using either approach (i.e., the same colour lines can look similar, but they are not the same), but the short- and long-term implications will differ markedly. These differences may be difficult to see if looking at estimates of survival rather than hazards.

Figure 3-4: Illustration of additive and multiplicative hazards models



The differences between the multiplicative and additive models are summarised in Table 3-12.

Table 3-12: Multiplicative versus additive relative survival models

	Multiplicative model	Additive model
Effect of covariates	Covariates have a multiplicative effect, meaning that they can be interpreted as having a relative effect on excess mortality	Covariates have an additive effect, meaning that they can be interpreted as absolute differences in excess mortality
Formula for total mortality rate: $\lambda(t; x_1)$	$\lambda^*(t; x_2) + \exp(\beta x_1)$ x_1 : covariate pattern in the study population; $\lambda^*(t; x_2)$: expected mortality rate in the general population at time t for covariate pattern x_2 (usually a subset of x_1); β : model coefficients	$\lambda^*(t; x_2) + \beta x_1$

One critical advantage of an additive covariate model within the context of relative survival was highlighted by Lambert *et al.*, namely that the excess mortality rate could be modelled as zero or even negative. This could be of relevance if patients are ultimately cured of their disease and subsequently adopt a ‘healthier’ lifestyle.¹⁵⁶

FP models are described earlier in Section 2.3.2. However, Lambert *et al.* considered the use of FP models to model relative survival, and more specifically focused on the use of third-order FP (FP3) models “because it is important to model the baseline excess mortality rate correctly as this is the reference category for covariate effects, and an FP3 model should provide increased flexibility over standard parametric models” (Lambert *et al.*, 2005, p.3877).¹⁵⁶ An FP3 model is defined as:

$$\ln(h(t)) = \begin{cases} \beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_2} + \beta_3 t^{p_3}, & p_1 \neq p_2 \neq p_3 \\ \beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_2} (\ln(t)) + \beta_3 t^{p_3}, & p_1 = p_2 \neq p_3 \\ \beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_2} (\ln(t)) + \beta_3 t^{p_3} (\ln(t))^2 & p_1 = p_2 = p_3 \end{cases} \quad \text{where } t^0 = \ln(t)$$

The authors present a motivating example, using data collected for all England and Wales cancer registrations between January 1971 and December 1990, with follow-up until 31 December 1995. The authors then restrict this data set to a total of 115,331 female breast cancer patients diagnosed between 1 January 1986 and 31 December 1990, and fit additive and multiplicative FP models for excess mortality.

To use the FP model to estimate excess mortality rates, the authors split the time scale of available data into specific intervals:

- For the first two months, time intervals of one week were used: this gave 8x 1-week intervals.
- After two months until one year, time intervals of two weeks were used: this gave 22x 2-week intervals.
- After one year, time intervals of 1 month were used for the remaining time horizon: this gave 48 x 1-month intervals.

Therefore, a total of 78 intervals were obtained for follow-up until 5 years. Then, two different FP3 models were fitted, adopting either a multiplicative or additive covariate structure:

- A *multiplicative* covariate FP3 model:
 - In this model, the log baseline excess mortality rate was modelled with time-dependent effects expressed as log mortality rate ratios.
- An *additive* covariate PF3 model:
 - In this model, the absolute baseline excess mortality rate was modelled with time-dependent effects expressed as the difference in excess mortality rates.

Key assumption(s)

The key assumption made within each of the different types of relative survival modelling approach taken are whether hazards are assumed be additive or multiplicative. This has implications for the interpretation of how disease-specific and other-cause hazards interact – for example, as highlighted for the additive model, excess mortality (i.e., disease-specific hazard) could be modelled as zero or even negative, but this cannot be modelled using multiplicative hazards. Furthermore, even though the method assumes that the overall excess mortality rate can vary over time, linear time-dependent effects were assumed for the deprivation groups (i.e., within each model fitted, each deprivation group was assigned a ‘dummy’ term and a linear term for time).

Strengths and limitations

The main advantage afforded by the methods presented is that ultimately a better fit could be obtained by specifying a more flexible model to represent excess hazards. However, in order to realistically fit these types of relative survival models, a very large data set is required in order to capture potential differences in outcome according to deprivation group. Furthermore, in order to fit the FP models, the authors needed to split the follow-up data into a number of different time intervals, requiring assumptions to be made in terms of where sensible delineations can be made to separate the data appropriately.

Rationale for inclusion in further investigation in the rest of the thesis

This method was proposed to allow for a more detailed exploration of deprivation effects, which are not usually considered within the context of OS extrapolation for HTA due to limitations in available data. Furthermore, the data set considered in the motivating example comprised of over 100,000 patients with breast cancer – a substantially greater number of patients than would be expected within clinical trial data typically available to inform HTA. The ability to capture time-varying effects of relative survival is expected to be challenging within the context of (relatively) small patient numbers, and so this specific application is not considered further. However, the relative survival aspect of this method (i.e., the approach to incorporating external evidence) features throughout the remainder of the thesis.

Demiris & Sharples, (2006)

Theme(s): Informative prior, general population.

In this study, Demiris & Sharples describe how additive and multiplicative hazard models can be used to synthesize information from multiple data sources to extrapolate survival estimates.¹⁵¹ Using the underlying principles of the multiplicative and additive hazards models (see description of Lambert *et al.*, (2005)), the authors undertook a Bayesian approach to synthesise data from multiple data sources. In this study, the authors present findings based on a *parametric* baseline hazard function (using a Weibull model), but also comment that a *semi-parametric* model could instead be used (which is later referred to by Benaglia *et al.* as a ‘Cox-like’ model).¹⁵²

In their motivating example, three data sources are acknowledged:

- Life table data (referred to as census data from the Government Actuary’s Department [GAD] within the original study, though these data are no longer hosted at the website provided in the study publication).
- Cohort data (in this case, IPD collected for a cohort of patients that received implantable cardioverter defibrillator [ICD] therapy. The data from this study are discussed in detail in a separate study by Buxton *et al.*, [2006]¹⁷⁶ within the context of a cost-effectiveness analysis).
- Meta-analysis output (here, this is based on a meta-analysis by Connolly *et al.*, [2000]¹⁷⁷ comparing the intervention: anti-arrhythmia drug [AAD] therapy, and the control: ICD therapy).

The original publication presents the methodology using advanced mathematical theory and notation, which may exceed the level of technical detail typically expected of health economics analysts working in HTA. For example, the prior processes for part of the semi-parametric model (‘Cox-like’) specification is described as follows (square brackets indicating references):

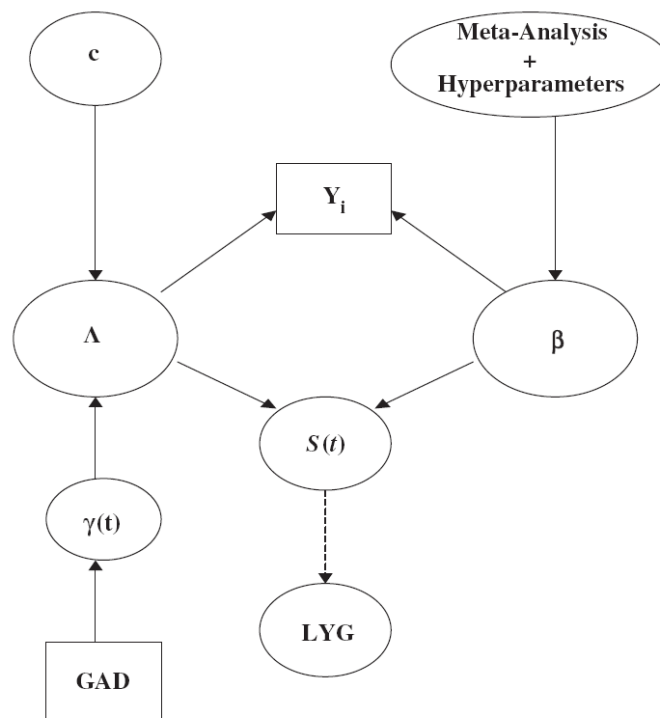
“We used the same $N(0, 10^4)$ prior for elements of β as in the parametric model. Independently of $\pi(\beta)$, we adopted the approach used by a number of authors that involves a neutral to the right [13] process for the cumulative hazard. We now describe a popular special case, the gamma process [14]. Let $Ga(a, b)$ denote the gamma distribution with shape $a > 0$ and scale $b > 0$. Let $\gamma(t)$ ($t \geq 0$) be an increasing (possibly with jumps) function such that $\gamma(0) = 0$ and let $U(t)$ ($t \geq 0$) be a stochastic process with the following properties: (i) $U(0) = 0$, (ii) the increments of $U(t)$ over disjoint intervals are independent and (iii) $U(t) - U(s) \sim Ga(c[\gamma(t) - \gamma(s)], c)$, $t > s$. Then the process $U(t): t \geq 0$ is called a gamma process and is denoted with $U(t) \sim \mathbb{GP}(c\gamma(t); c)$, where $\gamma(t)$ represents the mean and c can be thought of

as a precision parameter. For example, at time t a value of $c = 1$ sets the standard deviation to be equivalent to $\sqrt{\gamma(t)}$ and a value of $c = 100$ sets the standard deviation to $\sqrt{\gamma(t)}/10$. Other approaches exist for the choice of the parametric family of the increments, see Reference [3] for a review, including Hjort's [15] beta process (beta distributed increments) prior and the extended Gamma process prior, see References [16, 17] for more details." (Demiris & Sharples, 2006, p.1,967-1,968).¹⁵¹

The derivation of the baseline hazard function is not described further here as it is not critical to demonstrating how external evidence can be incorporated into the extrapolation of survival, but further details are presented in the study by Demiris & Sharples.¹⁵¹

The authors present a graphical model summarising how each of the data sources and parameters interact with each other, as shown in Figure 3-5.

Figure 3-5: The graphical model, as presented by Demiris & Sharples, (2006)



Key: c , precision parameter; GAD, Government Actuary Department; LYG, life-years gained; $S(t)$ survival function; Y_i , cohort data, β , covariate coefficients, Λ , underlying hazard, $\gamma(t)$, baseline hazard function.

Source: Excerpted and reprinted with permission from Statistics in Medicine. (Copyright © 2005, John Wiley and Sons). Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. Stat Med. 2006 Jun 15;25(11):1960-75.

The graphical model illustrates that the model is centred around the estimation of the cumulative baseline hazard function, Λ , which is influenced by both the cohort data (Buxton *et al.*, [2006]¹⁷⁶) and background mortality (GAD) data. The GAD data inform a general population hazard function, $\gamma(t)$, which contributes to the estimation of Λ , allowing the model

to account for expected survival (i.e., non-disease specific mortality). The cumulative baseline hazard, Λ , is modelled using a gamma process prior (as described in the excerpt presented previously), with its uncertainty determined by a hyperparameter, c . This parameter represents the uncertainty around the baseline hazard function, $\gamma(t)$, and can be considered a ‘precision parameter’. Large values of c translate to strong prior information for $\gamma(t)$, whereas smaller values for c represent diminishing influence of the prior information for $\gamma(t)$. Covariate effects, β , are also incorporated into the model, informed by a meta-analysis and associated hyperparameters. Together, the cumulative baseline hazard, Λ , and the covariate effects, β , are used to construct $S(t)$. Observed individual survival outcomes, Y_i , are modelled conditional on $S(t)$ and β , forming the likelihood component of the model. Then, once $S(t)$ has been estimated, this can be used to derive the life-years gained for AAD versus ICD.

Key assumption(s)

This method assumes PH hold for all causes of death between the index and reference populations – in this example, the cohort and life table data. Consequently, when considering a parametric baseline hazard function, the same underlying parameterisation is effectively assumed to apply for both the index and reference populations. In addition, the analyst is required to set a value for the precision parameter, c , which may be difficult to justify (and so a range of values may be worthwhile exploring).

Strengths and limitations

The framework set out by Demiris & Sharples provides a range of approaches that could be undertaken to incorporate prior information for the baseline hazard function, with alternative assumptions that “*can act as an informal sensitivity analysis*” (Demiris & Sharples, 2006, p.1973).¹⁵¹ As the proposed method adopts a Bayesian approach, all relevant data are combined within a single analysis.

The main limitation of this method is linked to the assumption of PH for death from all causes, acknowledging that in the long-term it may not be possible to test this assumption. The example considered by the authors refers to survivors of ventricular fibrillation or sustained ventricular tachycardia, as opposed to populations of patients initiating therapy following diagnosis or progression; and so an assumption of PH indefinitely may not be suitable for other populations that could receive additional therapies and/or experience rapid disease progression leading to complex hazard functions.

Rationale for inclusion in further investigation in the rest of the thesis

The method presented in this study is a frequently cited Bayesian approach to combine prior information as part of survival extrapolation linked to the general population, with a particular

focus on cost-effectiveness analysis – i.e., a type of Bayesian relative survival model. It is therefore included as a recognised method to incorporate external evidence into estimates of survival.

Lambert et al., (2007)

Theme(s): *General population.*

Lambert *et al.* present an extension to the NMCM method to allow for fitting such models in a relative survival framework.¹⁵⁸ Details of cure models are provided in Chapter 2 (see Section 2.3.2), and further context concerning relative survival aspect is captured in the discussion of Lambert *et al.*, (2005).¹⁵⁶

Key assumption(s)

The main assumption made within this analysis (a type of NMCM) is aligned to that of standard cure-based models; namely, that there exists a proportion of patients that ultimately do not experience and excess hazard of death linked to the disease of interest.

Strengths and limitations

As per the key assumptions, the main strengths and limitations of the approach presented in this study are essentially a combination of the respective features of both relative survival models and cure-based models (see discussion of Lambert *et al.*, (2005) for commentary related to relative survival methods, and Chapter 2 [Section 2.3.2] for commentary related to cure-based models). Accordingly, these are not presented again here for brevity.

Rationale for inclusion in further investigation in the rest of the thesis

This study builds upon a ‘standard’ parametric relative survival model by considering extensions for cure-based models. This extension, in addition to different extensions presented by Andersson *et al.*, (2013)¹⁶³ and Jakobsen *et al.*, (2019)¹⁶⁵, as well as a standard parametric relative survival model, are taken forward as applications of relative survival methodology using established survival extrapolation techniques.

Nelson et al., (2008)

Theme(s): *Piecewise.*

Nelson *et al.* describe a method to estimate the life expectancy (LE) of patients enrolled within large clinical trials.¹⁶⁹ Two methods are described in this study to inform longer-term survival extrapolation:

1. Use of an external data source to extrapolate beyond a ‘stability point’ for the cohort of interest, via an age-based model (‘External base case’).

2. Use of the observed experience of the cohort of interest to extrapolate beyond a 'stability point', via an age-based model ('Internal base case').

Here, a 'stability point' is defined as a point in time after which the hazard of dying can be described as a function of an individual's age and baseline covariates, regardless of the individual's age at baseline. The process of undertaking either method is similar, apart from the process to inform the longer-term extrapolation (i.e., whether external or internal data are used to produce the age-based model after the stability point). Based on the scope of this review, only the former method ('External base case') is considered further.

The key aspect of the method proposed by Nelson *et al.* is that long-term survival extrapolation can be estimated through the specification of a model which estimates the hazard of death as a function of a patients' *age*, rather than the *time elapsed since study entry*. The method may be considered a piecewise approach, given the development of a study model and a separate extrapolation model. However, given the approach to focus on different metrics for each model (time for the study model, age for the extrapolation model), it is described separately here.

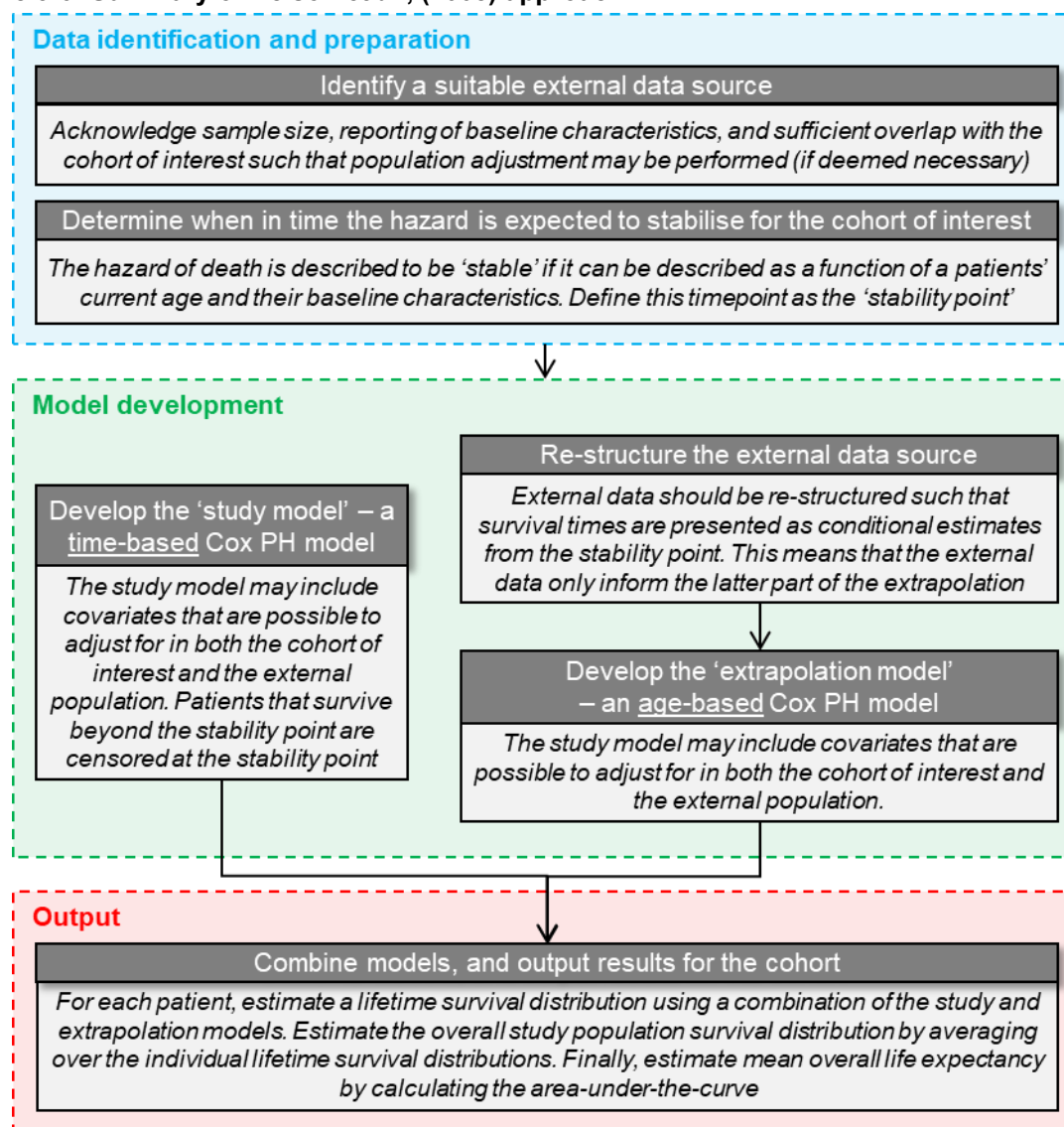
The authors consider an application of the external data method within the context of patients that have experienced a myocardial infarction (MI), taking data from two key sources:

- The GUSTO-I trial.¹⁷⁸ Data for 23,105 patients studied within GUSTO-I were considered in the authors' analysis, where this population represented the cohort of interest for survival extrapolation.
- The Duke Databank for Cardiovascular Disease (DDCD).^{179, 180} Data for 4,379 patients were identified from this database, for whom baseline characteristics were similar to GUSTO-I provided they had survived for a sufficient time period to be considered in the analysis.

In their application, the authors define the stability point to be at 1 year for three reasons: (1) the existence of a previous economic evaluation which used data available up until 1 year before extrapolation, (2) that it is commonplace to make use of "*all available primary trial follow-up*" (which is expected to refer to the aforementioned 1-year data), and (3) while the hazard may have stabilised prior to 1 year, it is considered "*conservative*" to select a later time point if sufficient follow-up data are available (Nelson *et al.*, 2008, p.5541).¹⁶⁹

A step-by-step process is presented by Nelson *et al.*, summarised in Figure 3-6 (for the 'External base case' analysis).

Figure 3-6: Summary of Nelson *et al.*, (2008) approach



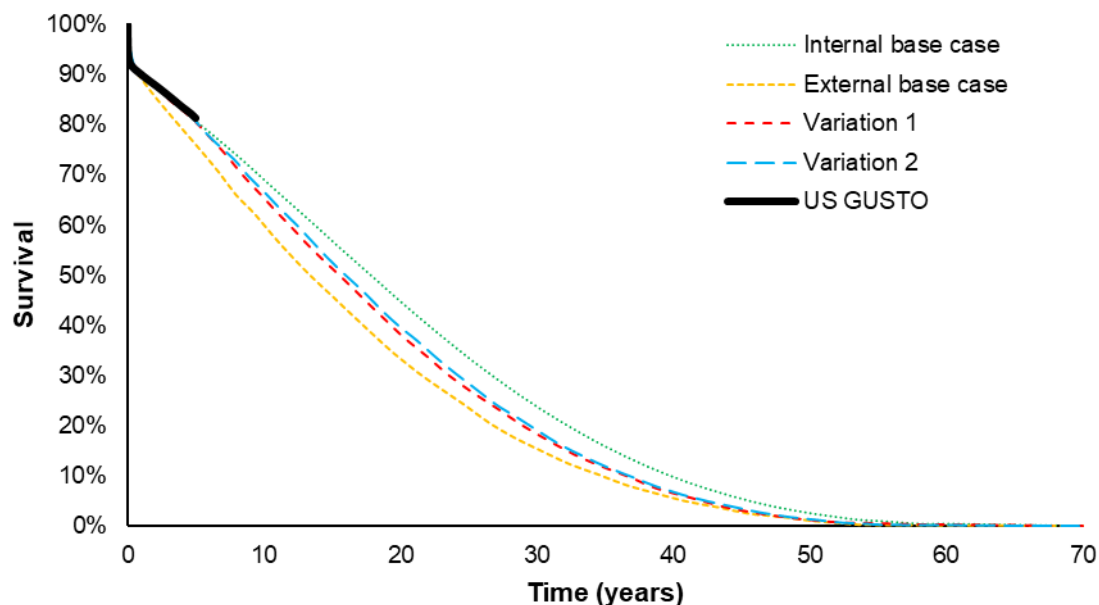
Key: PH, proportional hazards.

An illustration of the model fitted using external data by Nelson *et al.* is presented in Figure 3-7, alongside the internal base case and two other alternative models considered. The two alternative models are described as 'variations' on the internal and external base-case approaches, which include adjustments based on HRs:

- 'Variation 1' uses the GUSTO-I age-based extrapolation model to extrapolate survival between 1 and 5 years, but beyond 5 years the underlying hazard estimated via the GUSTO-I age-based extrapolation model is *increased* using an HR of 1.28 to consider a 'period effect' seen in DDCCD data.
 - In other words, an increased risk of death after 5 years in the age-based model still seen after adjustment versus before 5 years.

- In this variation, the external evidence is effectively limited to the specification of an HR to describe the differences in the age-based hazard of death before and after 5 years (assuming a similar pattern would apply for the GUSTO-I data, were follow-up data available beyond 5 years).
- 'Variation 2' uses the GUSTO-I age-based extrapolation model to extrapolate survival from 1 to 5 years and the DDCCD age-based extrapolation model for survival beyond 5 years. However, an HR of 1.70 is applied to *decrease* the underlying hazard projected by the DDCCD model to account for the better survival experience in GUSTO-1.
 - This approach 'uplifts' the DDCCD extrapolation based on the expectation that GUSTO-I patients will experience better survival outcomes in the longer term.

Figure 3-7: Predicted survival re-produced from Nelson *et al.*, (2008)



Note(s): This figure illustrates the estimates of overall survival using a range of alternative approaches by Nelson *et al.*¹⁶⁹ The solid, heavy-weighted line represents the Kaplan-Meier estimate of survival from the GUSTO-I study. The green, dotted line represents the 'Internal base case', which is not discussed further as it does not make use of external evidence. The yellow, dashed line represents the 'External base case', for which the estimation process is described in Figure 3-6. The dashed, red line represents 'Variation 1', and the dashed, blue line represents 'Variation 2'. Both of these variations make use of hazard ratio adjustments to account for potential differences in longer-term age-based extrapolations between the GUSTO-I and Duke Databank for Cardiovascular Disease cohorts.

Source: Copyright © 2008 Nelson *et al.* This figure was digitised and re-created for inclusion in this thesis. The original image is presented in Nelson CL, Sun JL, Tsiatis AA, Mark DB. Empirical estimation of life expectancy from large clinical trials: use of left-truncated, right-censored survival analysis methodology. *Stat Med.* 2008 Nov 20;27(26):5525-55.

Key assumption(s)

The main assumption made within this analysis is that for the aspects of the approach that rely on an age-based model, the hazard of death can be reliably estimated as a function of the patients' baseline and 'current' characteristics (i.e., age). If this assumption holds, this means that an age-based model could be derived from a larger data set to then inform long-

term survival. For the variations presented, any differences in hazards between groups were assumed to be constant over time (i.e., PH assumed to hold).

Strengths and limitations

As highlighted by the study authors, the main strength of the proposed analysis is the ability to make efficient use of potentially limited IPD (in terms of follow-up) by representing hazards as a function of *age* rather than *time*. By using age-based extrapolations, provided all necessary assumptions hold, extrapolations are largely based on empirical data, as opposed to assumptions linked to parametric modelling.

However, it is precisely this feature of the approach that comprises its main limitation – in order for the method to be suitable, it is necessary to robustly estimate an age-based extrapolation model. This requires a substantial number of patients across a range of age groups, and even if data exist it may not be appropriate to assume that long-term hazards are dependent only on current age and baseline characteristics (and importantly *not* time elapsed from baseline). Indeed, as illustrated through the variations presented by the authors, there is evidence even within the worked example that an age-based extrapolation may not be appropriate to consider without adjustment (as inspection of the DDCCD data suggested a higher risk of death by age for patients with a follow-up time after 5 years versus before 5 years [see Table IV of the Nelson *et al.* study]).¹⁶⁹

Rationale for inclusion in further investigation in the rest of the thesis

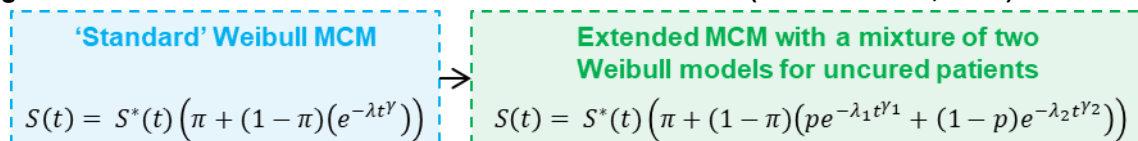
This method is included as a novel means of extrapolating survival outcomes using an age-based metric.

Lambert *et al.*, (2010)

Theme(s): *General population.*

Separately to the Lambert *et al.*, (2007) study concerning the use of NMCMs, in 2010, Lambert *et al.* published a study concerned with the use of a mixture of two Weibull models to reflect the survival experience of the uncured proportion of patients included within both a MCM and an NMCM.^{158, 160} The resultant formulae for the MCM using a mixture of two Weibull models compared with a standard Weibull MCM is provided in Figure 3-8.

Figure 3-8: Extension of MCM to include two Weibull models (Lambert *et al.*, 2010)



Key: γ , Weibull model scale parameter; λ , Weibull model shape parameter; p , mixture parameter; π , proportion cured; $S(t)$, overall survival function; $S^*(t)$, expected survival function for the general population; t , time.

Key assumption(s)

Please refer to discussion relating to MCMs and NMCMs more broadly for further information related to the key assumptions associated with such models fitted within a relative survival framework (see discussion of Lambert *et al.*, (2005) for commentary related to relative survival methods, and Chapter 2 [Section 2.3.2] for commentary related to cure-based models). However, specifically related to a mixture of two Weibull models, this method assumes that the ‘uncured’ proportion of patients experience disease-related mortality via a mixture of two Weibull models, demonstrated via the specification of two Weibull models shown in Figure 3-8.

Strengths and limitations

The strengths and limitations of this method are broadly similar to those that apply to standard MCMs and NMCMs. Please see discussion concerning the study by Lambert *et al.*, (2007) for more information. Specific to the assumption of a mixture of two Weibull models reflecting disease-specific survival, this method may ‘over-fit’ data, but implicitly assuming that there exist two groups of patients: (i) ‘cured’ patients, (ii) ‘uncured’ patients with a prognosis explained by a mixture of two Weibull models.

Rationale for inclusion in further investigation in the rest of the thesis

This study proposes an extension of MCMs and NMCMs to include a (separate) mixture of two Weibull models to reflect excess hazard. While this method may have its uses, for the purpose of this review, it was excluded from further consideration because it was not deemed to provide a substantially different approach to using external evidence versus a standard relative survival application (with or without a ‘cure’ element). This is because the incorporation of external evidence in the proposed approach is the same as an approach which may also make use of a single Weibull model to reflect disease-related survival. Despite this, the approach to incorporating external evidence used in this method and several others (i.e., relative survival) is taken forward in the remainder of the thesis.

Andersson *et al.*, (2013)

Theme(s): General population.

Andersson *et al.*, (2013) presented an extension of Royston & Parmar (2002)^{94*} spline-based models, to consider a relative survival framework.¹⁶³ To do this, the authors consider the all-cause survival function, $S(t)$, as the product of the expected survival function, $S^*(t)$, and the relative survival function, $R(t)$. The all-cause hazard function, $h(t)$, is therefore estimated as

* Royston & Parmar, (2002) spline-based models are described further in Chapter 2.

the sum of the expected and relative hazard functions ($\lambda_0(t)$ and $\lambda_E(t)$, respectively). The relative survival function, $R(t)$, is then estimated using a spline-based model (i.e., the spline-based part represents the novel part of this method). The authors also note that covariates could be included within each of these functions. Further details on relative survival in general is provided in Chapter 2 (Section 2.1.2).

This study by Andersson *et al.* was published after the lead author's earlier work, which considers the use of flexible parametric models within the context of a cure fraction model (Andersson *et al.*, 2011).¹⁶¹ The authors also acknowledged that two previous studies by Nelson *et al.*, (2007) and Lambert & Royston, (2009) first developed spline-based relative survival models.^{181, 182} However, for the purpose of this review, the study by Andersson *et al.* was considered to both (a) cover the methods proposed in these two earlier studies, and (b) provide a sufficiently novel estimation approach, given that the authors considered three possible variations of the relative survival modelling approach, which are discussed in turn below.

The three alternative assumptions considered for extrapolation of the spline-based relative survival model were:

- Assume that the log cumulative excess hazard beyond the last boundary knot follows the linear trend given by the estimated model parameters.
 - This means that the model “*behaves like a Weibull distributed excess hazard in the tail*” (Andersson *et al.*, 2013, p.5289).¹⁶³
- Assume statistical cure beyond the last boundary knot, by imposing constraints on spline parameters (i.e., assume no excess mortality risk in the long-term).
 - This effectively means that a ‘cure point’ is set at the position of the last boundary knot, and from this point onwards excess hazard is zero.
- Assume constant excess hazard beyond the last boundary knot, by imposing a constraint on a specific spline parameter.
 - This means that the excess hazard “*behaves like an exponential distribution in the tail*” (Andersson *et al.*, 2013, p.5290).¹⁶³

The choice of which assumption to be made in the tail is largely context dependent, though the authors present scenarios using all three models fitted to the same data. However, the authors urge caution in the use of these methods in the context of limited follow-up data. More specifically, the authors recommend at least 10 years of follow-up would be needed for at least some patients in order for the extrapolations to perform well.

Key assumption(s)

As noted above, the main assumption made within this method (outside of those that apply to any cure-based model analysis) is what happens to the excess hazard beyond the last boundary knot. Please refer to discussion of other cure-based methods for further details (see Chapter 2 [Section 2.3.2]), as well further information provided in Chapter 2 concerning flexible parametric modelling methods in general (also Section 2.3.2).

Strengths and limitations

The main strength of the approach presented by the authors is that different assumptions can be made for the long-term excess hazards experienced by the population under consideration, allowing for investigation of different modelling approaches and how influential the tail of the model is on overall results. However, this could also be highlighted as an important limitation of the overall approach. Three alternatives for modelling excess hazards in the long term are presented, each of which carries its own set of assumptions, and it may be challenging to choose between each of these.

Rationale for inclusion in further investigation in the rest of the thesis

This study builds upon a 'standard' parametric relative survival model by considering extensions for spline-based models. This extension, in addition to different extensions presented by Lambert *et al.*, (2007)¹⁵⁸ and Jakobsen *et al.*, (2019)¹⁶⁵, as well as a standard parametric relative survival model, are considered different applications of relative survival methodology using established survival extrapolation techniques.

Hawe *et al.*, (2014)

Theme(s): *Other.*

Note: *This method requires data from an RCT.*

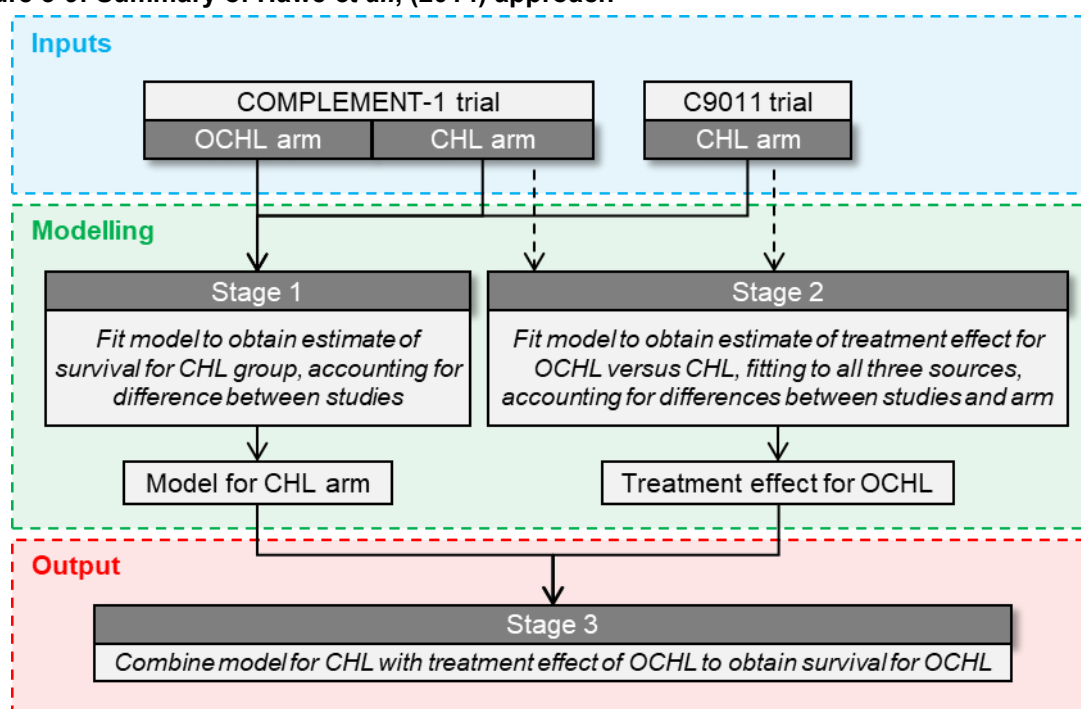
A conference abstract by Hawe *et al.*, (2014) describes an approach to make use of external data to inform long-term extrapolations of survival, in the context of patients with chronic lymphocytic leukaemia.¹⁷¹ The objective of the study was to produce extrapolations of survival for patients treated with chlorambucil + ofatumumab in the COMPLEMENT-1 trial ([NCT00748189](https://clinicaltrials.gov/ct2/show/study/NCT00748189)) that were 'guided' by long-term data available for the control arm (chlorambucil alone) in the C9011 trial (results published by Rai *et al.*, [2009]).¹⁸³ To achieve this, the authors considered a three-stage extrapolation approach (which is also illustrated in Figure 3-9):

- **Stage 1:** Fit a parametric survival model to data for patients treated with chlorambucil alone, by making use of data from the chlorambucil alone arms of COMPLEMENT-1

and C9011. This model therefore estimates survival for patients treated with chlorambucil alone, and includes a covariate to describe the differences between studies (but assumes the same shape parameter[s] and same parametric form). This stage produces a base survival model. The authors do not explicitly state which data source they deem to be most relevant for the base survival model, and instead present analyses using either option.

- **Stage 2:** Fit a parametric survival model to data for patients treated with chlorambucil alone in both studies, plus data for patients treated with chlorambucil + ofatumumab in COMPLEMENT-1. By including covariates for both study and arm, a treatment effect can be identified for chlorambucil + ofatumumab versus chlorambucil alone. This stage therefore produces an estimate of the treatment effect, through extracting the relevant parameter from the model.
- **Stage 3:** Apply the treatment effect (Stage 2) to the base survival model (Stage 1), to obtain a survival model for patients treated with chlorambucil + ofatumumab. In other words – plug in the treatment effect parameter from Stage 2 to the model from Stage 1 to obtain a new curve.

Figure 3-9: Summary of Hawe *et al.*, (2014) approach



Key: CHL, chlorambucil alone; OCHL, ofatumumab + chlorambucil.

Key assumption(s)

The primary assumption made in this analysis is that parameters obtained from each analysis are directly transferable into other analyses, and all models share the same parametric form.

More specifically, the treatment effect obtained in Stage 2 can be combined with the base survival model obtained in Stage 1 without any further adjustments to be considered. Therefore, it is implicitly assumed that the treatment effect associated with chlorambucil + ofatumumab from COMPLEMENT-1 would be hypothetically equivalent to a treatment effect in C9011 were it an RCT comparing chlorambucil + ofatumumab with chlorambucil alone.

Strengths and limitations

The main strength of the analysis lies within its ability to maximise the use of each information source based on what information it is able to provide – in this case, the COMPLEMENT-1 trial provides information on the relative effect of chlorambucil + ofatumumab versus chlorambucil alone, whereas the C9011 trial provides longer-term data but only for patients treated with chlorambucil alone. Relatedly, its main limitation is the simplistic nature in which the external evidence is included, which means that other model parameters, such as the shape parameter of the Weibull model, are completely unaffected by the specification of a different scale parameter.

Rationale for inclusion in further investigation in the rest of the thesis

This method takes a specific parameter from a model fitted to an external data source and applies this within a different model to estimate survival for the population of interest. It therefore represents a non-Bayesian means of making use of ‘prior’ information, and is included as a pragmatic, simplified approach to make use of external evidence.

Benaglia et al., (2015)

Theme(s): Informative prior, general population.

Benaglia *et al.* proposed the use of a two-component poly-hazard model to extrapolate survival, with particular focus on a poly-Weibull model (but some discussion is also presented for a poly-Gompertz model).¹⁵² In its simplest form, a poly-hazard model is a model which includes multiple hazard functions, $\lambda_1(t)$, $\lambda_2(t)$, ..., $\lambda_n(t)$, which when combined (i.e., summed) represent the overall hazard function, $h(t)$. The method proposed by Benaglia *et al.* posits that there are two potential causes of death (e.g., related to a given disease under consideration, or not), and that two data sources are available: study data (for which model fitting is ultimately being considered for), and an external data set to inform extrapolations.

The method assumes the non-disease-specific hazard of death experienced by the study population is identical to that of an external data set, whereas for the other cause of death (related to the disease), the hazards are proportional (but not necessarily the same) between the study data and an external data set. Based on these assumptions, the authors then

consider a poly-Weibull model, motivated by an example where disease-specific mortality is assumed to be proportional to an external data set, but other-cause mortality is identical.

The survival function for the poly-Weibull model is presented below:

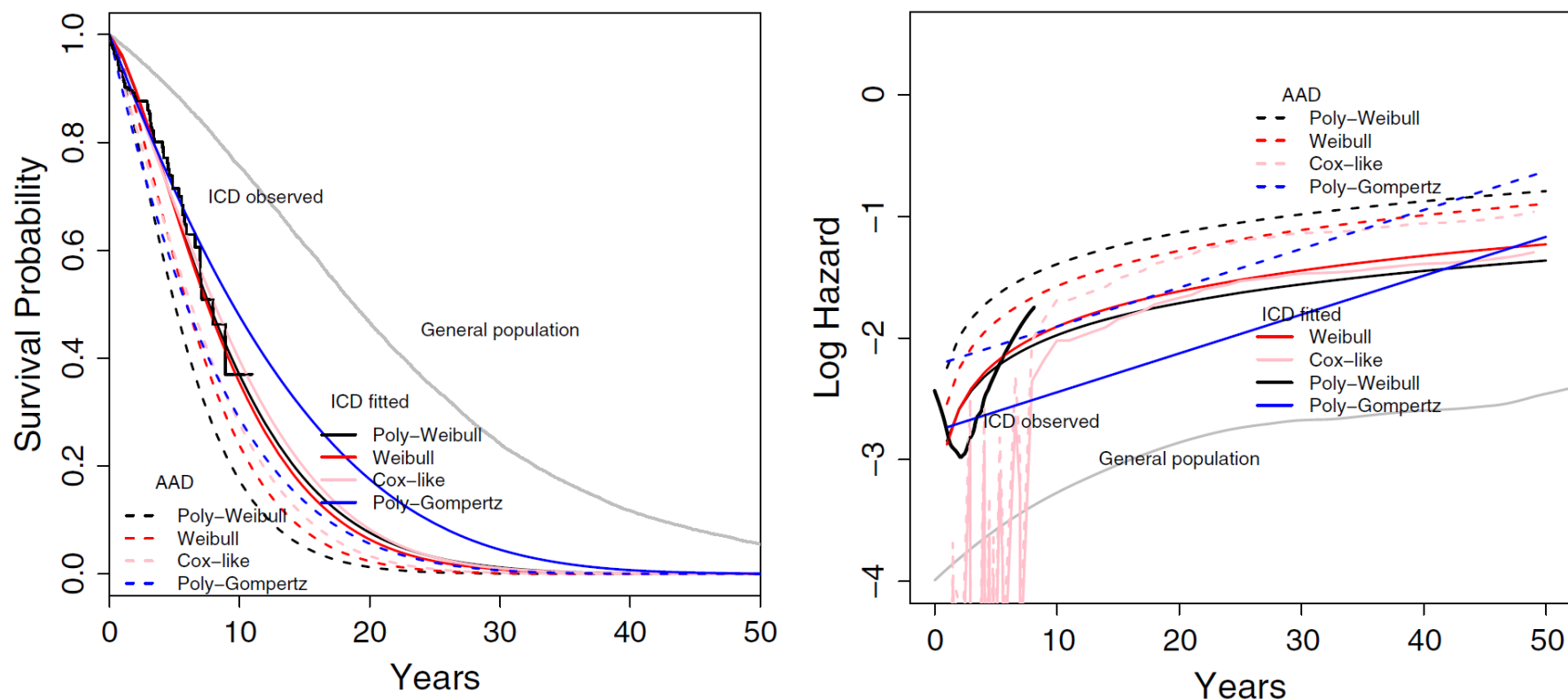
$$S(t) = \exp(-e^{\beta} \lambda_1 t^{\alpha_1} - \lambda_2 t^{\alpha_2})$$

In this equation, two causes of death are noted: 1 and 2. Cause 1 is disease-specific mortality in the example presented by Benaglia *et al.*, whereas cause 2 is other-cause mortality and assumed to be identical to the external data source. β is defined as the log-HR for the disease-specific hazard of death between the two groups of patients. λ_1 and λ_2 are rate parameters, whereas α_1 and α_2 are shape parameters, specific to each cause of death.

The authors note that causes of death are not observed in the study data – if these causes were recorded, other methods could be considered. Instead of fitting separate models to known causes of death, the poly-Weibull approach relies on the assumption that both causes of death contribute to an overall hazard of death, $h(t)$, and that these hazards are *additive*, and independent of each other.

While the initial description of the model is described with two components, the authors note that additional data could be considered to generate extrapolations for multiple treatment groups (e.g., by considering a third source of information describing the relationship between treatment arms). The authors presented an application of the poly-Weibull model for patients with cardiac arrhythmia given an ICD. The extrapolated survival curves using a poly-Weibull model (and other models) are presented in Figure 3-10, including a comparison to patients treated with AAD. To generate the comparison to AAD, prior distributions were set for an HR using the outputs from a published meta-analysis by Connolly *et al.*, (2000) – the same meta-analysis cited by Demiris & Sharples, (2006).^{151, 177}

Figure 3-10: Overview of extrapolations re-produced from Benaglia *et al.*, (2015)



Key: AAD, anti-arrhythmic drugs; ICD, implantable cardioverter defibrillator.

Note(s): This figure illustrates the poly-Weibull (and other model) fits within the context of a case study comparing patients given an ICD versus patients treated with AAD, re-produced from Benaglia *et al.*, (2015).¹⁵² The left-hand plot demonstrates the survival curves, whereas the right-hand plot shows the log hazards. The poly-hazard models (here, poly-Weibull and poly-Gompertz) make use of data from a group of patients given an ICD (shown via the black Kaplan-Meier estimate) as well as general population data (shown via the light grey line). Extrapolations for the AAD group were produced by setting a prior for a hazard ratio for arrhythmia-related deaths based on treatment assignment, using a published meta-analysis by Connolly *et al.*, (2000). The poly-Gompertz model provides a relatively poor fit to the available data, and the Bayesian semi-parametric Cox-like model was explored by the authors to investigate the potential impacts of allowing a more flexible baseline hazard for describing general population survival. The poly-Gompertz model is not discussed further in this sub-section as similar commentary applies per the poly-Weibull model. The 'Cox-like' model is also used in the study by Demiris & Sharples, (2006), but is not described further here.

Source: Excerpted and reprinted with permission from Statistics in Medicine (open access article distributed under the terms of the Creative Commons CC BY license). Benaglia T, Jackson CH, Sharples LD. Survival extrapolation in the presence of cause specific hazards. Stat Med. 2015 Feb 28;34(5):796-811.

The authors explain that based on Figure 3-10 (right-hand plot), it can be seen that the “empirical hazard rate fluctuated around an approximately constant value during the observed follow-up period” but that this is “unlikely to remain constant in the long term because we expect mortality rates to increase with age” (Benaglia et al., 2015, p.804).¹⁵² Considering both the log hazards and the survival extrapolations (left-hand plot in Figure 3-10), the Weibull and poly-Weibull produce similar extrapolations for the ICD arm, though the poly-Weibull predicts slightly lower hazards. Interestingly however, for the AAD group, hazards for the poly-Weibull are higher versus the Weibull model (as the poly-Weibull includes the specification of an HR for arrhythmia-related deaths only, whereas the Weibull model relies on an HR for all deaths).

Key assumption(s)

Unlike the approach presented by Demiris & Sharples, (2006)¹⁵¹ which assumed PH for all causes of death between the index (ICD) and reference (general population) populations, the authors of this study assume a higher (PH) risk of death for the disease of interest, but identical risk of death from other causes, between the index and reference populations. Relatedly, this method assumes that data for the index population provide no information in terms of the cause of death, but long-term reference population data provide information on cause of death.

Strengths and limitations

This method provides an intuitive means of combining information from multiple sources, through several simple assumptions (e.g., no difference in death from other causes, and a proportional increase in the risk of death related to the disease). As per the approach by Demiris & Sharples, (2006), all relevant data are combined within a single cohesive analysis.

However, in addition to the assumption of a proportionate change in the risk of death from the cause of interest (and corresponding information required), another limitation of this analysis is apparent when considering the information needed to compare to other treatments. To appropriately compare to another treatment arm, a summary outcome measure (e.g., an HR) is needed for disease-specific deaths (in this example, arrhythmia-related deaths). However, this limitation is only apparent when considering the need to derive a survival curve for a comparator intervention (assuming IPD are not available).

Rationale for inclusion in further investigation in the rest of the thesis

This method is included as an option to consider incorporating external evidence into survival extrapolations if sufficient information is available concerning competing risks of death in the general population. In principle, this links the method to a relative survival approach, given that the external evidence source comprises general population data. Poly-hazard models are considered in general (given that this study considers the use of both poly-Weibull and poly-

Gompertz models), though the poly-Weibull model is most commonly discussed in the literature.

Larkin et al., (2015)

Theme(s): *Piecewise.*

The authors of this study make use of a three-part piecewise modelling approach, to estimate the survival of patients with advanced melanoma treated with ipilimumab versus BSC.¹⁴⁶ The term ‘piecewise’ is used here to describe a survival function which is made up of multiple sub-functions, each of which applies to a given time interval. Given that ‘piecewise’ has been used extensively to describe models developed across a range of disciplines, it was not possible to ascertain which authors first developed a piecewise model for estimating survival while making use of external evidence, but this was the earliest example identified via the SLR.

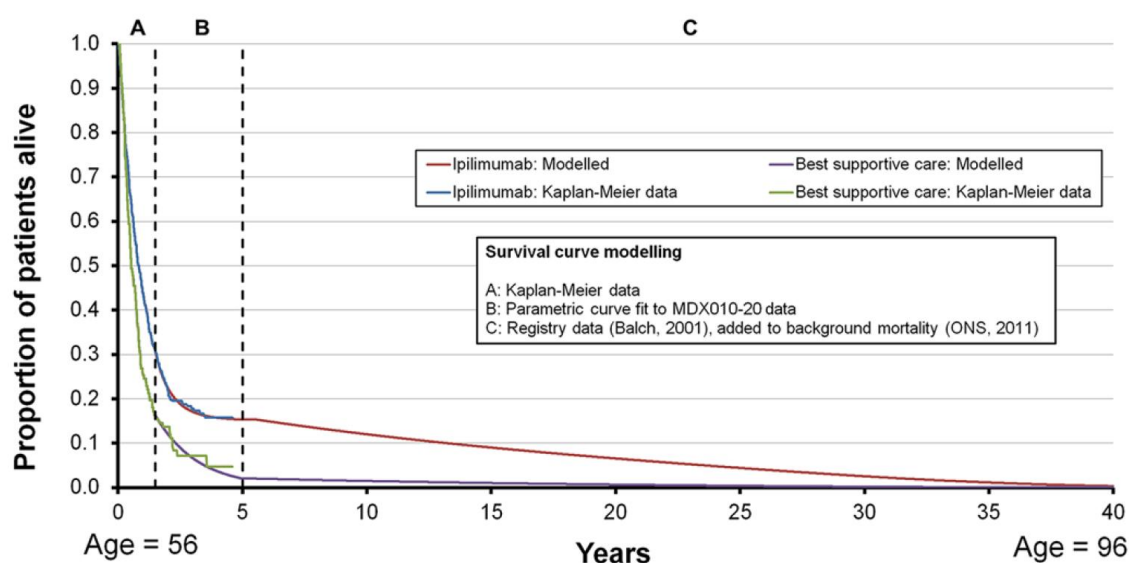
The three parts of the curve developed by Larkin *et al.* are described below:

- **A (<1.5 years):** The KM estimate of OS from the Phase III MDX010-20 study ([NCT00094653](#)) was taken and applied directly to estimate OS in this period. Note that this study comprised the main evidence base in NICE TA268 of ipilimumab for previously treated advanced (unresectable or metastatic) melanoma – the first cancer immunotherapy assessment published by NICE.¹⁸⁴
- **B (1.5-5 years):** Parametric models were fitted to data from MDX010-20 to reflect OS in this period.* 5 years was selected as this corresponded to the end of the available follow-up data from MDX010-20 at the time of analysis.
- **C (>5 years):** Data were identified from a published study by Balch *et al.*, (2001) for 1,158 patients with metastatic melanoma.¹⁸⁵ Absolute mortality rates estimated from these data were added to background mortality rates of the UK population taken from life tables (adjusted for age), and used to inform the remainder of survival curve.

The resultant models using the three-part curve fit are presented in Figure 3-11.

*While not explicitly stated, it is expected that these models were fitted to *re-based* survival data (i.e., re-setting the index point to 1.5 years, such that observations in the previous time period did not influence the survival models for the future time period) based on discussion included within the NICE TA268 committee papers (referenced by Larkin *et al.*, 2015).¹⁴⁶ Larkin J, Hatswell AJ, Nathan P, et al. The Predicted Impact of Ipilimumab Usage on Survival in Previously Treated Advanced or Metastatic Melanoma in the UK. *PLoS One* 2015; 10: e0145524. 2015/12/25. DOI: 10.1371/journal.pone.0145524.

Figure 3-11: Three-part curve fit re-produced from Larkin *et al.*, (2015)



Key: ONS, Office for National Statistics.

Note(s): This figure illustrates the three-part curve fit presented in Larkin *et al.*, (2015).¹⁴⁶ This is taken from panel C of Figure 1 from the original study publication.

Source: Copyright © 2015 Larkin *et al.* This figure is re-produced here under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited: Larkin J, Hatswell AJ, Nathan P, Lebmeier M, Lee D (2015) The Predicted Impact of Ipilimumab Usage on Survival in Previously Treated Advanced or Metastatic Melanoma in the UK. PLoS ONE 10(12): e0145524. <https://doi.org/10.1371/journal.pone.0145524>.

The decision to use a three-part piecewise modelling approach was justified primarily due to the shape of the KM estimate of OS in MDX010-20. The authors describe the shape of the KM estimate as “*unusual... as an initial steep decline is observed before the curve reaches a ‘plateau’*” (Larkin *et al.*, 2015, p.3).¹⁴⁶ This ‘plateau’ was explained to likely cause most simple parametric modelling approaches to produce unrealistic extrapolations.

Key assumption(s)

The key assumption made in this analysis (specific to the use of external evidence) was that the hazard of death beyond 5 years could be generalised to the experience of an external population plus the expected mortality of patients in the general population. This disregards entirely the projections that would have been made based on the shorter-term trial data, and overwrites these with extrapolations based on the longer-term, external data.

Strengths and limitations

An important advantage of this approach is that it appears logical to combine data from multiple sources, with a preference for using trial-based estimates up until a point, and then only switching to an extrapolation based on longer-term data from another source when necessary. However, this simultaneously captures the primary limitation of the approach – that the external data are only used from a specific point in time, with no ability to adjust for

similarities in populations, and limited ability to compare populations because only baseline characteristics were available for the external study. In principle, if IPD were available from Balch *et al.* it may have been possible to consider some form of adjustment to account for differences in patient characteristics, though this is not discussed by the authors.

Another limitation is that there is a risk of double counting deaths in the longer-term, as all-cause mortality from the external source is added to all-cause mortality from the general population. Furthermore, owing to the piecewise modelling approach taken without any attempt at 'smoothing' between each component, there is likely to be discontinuities present in the underlying hazard function. This can be inferred from Figure 3-11, where the hazard of death at approximately 5 years is near-zero (survival curve is flat), but after switching to the Balch *et al.* extrapolation the hazard is clearly greater than zero (survival curve no longer flat).

Rationale for inclusion in further investigation in the rest of the thesis

While this method has a number of limitations, it represents a relatively 'standard' piecewise model that has been used in a number of studies and company submissions to NICE. It is therefore included as a simple approach to incorporate external evidence within survival extrapolations.

Barlev *et al.*, (2017)

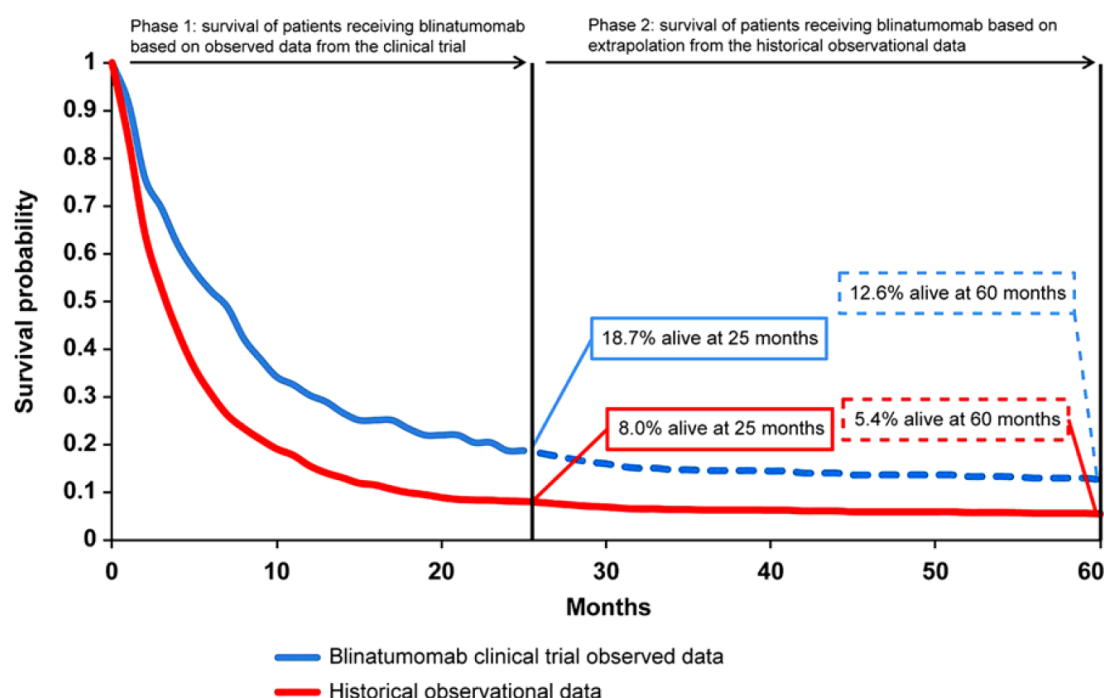
Theme(s): Piecewise.

The authors of this study describe a three-part piecewise modelling approach, undertaken to estimate the survival of adult patients with a type of acute lymphoblastic leukaemia treated with blinatumomab.¹⁴⁷ This approach is similar in concept to the approach of Larkin *et al.*, (2015)¹⁴⁶ but with different types of inputs to inform each of the three parts, as described below:

- **Phase 1 (<25 months):** The KM estimate of OS from the Phase II MT103-211 study ([NCT01466179](#)) was taken and applied directly.
- **Phase 2 (25-60 months):** Conditional survival estimates were estimated from an observational dataset of patients receiving salvage chemotherapy and applied without adjustment. This means that no additional benefit in terms of the relative hazard of death for blinatumomab versus salvage chemotherapy was assumed to apply.
- **Phase 3 (>60 months):** The authors assumed that patients surviving until 5 years are considered cured of their disease, and so conditional survival estimates were taken from life tables (though a specific source was not provided). It is assumed that these would be age- and sex-matched, though this is not explicitly stated.

An illustration of the complete OS curve was not presented in the article, but a restricted extrapolation up until 5 years (reflecting the first two phases) is presented in Figure 3-12.

Figure 3-12: 5-year overall survival curves re-produced from Barlev *et al.*, (2017)



Note(s): This figure illustrates the first two parts of the three-part curve fit presented in Barlev *et al.*, (2017).¹⁴⁷ This is taken from panel C of Figure 1 from the original study publication.

Source: Copyright © 2017 Barlev *et al.* This figure is re-produced under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0>), which permits use, duplication, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. Barlev A, Lin VW, Katz A, Hu K, Cong Z, Barber B. Estimating Long-Term Survival of Adults with Philadelphia Chromosome-Negative Relapsed/Refractory B-Precursor Acute Lymphoblastic Leukemia Treated with Blinatumomab Using Historical Data. *Adv Ther.* 2017 Jan;34(1):148-155.

Of note, the authors did not provide an explicit reason for why this piecewise approach was taken in favour of other approaches (including, for example, standard parametric survival modelling methods).

Key assumption(s)

This method assumes that from a given point in time, the estimated all-cause hazard of death can be directly taken from an external source and applied for a specified time period. This particular method makes use of two different external sources that apply at two different time points, and both external sources are assumed to reflect the same patient population as the Phase II MT103-211 study. In addition, the hazard of death beyond 5 years is assumed to be the same as the general population, effectively assuming cure beyond this time point.

Strengths and limitations

This piecewise approach is relatively simple to explain and makes use of decreasingly relevant data over time as the end of follow-up in each study is reached. However, it suffers from the need to rely on the generalisability of the external populations to the study used to inform the first component of the model, alongside the strong assumption of cure beyond 5 years.

Rationale for inclusion in further investigation in the rest of the thesis

This method is included for the same reasons as highlighted for the study by Larkin *et al.*, (2015).¹⁴⁶

Fisher et al., (2017)

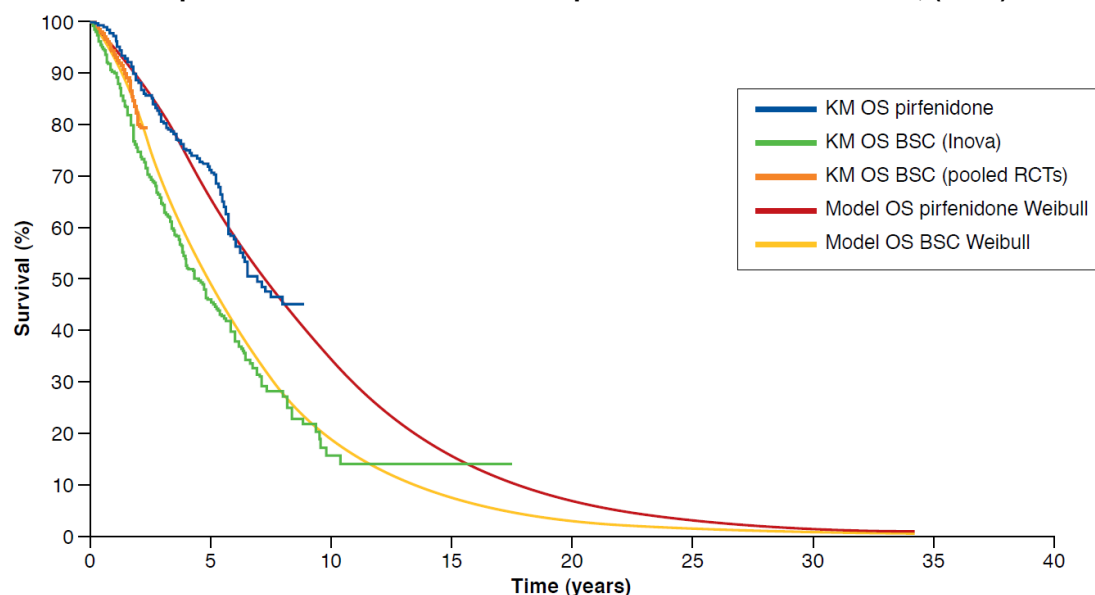
Theme(s): Piecewise.

This study adopts a similar approach to Barlev *et al.*, (2017)¹⁴⁶ to estimate OS for patients with idiopathic pulmonary fibrosis (IPF) managed with best supportive care (BSC), but is motivated by limitations of study design and limited follow-up data for BSC patients in particular.¹⁴⁸ A two-part model was considered in this study:

- **≤1 year:** Pooled data from three Phase III, international, RCTs: ASCEND ([NCT01366209](#)), CAPACITY 1 ([NCT00287729](#)), and CAPACITY 2 ([NCT00287716](#)) were used to generate a KM estimate of OS up until 1 year after initiation of treatment, for the BSC arms of these studies.
- **>1 year:** Data from the Inova registry were used to fit parametric survival models, which were used to describe OS for the remainder of the modelled time horizon. It is not clear if the authors defined a new baseline at 1 year when fitting the parametric models to the Inova registry data, or if the models were fitted from the original baseline.

The resultant extrapolation for the BSC arm is presented in Figure 3-13, alongside the fully parametric extrapolation used for the pirfenidone arm, and available KM estimates.

Figure 3-13: Extrapolations of overall survival re-produced from Fisher *et al.*, (2017)



Key: BSC, best supportive care; KM, Kaplan-Meier; OS, overall survival; RCTs, randomized controlled trials.

Note(s): This figure illustrates the extrapolations for pirfenidone and BSC presented in Fisher *et al.*, (2017).¹⁴⁸ This is taken from Figure 1 from the original study publication. The extrapolation of primary interest to this review is that of the BSC arm, where the KM estimate of OS from the pooled RCTs was considered to apply up until 1 year, after which a Weibull model fitted to data from the Inova registry was applied.

Source: Excerpted and reprinted with permission from the Journal of Managed Care & Specialty Pharmacy. Fisher M, Nathan SD, Hill C, Marshall J, Dejonckheere F, Thuresson PO, Maher TM. Predicting Life Expectancy for Pirfenidone in Idiopathic Pulmonary Fibrosis. J Manag Care Spec Pharm. 2017 Mar;23(3-b Suppl):S17-S24.

The authors explained that OS data from the pooled RCT data after 52 weeks were subject to several limitations, though a key limitation was that a large number of patients enrolled in an extension study (RECAP, [NCT00662038](#)) in which they received the intervention (pirfenidone 2,403 mg/day), meaning that long-term data for the placebo group that are not confounded by crossover are not available from the pooled RCT data.¹⁴⁸ 1-year OS across the pooled RCTs for the BSC arm was 93%, and so projecting lifetime survival estimates using traditional parametric survival analysis methods would be unlikely to yield robust extrapolations.¹⁴⁸

Key assumption(s)

This method assumes that after 52 weeks, the pooled RCT data no longer provide a suitable means of extrapolation, and so instead patients are assumed to experience the same all-cause hazard of death as per patients in the Inova registry. Therefore, 1-year survivors from the placebo arms of the pooled RCTs and the Inova registry are assumed to be generalisable, with no explicit adjustment undertaken to 'match' patient characteristics.

Strengths and limitations

This piecewise approach considers a straightforward 'switch' from the pooled RCT data to an external source, motivated by a feature of the RCT designs (crossover for the placebo arm). The rationale behind the approach taken is clear, and it is simple to understand and explain.

Nevertheless, it assumes that patient populations are completely generalisable, with no adjustment undertaken to address any (potential) imbalances. This particular application makes use of trial data combined with registry data, where it would be expected that the registry cohort comprises a comparably sicker population, and so trial-based extrapolations based using these data may be pessimistic.

Rationale for inclusion in further investigation in the rest of the thesis

This method is included for the same reasons as highlighted for the study by Larkin *et al.*, (2015).¹⁴⁶

Guyot *et al.*, (2017)

Theme(s): Piecewise, informative prior, general population, other.

Note: This method requires data from an RCT.

Guyot *et al.*, (2017) present a Bayesian approach which makes use of a multi-parameter evidence synthesis (MPES) method to estimate the survival function.¹⁵³ The article, published in 2017, expands upon work conducted by the authors originally presented at the 2014 ISPOR annual European meeting held in Amsterdam, The Netherlands.¹⁴¹

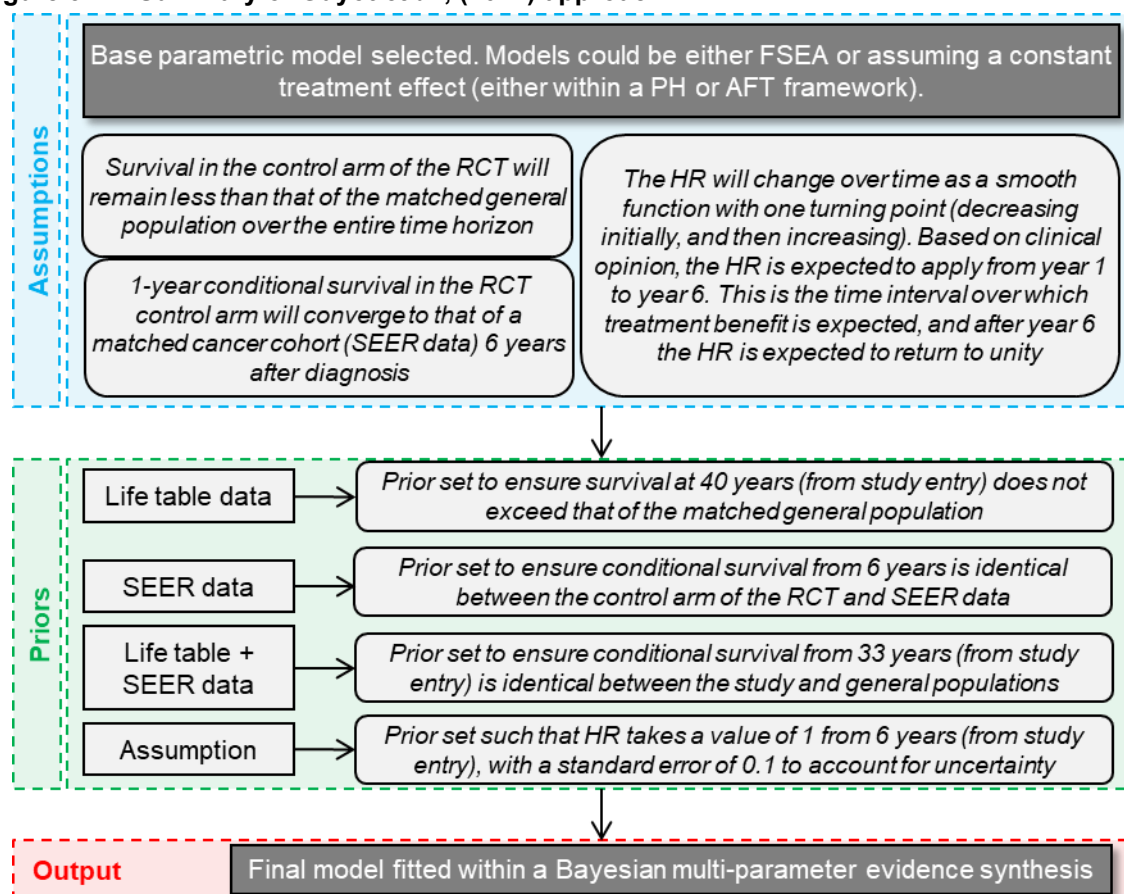
From a theoretical perspective, the approach is akin to that of Demiris & Sharples, (2006)¹⁵¹: multiple sources of information are synthesised within a Bayesian framework to estimate the survival function. However, rather than focusing on estimating a baseline hazard function and then the impact of covariates, Guyot *et al.* consider the overall hazard function and how different sources of information may be used to inform this over time. Therefore, using the approach of Guyot *et al.*, a breakdown of the components of the hazard function is not directly considered (as was the case for the studies by Demiris & Sharples¹⁵¹ and Benaglia *et al.*¹⁵²).

The authors focus on a motivating example based on an RCT of patients with head and neck cancer ([NCT00004227](#)), reported by Bonner *et al.*, (2006).¹⁸⁶ In addition to data from the Bonner *et al.* study¹⁸⁶, two additional sources of information were considered:

- Data for the survival of a matched cohort from the general population.
- Data from the SEER registry for patients with head and neck cancer.⁸⁶

A summary of the approach taken by Guyot *et al.*, (2017)¹⁵³ is provided in Figure 3-14. The approach can be described as making assumptions about the hazard function over time and translating these into prior distributions (which may be considered restrictions for model fitting). The base model could take several functional forms, though the most flexible approach presented by Guyot *et al.* was a spline-based model. Ultimately, the model estimation process is based on a common model, where all parameters are estimated simultaneously.

Figure 3-14: Summary of Guyot et al., (2017) approach



Key: AFT, accelerated failure time; FSEA, fitted separately to each arm; HR, hazard ratio; PH, proportional hazards; RCT, randomised controlled trial; SEER, Surveillance, Epidemiology and End Results cancer registry.

Note: The prior set to ensure identical conditional survival from 33 years was imposed retrospectively by the authors. This was because once the SEER data were included within the model, 1-year conditional survival estimates for the trial population 'crossed' the estimates for the general population at 33 years.

A more detailed explanation of this method in particular was prepared as part of the MPES tutorial paper which was developed as part of this thesis. Please refer to this paper for more information about this method, and how it compares to another MPES approach.⁴

Key assumption(s)

Several context-specific assumptions are made as part of the model fitting process, as described in Figure 3-14. In addition to these assumptions, the authors explain that as with any Bayesian analysis, it is implicitly assumed that the external evidence being used to specify priors are “providing unbiased estimates of the target parameters” (Guyot et al., 2017, p.363).¹⁵³ Finally, the authors also assume that the target population for analysis is that population in the index study (in this example, the RCT population), though in reality the target population for HTA decision-making would be the population of patients expected to receive the intervention in practice.

Strengths and limitations

The main strength of this approach is that it can combine multiple different sources of information into one coherent analysis. However, a principal limitation of the method proposed by Guyot *et al.* is that it involves the incorporation of several different sources of external evidence, and a number of assumptions to ensure the resultant model estimates exhibit face validity. While this may be viewed as a strength of the method (i.e., that it allows the analyst to combine multiple evidence sources), the downside is that this means the method relies on several assumptions and potential *post-hoc* adjustment. As the process of estimation is complex, the resultant estimates of survival are less transparent versus a simpler (though perhaps less technically robust) method.

Rationale for inclusion in further investigation in the rest of the thesis

The authors of this study present a method that combines information from multiple external evidence sources, and so this method is included as one of the only approaches that allows for several external evidence sources to be leveraged within one analysis.

Hwang *et al.*, (2017)

Theme(s): *General population.*

In 2017, Hwang *et al.* proposed a new extrapolation approach building on the framework of the Monte Carlo estimator¹⁶⁸, called *rolling extrapolation*.¹⁷² Details of the 'Monte Carlo estimator' method are provided in Appendix C. For completeness, the relevant aspects of this method that apply to the rolling extrapolation method are summarised below.

Both the Monte Carlo estimator and rolling extrapolation methods requires the estimation of a reference curve. The authors explain that as a minimum, patients should be matched on age and sex, if using life tables. Then, if life tables are to be used, a reference curve can be generated. Details of this are provided in Appendix C, but essentially this involves deriving a background mortality curve using relevant probabilities of death based on the characteristics of the cohort of interest.

Once a reference curve has been defined, both methods require the estimation of the ratio of OS between the index and reference populations, evaluated at time t . This ratio is termed $W(t)$, and is expected to either slowly decrease or remain approximately constant after t is larger than a 'critical' or 'unstable' time point, denoted T_S . $W(t)$ may also be considered as a measure of relative survival.

Since the hazard of death is expected to be poorer in the index population versus the reference population, it is expected that $W(t)$ takes a value between 0 and 1 for all values of t . To consider a regression model to explore how $W(t)$ changes as t increases, a logit

transformation is applied such that $\text{logit}(W(t))$ takes values on the scale $(-\infty, +\infty)$. The logit function is presented below for completeness:

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

In the Monte Carlo estimator paper, the authors explain that the tail ends of the logit transform of $W(t)$ are expected to be unstable, but the central part of the transform will be approximately linear. Using this linear aspect of the logit transformed $W(t)$, a linear regression can then be fitted:

$$\text{logit}(W(t)) = \log\left(\frac{W(t)}{1-W(t)}\right) = \alpha + \beta t + N_t \quad \text{for } T_S \leq t \leq T_E$$

In the equation above, α and β are intercept and gradient terms within a simple linear regression, and N_t represents a noise term which is independently and normally distributed with mean $\mu = 0$ and variance σ^2 . T_S represents a time point after which $W(t)$ is expected to be approximately constant, whereas T_E represents a time point less than or equal to the last recorded follow-up time, T_F . The determination of both T_S and T_E is not automated, but the authors describe several ways they may be determined, including reference to a study by Gelber *et al.*, (1993).¹⁸⁷ Alternatively, the authors explain that different values of T_S and T_E may be explored and edited based on the fit of the linear regression.

After determining the least squares estimates of the regression parameters ($\hat{\alpha}$ and $\hat{\beta}$), an estimate of OS curve for the index population for $t \geq T_F$ can be obtained using the following equation:

$$S_{\text{Index}}(t) = S_{\text{Reference}}(t) \left(\frac{\exp(\hat{\alpha} + \hat{\beta}t)}{1 - \exp(\hat{\alpha} + \hat{\beta}t)} \right)$$

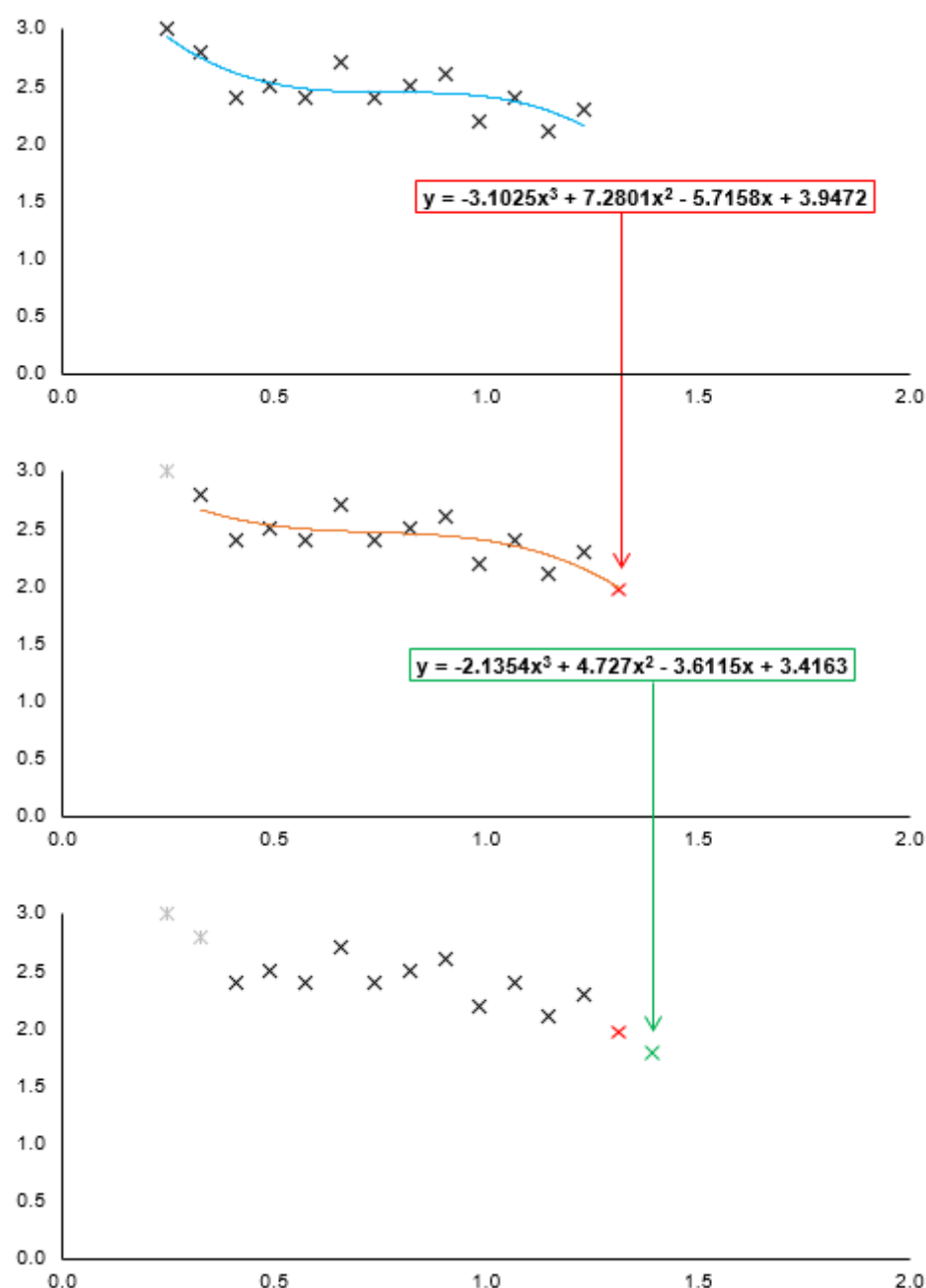
In the equation above, $S_{\text{Index}}(t)$ denotes the survival curve for the index population, and $S_{\text{Reference}}(t)$ denotes the survival curve for the reference population. Using this relationship, the survival curve for the index population can be extrapolated after T_F using a piecewise approach.

For the rolling extrapolation method, a key assumption made in the Monte Carlo estimator of a constant excess hazard is relaxed, by using a restricted cubic spline (RCS) model to extrapolate survival, as opposed to using a linear regression model. This change means that the selection of relevant time points between which to estimate a linear projection (i.e., selection of T_S and T_E) can be avoided, and instead the full follow-up period may be considered to estimate $\text{logit}(W(t))$.

In consideration of using spline-based models, the authors explain that previous studies (e.g., Andersson *et al.*, [2013]¹⁶³, which is discussed within the context of relative survival methods) have considered RCS models which are restricted to be linear beyond the final knot. However, the authors explain that if $\text{logit}(W(t))$ does not appear to be approaching a straight line, the longer-term extrapolation may not be accurate. To address this, the authors propose to iteratively fit and re-fit models of $\text{logit}(W(t))$ by dropping earlier observations and treating latter (predicted) estimates as if they were observations.

The principle behind the rolling extrapolation approach is illustrated in Figure 3-15, shown using a simplified, worked example (not provided by the study authors themselves). An iterative approach is taken, where in the first iteration, an equation is generated for $\text{logit}(W(t))$, from which it is possible to estimate the next point in the series (shown in red). After this point is estimated, the earliest point is dropped, and the line of best fit is re-estimated, allowing for generation of the next point (shown in green). This process could then plausibly be repeated for as long as necessary until the end of the modelled period.

Figure 3-15: Illustration of principle behind rolling extrapolation approach

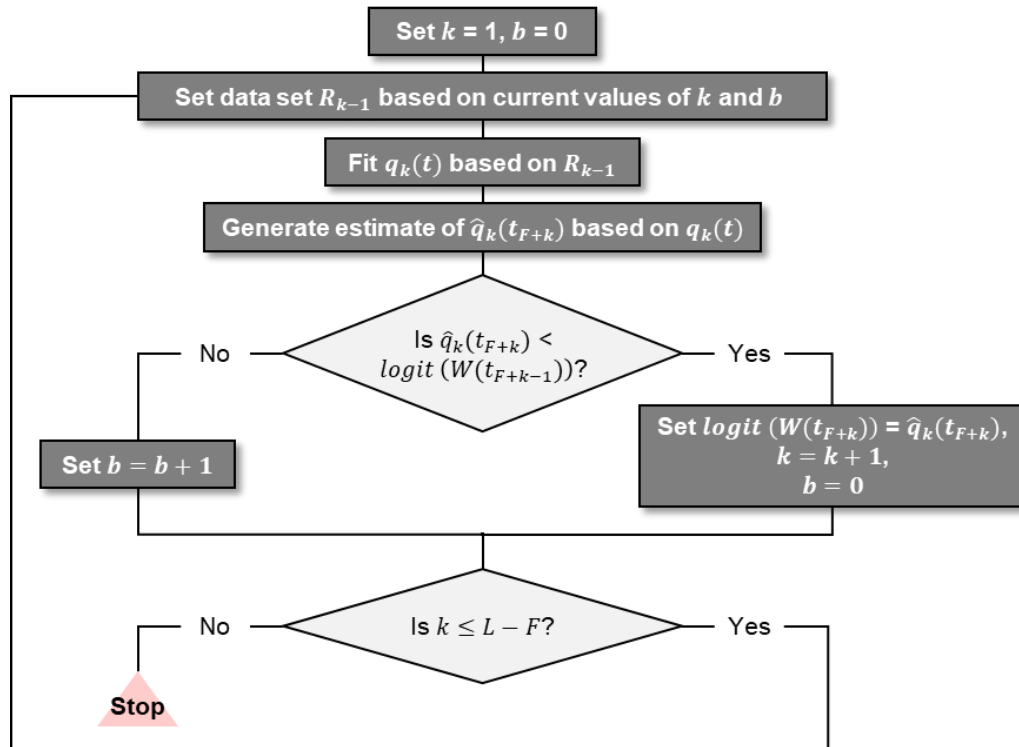


Note: These plots are purely illustrative – the x- and y-axes are not specific to survival analysis directly. Points shown in black are based on empirical data. The point in red is estimated using the line of best fit from all the data points in black (line shown in blue). The point in green is estimated from a line of best fit based on all the points in black except from the first point, plus the estimated point in red (line shown in orange).

The full rolling extrapolation algorithm is presented as a flowchart in Figure 3-16. The flowchart highlights an important step which is specific to the extrapolation of survival, where after the next data point has been estimated, it is compared to the previous time point value. If the predicted value for the next time point is lower than the value of the previous time point, the data set is reduced to remove the first time point value, and then the process is repeated.

However, if the predicted value for the next time point is greater than or equal to the value of the previous time point, this value is taken forward, and the next step of the iteration is calculated. The algorithm is followed up until the last time point required. The authors describe their method in terms of time in units of one month – for example, the authors state: “*With the newly updated logit $W(t)$ data series, the one-step-ahead prediction process would be repeated forward $L - F$ times, where L is the final month of extrapolation.*” (Hwang *et al.*, 2017, p.e355).¹⁷² In principle, the choice of time unit would need to be specified by the analyst.

Figure 3-16: Rolling extrapolation algorithm flowchart

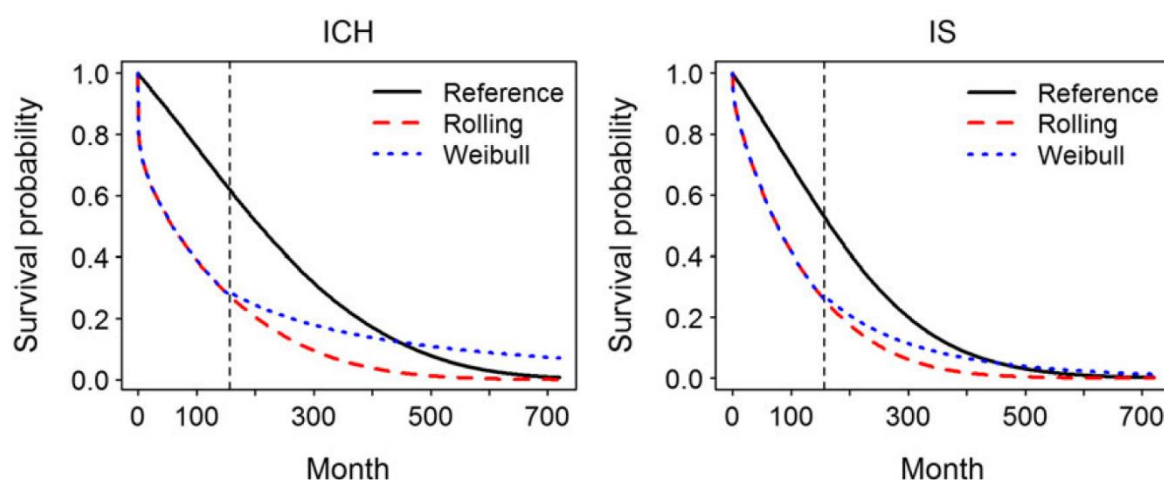


Key: k , iteration variable; b , offset variable; F , total number of data points; L , final data point to be predicted; $W(t)$, relative survival estimate, evaluated at time t ; R_{k-1} , data set containing estimates of $\text{logit}(W(t_j))$ where $j \in \{k+b, k+b+1, \dots, F+k-2, F+k-1\}$; $q_k(t)$, spline model fitted to R_{k-1} , evaluated at time t ; $\hat{q}_k(t)$, prediction of spline model, evaluated at time t .

Source: This diagram was produced based on information reported by Hwang *et al.*, (2017).¹⁷²

Hwang *et al.* compare their method to a standard Weibull model using data for patients that have had an intracerebral haemorrhage or an ischemic stroke. The resultant extrapolations are presented in Figure 3-17.

Figure 3-17: Rolling extrapolation example re-produced from Hwang *et al.*, (2017)



Key: ICH, intracerebral haemorrhage; IS, ischemic stroke.

Note(s): This figure illustrates the rolling extrapolation method versus a standard Weibull model in two cohorts of patients (left: ICH patients, right: IS patients), presented in Hwang *et al.*, (2017).¹⁷² This is taken from Figure 3 from the original study publication. The reference survival curves were estimated by the authors using data generated for age- and sex-adjusted referents with no history of stroke.

Source: Copyright © 2017 John Wiley and Sons. This figure is re-produced with permission from: Hwang JS, Hu TH, Lee LJ, Wang JD. Estimating lifetime medical costs from censored claims data. *Health Econ.* 2017 Dec;26(12):e332-e344. doi: <https://doi.org/10.1002/hec.3512>.

Key assumption(s)

This method shares the same assumptions per the Monte Carlo estimator, with the exception that the assumption of a constant excess hazard is relaxed, by using an RCS model to extrapolate survival. A key assumption shared by both the Monte Carlo estimator and the rolling extrapolation method is that $W(t)$ takes values between 0 and 1 for all timepoints, meaning that the reference population is expected to always have a lower hazard of death at all timepoints versus the index population. Please refer to Appendix C for further information concerning the Monte Carlo estimator.

Strengths and limitations

This method shares most of the strengths and limitations of the Monte Carlo estimator, excluding those related to the assumption of a constant excess hazard, and so are not described further here. Please refer to Appendix C for further information concerning the Monte Carlo estimator.

Rationale for inclusion in further investigation in the rest of the thesis

This method is included on the basis of it addressing a limitation of the Monte Carlo estimator and providing a unique means of incorporating data from a reference population to extrapolate survival.

Pennington et al., (2018)

Theme(s): *Other.*

Pennington *et al.* illustrate approaches to make use of summary external data to extrapolate survival.¹⁷³ While the authors also consider the potential role of external evidence to inform model selection, this is beyond the scope of the review and so only approaches that leverage external evidence to inform survival extrapolation are considered here.

The authors present worked examples comparing outcomes for patients that receive either a cemented or a cementless total hip replacement, based on data made available to the authors from the National Joint Registry (NJR, henceforth termed the index population).¹⁸⁸ External data were identified from the Hospital Episode Statistics (HES) database, which were considered to inform the extrapolation of survival beyond the point of maximum follow-up for the index population, referred to by the authors as t_{max} .¹⁸⁹

As two approaches were presented in this study, each method is presented separately in the sub-sections that follow.

'Summary data only' piecewise model

In this approach, the authors assumed that the external data may be available only as summary statistics. If data are only available as summary statistics, the authors assume that these provide estimates of survival at two time points: $t_1 < t_{max}$, and $t_2 > t_{max}$. Then, assuming a constant risk of the event of interest after t_1 , the authors estimate the failure probability per unit time (i.e., cost-effectiveness model cycle length [e.g., 1 week]) to inform the remainder of the survival curve. To estimate the failure probability per unit time, the authors provide the following formula:

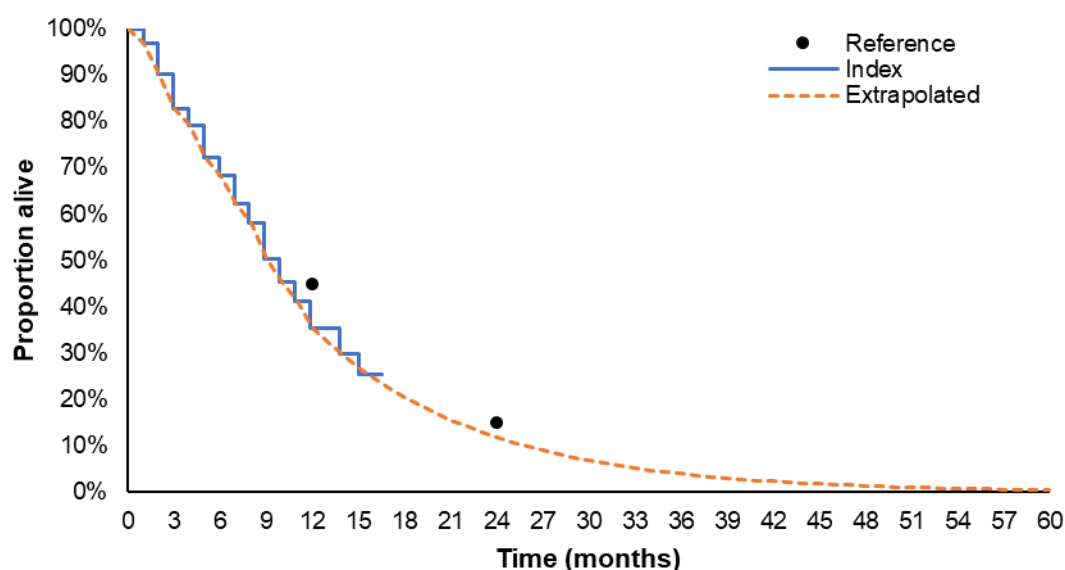
$$F = 1 - \sqrt[j]{\frac{S_2}{S_1}}$$

Here, j refers to the number of time periods between t_1 and t_2 ; whereas S_1 and S_2 refer to the corresponding survival proportions at times t_1 and t_2 , respectively. This approach may be described as a piecewise model, comprising of the KM estimate applied directly up until t_1 , after which a constant hazard is estimated and used to inform the remainder of the modelled time horizon.

An example of this method is presented in Figure 3-18. In this illustrative example, the maximum follow-up time, t_{max} , in the index population was approximately 16.5 months. External data were assumed to be available for the estimation of 1- and 2-year survival at 45% and 15%, respectively ($t_1 = 1 \text{ year}$, $t_2 = 2 \text{ years}$), which are labelled as 'Reference' in Figure

3-18. Using a monthly cycle length, the failure probability was estimated at 8.75% and used to inform the remainder of the extrapolation after t_1 (1 year).

Figure 3-18: Illustration of Pennington *et al.* method with only summary data



Note: In this figure, 'Index' refers to the Kaplan-Meier estimate of survival from the index population; 'Reference' illustrates the summary statistics available from an external source of information; and 'Extrapolated' shows the extrapolated estimates of survival combining data from the index and reference populations.

Parametric model with covariate(s) based on external data, with/ without HR adjustment

In this approach, the authors assume that IPD are available for the control group (instead of only summary data), and so an alternative method could be used. Recommendations made by authors of previous studies by Demirir & Sharples (2006)¹⁵¹ and Benaglia *et al.*, (2015)¹⁵² were acknowledged by Pennington *et al.* to inform the approach taken in their study.

To consider the use of data from the index population (NJR data) and the external population (HES database), Pennington *et al.* specify a dummy variable to distinguish between populations. Then, parametric models were fitted including covariates for age and sex, and models were stratified by the type of hip replacement. This approach bears similarities with the approach taken by Hawe *et al.*, (2014)¹⁷¹ to include a dummy variable for population, and fit models with covariates, and so a separate illustration is not presented here. The authors also highlight that this method using an HR “*assumes no interaction between treatment effect and subgroup.*” (Pennington *et al.*, 2018, p.823).¹⁷³ This is an important consideration when taking outputs from multiple models and combining them together (using the same ‘plugging in’ approach of by Hawe *et al.*, [2014]).¹⁷¹

The authors highlight that summary data could also be considered in addition to use of external IPD by calculating an HR between the intervention and control groups using the following formula:

$$\widehat{HR} = \frac{\ln \left(\frac{S_{2I}}{S_{1I}} \right)}{\ln \left(\frac{S_{2C}}{S_{1C}} \right)}$$

Here, S_{ij} refers to the proportion of patients alive on either the control (denoted by subscript C) or intervention (subscript I) arm, at times t_1 and t_2 (denoted by subscripts 1 and 2, respectively). Then, as per the previous approach, survival after t_1 may be informed through assuming a constant HR versus the control group modelled survival.

Key assumption(s)

‘Summary data only’ piecewise model

This approach assumes that the risk of death is constant between two summary data points, effectively meaning that an exponential model is assumed to apply between these time points, and for as long as the extrapolation is required.

Parametric model with covariate(s) based on external data, with/ without HR adjustment

Per the approach of Hawe *et al.*, (2014)¹⁷¹, parameters obtained from each component of the analysis are assumed to be directly transferable into other components of the analysis. Based on the availability of IPD, this approach allows for some adjustments to be made reflecting differences in the baseline patient characteristics across studies. However, this requires both data sets to contain the same information on patients, and for there to be sufficient overlap between the populations such that adjustments are possible to estimate appropriately. Finally, a constant HR is assumed to apply, meaning that any treatment effect is assumed to be independent of time.

Strengths and limitations

‘Summary data only’ piecewise model

The key strength of this method is to make use of very limited information in a transparent manner, though it is fundamentally limited by the strong and untestable assumption of a constant hazard that applies indefinitely.

Parametric model with covariate(s) based on external data, with/ without HR adjustment

This approach shares strengths with the approach of Hawe *et al.*, (2014)¹⁷¹, with the added benefit of having the ability to make some adjustments between populations. However, the

method is nevertheless limited in terms of assumptions of transferability of parameters and specification of a constant HR to describe differences in survival between groups.

Rationale for inclusion in further investigation in the rest of the thesis

Both methods were taken forward owing to their ability to incorporate different types of external evidence in an intuitive manner, based on availability.

Cope et al., (2019)

Theme(s): *Piecewise, informative prior.*

Cope *et al.*, (2019) demonstrated how expert opinion may be combined with trial data to inform long-term survival extrapolations.⁹¹ The authors focused on a case study of the ELIANA trial ([NCT02435849](#)) of tisagenlecleucel for children and young adults with relapsed or refractory acute lymphoblastic leukaemia. At the time the study was conducted, data for the outcome of OS in ELIANA were reported in a study by Grupp *et al.*, (2016), with follow-up to approximately 18 months.¹⁹⁰ Using these data, the authors were able to obtain estimates of the hazard function up until 18 months. After 18 months however, the authors made use of expert opinion from seven experts.

Expert opinion was obtained for the proportion of patients expected to still be alive at 2, 3, 4, and 5 years. The authors made use of an elicitation tool adapted from Version 3.0 of the SHeffield ELicitation Framework (SHELF).⁹⁰ The experts were asked to provide point estimates and plausible limits, so that uncertainty in their estimates could also be accounted for within model fitting. From the elicited opinion, discrete hazard estimates from the end of follow-up in ELIANA (approximately 18 months) up until 5 years (maximum time point from experts) could then be produced.

A dataset was then produced including both the expert-derived hazard estimates as well as the original set of hazards from the ELIANA trial. This differs to a traditional data set used to fit survival models, as the 'data' are hazard estimates, not observed survival times and censor indicators. Seven separate datasets were produced for the set of discrete hazards estimated from each expert. Using a combination of hazard estimates from the ELIANA trial and expert opinion, models were then fitted within a Bayesian framework – in other words, the expert opinion was considered 'additional data' to be used as part of overall model fitting, drawing parallels with the MPES approach of Guyot *et al.*, (2017).¹⁵³

The authors fitted second-order FP models (FP2 models) to extrapolate survival, meaning that two powers were required to specify (p_1 and p_2). Values of 0 or 1 were considered for p_1 , whereas values of $\{-1, -0.5, 0, 0.5, 1\}$ were considered for p_2 . This combination of powers for

use within FP2 models was considered to provide extensions to the Weibull and Gompertz models, allowing for arc- and bathtub shaped hazard functions.⁹¹

The formula for an FP2 model used by Cope *et al.*, is presented below:

$$\ln(h(t)) = \begin{cases} \beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_2}, & p_1 \neq p_2 \\ \beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_1}(\ln(t)) & p_1 = p_2 \end{cases} \quad \text{where } t^0 = \ln(t)$$

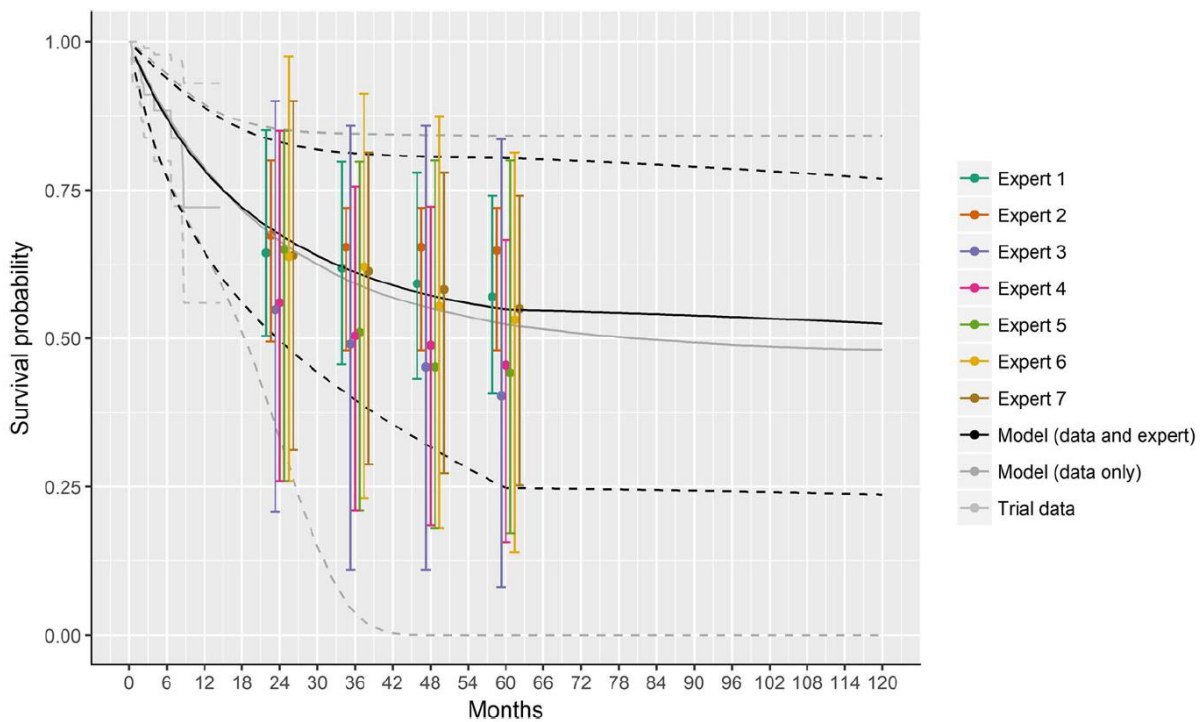
While a range of flexible approaches were considered (including the FP2 models described above), ultimately a Gompertz model was considered by the authors (and based on expert opinion) the most suitable for extrapolation.

A Gompertz model can be considered a special case of a first-order FP (FP1) model, where $p_1 = 1$, shown below:

$$\ln(h(t)) = \beta_0 + \beta_1 t^{p_1} = \beta_0 + \beta_1 t, \quad \text{where } p_1 = 1$$

Models fitted by Cope *et al.* with and without expert opinion are presented in Figure 3-19 (assuming a Gompertz distribution).

Figure 3-19: Estimates of overall survival reproduced from Cope *et al.*, (2019)



Note(s): This figure illustrates the estimates of overall survival using a Gompertz model with and without the incorporation of expert opinion by Cope *et al.*⁹¹ The circles and solid lines represent point estimates and plausible limits provided by the experts. Dashed lines represent 95% credible intervals for the fitted models, and 95% confidence intervals for the Kaplan-Meier estimate based on the trial data. The point estimates provided by the experts are offset slightly so that the range of estimates provided for each yearly interval can be easily seen. This is taken from Figure 4 from the original study publication.

Source: Copyright © 2019 Cope *et al.* This figure is re-produced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. Cope S, Ayers D, Zhang J, Batt K, Jansen JP. Integrating expert opinion with clinical trial data to extrapolate long-term

After the models were fitted separately for each expert, an ‘overall’ model was produced, though the authors do not describe the process followed to combine the results from each model (other than explaining that the approach they followed avoided pooling or model averaging).

Key assumption(s)

The approach presented in this study assumes that each expert provides the ‘same’ value of opinion but acknowledges their individual expression of uncertainty in the estimates provided. Therefore, regardless of the knowledge base of each expert, they are assigned equal weight in the overall estimation of survival (though this could be edited by the analyst, if there was [for example] rationale to suggest specific weights could or should be used for each expert). The authors in this example also assumed that beyond 5 years, all patients were long-term survivors and were consequently modelled based on age- and sex-matched general population data for all-cause mortality, with an adjustment for excess mortality linked to paediatric acute lymphoblastic leukaemia.

Strengths and limitations

The method provides an intuitive application of expert information where it is most valuable (i.e., beyond the duration of follow-up in the study of the index population). However, it is not entirely clear how expert estimates were combined due to limited reporting for how the results were ultimately produced. In addition, it is only possible to consider this type of analysis within the context of fitting models that are based on discrete hazards, as expert opinion cannot be converted into IPD with numbers at risk. This means that the exact approach presented could not be undertaken for other types of survival extrapolation that are based on traditional TTE data, because of the specific model specification used.

If used in practice, it is expected that this type of method would introduce several challenges owing to its reliance on expert opinion. For example, while this method may appear to generate plausible estimates of survival because it aligns with expert opinion, this is to be expected because the opinion is a model input. In addition, experts will most likely “anchor” their estimates to the earlier part of the survival curve based on seeing an initial KM estimate.

Rationale for inclusion in further investigation in the rest of the thesis

This method provides a combined analysis of empirical evidence and expert opinion, and so therefore represents a potentially useful method to incorporate external *information* if external *data* are not available.

Jakobsen et al., (2019)

Theme(s): *General population.*

The study by Jakobsen *et al.* presents three different approaches to predict relative survival using flexible parametric modelling methods (also known as spline-based models), two of which consider a cure fraction.¹⁶⁵ Each of the three methods made different assumptions mostly affecting the extrapolated tail of the survival curve:

- Assume that the log cumulative excess hazard beyond the last boundary knot follows the linear trend, without necessarily any explicit modelling of a cure fraction, assuming that excess hazards are modelled using extrapolations based on spline-based models.
 - This approach is attributed to a study by Nelson *et al.*, (2007).¹⁸¹ However, this method is considered to be captured as part of the study by Andersson *et al.*, (2013)¹⁶³, which is discussed separately.[†]
- Assume that the log cumulative excess hazard beyond the last boundary knot is a constant (i.e., zero slope, imposing the assumption of cure).
 - This is the approach taken by Andersson *et al.*, (2011)¹⁶¹, which is described by Jakobsen *et al.* as a strong assumption.
- Assume a ‘cure’ by restricting one of the parameters for the linear spline variable to be zero, but without imposing the same restriction on the log cumulative excess hazard as per Andersson *et al.*, (2011).¹⁶¹
 - This is a novel approach proposed by Jakobsen *et al.*, which is a different way of capturing a ‘cure’ using a MCM approach versus the approach of Andersson *et al.*, (2011), which used an NMCM approach.¹⁶¹ The approach of Jackobsen *et al.* “enables flexible modelling of the relative survival without the strong assumption of cure after the last knot while providing the more intuitive interpretation of [an MCM]” (Jakobsen *et al.*, 2019, p.3).¹⁶⁵ It may be argued, however, that any assumption of a ‘cure’ could be seen as ‘strong’, regardless of the specific mechanisms used to impose the assumption.

Key assumption(s)

As described above, the main assumption made within this approach (as with all other cure-based methods) is the existence of a cured proportion of patients. Please refer to discussion of cure-based in Chapter 2 for more details (Section 2.3.2).

[†] The study by Nelson *et al.*, (2007) was also later captured in the second iteration of this review.

Strengths and limitations

A key strength of the approach presented by Jakobsen *et al.* is that it addresses a possible limitation of a previous approach by Andersson *et al.*, (2011) by avoiding the assertion of cure beyond the final boundary knot, and otherwise fitting a similar type of model. However, this strength may be seen as a red herring, since this is still a type of cure assumption. Relatedly, a disadvantage identified by the authors was that in scenarios run where no cure was present (as part of a simulation study), any cure-based model was “*substantially biased*” (Jakobsen *et al.*, 2019, p.8).¹⁶⁵

Rationale for inclusion in further investigation in the rest of the thesis

This study builds upon a ‘standard’ parametric relative survival model by considering extensions for cure- and spline-based models. This extension, in addition to different extensions presented by Lambert *et al.*, (2007)¹⁵⁸ and Andersson *et al.*, (2013)¹⁶³, as well as a standard parametric relative survival model, are taken forward as applications of relative survival methodology using established survival extrapolation techniques.

Vickers, (2019)

Theme(s): Piecewise, other.

Note: This method requires data from an RCT.

Vickers, (2019) considered an application of six approaches to fit models to simulated data to extrapolate survival, based around a SEER database analysis by Ganti *et al.*, (2016) of patients with early-stage NSCLC.^{174, 191} This study builds upon previous work by Vickers & Hawe, (2018) presented at the 2018 ISPOR annual European meeting held in Barcelona, Spain.¹⁹² The six approaches considered were:

- 1) The most plausible parametric survival model, determined based on guidance presented in TSD 14.¹⁰
- 2) Approach 1) with external data used for extrapolation.
- 3) A bootstrapped hybrid model (an approach that was adapted from published studies by Gelber *et al.*, [1993]¹⁸⁷ and Bagust & Beale, [2014]¹⁹³).
- 4) Approach 3) with external data used for extrapolation.
- 5) An ensemble of parametric models with external data used for extrapolation.
- 6) A Bayesian, simultaneous, flexible, spline-based model (an approach adapted from the method of Guyot *et al.*, [2017]¹⁵³).

Approaches 1) and 3) do not make use of external evidence, and so were not considered further. Approach 6) is an application of the method proposed by Guyot *et al.*¹⁵³, with relatively minor modifications to align with the example considered by Vickers, and so is not considered

further. This leaves approaches 2), 4), and 5); however, each of these approaches incorporate the external data in the same way, so in essence represent one possible approach to incorporate external data within survival extrapolation.

In these approaches, Vickers denotes t_1 to denote the maximum follow-up time in the study of the index population (here, the study by Ganti *et al.*). To extrapolate beyond t_1 , a model is fitted to an external data source and general population data. The hazards projected from this model are then used to inform survival for the control group, whereas adjustments to these hazards to account for a potential treatment effect are used for the intervention group. A range of options were considered for estimating the short-term hazards in approaches 2), 4), and 5); however, as only the external data were used to inform extrapolation beyond t_1 , these methods are not discussed here.

To inform estimates of survival beyond t_1 , Vickers describes a multi-stage process to produce a model to generate long-term hazards:

- First, Vickers sought to identify a suitable long-term source to inform estimates of survival beyond t_1 . A search of the literature yielded a study by Bach *et al.*, (1999)¹⁹⁴ as the source that produced outcomes most similar to those reported by Ganti *et al.*, and was therefore considered appropriate for informing extrapolations.
 - Ganti *et al.* considered data for patients aged 80 years or older, diagnosed with clinical stages I and II NSCLC, between 1988 and 2007, identified via the SEER database.¹⁹¹
 - Bach *et al.* presented findings from a population-level study using data from the SEER database, originally motivated by an investigation into racial differences in the treatment of early-stage lung cancer for black and white NSCLC patients. Therefore, the data from this study comprised of “*black patients and white patients 65 years of age or older who were given a diagnosis of resectable [NSCLC] (stage I or II) between 1985 and 1993 and who resided in 1 of the 10 study areas of the [SEER] program (10,984 patients)*” (Bach *et al.*, 1999, p.1,198).¹⁹⁴ These data provided estimates of survival up until approximately 9 years following diagnosis (based on Figure 1 in Bach *et al.*, [1999]).
 - IPD from both Ganti *et al.* and Bach *et al.*, (1999) were re-created using the published algorithm by Guyot *et al.*, (2012).⁴³
 - The re-created IPD from Bach *et al.* were then adjusted by subtracting 1 month and removing negative values to align with the Ganti *et al.* data. The authors justify this based on the following rationale:

“[KM] estimates from the reconstructed data are presented in the online appendix... This chart shows a steep drop from 0 to 1 month for the no-treatment arm and high survival probability for the surgery plus radiotherapy arm. This result may be due to the data being observational and the possibility that some patients may have died before the intended treatment started or that patients may have had a short life expectancy and did not receive any treatment. Conversely, the fittest patients may have been considered suitable for radiotherapy plus surgery. A sudden drop in survival at the start of a study is rarely observed in oncology trials, as they typically have an inclusion criterion that patients have a life expectancy of 3 months. To mimic a trial in which such patients would likely not be included, 1 month was subtracted from patients’ survival times in all treatment arms and records with a negative time removed from the data.” (Vickers, 2019, p.927).¹⁷⁴

- The best-fitting AFT model (Weibull) was fitted to the combined data to elicit a time acceleration factor (intended to adjust the Bach *et al.* data to better ‘match’ the Ganti *et al.* data), which was then applied to the Bach *et al.* data to produce adjusted data.
 - Survival times exceeding the maximum follow-up of the un-adjusted Bach *et al.* data were censored at the original maximum follow-up time.
 - The process of fitting the AFT model was repeated by cutting the data at annual time points, to check if the time acceleration factor was approximately constant over time.
 - No clear guidance was presented by Vickers for what could be done if this was not approximately constant over time, though it could be argued that this would suggest the data sources are not compatible (at least in the context of the proposed survival extrapolation method).
 - The resultant adjusted data (from applying the time acceleration factor) were then used to fit a Royston & Parmar spline-based model, which was used to inform estimates of the hazard of death from t_1 until the end of follow-up in the Bach *et al.* data (t_2).
- PH parametric models were also fitted to age- and sex-matched general population data, of which a 3-knot Royston & Parmar spline-based model was selected.
 - The general population model was adjusted via an HR which was estimated to adjust for differences in the estimated hazard of death at t_2 in the model based on the adjusted Bach *et al.* data and the general population.
 - The adjusted general population model was used to inform estimates of the hazard of death from t_2 until the end of the modelled time horizon (t_{max}).

This process yielded estimates of the long-term hazards for the control group. To account for the impact of a given intervention, further adjustments were made to the long-term hazards,

described by Vickers as “[HR] *tapering*” (Vickers, 2019, p.931)¹⁷⁴, for which a published cost-effectiveness study by Herring *et al.*, (2016) was cited.¹⁹⁵ While indirect comparison methods fall beyond the scope of this review, the HR tapering approach included within the approach by Vickers is described here for completeness.

In the study by Herring *et al.*, a scenario was considered in their cost-effectiveness analysis wherein the implied HR for OS between the intervention and the control gradually increases to unity after the end of follow-up in the clinical trial over a period of 5 years.¹⁹⁵ This is also referred to by Herring *et al.* as a “*waning treatment effect*” (Herring *et al.*, 2016, p.86).¹⁹⁵

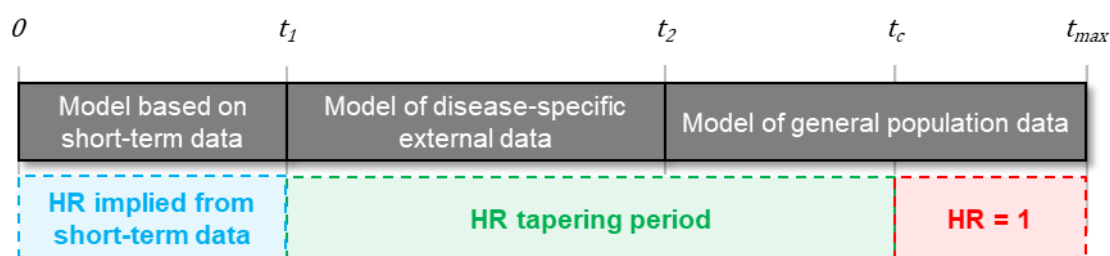
Vickers describes the application of the HR tapering based on three distinct time intervals, as shown in the equation below:

$$HR = \begin{cases} \text{Implied via trial data,} & \text{if } t < t_1 \\ \frac{h_B(t_1)}{h_A(t_1)} + \left(\frac{1 - \frac{h_B(t_1)}{h_A(t_1)}}{t_c - t_1} \right) t, & \text{if } t_1 \leq t \leq t_c \\ 1, & \text{if } t > t_c \end{cases}$$

In this equation, the HR for the period before the end of follow-up is implied via analysis of the trial data. Depending on the choice of model to inform the period of the extrapolation, this could be a constant or a function over time. After a specified time, t_c , the HR returns to unity. In the interim, the HR is estimated based on linear interpolation between the implied HR at t_1 , and the assumed HR of 1 at t_c . The choice of t_c is separate to the determination of t_2 , but in principle these values could be equal. The choice of t_c is selected based on the expected duration of a given treatment effect, which in the example presented by Vickers was simulated according to a Normal distribution with a mean of 10 years and a standard deviation (SD) of 2 years.¹⁷⁴ Therefore in the example by Vickers, no explicit external evidence is used to inform the HR tapering, except that an assumed distribution was used.

A diagram illustrating the method proposed by Vickers is presented in Figure 3-20. This diagram illustrates at which time point data from a given source are used to inform extrapolation, and how the HR comparing treatments is adjusted based on the HR tapering approach.

Figure 3-20: Illustration of Vickers method



Key: HR, hazard ratio.

Note: In this diagram a number of time points are shown. t_1 refers to the end of follow-up in the short-term data; t_2 refers to the end of follow-up in the disease-specific external data (which is assumed to be longer than the short-term data); t_c refers to a point in time at which the HR between treatment arms is assumed to return to unity; and t_{max} refers to the time point at which extrapolations are required over (e.g., lifetime).

Key assumption(s)

Owing to the lack of IPD for the disease-specific external data, this method assumes that the external data comprise a population similar to the index population which is used to estimate the short-term model. Similar to a number of other methods that make use of general population data, beyond a given time point patients are assumed to have zero excess hazard related to their disease (or in other words, could be considered as effectively 'cured' of their disease in terms of their LE).

Strengths and limitations

This method makes use of data from a given source for a period over which they are available, before switching to less generalisable but longer-term data. The method therefore is grounded in the logic of using the best available data as much as possible. However, in this method, no consideration is given to potential disagreements with extrapolations based on empirical data from the index population if they were hypothetically continued after a given cut point versus the external data (which is assumed to provide a better fit versus a standard extrapolation without any external evidence being considered). Furthermore, given the use of multiple cut points at which hazards change depending on the model being used, there are several discontinuities in the underlying hazard function across the whole modelled time horizon (see Chapter 2 for discussion related to possible issues that stem from hazard function discontinuity [Section 2.1.1]).

Rationale for inclusion in further investigation in the rest of the thesis

This method is a specific type of piecewise modelling approach with several adjustments, and so is included separately to the 'standard' piecewise model for completeness. However, it bears substantial similarity with other piecewise modelling methods.

Summary of first iteration

In the first iteration, 12 distinct methods were identified, reported in 15 studies.^{91, 146-148, 151-153, 158, 163, 165, 169, 171-174} Upon review of the identified methods, a number of areas for improvement in the original search strategies were identified, including:

- Four methods were identified that were described using Bayesian terminology, and so some other Bayesian studies may have been missed.
- Some studies referred to external evidence as 'historical', which was not included as a term within the original search strategies.
- Some studies described background mortality using terms such as 'baseline survival' or 'baseline hazards', which were also not included as terms in the original search strategies.

In addition, a relatively large number of studies were identified which reported on nomograms – a type of diagram which can be used to represent the relationship between different variables, and ultimately allow for the approximate calculation of a particular mathematical function. As nomograms are outside of the scope of the review, the search strategies could reasonably exclude nomograms if mentioned within the title or abstract (on pragmatic grounds).

Therefore, based on the points noted above, a second iteration was undertaken with some edits made to the search strategies.

3.3.2 Second iteration

Edits made to search strategy

On the basis of the areas for improvement identified as part of the first iteration, the search strategies were updated, as presented in Table 3-13 for searching Ovid. Edited lines are marked with a letter, with a footnote to explain the change made.

Table 3-13: Ovid searching (second iteration, Medline and EconLit)

#	Search term(s)
1	((extrap* or estimat* or project* or predict* or expect* or express*) adj3 survival).tw.
2 ^a	((parametric adj2 survival) or (surviv* adj2 (model* or curve*))).tw.
3	(extrap* adj2 (trial* or mortality)).tw.
4	((estim* or express*) adj2 life-expectancy).tw.
5 ^b	((Bayesian adj (model* or framework or surviv*)) or (prior* adj3 (data or information or hazard*))).tw.
6 ^c	((real-life or real-world) adj3 (data or evidence)) or (RWE or RWD)).tw.
7	(external adj (information or data)).tw.
8 ^c	(background mortality or life-table).tw.
9	competing risk*.tw.
10	(excess adj (mortality or hazard*)).tw.
11 ^d	(historical and (data or trial or control or evidence)).tw.
12	(expert* adj (information or opinion)).tw.
13	(observational adj2 data*).tw.
14 ^e	((disease or cause) adj (related or specific)) or baseline) adj (survival* or hazard*)).tw.
15	(registry* or registries).tw.
16	(reference adj2 population).tw.
17	or/1-4
18	or/5-16
19	17 and 18
20 ^c	(letter or historical article or editorial or conference abstract).pt.
21 ^c	19 not 20
22	exp animals/ not humans.sh.
23	21 not 22
24	limit 23 to english language
25 ^f	24 not nomogram.tw.

Note: ^a Line added in acknowledgement of some studies referring to survival models or survival curves, without saying 'parametric'. Some papers described extrapolating survival curves, or even abbreviated this to just extrapolating curves. ^b The previous strategy did not include any specific terms related to Bayesian terminology, and therefore was likely to have missed some relevant studies. This line was added to capture studies using these terms. ^c Some lines were combined and condensed in the interest of brevity, but the content of the search terms themselves was left unchanged. ^d This line was added in to address the fact that the term 'historical' was sometimes used to described external evidence. ^e This line was edited to acknowledge some studies referring to 'baseline survival' or 'baseline hazards'. ^f Added to remove nomogram studies that are not relevant. This increased specificity without compromising sensitivity of the search.

After running the second iteration for the three databases in Ovid, the previous animal filter (line 22 in Table 3-13) was found to not appropriately filter studies identified via EMBASE. Therefore, the Ovid search was run separately for Medline and EconLit versus EMBASE. The EMBASE search was identical to the search shown in Table 3-13, with the exception of line 22 which was changed from 'exp animals/ not humans.sh.' to '(exp animal/ or nonhuman/) not exp human/'.

These changes were carried through to the WOS search, with one additional edit. The original search made use of title and abstract fields for searching, but for simplicity the updated search was based on the topic field. The topic field searches the title, abstract, author keywords, and Keywords Plus® (a set of key words unique to the WOS database, which includes words and phrases that have been harvested from the titles of cited articles). This edit was made in the interest of efficiency in case the search was to be repeated in the future. The updated WOS search strategy for the second iteration is shown in Table 3-14.

Table 3-14: Web of Science searching (second iteration)

#	Search term(s)
1*	TS = (((extrap* OR estimat* OR project* OR predict* OR expect* OR express*) near/3 survival))
2*	TS = (((parametric near/2 survival) OR (surviv* near/2 (model* OR curve*))))
3*	TS = (extrap* near/2 (trial* or mortality))
4*	TS = ((estim* or express*) near/3 (life-expectancy))
5*	TS = (((Bayesian near/1 (model* or framework or surviv*) or (prior* near/3 (data or information or hazard*))))
6*	TS = (((real-life or real-world) near/3 (data or evidence)) or (RWE or RWD))
7*	TS = ((external near/1 (information or data)))
8*	TS = (background-mortality or life-table)
9*	TS = (competing risk*)
10*	TS = (excess near/1 (mortality or hazard*))
11*	TS = (historical and (data or trial or control or evidence))
12*	TS = (expert* near/1 (information or opinion))
13*	TS = (observational near/2 data*)
14*	TS = (((disease or cause) near/1 (related or specific)) or baseline) near/1 (survival* or hazard*)
15*	TS = (registry* OR registries)
16*	TS = (reference near/2 population)
17*	#4 OR #3 OR #2 OR #1
18*	#16 OR #15 OR #14 OR #13 OR #12 OR #11 OR #10 OR #9 OR #8 OR #7 OR #6 OR #5
19	#18 AND #17 Refined by: [excluding] databases: (MEDLINE) AND databases: (WOS) <i>Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English</i>
20†	ANIMAL STUDIES FILTER (see Appendix A)
21	(#19 not #20) Refined by: document types: (ARTICLE OR EARLY ACCESS OR BOOK OR REVIEW) <i>Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English</i>

Note: *Each of these lines also include the following additional information: 'Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English'.

†Owing to its length, the filter applied to remove animal studies from the search is presented separately in Appendix A for simplicity.

Number of identified studies

The search strategies presented in Table 3-13 (Ovid, accounting for the edit made for EMBASE which was searched separately) and Table 3-14 (WOS) were run on 1 July 2021. Searches were run to the latest date available as of the search date for each database. The date limits for each database are provided below:

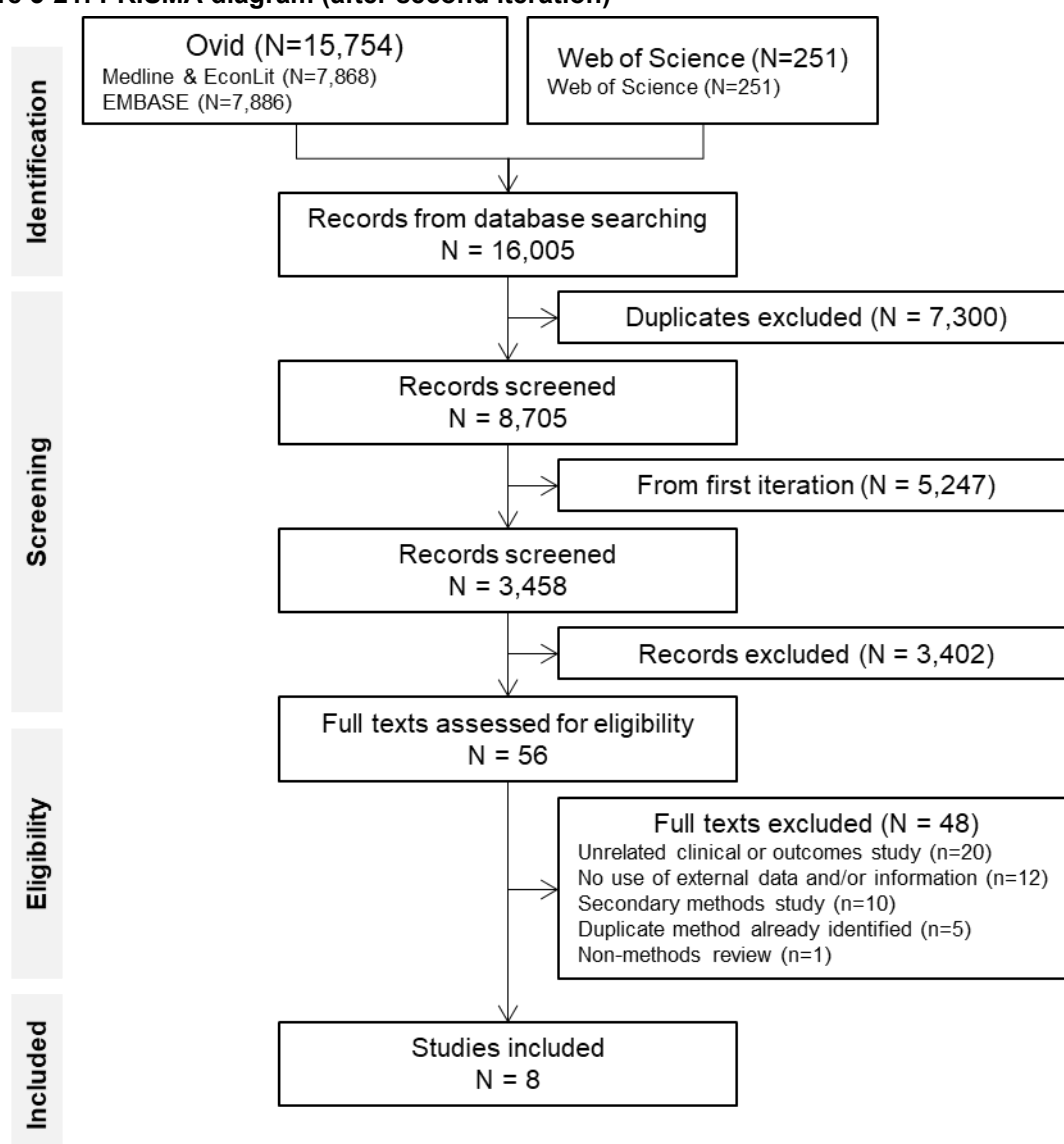
- Ovid
 - Econlit <1886 to 24 June 2021>,
 - Embase <1974 to 30 June 2021>,
 - Ovid MEDLINE® and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions® <1946 to 30 June 2021>
- WOS
 - WOS (All databases, later filtered) <1864 to 1 July 2021>

For the second iteration, a total of 7,868 records from the Ovid search of Medline and EconLit were identified, plus 7,886 records from the Ovid search of EMBASE, with an additional 251 records from the WOS search, giving a grand total of 16,005 records prior to de-duplication.

After de-duplication (using the built-in functionality of EndNote), 7,300 duplicate records were removed, leaving 8,705 remaining.

The records from the first iteration were then compared to those obtained via the second iteration, with those already considered in the first iteration removed. This led to 5,247 additional records being removed which were identified in the previous search, leaving a total of 3,458 records for primary screening. After primary screening based on titles and abstracts, 3,402 records were excluded, leaving a total of 56 records eligible for secondary screening. Following full-text screening, 48 records were excluded, which resulted in a total of 8 records suitable for data extraction.^{181, 196-202} The process of identifying suitable studies is presented in Figure 3-21, as a PRISMA diagram.¹⁷⁵

Figure 3-21: PRISMA diagram (after second iteration)



Tabulation of identified methods

Each of the 8 identified studies were grouped into one of the categories previously established as part of the first iteration. Further explanation of these methods is provided when discussing each method in turn (see 'Presentation of individual methods').

Methods taken forward for further evaluation in later chapters

A tabulated summary of the methods that are considered further in later thesis chapters is provided in Table 3-15. In total, 5 distinct methods were taken forward (plus 1 alternative version of a previously-identified method [relative survival]), reported in 4 studies.

Table 3-15: Methods identified in second iteration taken forward for further evaluation in later chapters

Reference	Method(s)	Theme(s)	Rationale for inclusion
Nelson <i>et al.</i> , (2007)	Relative survival spline-based model	General population	This method was acknowledged in a number of studies identified as part of the first iteration. Therefore, this study was included in addition to the method identified in the first iteration for completeness (though they describe the same overall approach to incorporating external evidence).
Soikkeli <i>et al.</i> , (2019)	Informative prior on shape parameter	Informative prior	This method represents a formalisation of a previous approach identified in the first iteration (Hawe <i>et al.</i> , 2014) ¹⁷¹ and is considered a novel means of using external data to inform the shape parameter of a given parametric survival model.
Felizzi <i>et al.</i> , (2021)	Informed mixture-cure model	- General population - Other	The informed mixture-cure model provides an alternative, novel means of drawing from external evidence sources to inform the estimation of a cured proportion. This approach is therefore included as an alternative to the 'uninformed', standard mixture-cure model wherein the cure fraction is estimated as part of the overall model fitting process.
van Oostrum <i>et al.</i> , (2021)	General population adjustment – external additive hazards	General population	This study presents a range of different approaches to make use of general population data to inform (or adjust) survival extrapolations. While the methods have been used in practice in previous economic evaluations, the authors of this study consolidate the approaches into one study, and therefore this group of methods is considered further as part of this thesis. The internal additive hazards approach is captured previously as part of the relative survival modelling methods discussed in the first iteration.
	General population adjustment – convergence of hazards	General population	
	General population adjustment – proportional hazards model	General population	

Methods not taken forward

A tabulated summary of the methods that are not considered further is provided in Table 3-16.

Table 3-16: Methods identified in second iteration not taken forward but included in the review

Reference	Method(s)	Rationale for exclusion
Jackson <i>et al.</i> , (2010)	Semi-parametric Bayesian model with constant hazard extrapolation	This approach does not yield a final model for OS extrapolated over a lifetime horizon. Instead, the method gives a means of estimating probabilities for inclusion within a state transition modelling framework, which falls outside the remit of this review.
Murray <i>et al.</i> , (2014)	Semi-parametric Bayesian commensurate survival model	This method does not itself inform extrapolations of survival. Rather, it focuses on estimating survival within the observed time period, and does not provide a mechanism by which extrapolations of lifetime survival could be obtained. Modifications would be needed in order for this approach to be suitable for use in survival extrapolation, and so is excluded on the basis of the method proposed (in its current form) not being suitable for extrapolation purposes.
Lewis <i>et al.</i> , (2019)	Bayesian hierarchical survival model	The broad approach of fitting a Bayesian hierarchical model provides a novel means of borrowing from historical data to inform estimation of survival outcomes. While this method may be considered a novel means of using historical data to inform survival estimation, the specific example presented by the authors of this study focuses predominantly on relative treatment effects, and therefore the approach is excluded.
Mba <i>et al.</i> , (2021)	Relative survival regression model correcting for potentially inaccurate background mortality rates	This method does not consider extrapolation of OS – instead, it aims to adjust background mortality rates used in relative survival models using data-driven, age-specific correction factors to reduce the bias in excess hazard estimates when life tables are misaligned with the true population characteristics. Therefore, it is unclear how the approach could practically improve extrapolations for HTA beyond potentially increasing confidence in differences in survival within given subgroups of patients.

Presentation of individual methods

Each of the identified method studies taken forward for the remainder of the thesis are discussed in this sub-section. Method studies that were discarded from further consideration are presented in Appendix C (with reasons provided for why specific methods were discarded).

Nelson *et al.*, (2007)

Theme(s): *General population.*

Nelson *et al.* were the first to propose extending Royston & Parmar flexible parametric models to be fitted within a relative survival setting.¹⁸¹ The authors explored the specification of an additive hazards, flexible, spline-based parametric survival model fitted in a relative survival

framework. Please refer to discussion of the relative survival studies identified as part of the first iteration for further information concerning the relative survival framework.

Key assumption(s)

The main assumption made within the approach proposed by the authors is that a flexible parametric model is needed to appropriately reflect the underlying disease-specific hazard of death. Otherwise, all other assumptions are linked to those related to any model fitted in a relative survival framework assuming additive hazards.

Strengths and limitations

As per the key assumptions, the main strengths and limitations of the approach presented in this study are essentially those that apply to all relative survival models (see Chapter 2, Section 2.1.2). Accordingly, these are not presented again here for brevity.

Rationale for inclusion

For completeness, this method is included as part of the range of approaches captured to reflect relative survival approaches based on ‘standard’ parametric models, spline-based models (as described in this study), cure-based models, and possible combinations of these.

Soikkeli et al., (2019)

Theme(s): Informative prior.

In this study, Soikkeli *et al.* describe the use of historical trial data to set *a priori* distributions when fitting parametric survival models.¹⁹⁹ The data available for consideration by Soikkeli *et al.* were:

- The VISTA study ([NCT00111319](#)): A Phase III, international, RCT of bortezomib + melphalan + prednisone (VMP) versus melphalan + prednisone (MP) for the treatment of patients with previously untreated MM.²⁰³ This comprises the study from which Soikkeli *et al.* intended to produce extrapolations, using a preliminary data cut comprising follow-up data until a maximum of 30 months.
- The MY.7 study ([NCT00002678](#)): A Phase III, Canadian, RCT comparing induction therapy with melphalan plus dexamethasone to MP, and maintenance therapy with dexamethasone to observation, in previously untreated MM patients.²⁰⁴ This comprises the historical study data used by Soikkeli *et al.*

The approach considers the use of historical trial data (i.e., the MY.7 data) in two ways. First, the authors determined the most suitable parametric model on the basis of the MY.7 study, choosing a Weibull model (though the informal use of historical data to identify suitable models

is outside the scope of this review). Then, the shape parameter of the Weibull model was used to define an informative prior distribution for the shape parameter of a different Weibull model fitted to the VISTA study data.

The authors of the study did not have access to the original IPD from either study, and so IPD were recreated via digitisation and use of the Guyot *et al.* algorithm.⁴³ To specify the prior on the shape parameter of the Weibull model, the authors consider the following notation for the survival function of a Weibull model:

$$S(t_1) = e^{(-\lambda_1 t_1^{\rho_1})}$$

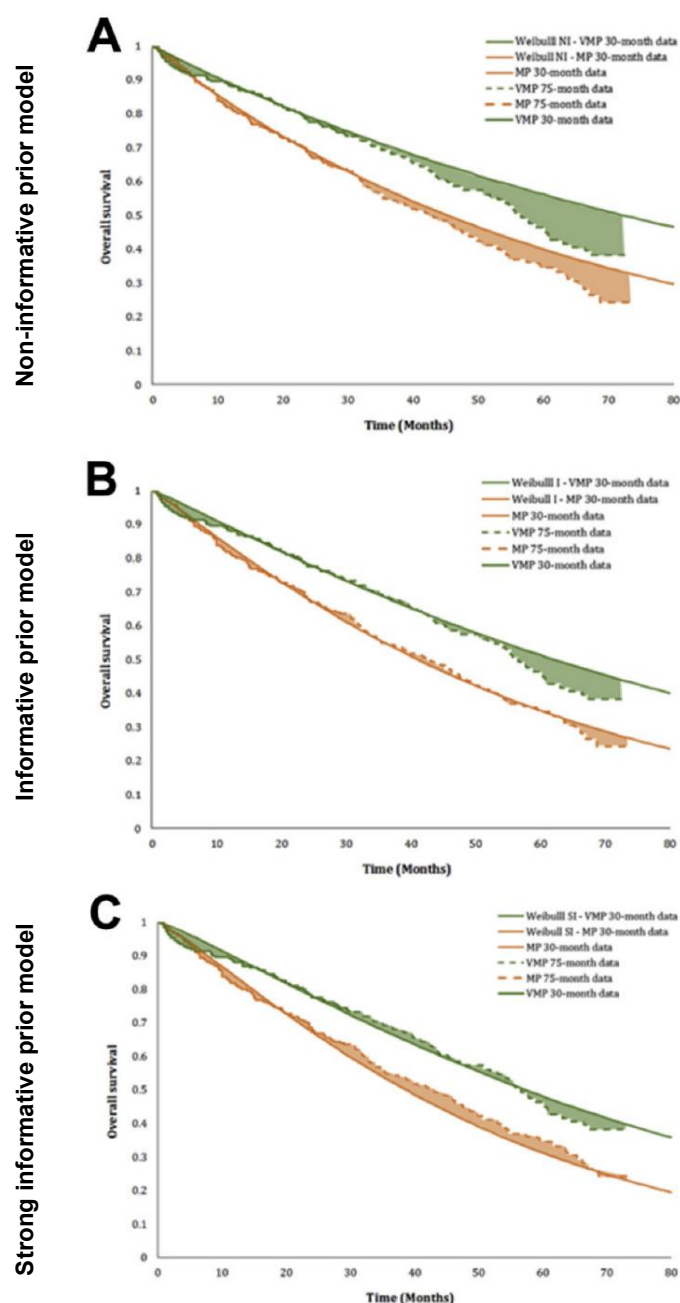
Here, λ_1 represents the scale parameter of the MY.7 study, and ρ_1 represents the shape parameter of the MY.7 study. In addition, t denotes time, and i refers to an individual patient in the MY.7 study. The authors then fitted a Weibull model to the MY.7 study with non-informative priors. Two prior distributions were specified. For the scale parameter, a non-informative normal distribution was specified, with mean $\mu = 0$ and precision $\tau = 0.000000001$, where precision is defined as the reciprocal of the variance. For the shape parameter, a non-informative gamma distribution was specified, with shape $\alpha = 1000$ and rate $\beta = 1000$.

After fitting the Weibull model to the MY.7 study, the authors then proceeded with fitting a Weibull model with an informative prior for the shape parameter to the VISTA study. Taking the mean (ρ_1) and SD (SD) of the shape parameter from the MY.7 study model, an informative gamma distribution was specified for the VISTA study model, with shape $\alpha = \rho_1 \beta$ and rate $\beta = \frac{\rho_1}{SD^2}$. For the scale parameter, a non-informative normal distribution was specified, with mean $\mu = 0$ and precision $\tau = 0.000000001$. The authors also fitted the Weibull model with a covariate for treatment arm, allowing for survival models for VMP and MP to be fitted within one process, and the prior for this covariate was established as per the scale parameter (i.e., non-informative normal).

Ultimately, the final analysis was described as the 'informative prior' scenario. In addition to the 'informative prior' scenario, the authors presented a model with non-informative priors ('non-informative prior' scenario) and explored a sensitivity analysis wherein the shape parameter was effectively 'fixed' at the value obtained from the analysis of the MY.7 study, labelled as the 'strong informative prior' scenario. Therefore, a total of three scenarios were presented. The authors repeated the analysis using a lognormal model instead of a Weibull model, but the approach to incorporate the data from the MY.7 study was essentially identical, and so the results of the lognormal analysis are not described here.

The results of the three models (non-informative prior, informative prior, and strong informative prior) were validated against a later data cut of the VISTA study (maximum follow-up of 75 months, versus 30 months used to fit the models). The resultant survival extrapolations produced by each model is presented in Figure 3-22.

Figure 3-22: Models with priors on shape parameter re-produced from Soikkeli *et al.*, (2019)



Key: I, informative prior; MP, melphalan + prednisone; NI, noninformative prior; SI, strong prior; VMP, bortezomib + melphalan + prednisone.

Note(s): This figure illustrates the non-informative, informative, and strong informative prior Weibull models presented in Soikkeli *et al.*, (2019).¹⁹⁹ This is taken from panels A-C of Figure 2 from the original study publication.

Source: Copyright © 2019 Soikkeli *et al.* This figure is re-produced according to the [Creative Commons Attribution-NonCommercial-No Derivatives License \(CC BY NC ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/). Soikkeli F, Hashim M, Ouwers M, Postma M, Heeg B. Extrapolating Survival Data Using Historical Trial-Based a Priori Distributions. Value Health. 2019 Sep;22(9):1012-1017. doi: 10.1016/j.jval.2019.03.017.

Key assumption(s)

The primary assumption made in this analysis is that the shape parameter obtained from the model fitted to the MY.7 study is at least in principle ‘similar’ to the expected shape parameter of the model to be fitted to the VISTA study. Further to this, the approach assumes the same model can be assumed to apply for both studies (e.g., in this case, a Weibull model is the best model for both studies), and that the studies themselves are reasonably similar such that the external source provides relevant information as no formal adjustment of the MY.7 study data to account for potential differences in baseline patient characteristics was undertaken as part of this analysis. However, the authors stated that *“In case the patient characteristics between the trials differ, population-adjusted indirect comparison methods can be used to correct for these differences”* (Soikkeli *et al.*, 2019, p.1016)¹⁹⁹, and cited a study by Signorovitch *et al.* detailing how to perform a MAIC.¹²¹

Strengths and limitations

This method provides a simple yet intuitive application of prior information to inform survival extrapolation, within a consistent framework. It benefits mostly from the transparency in the approach, and the ability to determine how influential the prior information is on model results through the specification of the precision parameter of the prior distribution (i.e., the value of the *SD* parameter could be arbitrarily increased or decreased to explore the impact on results).

However, this in turn highlights a key limitation of the analysis – that it requires the user to specify the precision parameter, which if small may effectively yield near-identical results to the non-informative model, which (as highlighted by the authors) may, in general, provide *“similar results as standard frequentist approaches”* (Soikkeli *et al.*, 2019, p.1014).¹⁹⁹ In addition, while the focus of this review is not on the estimation of a relative treatment effect in general, the authors highlighted that the ability to specify a prior on the shape parameter within a model fitted within a PH framework (such as the Weibull in this illustrative example) relies on the PH assumption.

Rationale for inclusion in further investigation in the rest of the thesis

This approach formalises the use of prior information on the shape parameter of a parametric model, previously described via the study by Hawe *et al.*, (2014). This method is therefore included as an alternative approach to do this, within a Bayesian framework.

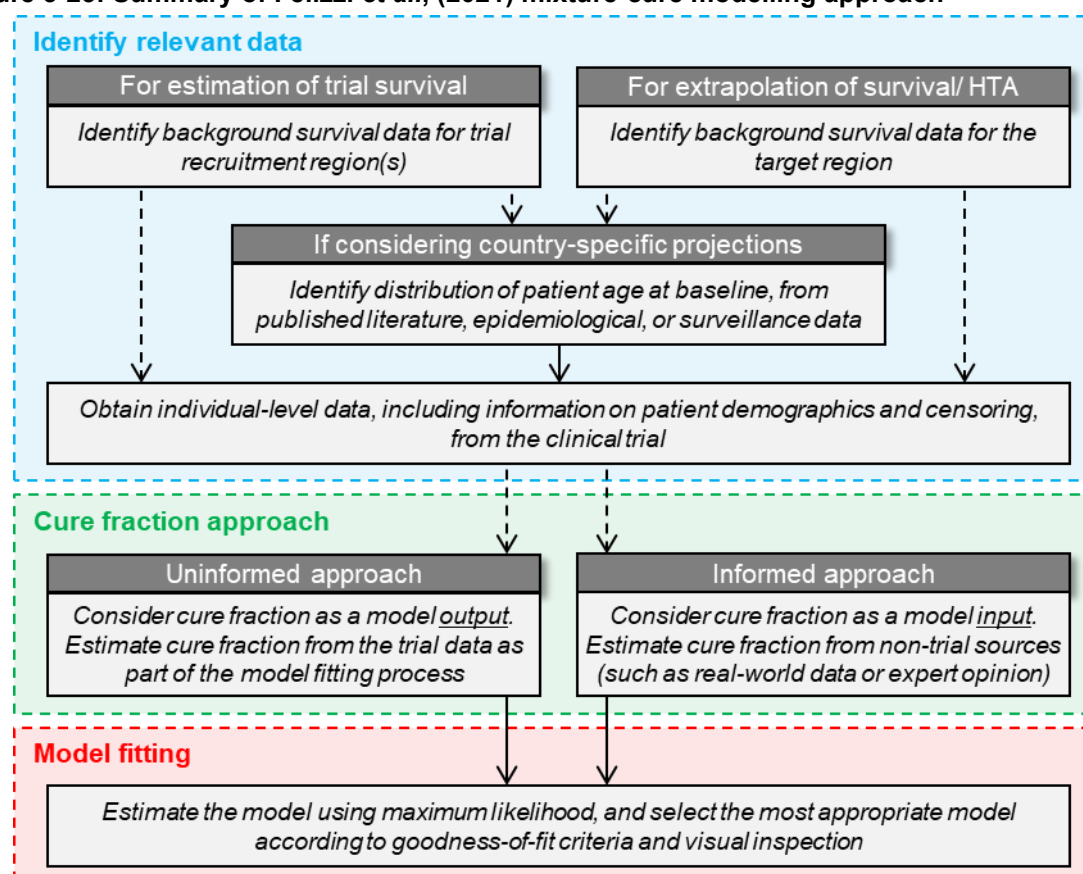
Felizzi et al., (2021)

Theme(s): *General population, other.*

Felizzi *et al.* developed a tutorial paper on how to fit two different types of MCM.²⁰¹ The first of these is described as an ‘uninformed’ approach, in which the cure fraction (π , based on standard notation) is estimated from trial data; whereas the second is described as an ‘informed’ approach, in which π is “*obtained from an external source (e.g., real-world data) used as an input to the model*” (Felizzi *et al.*, 2021, p.143).²⁰¹ The first (uninformed) approach has been discussed previously in Chapter 2 (Section 2.3.2), and within the context of several other cure-based modelling approaches identified in the first iteration of the search, whereas the second (informed) approach was not previously identified. The remainder of the discussion on this study is focused on the informed approach, but some aspects also apply to the uninformed approach.

In setting up the informed MCM, the authors describe the need to follow a staged process to identify relevant data to inform the model, then the approach to take to informing π , and then ultimately fitting the model. This process is described in Figure 3-23, and illustrates that the main difference between the informed and uninformed approaches is if π is considered a model *input* (informed approach) or a model *output* (uninformed approach).

Figure 3-23: Summary of Felizzi et al., (2021) mixture-cure modelling approach



Key: HTA, health technology assessment.

The authors explain that in the informed approach, it is necessary to obtain π to include as a model input (which is treated as a fixed parameter as part of the model fitting process). While specific examples are not highlighted, the authors explain that if the external data are available at the individual level, it may be necessary to fit a 'helper' MCM and use the outputted π from this MCM as an input to the informed MCM to be fitted to the trial data of interest.

Key assumption(s)

In the informed MCM approach, there is one critical, additional assumption imposed in addition to standard assumptions made within cure modelling – that there exists an external evidence source which can be used to obtain an estimate of π , which is subsequently considered a 'true' value, and is fixed as part of the subsequent model fitting process.

Strengths and limitations

The key strength of this approach over a more traditional, uninformed MCM approach is that a clinically plausible MCM could still be obtained with limited follow-up data through estimating π with external evidence (such as real-world data or clinical expert opinion). The informed

approach therefore provides an alternative means of estimating the parameters for an MCM making use of all potentially relevant information.

In terms of limitations, the authors highlight that the ability to fit both MCMs using the methods described requires the availability of IPD, including demographic information, which may not always be available for HTA purposes (for example, in the context of a non-trial comparator treatment, comparisons may be based on published summary data only). In addition, as noted above, the informed π is considered 'fixed', and so any uncertainty in this parameter may not be adequately reflected within a given analysis. The value of π may also need to be re-considered in light of the parametric model chosen to represent the uncured proportion. Furthermore, the estimation of π may be especially challenging for new interventions, given that long-term data are unlikely to exist, and so it may not be feasible to use this method in this context.

A further limitation is implied through the specification of background mortality data to inform the model. The authors explain: *"If the focus of the model is survival estimation within the trial, target countries are those from which trial participants were recruited. In contrast, if the focus is on extrapolation beyond the trial's geographical and time scope (e.g., as part of an HTA assessment), the target country is the country for which the extrapolation is to be conducted"* (Felizzi *et al.*, 2021, p.145).²⁰¹ This highlights a potential discordance between the trial and target populations, which may or may not be possible to address in the way described by the authors (if, for example, disease-specific survival varies by region).

Rationale for inclusion in further investigation in the rest of the thesis

This approach is included as an alternative means of fitting an MCM using external evidence to inform π .

van Oostrum *et al.*, (2021)

Theme(s): General population.

van Oostrum *et al.* describe a range of different approaches that could be undertaken to incorporate general population mortality into survival extrapolations methods, focusing on HTA of cancer drugs.²⁰² The authors explore a range of four alternative approaches, which are described in turn below, in addition to an approach which includes no general population adjustment (i.e., no use of external evidence).

Internal additive hazards model

This approach is described as adding general population mortality hazards into the log-likelihood function of a parametric distribution as part of the model fitting process. Therefore, the approach could be considered as a standard application of a relative survival model, with

additive (as opposed to multiplicative) hazards. The additive hazards approach has been previously described in general, and so is not discussed further here for brevity. Further details can be found in Chapter 2 and in discussion of a study by Demiris & Sharples, (2006).¹⁵¹ The authors found that this method “*had the highest face validity for all data sets*” compared with the other approaches (van Oostrum *et al.*, 2021, p.1294).²⁰² For clarity, this approach does not assume data are available that determine the cause of death at the individual patient level.

External additive hazards model

The external additive hazards approach is similar to the internal approach but assumes that the general population hazards are negligible for the duration of follow-up available for the trial population, and so these are ‘added’ on top of the hazard of death estimated by fitting a standard survival model to the trial data. The authors explain that over time, the general population mortality hazards are expected to increase and so assuming these are negligible during the trial follow-up means that they only materially impact longer-term extrapolations. However, if general population hazards cannot be assumed to be negligible (e.g., if considering an elderly trial population) “*this approach will overestimate the probability of an event during the trial follow-up period.*” (van Oostrum *et al.*, 2021, p.1297).²⁰² This means that the total hazard of death may be affected by some degree of double counting of non-disease-related deaths.

Converging hazards model

In this approach, the hazards estimated from an unadjusted model are applied until the general population mortality hazards exceed the unadjusted model. After this time point, general population mortality hazards are applied for the remainder of the modelled time horizon (effectively ‘capping’ the hazards projected by the model). Elements of the ‘converging hazards’ approach have been applied in a number of approaches previously described as piecewise methods identified in the first iteration, but this method is considered separately here for completeness. One possible interpretation of the converging hazards model is that after the time point at which general population hazards are applied, patients are assumed to be ‘cured’ as the overall hazard of death experienced is per the general population.

Proportional hazards model

In the PH model, the total hazard of death is estimated according to general population mortality hazards multiplied by an HR representing excess mortality experienced by the trial population versus the general population. The authors do not explain how the HR is estimated, but it is speculated that this may be using a Cox PH model, given that this is a standard approach to estimate an HR in survival analysis.⁴⁵ This method relies on the PH assumption holding between the trial and general populations across the entire time horizon being

considered, which may be challenging to validate in the context of some populations or may simply be unsuitable in some situations.

Key assumption(s)

A summary of the key assumptions made by each approach is presented below:

- **Internal additive hazards model:** Hazards related to disease and other causes are additive (not multiplicative, as has been explored in some other previous relative survival modelling approaches).
- **External additive hazards model:** In addition to the assumptions made for the internal additive hazards model, the external model assumes that general population hazards are negligible during trial follow-up (i.e., the impact of double-counting non-disease related deaths during the trial follow-up period is negligible).
- **Converging hazards model:** Without adjustment, the parametric model would yield estimates of long-term hazards that fall below those for the age- and sex-adjusted general population. This method assumes it is reasonable to effectively 'ignore' the implausible long-term extrapolation in favour of applying general population hazards. Indirectly, this method also assumes no disease-related mortality after convergence, which is analogous to assuming 'cure'.
- **Proportional hazards model:** As described previously, the relationship between general population survival and the survival of the group of interest can be adequately represented by the specification of a single HR.

Strengths and limitations

As general population mortality data are routinely available in the UK, the approaches described by the authors could be readily applied with limited concerns regarding availability of data. Each approach considers a transparent and intuitive application of general population data to inform extrapolations, and as a range of options are presented it can be determined how influential the introduction of the data could be on extrapolations versus no inclusion at all, and versus other ways of incorporating it.

The authors highlight that the internal additive hazards approach appears to generally outperform the other approaches, suggesting that the other methods may be obsolete (though this is not explicitly stated by the authors). In addition, the external evidence used comes from the general population, and so this type of information may not be influential on survival extrapolations in some circumstances (e.g., for populations with terminal diseases). Another limitation is that the authors fitted models using each approach via a Bayesian modelling framework, which may limit the ability of analysts to make use of standard software currently employed routinely to produce survival extrapolations to inform HTA.

Rationale for inclusion in further investigation in the rest of the thesis

The approaches described above each provide alternative methods to incorporate general population mortality hazards within survival extrapolations. Even though elements of each method are considered in different approaches, each method is taken further as a novel approach to specifically incorporate only general population hazards into survival extrapolations.

Summary of second iteration

In the second iteration, 7 distinct methods were identified and taken forward, reported in 5 studies.^{181, 198, 199, 201, 202} Upon review of the identified methods, no additional search terms were identified that could plausibly yield additional hits through additional database searching, and so no further database searches were considered. Of the 5 studies identified in the second iteration, two were published after the first iteration was completed, and the remaining three were each identified as a result of including the additional Bayesian search terms. The search then proceeded to citation and reference checking, a review of conference proceedings, and verification of the final set of identified methods.

3.3.3 Additional searching methods

Citation searches (forward citation searching)

Each of the studies included as a 'pearl' in either the first or second iteration were included within the citation searching step. Citation tracking was performed using Google Scholar (<https://scholar.google.co.uk/>). The record for the identified paper was found, and then the built-in 'cited by' functionality was used to determine how many papers cited a given study. Papers were then searched to identify whether any additional studies were missed by the formal database searching. Citation tracking was undertaken in the period October to November 2021, with the findings summarised in Table 3-17.

Table 3-17: Citation tracking results

Study	#	Comments
Jackson <i>et al.</i> , (2016)	61	No additional studies identified
First iteration		
Piecewise methods		
Larkin <i>et al.</i> , (2015)	14	No additional studies identified
Barlev <i>et al.</i> , (2017)	10	No additional studies identified
Fisher <i>et al.</i> , (2017)	75	No additional studies identified
Bayesian methods		
Demiris & Sharples, (2006)	31	No additional studies identified
Benaglia <i>et al.</i> , (2015)	15	No additional studies identified
Guyot <i>et al.</i> , (2017)	35	No additional studies identified
Cope <i>et al.</i> , (2019)	9	No additional studies identified
Relative survival methods		
Lambert <i>et al.</i> , (2007)	238	One study identified: Botta <i>et al.</i> , (2021).
Andersson <i>et al.</i> , (2013)	94	No additional studies identified
Jakobsen <i>et al.</i> , (2019)	6	No additional studies identified
Other methods		
Nelson <i>et al.</i> , (2008)	30	No additional studies identified
Hawe <i>et al.</i> , (2014)	1	No additional studies identified
Hwang <i>et al.</i> , (2017)	28	No additional studies identified
Pennington <i>et al.</i> , (2018)	3	No additional studies identified
Vickers, (2019)	5	No additional studies identified
Second iteration		
Bayesian methods		
Lewis <i>et al.</i> , (2019)	11	No additional studies identified
Soikkeli <i>et al.</i> , (2019)	4	No additional studies identified
Relative survival methods		
Nelson <i>et al.</i> , (2007)	231	No additional studies identified
Other methods		
Felizzi <i>et al.</i> , (2021)	1	No additional studies identified
van Oostrum <i>et al.</i> , (2021)	0	No additional studies identified

Botta *et al.*, (2021)

This study builds on a standard MCM by including a relative risk, α , to reflect ‘cured’ patients maintaining an extra risk of dying versus the general population which is assumed to be due to causes other than the cancer.²⁰⁵ Therefore, this method appears similar in concept to the approach of Hinchcliffe *et al.*, (2012)¹⁶² (see Appendix C), which was excluded due to a combination of a lack of practical use within the context of HTA (since adjusting for a large proportion of deaths observed in the general population is unlikely to be consequential for decision-making). However, this method considers a differential risk in terms of the overall risk of death (as opposed to the proportion of deaths related to the cancer in the general population, per the method of Hinchcliffe *et al.*), which may be more applicable to an HTA context. Nevertheless, as only an abstract was available, this method was excluded on the basis of it not being possible to understand precisely how the method would be re-produced in other contexts without additional information concerning the model fitting process. Also, it should be noted that the concept of an SMR is commonly used in HTA to account for potential

differences in survival relative to the general population. For further details about the use of SMRs, please refer to Chapter 1 (Section 1.3) and Chapter 2 (Section 2.1.2).

Reference checking (backward citation searching)

Each of the studies included as a 'pearl' in either the first or second iteration were included within the reference checking step. Reference checking was performed searching the reference list of each study to identify whether any additional studies were cited that were missed by the formal database searching. The process of reference checking was undertaken performed in the period October to November 2021, with the findings summarised in Table 3-18. Please note that a 2024 update search was performed following conclusion of the initial SLR.

Table 3-18: Reference checking results

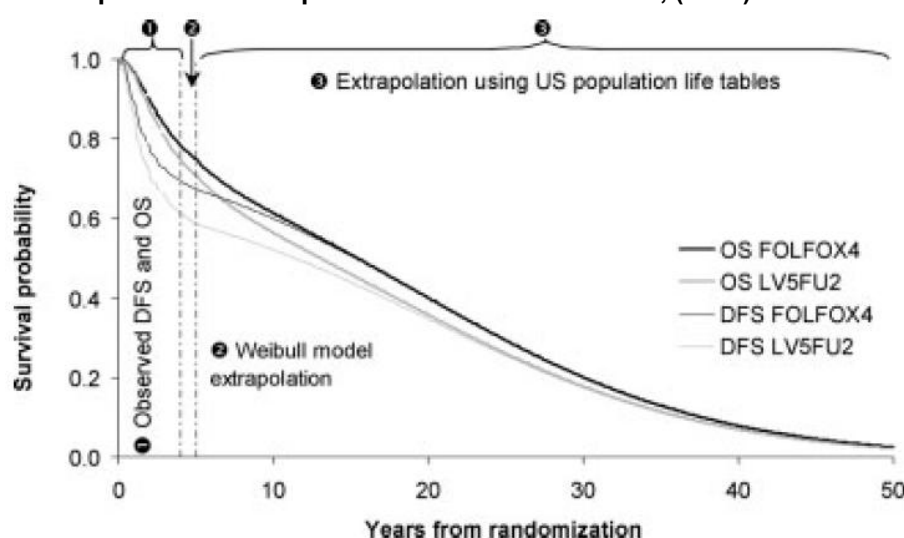
Study	#	Comments
Jackson <i>et al.</i> , (2016)	73	Three economic evaluations identified by Aballéa <i>et al.</i> , (2007), Hisashige <i>et al.</i> , (2008), and Howard <i>et al.</i> , (2009). These studies were initially excluded (due to study design) but are included here for completeness.
First iteration		
Piecewise methods		
Larkin <i>et al.</i> , (2015)	30	No additional studies identified
Barlev <i>et al.</i> , (2017)	26	No additional studies identified
Fisher <i>et al.</i> , (2017)	35	No additional studies identified
Bayesian methods		
Demiris & Sharples, (2006)	24	No additional studies identified
Benaglia <i>et al.</i> , (2015)	10	No additional studies identified
Guyot <i>et al.</i> , (2017)	56	No additional studies identified
Cope <i>et al.</i> , (2019)	31	No additional studies identified
Relative survival methods		
Lambert <i>et al.</i> , (2007)	35	No additional studies identified
Andersson <i>et al.</i> , (2013)	37	One study identified by Caro <i>et al.</i> , (2004)
Jakobsen <i>et al.</i> , (2019)	23	No additional studies identified
Other methods		
Nelson <i>et al.</i> , (2008)	32	One study identified by Caro <i>et al.</i> , (2004)
Hawe <i>et al.</i> , (2014)	7	No additional studies identified
Hwang <i>et al.</i> , (2017)	25	No additional studies identified
Pennington <i>et al.</i> , (2018)	48	No additional studies identified
Vickers, (2019)	41	No additional studies identified
Second iteration		
Bayesian methods		
Lewis <i>et al.</i> , (2019)	19	No additional studies identified
Soikkeli <i>et al.</i> , (2019)	18	No additional studies identified
Relative survival methods		
Nelson <i>et al.</i> , (2007)	29	No additional studies identified
Other methods		
Felizzi <i>et al.</i> , (2021)	59	No additional studies identified
van Oostrum <i>et al.</i> , (2021)	42	No additional studies identified

Aballéa et al., (2007)

This is an economic evaluation identified via the JR.²⁰⁶ This method is excluded from further consideration in later chapters of the thesis, as the method used is essentially a three-component piecewise model, similar to the approach taken by Larkin *et al.*, (2015).¹⁴⁶ The extrapolation is presented in Figure 3-24, and comprises three components:

- OS ≤ 4 years based on the KM estimate.
- $4 < \text{OS} \leq 5$ years inferred via a Weibull model extrapolation.
- OS > 5 years inferred via extrapolation using general population estimates.

Figure 3-24: Three-part curve fit re-produced from Aballéa *et al.*, (2007)



Key: DFS, disease-free survival; FOLFOX4, oxaliplatin/5-fluorouracil/leucovorin; LV5FU2, 5-fluorouracil plus leucovorin; OS, overall survival.

Note(s): This figure illustrates the three-part curve fit presented in Aballéa *et al.*, (2007). This is taken from Figure 2 from the original study publication.

Source: Re-produced from Aballéa *et al.*, (2007): Aballéa S, Chancellor JV, Raikou M, Drummond MF, Weinstein MC, Jourdan S, Bridgewater J. Cost-effectiveness analysis of oxaliplatin compared with 5-fluorouracil/leucovorin in adjuvant treatment of stage III colon cancer in the US. *Cancer*. 2007 Mar 15;109(6):1082-9. Figure presented here with permission from John Wiley and Sons. Copyright for this figure is retained by John Wiley and Sons.

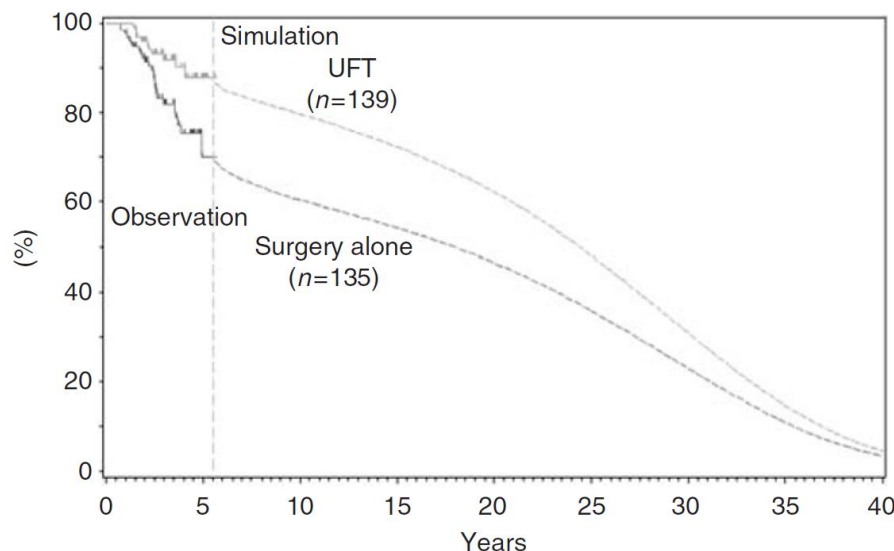
Hisashige et al., (2008)

This is an economic evaluation identified via the JR.²⁰⁷ This method is excluded from further consideration in later chapters of the thesis as the method used is represented by a combination of a piecewise method (KM estimate followed by extrapolation), and the competing-risks model by Maetani *et al.*, (2004).¹⁵⁵ The study by Maetani *et al.* was specifically cited by the authors, but was excluded in the systematic review because it is essentially a relative survival application of a cure model – please see Appendix D for details.

In brief, Hisashige *et al.*, estimated survival through use of the KM estimate until 5.6 years, followed by a 'competing-risks' model (as defined by Maetani *et al.*).¹⁵⁵ The authors do not

provide detail on exactly how the model was fitted, but they explain that the model has two components: a disease-specific survival curve and a disease-independent survival curve. The disease-specific survival curve was estimated using a log-logistic model, and the disease-independent survival curve was estimated by “the survival curve of the general population matched for age and sex of the subjects” (Hisashige *et al.*, 2008, p.1,234).²⁰⁷ The resultant survival curves are presented in Figure 3-25.

Figure 3-25: Survival extrapolation re-produced from Hisashige *et al.*, (2008)



Key: UFT, uracil-tegafur.

Note(s): This figure illustrates the survival extrapolation presented in Hisashige *et al.*, (2008). This is taken from Figure 1 from the original study publication.

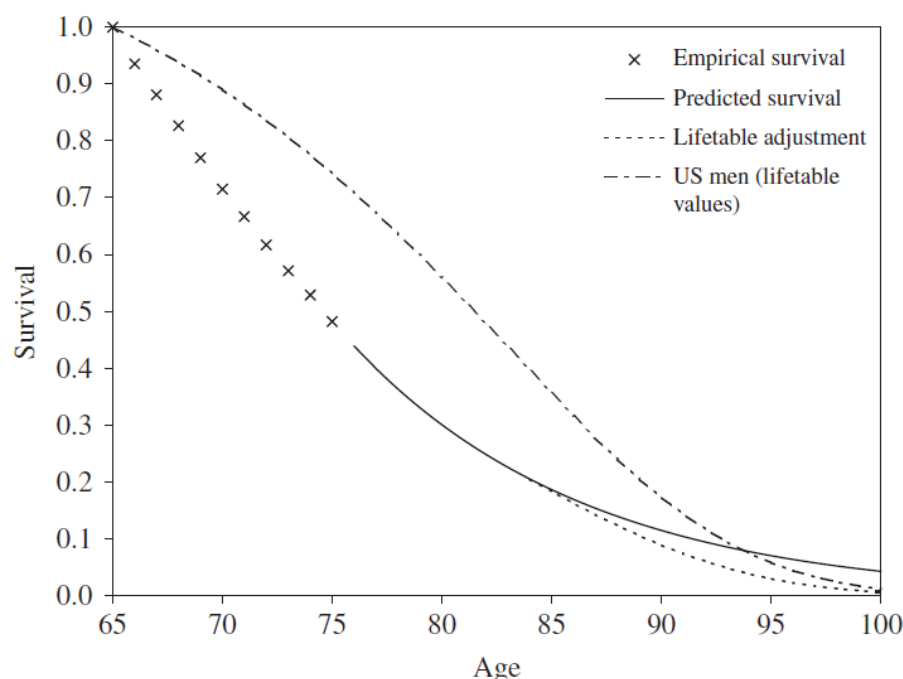
Source: Copyright © 2008 Hisashige *et al.* This figure is re-produced according to the [Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License](https://creativecommons.org/licenses/by-nc-sa/3.0/). Hisashige A, Yoshida S, Kodaira S. Cost-effectiveness of adjuvant chemotherapy with uracil-tegafur for curatively resected stage III rectal cancer. *Br J Cancer*. 2008 Oct 21;99(8):1232-8. doi: 10.1038/sj.bjc.6604666.

Howard *et al.*, (2009)

This is an economic evaluation identified via the JR.²⁰⁸ This method is excluded from further consideration in later chapters of the thesis as the method used is very similar to the three-component piecewise model approach taken by Larkin *et al.*, (2015).¹⁴⁶ The method makes use of empirical data for the first 10 years, after which a parametric model is fitted. However, if the extrapolated survival exceeded general population estimates, the curve was ‘adjusted’ to ensure survival is no higher than expected in the general population.

An example extrapolation is presented by Howard *et al.* and is re-produced in Figure 3-26 for completeness. In this illustration, the crosses represent the estimates based on observed data, and the solid line represents a Weibull model extrapolation. The long-dashed line represents the survival experience of the age- and sex-adjusted general population, with the dotted line showing the final model after adjusting for background mortality.

Figure 3-26: Example survival extrapolation re-produced from Howard *et al.*, (2009)



Note(s): This figure illustrates an example survival extrapolation presented in Howard *et al.*, (2009), for a 65-year-old male with early-stage cancer. This is taken from Figure 2 from the original study publication.

Source: Re-produced from Howard *et al.*, (2009): Howard DH, Tangka FK, Seeff LC, Richardson LC, Ekwueme DU. The impact of detection and treatment on lifetime medical costs for patients with precancerous polyps and colorectal cancer. *Health Econ.* 2009 Dec;18(12):1381-93. Figure presented here with permission from John Wiley and Sons. Copyright for this figure is retained by John Wiley and Sons.

Caro *et al.*, (2004)

Caro *et al.* considered an analysis of survival for patients diagnosed with peripheral arterial disease, myocardial infarction, or ischemic stroke, focusing on the population of the CAPRIE clinical trial.^{209, 210} However, instead of extrapolating estimates of survival using data collected from the CAPRIE study, the authors collected and analysed data from an external data source: the Saskatchewan Health administrative database, in order to produce survival estimates for a similar population to the CAPRIE study. The external data source had substantially longer follow-up versus the CAPRIE trial: *“Follow-up data [from the Saskatchewan Health administrative database] were obtained for these analyses up to December 31, 2000 and history back to January 1, 1980.”* (Caro *et al.*, 2004, p.629).²⁰⁹

The approach presented within this study makes use of an external data source to produce estimates of survival in a similar population; however, the method itself does not consider incorporating the external data with the data available for the population of primary interest (in this case, the CAPRIE clinical trial cohort). In other words, the method essentially ignores the trial data, except to consider a ‘similar’ cohort in terms of their baseline characteristics). Therefore, this approach is not considered further.

Conference proceedings search

The ISPOR Presentations Database does not allow for an in-depth search strategy to be constructed and executed, and it was not possible to search for one overarching string of possible terms. Therefore, keyword searching was instead used to identify potentially relevant studies. Keywords were used to identify studies that included the words or phrases presented in Table 3-19 within the title or abstract.

Table 3-19: ISPOR Scientific Presentations Database search strings

Search term(s)	Hits
"survival" AND extrapolat*	248
"survival" AND ("curve" or "curves")	26
"long-term survival"	256
(historical OR external) AND (data OR information)	244
Total number of abstracts and titles screened	774

The keywords searched were identified on the basis of those included in the search strategies used for formal database searching. However, in addition, the phrase 'long-term survival' was included as a separate string owing to the fact that the ISPOR Presentations Database allows for a more targeted search of studies referencing this particular phrase (as opposed to all published literature). Conference proceedings were excluded if they were presented at an ISPOR meeting held before 2020, based on the expectation that methods presented earlier than this time are expected to have subsequently been published in a journal. This captured presentations from the following ISPOR meetings:

- ISPOR Europe 2021 (virtual conference).
- ISPOR 2021, Montreal, Canada.
- ISPOR Europe 2020, Milan, Italy.
- ISPOR Asia Pacific 2020, Seoul, South Korea.
- ISPOR 2020, Orlando, Florida, USA.

Only meetings up until the end of 2021 were considered owing to the timeframe when this search was conducted. As conference proceedings typically provide a more limited description of the methodological approach(es) taken versus a full manuscript publication, a brief summary of the identified methods from conference proceedings is shown below in lieu of attempting to complete the data extraction table used for the published studies.

In total, three potentially relevant methods were identified, all contained within presentations from the ISPOR Europe 2021 virtual conference.²¹¹⁻²¹³ Of these presentations, only one (by Che *et al.*) provided sufficient information to allow the reader to understand how the method may be practically applied, yet even in this study full details were not presented.²¹¹ The method

of Che *et al.*²¹¹ may be considered an extension to a more traditional piecewise analysis where there is a gradual as opposed to immediate change in projected hazards.

Details of the other two methods are provided in Appendix D. However, in brief, the approach of Briggs *et al.*²¹² considered a Bayesian MPES, drawing parallels to the approach of Guyot *et al.*, (2017)¹⁵³ but with limited details presented. Finally, the approach of Pham *et al.*²¹³ considers using an external data source as a ‘third arm’, which appears similar to the approach of Hawe *et al.*, (2014)¹⁷¹ or Pennington *et al.*, (2018)¹⁷³, though it is not possible to confirm this from the available information.

Owing to reporting limitations, only the method of Che *et al.*, (2021) was considered further as part of my research.²¹¹

Table 3-20: Methods identified in conference proceedings search

Reference	Method(s)	Theme(s)	Rationale for inclusion or exclusion from later thesis chapters
Che <i>et al.</i> , (2021)*	Blended survival model	- Piecewise - Informative prior	Include. Represents a novel extension of a standard piecewise approach, by considering a gradual re-weighting of hazards from two different models.
Briggs <i>et al.</i> , (2021)	Bayesian multi-parameter evidence synthesis	N/A	Exclude. Insufficient reporting.
Pham <i>et al.</i> , (2021)	Use of historical trial data as a third arm	N/A	Exclude. Insufficient reporting.

Key: N/A, not applicable.

Note: *This method was later published as a full paper – see Section 3.3.5.

Further details of the method of Che *et al.* are presented below, as well as details of the other two excluded methods. Please note that in the 2024 update to this review (see Section 3.3.5), the full paper by Che *et al.*, (2023)²¹⁴ was identified – this is the same method, but the paper provides further details.

Che *et al.*, (2021 and 2023)

Theme(s): Piecewise, informative prior.

Note: This method was originally published as a conference abstract in 2021, but later as a full paper in 2023. The description below is based on a combination of both sources of information about the method.

Che *et al.* consider a piecewise approach where two different survival curves are available: one from the trial population of interest (denoted 'obs'), and another from an external source (denoted 'ext'). The blended survival curve (denoted 'ble') is estimated as a weighted average of the 'obs' and 'ext' curves:

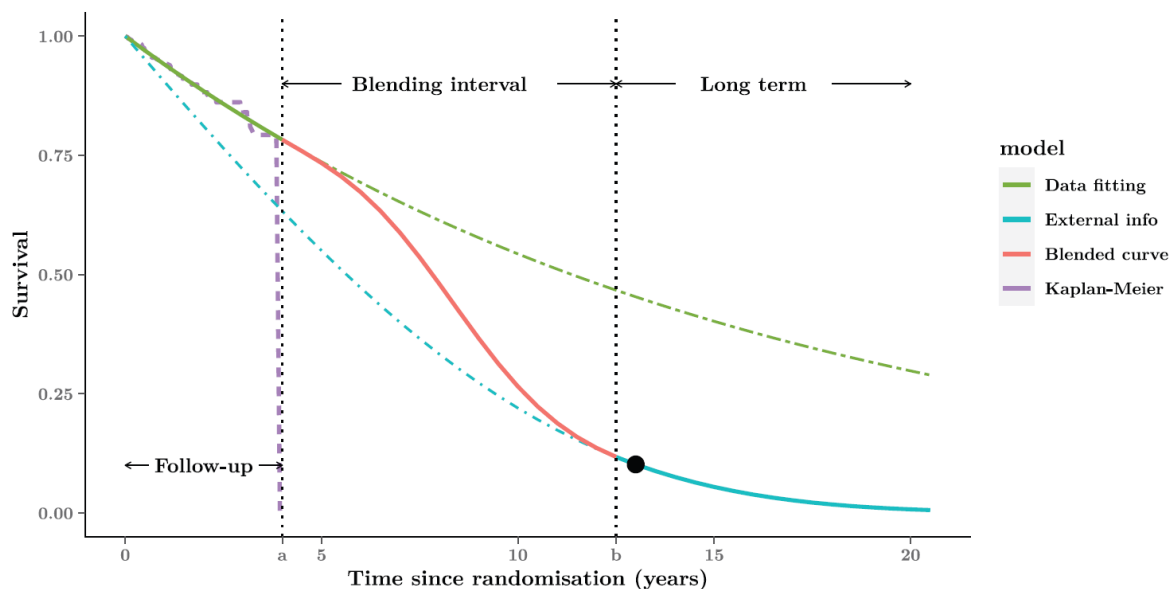
$$\log S_{ble} = (1 - p) \cdot \log S_{obs} + p \cdot \log S_{ext}$$

Here, p is defined as the weight applied to the survival function for the 'ext' group (S_{ext}). In the example presented by the authors, p takes a value of 0 until a specified point in time (a) and takes a value of 1 after a different (later) point in time (b). This means that prior to a , the external curve has no influence on survival, and beyond b it is fully driving estimates of survival.

Between time points a and b , p takes a value of between 0 and 1, based on a Beta distribution (calibrated based on the values of a and b , as well as some assumptions about the shape of the Beta distribution, though further details regarding these assumptions were not presented by the authors). In this interim period, the estimated proportion of patients alive is a blend of estimates from both the 'obs' and 'ext' groups, though the weight is a function of time (i.e., there is no fixed weight that applies in this time period).

For the avoidance of doubt – the blended survival method 'blends' estimates of *survival*, not *hazards*. This can be seen in Figure 3-27, which shows that after the blending interval, survival is equal to the estimate of survival for the external information source (i.e., $S_{ble} = S_{ext}$).

Figure 3-27: Graphical representation of the blended curve method



Note(s): This figure illustrates how the blended survival method works. This is taken from Figure 2 from the original study publication.

Source: Copyright © 2023 Che *et al.* This figure is re-produced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. Che Z,

While not explicitly stated, it is implied in the original conference presentation that the model is fitted in a Bayesian framework based on a different presentation from the same authors on the same topic at the same conference.²¹⁵ In the 2024 SLR update, the full paper was identified which provided more details on the method, confirming the method is fitted in a Bayesian framework.²¹⁴

Key assumption(s)

The key assumptions made in this method are essentially the same as those that apply for other piecewise approaches – i.e., that there is a generalisable external population which can be used to provide meaningful estimates of longer-term survival. Moreover, in this particular example, it is necessary to specify two timepoints (a and b) between which the external data are partially relevant, and later the only relevant source of information for long-term survival. It is also necessary to determine the parameters that are used to inform the value of p , though the worked example provided in the full paper proposes some default settings.²¹⁴

Strengths and limitations

The main strength of this approach over a more traditional piecewise model is that a sudden (and likely clinically implausible) change in hazards can be avoided by imposing the assumption that survival gradually approach those of the external group. However, this in turn introduces added limitations in the form of assumptions that need to be imposed to control the time points over which survival gradually changes from being informed by one source to the other, as well as the shape of the distribution controlling how the weight changes from 0 to 1.

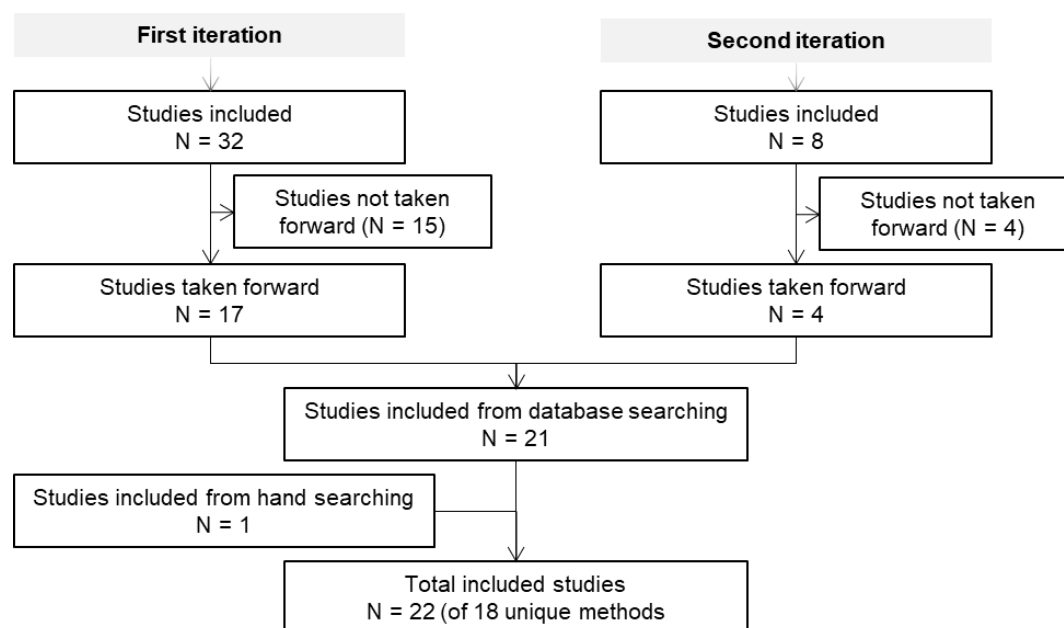
Rationale for inclusion in further investigation in the rest of the thesis

This approach is included as an alternative approach to a standard piecewise modelling approach, avoiding the potential issue of an abrupt change in projected hazards, but weighting on survival rather than hazards.

3.3.4 Summary and verification of identified methods

A summary of the final set of identified methods is presented in Figure 3-28, demonstrating the number of studies identified at each stage of the searching process.

Figure 3-28: Summary of final set of methods



Prof. Nicholas Latimer reviewed the final list of identified studies in January 2022 to ascertain whether any known studies of relevance have been missed as part of the literature review. Prof. Nicholas Latimer was not aware of any that appeared to be missing from the collection of included studies identified by the searches. However, a number of methods anticipated to be reflected within the NICE appraisals review (discussed further in Chapter 4) were highlighted by Prof. Nicholas Latimer as these are expected to have been used in HTA reports:

- Use of external data within a response-based modelling framework.
- Use of post-progression survival (or a related measure) from an external source combined with pre-progression survival from the main study.

In addition, the lead author of the previous review of methods, Dr. Chris Jackson, reviewed the final list of included methods, and confirmed that he was unaware of any other methods that were not identified.

3.3.5 2024 update

Following publication of the initial SLR findings in *Medical Decision Making* (Bullement *et al.*, 2023)³, two notable studies were published:

- Che *et al.* published their ‘blended survival’ method as a full paper, building on the conference abstract identified in the SLR.²¹⁵ I was made aware of this paper as one of my supervisors (Gianluca Baio) is a co-author.
- Jackson published an alternative MPES approach to that of Guyot *et al.*, and developed an R package called ‘*survextrap*’.²¹⁶ I was made aware of this following a

discussion with the author (lead author of the JR), and I am acknowledged in the paper accordingly.

For completeness, an SLR update was undertaken in late 2024 to formally identify these studies (both published in the MEDLINE-indexed journal *Medical Decision Making*), alongside any other studies of potential relevance that were published after the search date of the SLR (last ran on 1 July 2021 – see sub-section 3.3.2). No formal data extraction was undertaken as part of the 2024 update; instead, only relevant methods (i.e., those meeting the criteria to be considered further in thesis, as described earlier in this chapter) were described.

No edits were made to the search strategies used for the second iteration. This comprised of two Ovid searches (one for Medline and EconLit, and one for EMBASE). In the previous searches, WOS was also searched, but this was not considered in the 2024 update as all relevant studies were already indexed in either Medline, EconLit, or EMBASE. The Ovid search strategy is shown in Table 3-13, though as described previously, line 22 was changed from 'exp animals/ not humans.sh.' to '(exp animal/ or nonhuman/) not exp human/' for the EMBASE search. In addition, studies were limited to those published after the previous search cut-off by only including studies published 2021 or later.

Number of identified studies

The searches were executed on 1 September 2024, run to the latest date available as of the search date for each database. The date limits for each database are provided below:

- Ovid
 - Econlit <1886 to 22 August 2024>,
 - Embase <1974 to 2024 Week 35>,
 - Ovid MEDLINE® and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions® <1946 to 29 August 2024>

For the 2024 update, a total of 3,120 records from the Ovid search of Medline and EconLit were identified, plus 3,085 records from the Ovid search of EMBASE, giving a grand total of 6,205 records prior to de-duplication. After de-duplication (using the built-in functionality of EndNote), 2,637 duplicate records were removed, leaving 3,568 records remaining for primary screening. After primary screening based on titles and abstracts, 3,557 records were excluded, leaving a total of 11 records eligible for secondary screening. After full-text screening, 5 records were excluded, which resulted in a total of 6 records. No attempt was made to automatically remove any duplicate entries that were published in 2021 that may have been identified by the first or second iterations, but these were excluded manually at either primary or secondary screening if identified.

Methods taken forward for further evaluation in later chapters

A tabulated summary of the methods that are considered further is provided in Table 3-21. In total, two distinct methods were taken forward, reported in two studies.

Table 3-21: Methods identified in 2024 update taken forward for further evaluation in later chapters

Reference	Method(s)	Theme(s)	Rationale for inclusion
Che <i>et al.</i> , (2023) ²¹⁴	Blended survival model	• Piecewise • Informative prior	Full paper version of previously identified conference materials.
Jackson (2023) ²¹⁶	M-spline MPES	• Piecewise • Informative prior • General population • Other	Alternative specification of MPES which aims to address some limitations of the MPES approach presented by Guyot <i>et al.</i> , (2017). ¹⁵³

Key: MPES, multi-parameter evidence synthesis.

Methods not taken forward

A tabulated summary of the methods that are not considered further is provided in Table 3-22.

Table 3-22: Methods identified in 2024 update not taken forward but included in the review

Reference	Method(s)	Rationale for exclusion
Gallacher (2024) ²¹⁷	Point-estimate common model	This paper presents a user-friendly tool to fit a common parametric model between user-specified time points. However, the approach does not involve fitting a model to patient-level data from a clinical trial, so it is excluded.
Langford <i>et al.</i> , (2022) ²¹⁸	Informative prior with proxy disease	This approach builds on the informative prior method of Soikkeli <i>et al.</i> , (2019) ¹⁹⁹ by using data from a different disease area as a proxy, with some re-weighting. However, the underlying integration of the external evidence is unchanged, and so as this method is already captured it is excluded.
Sanchez Alvarez <i>et al.</i> , (2023) ²¹⁹	Informative prior with meta-analysis	As for the row above, this approach also builds on the informative prior method of Soikkeli <i>et al.</i> , (2019) ¹⁹⁹ , but uses meta-analysed data to derive an informative prior, rather than one data source. It also considers setting the prior on the scale or rate parameter, rather than the shape parameter of a given distribution. However, the underlying integration of the external evidence adopts the same overarching principle of the approach of Soikkeli <i>et al.</i> , so this method is also excluded.
Traore <i>et al.</i> , (2023) ²²⁰	Bayesian survival with PFS surrogacy	This method uses data for PFS to support extrapolations of OS, and therefore it is excluded on the grounds that it requires non-OS data.

Key: PFS, progression-free survival; SLR, systematic literature review.

Presentation of individual methods

Each of the identified method studies taken forward for the remainder of the thesis are discussed in this sub-section. Method studies that were discarded from further consideration are presented in Appendix C (with reasons provided for why specific methods were discarded).

Che et al., (2021 and 2023)

Theme(s): *Piecewise, informative prior.*

Che *et al.* present a more detailed description of the blended survival method in 2023, which was originally identified as a conference abstract published in 2021 (see Section 3.3.3 for a description of the method itself), including code to execute the method.²¹⁴ In the remainder of the thesis, the blended survival method is attributed to the full paper citation, for completeness.

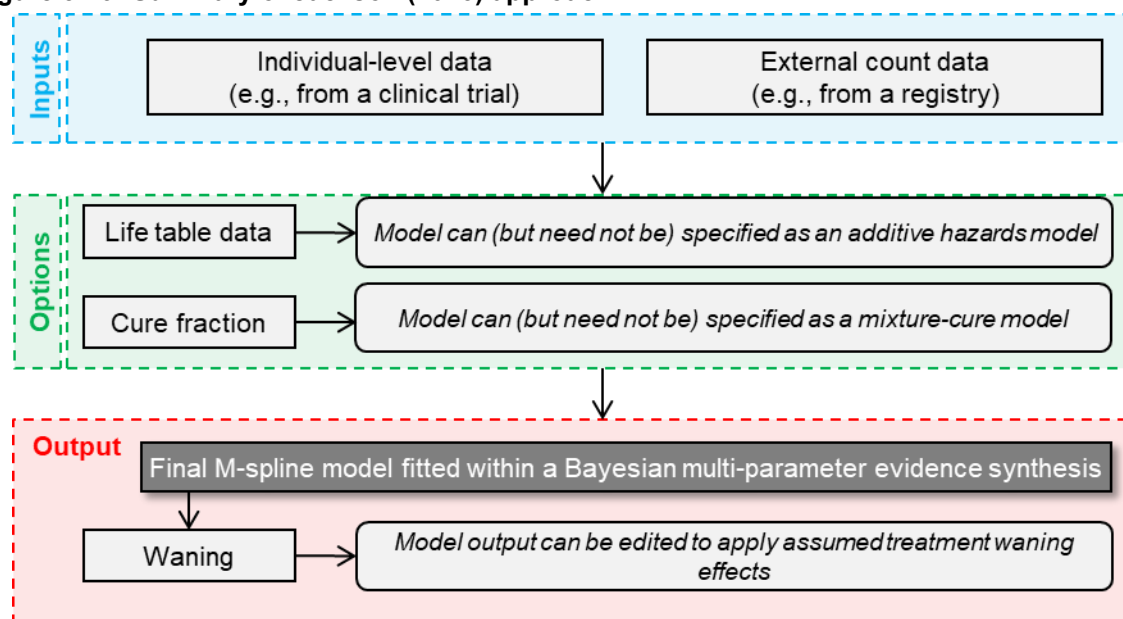
Jackson (2023)

Theme(s): *Piecewise, informative prior, general population, other.*

Jackson (2023)²¹⁶ presents an alternative MPES method to the approach of Guyot *et al.*, (2017)¹⁵³ to estimate the survival function. A detailed description of Jackson's MPES is provided in the original study publication, and an introduction to the method (alongside a comparison to Guyot's MPES) is provided in a tutorial paper (written as part of the PhD work reported in this thesis, and discussed further in Chapter 5).⁴ However, in brief, Jackson's MPES uses an M-spline, which is a different type of spline-based model compared with other modelling approaches that have been used in HTA previously (most notably, RCS models which have been used in previous STAs conducted by NICE).⁵² Jackson highlights that M-splines "*have some computational advantages over other kinds of splines and flexible models*" (Jackson, 2023, p.3).²¹⁶ In addition, Jackson's MPES is designed such that it can consider multiple sources of external count data.

Jackson describes three "*special mechanisms*" that can be applied to the MPES model to ideally improve the plausibility of the survival extrapolations. These include accounting for general population mortality within a relative survival framework, accounting for a potential 'cured' proportion of patients via a mixture-cure modelling approach, and accounting for potential treatment effect waning via a *post-hoc* adjustment. A summary of the approach taken by Jackson (2023)²¹⁶ is provided in Figure 3-29.

Figure 3-29: Summary of Jackson (2023) approach



Key assumption(s)

As with Guyot's MPES, an important, implicit assumption made by Jackson's MPES is that the external data represent the conditional survival experience of the trial population (i.e., the population for whom patient-level data are available).

Strengths and limitations

The main strength of this approach is that, like Guyot's MPES, it can combine multiple different sources of information into one analysis. Unlike Guyot's MPES, it is executable via an R package 'survextrap', with additional features allowing for consideration of a cure fraction and background mortality via an excess hazards model. However, as with Guyot's MPES, the process of estimation is complex, and therefore the resultant estimates of survival are less transparent versus a simpler (though perhaps less technically robust) method.

Rationale for inclusion in further investigation in the rest of the thesis

The author of this study presents a method that addresses some of the limitations of Guyot's MPES, and therefore is included as an alternative form of MPES for survival analysis.

Summary of 2024 update

In the 2024 update, two distinct methods were identified and taken forward, reported in two studies (one of which was previously identified as a conference presentation).^{214, 221} Both of these studies were known outside the context of the SLR update (via personal communication with Dr. Chris Jackson, and via identification of the Che *et al.* conference proceedings in an earlier iteration, of which one of my supervisors Prof. Gianluca Baio is also a co-author).

However, the identification of these studies through the execution of the second iteration search strategies confirmed that these search strategies are likely to be suitable to identify any method studies that could be published in the future.

3.4 Discussion

3.4.1 Summary

This SLR was undertaken to identify methods of estimating survival that incorporate an element of external evidence within the estimation process. This review therefore builds on the previous work of Jackson *et al.*, (2017)²⁶ by: (i) expanding the scope of methods to include those that make use of non-data sources of external evidence (e.g., clinical expert opinion), (ii) adopting SLR methods to increase transparency of how studies were identified and how searches could be reproduced, and (iii) updating the review to include more recently published studies. To the best of my knowledge, this SLR represents the first SLR of methods literature concerning the use of external evidence sources to inform survival extrapolation estimates.

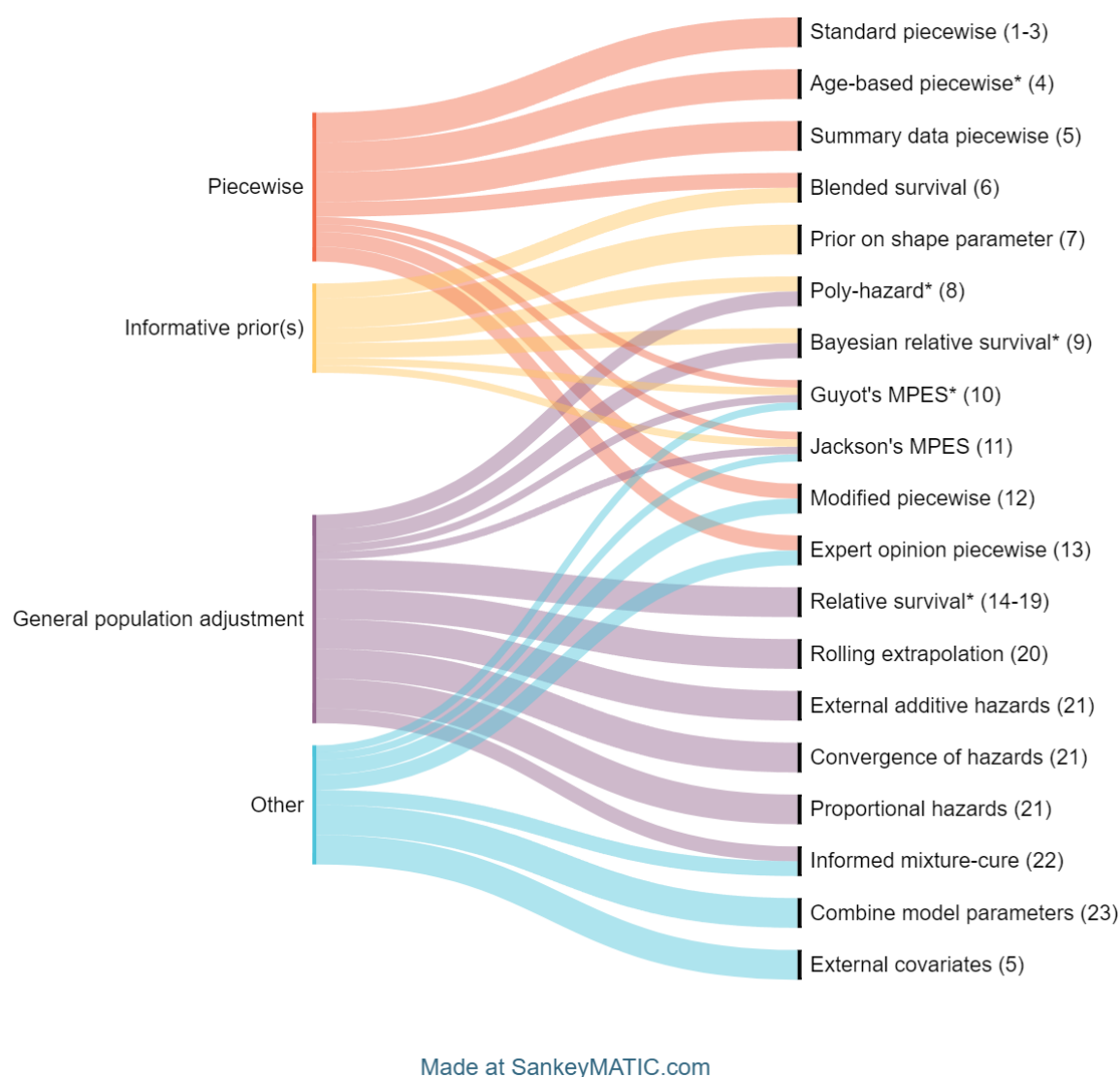
3.4.2 Searching

The CPG searching approach was used with the expectation that additional search terms may later be revealed after the first iteration, which was the case in this search. However, after the second iteration, no additional search terms were identified that were expected to lead to the identification of further relevant studies. The CPG method also facilitated transparent use of studies identified by the JR, by considering these studies as pearls for which the initial search strategy could be designed to retrospectively identify. Forwards and backwards citation searching did not yield any further studies, as well as provision of the final set of studies to both Prof. Nicholas Latimer and Dr. Chris Jackson (experts in the field of survival extrapolation). This increases confidence that the formal database searches are unlikely to have missed any important studies. Despite this, it should be noted that the original JR did not consider methods that used external *information*, hence the addition of specific search terms in the developed search strategy focusing on information sources (such as expert opinion).

3.4.3 Overall findings

A finding from the SLR of immediate interest was the number of different approaches to make use of external evidence within survival extrapolation that were identified. In total, across each of the searching iterations and additional searching methods, 19 unique methods to incorporate external evidence into survival extrapolations were identified. For completeness, a Sankey diagram showing the full set of the identified methods is presented in Figure 3-30. This diagram aims to illustrate how some methods may fall into several different categories or overall strategies to incorporate external evidence.

Figure 3-30: Sankey diagram of methods identified



Key: MPES, multi-parameter evidence synthesis.

Note: This is an updated version of a diagram originally presented by Bullement et al., (2023). © The Author(s) 2023. Diagram created using SankeyMATIC, available at: <https://sankeymatic.com/build/>. Methods marked with an asterisk (*) denote studies referenced in the original review by Jackson *et al.*, (2017).²⁶

Sources: (1) Larkin et al., 2015¹⁴⁶; (2) Barlev et al., 2017¹⁴⁷; (3) Fisher et al., 2017¹⁴⁸; (4) Nelson et al., 2008¹⁶⁹; (5) Pennington et al., 2018¹⁷³; (6) Che et al., 2023²¹⁴; (7) Soikkeli et al., 2019¹⁹⁹; (8) Benaglia et al., 2015¹⁵²; (9) Demirir & Sharples, 2006¹⁵¹; (10) Guyot et al., 2017¹⁵³; (11) Jackson, 2023²¹⁶; (12) Vickers, 2019¹⁷⁴; (13) Cope et al., 2019⁹¹; (14) Lambert et al., 2005¹⁵⁶; (15) Lambert et al., 2007¹⁵⁸; (16) Lambert et al., 2010¹⁶⁰; (17) Nelson et al., 2007¹⁸¹; (18) Andersson et al., 2013¹⁶³; (19) Jakobsen et al., 2019¹⁶⁵; (20) Hwang et al., 2017¹⁷²; (21) van Oostrum et al., 2021²⁰²; (22) Felizzi et al., 2021²⁰¹; (23) Hawe et al., 2014¹⁷¹.

The authors of the original JR categorised methods by the relationship of hazards between the population of primary interest (the 'index' population), and the external evidence source (the 'reference' population). However, this approach to grouping methods is not fit for purpose when considering methods that may allow for a broad range of different relationships to hold over a lifetime horizon, methods that combine approaches, and/or methods that do not make

use of external data (but instead make use of other information sources, such as clinical expert opinion).

Despite challenges in grouping methods into categories for ease of understanding, it was possible to identify clear themes that distinguish aspects of the methods from one another:

Piecewise methods involved external evidence which was mostly used only after a given point in time. This type of approach (generally) restricts use of external evidence to the portion of a survival curve that may be associated with the most uncertainty (i.e., the longer-term extrapolated portion). Piecewise methods are arguably the simplest of the approaches identified, owing to their intuitive application and because they can be implemented without specialist software. Nevertheless, their simplicity may also be viewed as their most important limitation as each piecewise approach makes several strong (and often clinically implausible) assumptions about the pattern of survival in the relevant population (e.g., a sudden increase or decrease in the hazard of death where two different curves meet).

A range of different **Bayesian methods** demonstrated how external evidence could be factored into model estimation with varying degrees of influence, by controlling how informative the external evidence is on prior distributions. A number of different Bayesian methods were identified, the most comprehensive of which comprised MPES approaches (Guyot *et al.*, 2014¹⁴¹ and Jackson 2023²¹⁶) making use of several different external evidence sources. In spite of their advantages, these methods require the use of specialist Bayesian analysis software (such as WinBUGS²²² and JAGS¹¹²), and therefore involve additional complexity in terms of producing estimates of survival, which may explain in part their seemingly limited uptake in the literature (given that several recent method studies adopted a non-Bayesian approach), even if the overall approach is logical.

After identifying and categorising the identified methods, **general population adjustment methods** seemed most logical to group as a category by itself. These methods were separated as either the main or the only source of external evidence comes from general population life tables. Relative survival methods have historically been used in larger studies of cancer populations to describe the relative experience of a group with a disease versus an external population but have intuitive appeal for generating plausible long-term survival estimates for older populations in particular. These approaches refer to the case where the source of external evidence comes from estimated survival probabilities from another population, which can be assumed similar, though often this appears to involve the use of general population life tables (though this is not an explicit requirement of this type of method). For example, a disease registry could be used as the 'baseline' group from which to derive relative estimates. Within the general population adjustment methods identified, life table

estimates are assumed to be certain (i.e., that there is no uncertainty around estimates taken from life tables), which separates this source of external evidence from other sources (such as observational study data, or clinical expert opinion) which will be subject to uncertainty that should ideally be reflected within the survival estimation approach used. Relative survival methods have also been combined with other ‘novel’ survival extrapolation methods. For example, Kearns *et al.*, (2022) proposed the use of dynamic relative survival models for use in HTA, though these methods were not captured as part of this SLR since the ‘novel’ aspects are not related to the incorporation of the external evidence.²²³

Finally, a number of **other methods** were also identified which included several approaches that were not possible to include within the aforementioned categories. Some of these methods could be considered alternative approaches to some of the Bayesian methods identified, but in a non-Bayesian framework; or could also be deemed minor modifications to established methods outside a ‘use of external evidence’ context.

3.4.4 Use of methods in remainder of the thesis

Each of the methods identified in this chapter is categorised as one of the following:

- **Methods retained:** Any method which is not discarded is taken forward throughout the remainder of the thesis, alongside a specific name for the method which is then used throughout the remainder of the thesis for ease of reference.
- **Methods discarded:** Any method which is not considered further for the remainder of the thesis. Each discarded method is described alongside rationale for its omission from further consideration.

The methods retained throughout the remainder of the thesis are presented in Table 3-23. The discarded methods are summarised in Table 3-24.

Table 3-23: Summary of retained methods for method testing

Name of method	Citation	Identification	Page
Bayesian relative survival	Demiris & Sharples, (2006) ¹⁵¹	SLR (1 st iteration)	83
Relative survival	Lambert <i>et al.</i> , (2007) ¹⁵⁸	SLR (1 st iteration)	90
Combine model parameters	Hawe <i>et al.</i> , (2014) ¹⁷¹	SLR (1 st iteration)	94
Standard piecewise	Larkin <i>et al.</i> , (2015) ¹⁴⁶	SLR (1 st iteration)	103
Original MPES	Guyot <i>et al.</i> , (2017) ¹⁵³	SLR (1 st iteration)	109
Summary data piecewise	Pennington <i>et al.</i> , (2018) ¹⁷³	SLR (1 st iteration)	117
Expert opinion	Cope <i>et al.</i> , (2019) ⁹¹	SLR (1 st iteration)	120
Informative prior on shape	Soikkeli <i>et al.</i> , (2019) ¹⁹⁹	SLR (2 nd iteration)	135
Blended survival	Che <i>et al.</i> , (2023) ²¹⁴	SLR (2024 update)	151
M-spline MPES	Jackson (2023) ²¹⁶	SLR (2024 update)	157

Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

Table 3-24: Summary of discarded methods

#	Name of method	Citation	Identification
1	Age-based piecewise	Nelson <i>et al.</i> , (2008) ¹⁶⁹	SLR (1 st iteration)
2	Modified piecewise	Vickers (2019) ¹⁷⁴	SLR (1 st iteration)
3	Informed mixture-cure	Felizzi <i>et al.</i> , (2021) ²⁰¹	SLR (1 st iteration)
4	External covariates	Pennington <i>et al.</i> , (2018) ¹⁷³	SLR (1 st iteration)
5	Alternative general population adjustment	van Oostrum <i>et al.</i> , (2021) ²⁰²	SLR (1 st iteration)
6	Rolling extrapolation	Hwang <i>et al.</i> , (2017) ¹⁷²	SLR (1 st iteration)
7	Poly-hazard with cause-specific data	Benaglia <i>et al.</i> , (2015) ¹⁵²	SLR (1 st iteration)

Key: SLR, systematic literature review.

The following sub-sections describe the reason(s) why some of the methods were ultimately discarded from the remainder of the thesis (i.e., further methods testing [Chapter 5] and inclusion in the NCS [Chapter 6]). It should be noted that these methods were previously described as being included in further investigation in the rest of the thesis; however, upon further investigation into how the method works, these methods were not deemed suitable to consider in later chapters. Despite this, it was deemed appropriate to include these methods in the overall synthesis of findings from the SLR, which is why these methods feature in the summary of methods (including, for example, the Sankey diagram shown in Figure 3-30).

Age-based piecewise

Presented by Nelson *et al.*, (2008).¹⁶⁹ This method involves creating an age-based prediction model to extrapolate survival beyond the end of follow-up. There are specific cases where this approach may provide credible extrapolations (e.g., for a population with cardiovascular disease where an age-based model for survival may be possible to derive). However, this method is unlikely to be suitable within the context of HTA, given small sample sizes, limited data, and a lack of age-based mortality models for specific cancer populations. It is therefore not considered further in this thesis.

Modified piecewise

Presented by Vickers (2019).¹⁷⁴ This method builds on a piecewise model, making use of an HR to account for a treatment effect, but apart from this modification is effectively a standard piecewise model. As the only external evidence here refers to how one treatment group may compare to another, and that the focus of this thesis is not on the relative effectiveness of alternative treatments, this method was not considered further in the thesis.

Informed mixture-cure

Presented by Felizzi *et al.*, (2021).²⁰¹ This method addresses a specific context where there is plausibly a statistically cured proportion of patients, and that this proportion could be informed through external evidence (either clinical opinion, or external data). While this may be relevant to consider under some circumstances, this context does not reflect the majority

of HTAs of cancer medicines and relies upon two things being true to be a plausible method (i.e., existence of a cured proportion, and reliable estimation of said cure proportion). Where this method is relevant, it is likely the case that many other methods would be inappropriate to consider (i.e., any method that does not consider a cured proportion), and *vice versa* (i.e., where there is no cure proportion, this method will not be appropriate to use). As such, this method was not considered further as part of the thesis, given that ultimately aims to address a specific, uncommon research question that many of the other methods are not capable of addressing.

External covariates

Presented by Pennington *et al.*, (2018).¹⁷³ This method may be considered a blend of the summary data piecewise method from the same authors, which is retained in the thesis; and the combine model parameters approach of Hawe *et al.*, (2014), which is also retained in the thesis. This method was not considered further in the thesis as while the combination of these approaches may be considered novel, it may be theoretically possible to combine many of the identified methods, yet this would complicate the process of exploring the performance of individual methods. Therefore, to preserve the unique features of each method in the thesis for comparison purposes, this method was not explored further.

Alternative general population adjustment

Presented by van Oostrum *et al.*, (2021).²⁰² As a part of their exploration of general population adjustment models, van Oostrum *et al.* demonstrated that the internal additive hazards model outperformed three alternative approaches: (i) an external additive hazards model; (ii) a model which assumed convergence of hazards from the trial population with those from the general population, and (iii) a proportional hazards-based approach. These methods are therefore not considered further since they are superseded by the internal additive hazards model.

Rolling extrapolation

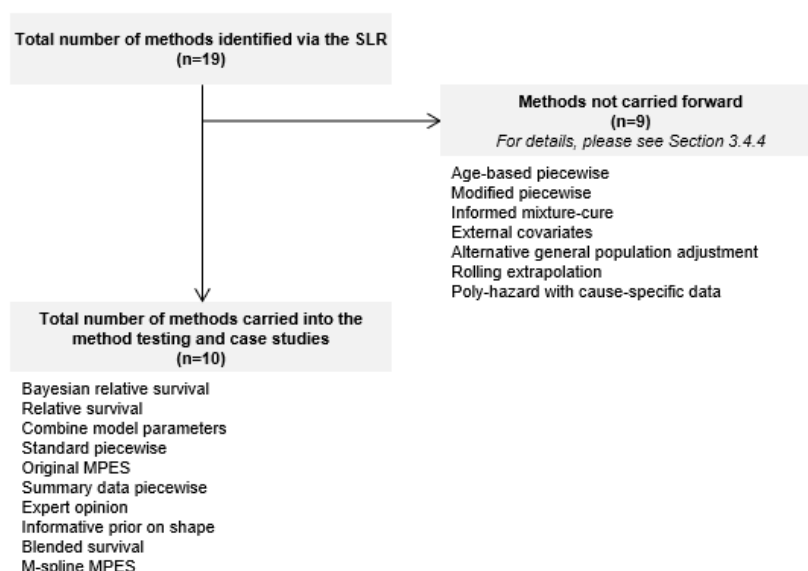
Presented by Hwang *et al.*, (2017).¹⁷² This method requires a custom install of an *R* ‘add-on’ called ‘iSQoL2’, with a link provided in the paper to a website which is no longer hosted (though is possible to find online via a website such as web.archive.org). In addition to issues that may arise with users needing to install custom add-ons to institutional computers, this also means that the method cannot be explored thoroughly as it cannot easily be used in conjunction with other standard *R* package installations. The add-on also makes use of pop-up boxes to link file paths, which is impractical for simulation exercises. The method is therefore not considered further in this thesis for these practical reasons.

Poly-hazard with cause-specific data

Presented by Benaglia *et al.*, (2015).¹⁵² This approach requires the user to select a parameterisation for general population mortality which adopts the same functional form as per disease-related death (e.g., Weibull). In the context of this thesis, this is highly restrictive since it is expected that the same functional form may not be suitable for extrapolation for HTA. There is also no standardised code to execute this specific approach, and while other methods to consider a poly-hazard model exist (such as using the *survHE R* package¹¹⁵) it is not possible to determine if the methods are the same or similar. This method is not considered further in the thesis; however, the relative survival approach is maintained as a method to consider cause-specific survival which does not require such an assumption to be made.

A summary of method identification and determination of which methods are carried forward in the remainder of the thesis is presented in Figure 3-31.

Figure 3-31: Methods summary following completion of SLR



Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

3.4.5 Conclusion

In this chapter, the findings of an SLR on methods for incorporating external evidence into survival extrapolation has been presented. The review identified and synthesised 19 unique methods, classified into four key themes, highlighting the range of different ways external evidence can be used to inform survival estimates. A structured search strategy, including an update in 2024, ensured comprehensive identification of relevant studies.

The SLR revealed that while numerous methods exist, based on my personal knowledge of HTAs of cancer medicines, relatively few appear to have been widely adopted in contemporary HTA, and newly proposed approaches often lack direct comparisons to existing methods. This

presents challenges in selecting and justifying methods for HTA applications, a topic explored further in Chapter 4. By systematically identifying these methods, the SLR provides reassurance that key approaches have been captured transparently, establishing a clear basis for comparison in the remainder of the thesis. In conclusion, by undertaking the SLR described here within, a comprehensive set of methods used to incorporate external evidence into survival estimates has been obtained.

4. Single Technology Appraisal (STA) review

Chapter synopsis

In this chapter, a second review is presented comprising a search of Single Technology Appraisals (STAs) conducted by the National Institute for Health and Care Excellence (NICE) of cancer drugs over a one-year period. This review of health technology assessments (HTAs) aimed to identify approaches that had been used in past appraisals, how these were perceived by reviewers, and ultimately how the methods affected decision-making. While all approaches were considered within this review, those in which external evidence was combined with trial data to make long-term projections of overall survival (OS) were of primary interest.

In total, 36 completed STAs of cancer drugs published by NICE in 2021 were included in a narrative synthesis. A variety of approaches were identified which demonstrated how companies have previously included external evidence within survival extrapolations. Examples include the estimation of survival after a particular event of interest (e.g., after disease progression), after a particular time point (e.g., after 5 years), and to represent a surrogacy relationship between outcomes (e.g., progression-free survival and OS). The uses of external evidence most often represented situations where data from the pivotal trial were 'missing' (either entirely [e.g., no data available from the trial] or in part [e.g., no data available after relapse]), rather than where the external evidence could supplement or strengthen estimates primarily driven by the pivotal trial data. Relatively few appraisals included use of external evidence directly within the process of estimating a survival model (beyond incorporation of background mortality data, or to allow a comparison to a treatment not considered within the pivotal clinical trial for the intervention). Instead, most submissions included informal reference to external evidence as a means of validation or comparison.

The External Assessment Groups and the appraisal committees appeared to be generally receptive to methods that aimed to leverage external evidence, as no proposed method was rejected outright. However, several criticisms were frequently raised, including conflicting clinical expert opinion, comparability of sources of evidence, and issues in the use of complex methodology to address uncertainty. Detailed technical descriptions of methodology used, both in general and when external evidence was leveraged, was found to be limited (even when the use of external evidence appeared seemingly straightforward, such as adjusting for background mortality only). Therefore, an important limitation of this review was the need to rely on information that was reported in the publicly available appraisal documentation.

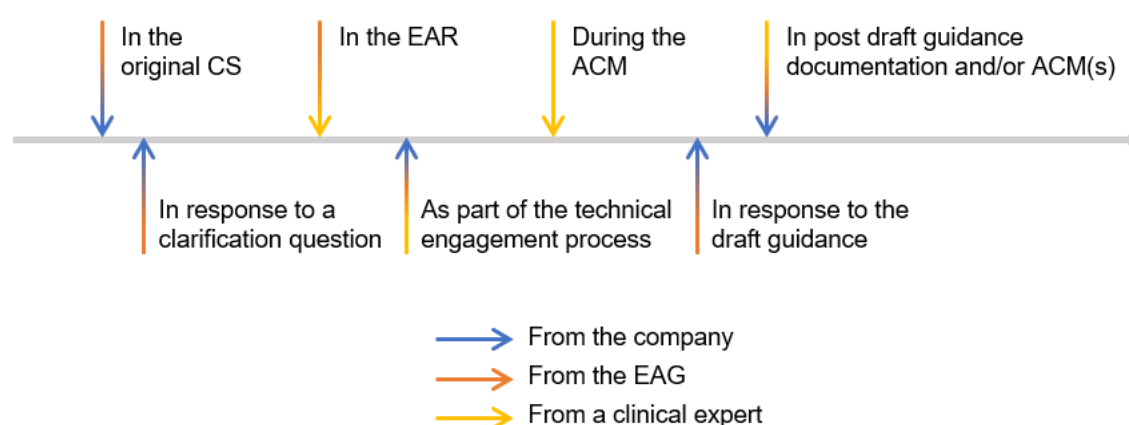
Findings from this review were presented at the ISPOR Europe 2022 meeting, and at the SchARR PGR conference 2023.^{6, 224} A later update to the review undertaken in collaboration with Sanofi was published in the Journal of Medical Economics (Petitjean et al., 2025).⁵

4.1 Introduction

To complement the SLR undertaken concerning methodological literature, a separate review was also undertaken to focus on published NICE appraisal guidance of cancer drugs.* This review aimed to ascertain which methods have been used in contemporary HTAs, so that it may be possible to understand how often and in what context(s) specific methods have been used, if there are any in particular that are commonly used, as well as the criticisms raised by the independent EAGs† and/or NICE appraisal committees. In addition, it is possible that some methods may have been used in NICE Technology Appraisals (TAs) but were not identified via the SLR (see Chapter 3), such as unpublished methods or methods that have only been considered within the context of an economic evaluation.

The availability of NICE guidance in the public domain facilitates an assessment of how external evidence has been used in ‘real’ decision making; including if, how, and when published methods have been practically used as part of an HTA process. The NICE TAs also shed light on how different pieces of evidence may be introduced into an appraisal over the course of an assessment, based on a standardised process (set out by NICE via its manual for health technology evaluations).³¹ For example, evidence could be included in the original company submission (CS), introduced by the EAG in its External Assessment Report (EAR), or perhaps provided by the company later in the appraisal process, such as during the technical engagement period, or following an initial negative recommendation issued via a draft guidance document (previously known as an Appraisal Consultation Determination). Figure 4-1 illustrates various different timepoints where external evidence could be introduced within the STA process.

Figure 4-1: Introducing external evidence into NICE STA process



Key: ACM, appraisal committee meeting; CS, company submission; EAG, External Assessment Group; EAR, External Assessment Report; NICE, National Institute for Health and Care Excellence; STA, single technology appraisal.

* See Chapter 1 for reasons why cancer drugs were the focus of both this review and the thesis.

† Historically, EAGs were called the Evidence Review Groups (ERGs), but this changed to EAGs following an update to NICE methods for conducting appraisals in 2022. These acronyms (EAGs and ERGs) are considered interchangeable throughout the thesis. As part of the STA process, an EAG produces an EAR, which was previously called an ERG report – again, both EAR and ERG report are considered interchangeable throughout the thesis.

Previous studies have considered which methods have been typically undertaken to inform survival extrapolations in NICE STAs, most notably TSD 14 which was developed to provide *“recommendations for how survival analysis can be undertaken more systematically”* (Latimer, 2011, p.5), compared with how survival analysis was undertaken for submissions to NICE prior to its publication.¹⁰ More recently, TSD 21 was published which describes flexible methods for survival analysis, including some reference to external evidence (e.g., through excess mortality or relative survival methods).⁵² However, unlike TSD 14, TSD 21 does not aim to set out a systematic approach to determine the choice of survival extrapolation technique – more specifically, the authors state: *“... the approaches detailed herein should not be considered as an extended list of survival methods to “try out” on data. Instead, care should be taken to think through the underlying mechanisms likely to be dictating short and long-term hazard/survival functions.”* (Rutherford *et al.*, 2020, p.17).⁵²

Of note, outside of formal NICE guidance, a number of studies have been published on approaches taken to estimating OS for use in NICE appraisals (and HTA submissions internationally).^{27, 53, 225-227} These studies have considered a range of objectives, including explaining how survival extrapolations are typically undertaken to inform HTA, understanding which extrapolation methods have been historically used (particularly in the context of specific classes of therapy, such as cancer immunotherapy), and evaluating the performance of specific extrapolation methods against latterly available data from the same clinical trial(s). However, none of these published studies focus on external evidence specifically, except for the previously discussed JR (Jackson *et al.*, [2017], see Chapter 3).²⁶

At the ISPOR Europe 2021 virtual conference, a poster by Reitsma and colleagues was presented on the topic of how external *data* (though not external *evidence* more broadly) had been used in previous NICE appraisals in the context of survival extrapolation.²²⁸ This presentation considered a review of NICE TAs published from December 2001 (no end date documented), adopting a relatively strict definition of which TAs were deemed to have used *“external data sources to quantitatively inform survival extrapolation”* (Reitsma *et al.*, 2021).²²⁸ The authors provided a tabulated overview of how external data had been used in published TAs from 2016 onwards. The authors concluded that *“Only rarely is external data used to inform survival extrapolation beyond clinical trial observations”* (Reitsma *et al.*, 2021).²²⁸

In contrast to Reitsma *et al.*, (2021),²²⁸ the NICE appraisals review described henceforth considers a more detailed assessment of external evidence use related to survival extrapolation, focusing on a smaller number of cancer drug TAs (published in 2021). The aim of this review was to provide a summary of how external evidence has been used to inform

survival estimation within recent cancer submissions made to NICE via the STA process. In particular, this review aims to identify where lesser-known methods may have been used, and how well these were received by the NICE appraisal committee.

4.2 Methods

4.2.1 Searching methods

Inclusion and exclusion criteria

A summary of the inclusion and exclusion criteria is provided in Table 4-1.

Table 4-1: Inclusion and exclusion criteria

PICOS	Inclusion	Exclusion
Population	People with a confirmed diagnosis of any cancer	People without a confirmed diagnosis of a cancer (e.g., people identified as having a high-risk of developing a particular cancer)
Intervention	Any pharmaceutical intervention aimed at the treatment of people with cancer	Any intervention for other health conditions, including treatment of cancer-related complications (e.g., chemotherapy-induced toxicity) or any non-pharmaceutical intervention (e.g., lifestyle modification)
Comparison	Any comparator	-
Outcome(s)	Any NICE STA which describes an approach to estimate lifetime OS	An economic evaluation that does not consider a lifetime horizon over which OS is extrapolated (e.g., a time-agnostic decision tree)
Study type	Completed and non-terminated NICE STA guidance published in 2021 including an economic evaluation	Any non-STA guidance (e.g., an HST appraisal) and any STA guidance published before or after 2021

Key: HST, Highly Specialised Technology; NICE, National Institute for Health and Care Excellence; OS, overall survival; STA, single technology appraisal.

The following documentation associated with each included TA was considered within the scope of this review:

- The original CS (excluding appendices, as these are not routinely published by NICE).
- The company's responses to the EAG's clarification questions.
- The EAG's report (including any subsequent reports and/or addenda).
- Materials prepared throughout the technical engagement period (if this took place).
- For appraisals with an interim decision, the draft guidance (and where relevant, multiple draft guidance documents).
- The final decision, presented in the final guidance issued by NICE.

Search approach

The NICE website (<https://www.nice.org.uk/guidance/published>) was searched on 11 February 2022 to identify non-terminated STAs of cancer drugs published in 2021. Guidance

from 2021 was selected to provide a recent overview of external evidence use in HTA decision making, for completed STAs.

4.2.2 Screening and eligibility

Committee papers were searched for relevant information covering the survival modelling approach taken by the company for the intervention and its comparator(s); EAG critique; and related commentary from the appraisal committee. Consideration was also given to the timing of when external evidence was introduced into the decision-making process (for example, provided in the original CS versus introduced at the technical engagement stage). Related information, such as the choice of model structure, was also extracted for context.

4.2.3 Data extraction

A data extraction table was developed to extract relevant information from each appraisal, in order to ascertain the extent to which external evidence was used to inform the estimation of survival as a part of the company's cost-effectiveness analysis. The data extraction table contains the following key elements:

- **Appraisal information:** Appraisal ID number, population, intervention, comparator(s), date published, recommendation type, submitting company, EAG, committee, whether or not the appraisal was a CDF exit appraisal, and the type of recommendation issued.
- **Evidence:** Main evidence source(s) for survival estimation in the CS, specific features of interest (e.g., crossover design); type(s) of external evidence used, background mortality adjustment, as well as any informal or qualitative use of external evidence.
- **Methods:** Economic model structure and how survival is estimated in it (directly or indirectly via intermediate health states or events), approach(es) to estimating survival (including sensitivity analyses), justification for the base-case approach taken.
- **Review:** EAG's perspective on the company's approach to estimating survival, EAG preferred approach (if this differs from the approach taken by the company), and if the EAG highlighted the estimation of survival as a 'key issue' for the appraisal committee.
- **Technical engagement:** Changes made by the company to its approach for estimating survival, including if additional external evidence was provided, and if the EAG then changed its preferred approach.
- **Appraisal:** Discussion contained within NICE guidance (both interim and final) with respect to survival estimation and the role of external evidence as part of this.

4.2.4 Limitations

The two main limitations of this review are (i) that the documentation being reviewed may be redacted in part, and (ii) that a relatively narrow set of inclusion criteria were specified to identify relevant appraisals. While some information may be redacted, this in itself may reveal some important implications for HTA best practice (e.g., if details of methods are redacted, this could explain, perhaps in part, their limited uptake in practice). The narrow inclusion criteria (covering only those STAs published in 2021) may mean that older appraisals that made use of relevant methods are missed from this review. However, as highlighted previously, it is expected that most of these will have been previously identified and discussed by Reitsma *et al.*, (2021).²²⁸ A discussion of the findings from Reitsma *et al.* is provided in the Discussion sub-section of this chapter.

4.2.5 Synthesis

In keeping with the aim of this review, a narrative synthesis is presented focusing on emergent themes with respect to how often external evidence has been used in submissions made to NICE as part of the STA process, including which data and/or evidence sources are most frequently used and which approaches have been leveraged. Where relevant, a number of case studies in which external evidence was presented are described.

While the main focus of the review was placed on quantitative use of external evidence (as this can be objectively assessed as whether or not it informed the appraisal committee's decision), use of external evidence on a qualitative basis is also described (for example, where an external evidence source is acknowledged when choosing between different survival extrapolations). Qualitative use of external evidence is described primarily with respect to the presentation of case studies.

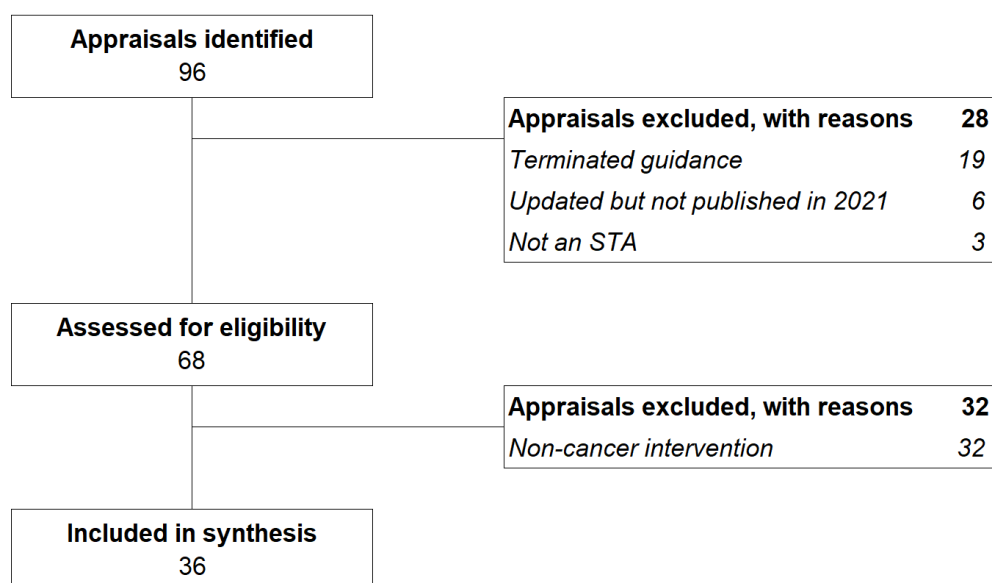
4.3 Results

4.3.1 Summary of findings

From a total of 96 appraisals published in 2021, 36 met the inclusion criteria for the review.[‡] Figure 4-2 illustrates the identification of appraisals, and reasons for exclusion.

[‡] Full documentation for each identified STA can be accessed via the NICE website by replacing the 'XXX' in the following hyperlink with the associated appraisal identifier: <https://www.nice.org.uk/guidance/taXXX>. The NICE STAs included in the synthesis were: 668, 669, 670, 673, 677, 680, 683, 684, 687, 689, 691, 692, 693, 695, 704, 705, 707, 709, 712, 713, 716, 720, 721, 722, 724, 725, 728, 736, 737, 739, 740, 741, 742, 746, 754, 756.

Figure 4-2: Identification of appraisals included in synthesis



Key: STA, single technology appraisal.

Source: This figure was originally presented with different formatting as part of a conference presentation.⁶

The number of STAs included in the synthesis by cancer type is summarised in Table 4-2. The most common cancer types reflected appraisals of treatments for lung, gastro-oesophageal, and prostate cancers.

Table 4-2: Included appraisals by cancer type

Cancer type	Number of STAs
Lung	5
Gastro-oesophageal	4
Prostate	
Blood/bone	3
Breast	
Colorectal	
Lymphoma	
Bladder	2
Leukaemia	
Ovarian	
Skin	
Bile duct	1
Head and neck	
Thyroid	

Key: STA, single technology appraisal.

Note: Cancer types determined based on simplified grouping.

The remainder of this results section presents a narrative summary of the included STAs, including reference to specific case studies. For full details regarding specific appraisals, please refer to the data extraction provided in Appendix E.

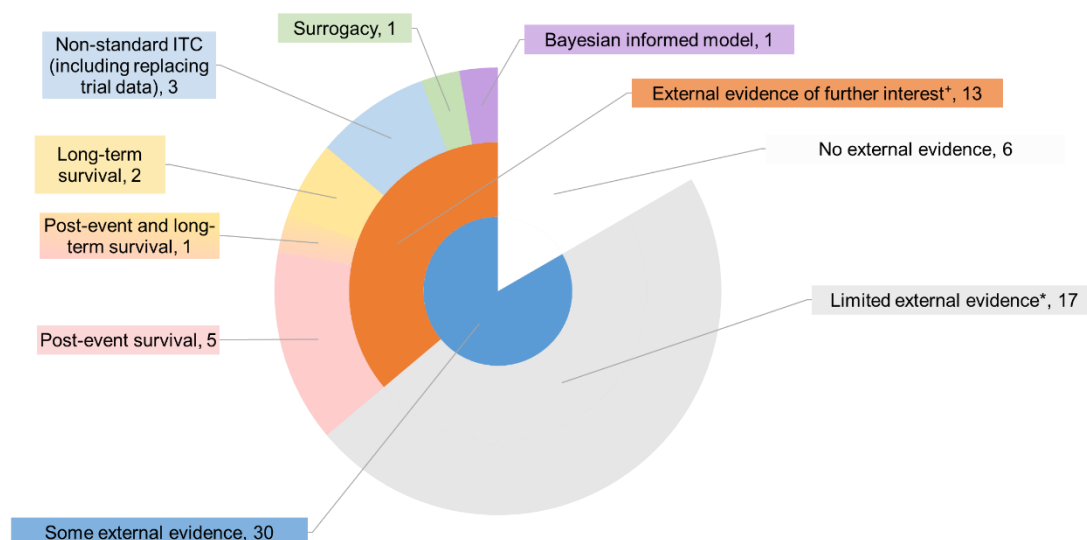
A breakdown of the included appraisals by how external evidence was used is presented in Figure 4-3. The three concentric circles illustrate the following:

- **Inner circle:** How many appraisals reported using some form of external evidence for survival extrapolation versus how many did not report any use of external evidence for survival extrapolation? This refers to the number of appraisals in which a quantitative application of external evidence affects the estimation of survival for either the intervention or at least one comparator within the cost-effectiveness analysis.
- **Middle circle:** How many appraisals considered external evidence use for survival extrapolation limited only to either:
 - (i) Background mortality adjustment (e.g., life table data used to adjust long-term estimates), and/or
 - (ii) A standard NMA (see Chapter 2, Section 2.3.4 for details) to produce a survival curve for a given treatment (e.g., applying an HR from an NMA to a reference curve to produce a survival curve).

Both of these examples qualify as uses of external evidence, but are well established methods used in HTA and therefore not of greatest interest for this review. Background mortality adjustment was not considered to be of great interest for the review, as such an approach is unlikely to be controversial or highly impactful on survival estimates. In addition, as the focus for this review was on the estimation of OS, use of a 'standard' NMA does not in and of itself constitute an example of where external evidence has been used to influence extrapolations of survival. Appraisals that considered other uses of external evidence (n=13) were denoted as examples of further interest for the purpose of this review.

- **Outer circle:** Of the appraisals deemed to be examples of further interest, how was the external evidence leveraged to inform the estimation of survival within the cost-effectiveness analysis? A top-level summary of how the external evidence influenced the estimation of survival (noting that depending on the structure of the cost-effectiveness model used, external evidence may influence the estimation of different survival-related endpoints).

Figure 4-3: Breakdown of appraisals by use of external evidence



Key: ITC, indirect treatment comparison.

Notes: *'Limited external evidence' refers to appraisals where evidence was used either to adjust for background mortality and/or to inform a standard ITC; *'External evidence of further interest' refers to appraisals that included applications of external evidence beyond background mortality adjustment and/or standard ITCs.

Source: This figure was originally presented with different formatting as part of a conference presentation.⁶

As a minimum, all of the 36 identified appraisals referred to either clinical expert opinion or historical evidence in an informal way in relation to the survival experience of the population being considered within each appraisal. The approach taken to eliciting clinical expert opinion could be based on seemingly informal input related to the clinical plausibility of the chosen survival extrapolation. For example, in TA742, the company states: “... *the choice of curve for the base case analysis was based on visual fit, clinical plausibility and external validation of the resulting clinical outcomes with the outcomes observed in the LIBRETTO-001 and SELECT trials.*” (TA742, CS, p.142), with no further detail provided in terms of how clinical plausibility was determined.²²⁹ However, in other appraisals, the approach taken to elicit expert opinion was more detailed – for example, in TA746 the CS explains that patients who are disease free after three years are assumed to “*experience recurrence and death events at the rate of the general population*”, and that this assumption was “*validated by UK clinicians*” (TA746, CS, p.65).²³⁰ A citation is provided to a clinical advisory board report, though it is stated that this source constitutes ‘data on file’, and so details are not publicly available.

Full details for all appraisals are provided in the data extraction (see Appendix E). Thirty appraisals included at least one source of external evidence as part of the estimation of survival for either the intervention or at least one comparator, leaving six with no use of external evidence in this way. Of the 30 appraisals, 17 included limited external evidence, either in the context of a standard ITC or to adjust extrapolations to account for background mortality. In terms of external *data* sources, the appraisals cited a variety of sources, including national population-level statistics (i.e., life tables), registry data, observational studies, and

data from other clinical trials (typically in a 'similar' population). The remaining 13 appraisals of further interest are discussed in the next section.

4.3.2 External evidence of further interest

The notable examples of appraisals that made use of external data are described below. Please note that due to copyright law, images cannot be routinely extracted from NICE appraisals without contacting individual companies, and so no plots are provided in the remainder of this chapter. However, relevant plots can be found (where not redacted) in the public appraisal documentation on the NICE website.

Post-event survival (n=5)

- **TA689 (acalabrutinib for treating chronic lymphocytic leukaemia):** The company's model included separate transitions for progression, pre-progression death, and post-progression death. Survival beyond progression is the transition which used external evidence, and was informed by data from one of two sources: the RESONATE or MURANO clinical trials. The clinical trial used for earlier transitions, ELEVATE-TN, was deemed to provide insufficient data to specifically model post-progression survival (though exact numbers of events were redacted in the CS). To inform this transition, the company generated survival curves for post-progression survival by fitting parametric models to pseudo-IPD obtained via published KM estimates from each of the studies. In its base-case analysis, the company chose the modelled estimates from RESONATE and MURANO to inform the PPS for the comparator and the intervention within the model, respectively. These sources were chosen based on clinical feedback provided to the company concerning the treatment pathway for CLL – in RESONATE, patients went on to receive a subsequent Bruton's tyrosine kinase inhibitor (such as ibrutinib), whereas in MURANO, patients received a venetoclax-based regimen. The EAG raised concerns with various aspects of the model which influenced the (indirect) estimation of OS. Notably, the EAG explained that the company's base-case approach modelled OS projections for the acalabrutinib arm that were *"very similar to that of the general population without CLL."* (TA689, EAR, p.125).²³¹
- **TA709 (pembrolizumab for untreated metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency):** In the company's model, most transitions were based on the pivotal KeyNote-177 study. However, post-surgery PFS and OS data were obtained from the literature (due to small numbers of patients and limited follow-up). Two studies were considered: Adam *et al.*, (2004) and Cucchetti *et al.*, (2015), with the latter used in the company's base-case analysis.^{232, 233} Both studies were observational studies that provided data for PFS and OS after surgery with curative intent in a population of patients with Stage IV colorectal cancer. While

not explicitly stated, it is expected that pseudo-IPD from Cucchetti *et al.* were generated in order to fit parametric models for PFS and OS following surgery. The company selected a Gompertz model for both PFS and OS. Post-surgery was modelled as a 'mini' PartSA model (i.e., patients would enter a 'post-surgery, progression-free' health state, and then transitions to 'post-surgery, progressed' and 'dead' were inferred via a PartSA model).

- **TA721 (abiraterone for treating newly diagnosed high-risk hormone-sensitive metastatic prostate cancer):** Survival for patients that have castration-resistant disease was informed by external trial data (via the COU-AA-302 study - a population with metastatic castration-resistant prostate cancer, mCRPC), whereas prior to progression to castration-resistant disease, patients were considered to have metastatic hormone-sensitive prostate cancer (mHSPC) and transitions to death were based on data from the pivotal LATITUDE study. To do this, the company took a three-step approach – first, the company extracted survival curves from a past appraisal (TA387) for which the COU-AA-302 study was the main study of relevance – one curve for the intervention (an active, life-extending treatment), and one for the comparator (BSC). The company then assumed any treatment which was considered “*life-extending*” would have the same outcomes as per the intervention in COU-AA-302, which was “*validated by clinical experts in prostate cancer*” (TA721, CS, p.28). Then, the company used expected market share data to produce a weighted average post-progression survival curve. Finally, the company also introduced a ‘calibration factor’ to survival estimates for mCRPC patients “*to adjust for the population differences between LATITUDE and COU-AA-302*” (TA721, CS, p.28).²³⁴ Further details are not available from the public submission materials as only a summary submission was published.
- **TA746 (nivolumab for adjuvant treatment of resected oesophageal or gastro-oesophageal junction cancer):** OS data from the pivotal trial (CheckMate 577) were deemed to be too immature to inform estimates of survival after patients experience disease recurrence, and so the company sought to inform transitions from ‘post-recurrence’ to ‘death’ from the literature. Based on the demographics and treatment history for patients in a retrospective cohort study identified in the literature (Lou *et al.*, [2013]),²³⁵ versus CheckMate 577, the CS states that “*the patient population was considered to align well to CheckMate 577 and the proposed indication*” (TA746, CS, p.95).²³⁰ This published survival estimate was used by the company to produce pseudo-IPD and fit survival curves, which were used for both treatment arms once patients had experienced disease recurrence. The company further explained that the “*application of the same post recurrence survival to both arms in the model is a*

conservative assumption that may underestimate the benefit of nivolumab [the intervention]” (TA746, CS, p.95).²³⁰ Of note, the external evidence considered a population of patients that did not receive the intervention, and so this was highlighted as a limitation of the approach taken by the company within its submission.

- **TA754 (mogamulizumab for previously treated mycosis fungoides and Sézary syndrome):** As part of the company’s model, some patients could undergo potentially-curative allogeneic stem cell transplantation. To inform subsequent survival outcomes, an approach attributed to a previous NICE appraisal (TA577) was implemented. In brief, the company re-created data from published KM estimates based on real-world data collected from the ‘London supra-regional centre’, and fitted survival curves to estimate DFS following allogeneic stem cell transplant. The EAG queried the choice of parametric model, but did not challenge the approach nor data source used by the company.

Long-term survival (n=2)

- **TA695 (carfilzomib with dexamethasone and lenalidomide for previously treated multiple myeloma):** In its model, the company used external data for long-term survival by fitting a ‘hybrid’ (piecewise) model. The model used data from the ASPIRE trial until 72 months, then data from a registry study (MyelomaToul) for the remainder of the time horizon. The registry data were reflective of patients treated with the control treatment only, and included a relatively large sample size (n=1,890 patients that received the relevant treatment in the relevant line). The ‘hybrid’ model is essentially a piecewise model, where the hazard of death from 72 months was determined from a model fitted to the MyelomaToul registry. However, the company sought to account for differences in the mortality between ASPIRE and MyelomaToul by fitting a Cox model to estimate time-dependent HRs before and after 10 months. The choice of 10 months was explained as: *“Based on visual assessment, it was found that the OS curves overlapped during the first 10 months after which they started to diverge.”* (TA695, CS, p.86).²³⁶ However, due to redaction of plots, it is unclear based on the CS if the company re-baselined the data from MyelomaToul or not (and if they did, at which time point the data were re-baselined at). The EAG raised issues with the clinical plausibility of the company’s preferred extrapolations of OS, suggesting that the overall approach taken led to over-inflated estimates for the intervention. The EAG also noted that had the company simply selected the model fitted to the ASPIRE data that yielded results in keeping with the MyelomaToul registry data, then the exponential model might have been chosen which produced estimates that were deemed similar to the base-case hybrid approach. The EAG preferred use of exponential models fitted to the ASPIRE

data only to inform its base-case analysis. The committee ultimately aligned with the EAG's preferred approach, stating in the final guidance: *"The committee preferred the exponential model for estimating overall survival for both treatment arms because it used data entirely from the ASPIRE trial."* (TA695, Final Guidance, Section 3.7).²³⁶

- **TA724 (nivolumab with ipilimumab and chemotherapy for untreated metastatic non-small-cell lung cancer):** To estimate OS, the company used data from the pivotal trial (CheckMate-9LA) and data from another study (CheckMate 227 - Part 1) to inform a piecewise approach. The approach taken was to take estimates from the KM estimate from CheckMate-9LA until 13 months, and then switch to a parametric extrapolation based on CheckMate 227 - Part 1 data. The EAG was broadly satisfied with the company's approach, explaining that *"... the [EAG] considers that after 13 months, the hazards in the nivolumab + ipilimumab arm of CheckMate-227 can be considered equivalent to that in CheckMate-9LA."* (TA724, EAR, p.123).²³⁷

Post-event and long-term survival (n=1)

- **TA684 (nivolumab for adjuvant treatment of completely resected melanoma with lymph node involvement or metastatic disease):** A patient-level meta-regression was conducted to compare the intervention (nivolumab) to the comparator (routine surveillance), using data from two studies (CheckMate-238 and CA184-029, where ipilimumab served as a link between these studies). This yielded parameters for a survival model which included a multi-dimensional treatment effect (i.e., did not assume proportional hazards or accelerated failure times). To then implement the resultant survival curves within a cost-effectiveness analysis, the company used the corrected group prognosis (CGP) method to weight the curves to ensure they were aligned with the population of interest.^{§238, 239} In addition, the CS states: *"long-term survival estimates were estimated from the AJCC 8th edition registry data applied at 10 years... The 10-year time point was selected as this is approximately when long-term melanoma survival outcomes begin to plateau, suggesting that the risk of death due to melanoma is reduced at this time"* (TA684, CS, p.38).²⁴⁰ Finally, the company also adjusted its projections of OS to account for background mortality. Therefore, in total, the company's approach for estimating survival used four different sources of data.

[§] This method was first proposed independently by Makuch (1982) and Chang *et al.*, (1982), with both articles appearing in the same issue of the same journal (see references in main text). The CGP method aims to adjust survival to account for confounders, by essentially producing a weighted average of survival curves that account for each possible combination of covariates within the sample of interest. This method is sometimes also called 'standardisation'. This approach differs from an 'average covariate method' approach, wherein the average values for each covariate are used to produce one survival curve for the population.

Non-standard ITC (n=3)

- **TA704 (trastuzumab deruxtecan for treating HER2-positive unresectable or metastatic breast cancer after 2 or more anti-HER2 therapies):** Data from the pivotal DESTINY-Breast01 study were deemed immature by the company in its submission: “[Compared to PFS] OS data are less mature, with a KM estimator of approximately 80% patients alive at the last data cut.” (TA704, CS, p.115).²⁴¹ Therefore, instead of fitting models to these data, for the intervention (trastuzumab deruxtecan), the company applied an HR to third-line data (with longer follow-up) for a different treatment with a similar mechanism of action (trastuzumab emtansine). For comparator treatments, the company produced survival estimates by fitting parametric survival curves to pseudo-IPD from what it deemed to be ‘relevant studies’. The EAG considered modelling of OS for both the intervention and comparators to not be robust; and in fact, OS modelling was the only key issue raised by the EAG in its report. Ultimately, this submission concluded with a recommendation that the intervention be made available via the CDF, pending further OS data collection. As of May 2025, the post-CDF appraisal (NICE ID5121) is marked as ‘awaiting decision’.²⁴²
- **TA720 (chlormethine gel for treating mycosis fungoides-type cutaneous T-cell lymphoma):** In this appraisal, the company did not claim an OS benefit for the intervention versus the comparator, and so the company derived one estimate of OS which applied for both treatment groups: “Transition probabilities were treatment independent and were based on disease stage only (i.e. independent of skin burden).” (TA720, CS, p.123).²⁴³ The data came from a published single-centre, retrospective cohort study of 1,502 patients by Agar *et al.*, (2010).²⁴⁴ The company took median survival estimates from the study by Agar *et al.* by disease stage to derive transition probabilities for use in the model, assuming an exponential distribution. The EAG made one small change to modelling OS, by assuming that the risk of death for patients with Stage IA disease would be the same as the general population, rather than deriving a transition probability based on the median estimate reported by Agar *et al.*
- **TA725 (abemaciclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy):** In this appraisal, an FP-NMA was undertaken for a range of studies. As part of this, a 'base curve' for a non-scoped comparator (but a treatment in the network) was produced using data from MONARCH 2 and CONFIRM studies. The exact methodology to combine data from MONARCH 2 and CONFIRM studies was unclear based on the CS, but it was implied that these data were synthesised as part of the broad FP-NMA approach. Based on the EAG's report, the EAG appeared to be mostly concerned with the population

informing the FP-NMA, which impacted all outcomes (including OS) but was not a criticism specific to OS alone.

Bayesian informed model (n=1)

- **TA741 (apalutamide with androgen deprivation therapy for treating hormone-sensitive metastatic prostate cancer):** The company followed a published ‘informed fits’ approach by Pennington *et al.*, (2018) to estimate OS.¹⁷³ In this approach, the company explained that “*historical data are used to inform the survival extrapolations*” (TA741, CS, p.39), though relatively little additional explanation was provided in the CS.²⁴⁵ The EAG highlighted that this approach assumes that the ‘shape’ parameter of a given parametric distribution is “*a study independent parameter and could be used from external data to inform the shape parameter for the new clinical trial*” (TA741, EAR, p.118).²⁴⁵ The EAR also implies that a mixture of trial and non-trial evidence was used as the ‘historical data’.

Surrogacy (n=1)

- **TA673 (niraparib for maintenance treatment of advanced ovarian, fallopian tube and peritoneal cancer after response to first-line platinum-based chemotherapy):** Based on immature OS data, the company explained in its submission that it was “*deemed inappropriate to use the highly immature PRIMA OS data as a means of estimating OS for niraparib within the model*” (TA673, CS, p.130).²⁴⁶ Instead, the company derived a relationship between the modelled treatment arms for the intermediate endpoint of PFS, to then inform a modelled benefit in terms of OS. More specifically, the company assumed that for each improvement of 1 month in PFS, there would be a 2-month gain in OS. The company explained that this approach had been “*adopted in past submissions (TA528 and TA611) and was used as evidence in a recent ovarian cancer appraisal (TA598)*” (TA673, CS, p.132).²⁴⁶ However, the EAG did not consider there to be enough robust evidence to justify the use of PFS as a surrogate for OS, and this approach to modelling OS was identified as the key model driver. The committee ultimately determined that OS modelling was an area of outstanding uncertainty, and this appraisal resulted in the intervention being recommended via the CDF. As of May 2025, the post-CDF appraisal (NICE ID6403) is marked as ‘awaiting decision’.²⁴⁷

Additional examples and themes

In addition to appraisals which sought to specifically leverage data collected from a source other than the pivotal clinical trial programme, some appraisals also made use of clinical opinion or assumptions based on previous STAs and/or published literature to influence

survival extrapolations (or inform the selection of specific parametric survival models) – i.e., external *information*. For example, in TA677, an MCM was used to extrapolate OS for the intervention, meaning that a proportion of patients experienced general population mortality (i.e., the ‘cured’ proportion). However, the CS states that “*general population mortality rates may be inappropriate given the impact of prior treatments on survival in these patients*” (TA677, CS, p.117), and so the company applied an SMR of 1.09 to adjust the survival trajectory of long-term survivors within its cost-effectiveness analysis (i.e., the inclusion of the SMR increases mortality risk for long-term survivors relative to the general population, *ceteris paribus*).²⁴⁸ The company explains that the SMR was taken from a study of patients with a different cancer (DLBCL), but that the overall approach was consistent with a previous appraisal for the same intervention in DLBCL (TA559). The specification of an SMR in itself may be considered an additional source of external evidence to inform survival, in addition to using life table data. This is an important consideration for the method testing and comparison chapters discussed later in the thesis (see Chapters 5 and 6).

The majority of appraisals factored background mortality estimates into extrapolations. However, owing to limited reporting, it was not always clear whether or not extrapolations were adjusted to account for background mortality. More specifically, it was only possible to establish with absolute certainty that at least 23 appraisals included some form of adjustment to account for background mortality. For the remaining 13 appraisals, it is unclear if background mortality was accounted for. For example, in TA737 the CS includes a single reference to ‘background mortality’ in the context of which variables were considered within a deterministic sensitivity analysis, but no other description is provided in terms of if or how background mortality data were used to adjust survival extrapolation.²⁴⁹

Another commonly cited reason for including external evidence was based on the need to derive an ITC for the intervention versus at least one comparator which was not reflected by the pivotal clinical trial. Sixteen of the total 36 appraisals included an ITC, which by definition involved some form of external evidence for at least one treatment arm being modelled. The focus for this review was on the estimation of OS, rather than relative effects between treatments, and so use of a ‘standard’ NMA does not in and of itself constitute an example of where external evidence has been used to influence a survival model *per se*. However, appraisals that included a non-standard ITC (either in terms of the methodology used which was more closely linked to the estimation of survival, or the motivation for using the ITC) were explored further (full details in Appendix E). For example, in TA725 the company estimated OS for the intervention and all comparators “*based on the survival output for the FP NMA*”, which involved undertaking an NMA on “*the beta parameters of the FP for the reference*

*treatment, fulvestrant, which was modelled using data from MONARCH 2 and CONFIRM as both include FUL 500 mg” (TA725, CS, p.24).*²⁵⁰

Technical descriptions of how external evidence was incorporated into the survival extrapolations for each appraisal were often limited. However, it should be noted that additional technical detail may have been contained within submission appendices that are not routinely published on the NICE website. For example, in TA741, the company briefly describes its ‘informed fits’ approach to estimate survival within its submission summary (i.e., Document A), but the technical detail is provided only within documentation which is not publicly available (which in this case includes Document B [the full company submission]).²⁴⁵

Based on the reporting contained within the appraisal documentation for those appraisals included in this review, neither the EAG nor the committee appeared to reject any proposed method in its entirety (i.e., no method was rejected based on the method itself being deemed unsuitable to inform decision making). However, the estimation of survival was often highlighted by the EAG as a ‘key issue’ and was discussed by the appraisal committee. This is to be expected, since the estimation of long-term OS is likely to be a key area of uncertainty for any treatment expected to lead to an improvement in survival. In addition, it should be noted that it will generally be in the company’s best interest to justify an approach to modelling OS which produces a sizeable benefit versus the comparator, as this will directly impact the ICER used in decision making. By extension, methods that leverage external evidence are likely to draw criticism if they are seen to potentially ‘improve’ survival estimates versus a simpler approach that does not leverage external evidence.

Common criticisms raised linked to the use of external evidence included the potential for clinical experts to give conflicting views regarding the most plausible estimates of survival (including as well, a timepoint at which ‘treatment effect waning’ should be applied, which was of particular importance for treatments given for a maximum specified period of time [i.e., where a ‘stopping rule’ was relevant]), differences in the evidence sources being used to inform survival (e.g., availability of subsequent therapies), and the use of complex methodology to address the uncertainty associated with a particular source of evidence but introduce another (e.g., potential over-specification of a complex model).

4.4 Discussion

4.4.1 Summary

This review found that several different types of external evidence have been used to inform the estimation of survival in recent NICE STAs of cancer drugs. In total, 13 appraisals factored external evidence from either a clinical trial or a non-trial source into the survival model fitting process. However, of the remaining n=17 STAs that included some limited use of external

evidence, this was used only to either retrospectively adjust extrapolations for face validity reasons or to derive an estimate of survival for a given treatment for which direct comparisons of survival were unavailable.

4.4.2 Searching

Since its first STA published in 2006, NICE has published hundreds of STAs – a large proportion of which have been appraisals of cancer drugs.²⁵¹ In this review, a relatively small number of STAs were targeted in order to feasibly provide a detailed review of external evidence use in appraisals that were conducted recently (i.e., those published in 2021). This intentionally focused search was considered more relevant to determine the *status quo* with respect to methodology employed and perspectives reflected by the EAG and the NICE appraisal committee. This was considered especially important taking into consideration changes in perceptions of external evidence over time – for example, the NICE real-world evidence framework which was published in 2022, which describes NICE’s desire to “*use real-world data to resolve gaps in knowledge and drive forward access to innovations for patients*” (NICE, 2022, p.4).²⁵² In addition, the previous research by Reitsma *et al.* provides a summary of case studies of particular note that were published before 2021.²²⁸

In order to minimise the risk of omitting any potentially interesting findings, broad definitions were used for what constitutes external evidence and what impacts survival estimation (either directly or indirectly), including informal use of external evidence. This revealed a broad range of different approaches to including external evidence within NICE STAs of cancer drugs, introduced by various parties throughout the appraisal process (as described for a number of case studies highlighted previously).

This review made use of published information contained within NICE appraisal documentation, available via the NICE website. Owing to the fact that most STA documentation includes some redacted material (such as confidential pricing information and unpublished study data), it is possible that some approaches to incorporating external evidence may have been missed. This also includes the appraisal committee’s perception of any redacted methodology (given limited reporting in guidance issued by NICE). It is also possible that an appraisal committee did not form a particular view on a particular method if it was not considered to be an issue of high importance within the context of a given appraisal. Furthermore, there were cases where reporting of specific features of a company’s method was insufficient to determine precisely how external evidence was used (e.g., how/whether background mortality was accounted for).

4.4.3 Overall findings

Further research is required to understand how, when, and why external evidence should, or should not, be used. However, even in the absence of such research, the findings from this review demonstrate how external evidence is commonly factored into contemporary HTA decision making, but often with limited technical descriptions. Without clear descriptions of approaches taken, transparency in HTA processes is limited, and the ability for future HTAs to apply accepted methodology is hindered. External evidence can reasonably be expected to feature in future HTAs, and thus reporting standards should be improved to facilitate standardised approaches and understanding of the use of relevant-yet-complex statistical methods for survival extrapolation.

As noted earlier, as documented in a conference poster from 2021, Reitsma *et al.* set out to review of the use of external data sources to quantitatively inform survival extrapolation in STAs, and assess the acceptance of such sources by EAGs.²²⁸ The authors of this study found that out of 528 STAs published from December 2001 to cut-off date (though this was not stated), 28 reported use of external data to quantitatively inform survival extrapolations. Data sources included historical clinical trial data (21 STAs) and registry data (12 STAs). The authors described main criticisms by the EAGs concerning the use of external data, summarising these as follows:

- The external data were not considered representative of the trial population.
- Where performed, adjustment of the external data to reflect the population was subject to limitations.
- The implementation of the external data within the modelling was not considered transparent.
- Incorrect methods were used (though details of this were not provided by the authors).
- Extrapolations of survival were deemed to be clinically implausible.

Overall, Reitsma *et al.* concluded that on the basis of appraisals published prior to the end of 2021 (given the poster was presented in November 2021), external data were rarely used to inform survival extrapolations. In addition, when external data were used to inform survival extrapolations, the data, methods, and justification of extrapolations were “*heavily criticised*” (Reitsma *et al.*, 2021, poster conclusions).²²⁸

The findings from Reitsma *et al.* are broadly aligned with the findings from this later STA review, except depending on how external evidence is defined, one could argue that such evidence is used regularly. In addition, the case for using external evidence is likely to become increasingly relevant over time as older STAs can provide useful evidence for decision making

(e.g., as was the case for TA754, which cited an approach from TA577 to aid survival estimation for the cost-effectiveness model.

This review set out to identify which methods have been previously used in contemporary HTAs. Of those methods identified in the literature (see Chapter 3), excluding adjustments for background mortality, only two methods were used in the appraisals investigated: a piecewise model, and an informed fits model (though the precise methodology for this second approach was not clear from the available published appraisal documentation). No additional methodological approaches to estimate OS using a combination of trial data and some external evidence were identified, though based on the scope of the STA review, some alternative approaches were identified which aimed to use other endpoints (such as PFS) to aid with OS extrapolations. While these approaches are beyond the scope of the thesis, it is noteworthy that such approaches have been used where OS data were limited – this may help in situations where there is a dearth of external evidence for the outcome of OS specifically.

Criticisms raised by EAGs frequently involved general issues with uncertainty in extrapolations, choices of data sources, and robustness of the estimates produced. The sentiments raised by the EAG were often echoed by the committee, and in cases where OS extrapolations were particularly uncertain, this was likely a key determinant of the decision to recommend the treatment via the CDF. Of the thirteen STAs identified of ‘further interest’, two led to CDF recommendations (TA673 and TA704), and in both cases OS was highlighted as a key area of uncertainty.^{241, 246} While re-assessments of these two STAs have not yet been conducted (as of late 2025), in due course they will offer a unique opportunity to evaluate how successful the approaches to estimate longer-term OS were, once updated data become available.

4.4.4 Conclusion

In this chapter, a review of STAs of cancer drugs conducted by NICE has been presented, focusing on how external evidence has been incorporated into survival extrapolations. The review identified a range of approaches, most commonly used to address gaps in trial data rather than to directly supplement survival model estimation. While external evidence was often referenced informally (either for validation or comparison to other data), direct (or ‘formal’) incorporation into survival models was relatively rare. Despite this, use of external data in the cost-effectiveness model featured in several of the STAs reviewed, particularly for specific transition probabilities (e.g., survival after surgery).

These findings suggest that although EAGs and NICE appraisal committees were generally receptive to methods leveraging external evidence, concerns were frequently raised regarding comparability of evidence sources, methodological complexity, and limitations in reporting.

The review was constrained by the level of detail available in public documentation, highlighting the challenges in fully understanding and assessing the methodologies used in contemporary HTA practice. Nevertheless, this analysis provides valuable insights into the practical application of external evidence in HTA, informing subsequent discussions in the thesis and further emphasising the relevance of the topic.

5. Method testing and case studies

Chapter synopsis

In this chapter, the methods identified previously were tested via case studies that represented plausible real-world scenarios where such methods may be used. The method testing aimed to: (i) verify that the methods could be executed using *R*, (ii) better understand how the methods work (e.g., which specific input data or settings are needed), and (iii) determine which methods were suitable for consideration in the neutral comparison study (NCS), presented in Chapter 6.

In total, three case studies were considered for three different cancer types: squamous cell carcinoma of the head and neck (SCCHN), melanoma (a type of skin cancer), and non-small-cell lung cancer (NSCLC). Data from each case study were identified reflecting a target population (assumed to be the clinical trial population), some external evidence (either from a longer-term clinical trial or a disease registry), and some assumptions (as required by some of the methods explored).

Of the approaches included for testing, the multi-parameter evidence synthesis (MPES) method proposed by Guyot *et al.* could not be readily executed using the publicly available code provided by the original study authors. Therefore, part of this chapter included the development of an executable version of Guyot's MPES using the statistical software package *R* (along with Stan probabilistic programming). For the remaining approaches, either off-the-shelf code was used (if, for example, the method was developed as an *R* package), or bespoke code was prepared to run the method based on the description provided by the study authors.

The method testing identified issues with some of the methods concerning how they could be practically used to inform HTAs. These issues included practical problems with setting up software and inputs (e.g., one method requires specifying inputs via a .txt file) and face validity concerns with respect to estimated survival curves (e.g., some methods produced estimates with abrupt changes in the estimated hazard of death, which is unlikely to be realistic). From a practical perspective, the investigations revealed that some method features appeared relatively unimportant (e.g., reference region for methods that used life tables), versus other features which were critical to producing plausible estimates (e.g., duration of the blending interval for the blended survival method). In addition, the method testing identified some important considerations for the NCS (e.g., the possibility of generating error messages).

The development of the R version of the MPES approach by Guyot et al. was published in Pharmacoconomics as part of a tutorial concerning MPES methods for HTA (Bullement et al., 2024).⁴ Testing of some survival models using relative survival methodology was included in a publication I co-authored as part of my work with industry (Vadgama et al., 2022).²

5.1 Introduction

In Chapters 3 and 4, a variety of different methodological approaches to leverage external evidence for survival extrapolations were identified and described. Some of these methods were identified alongside readily executable statistical analysis code and detailed descriptions of how others may apply the relevant methodology to their own research. However, some of the methods were found to be less easily used in other contexts, due to limited reporting and/or lack of executable statistical analysis code (in ‘standard’ software).

The purpose of this thesis is to explore different methods to incorporate external evidence into survival modelling in oncology, with a focus on application to HTA decision making. Therefore, it was deemed necessary to ascertain which methods could feasibly be used in HTA, from both an applicability and practicality perspective. This involves asking questions along the lines of: *“Does this method work with the type of data typically available to inform HTA for a cancer treatment?”* and *“Can this method be executed using standard software packages without extensive programming expertise?”*. It was also considered important to determine if any methods may produce unusual results, and if so, under what circumstances.

As a reminder, the methods retained following the method identification to carry forward into method testing are presented in Table 5-1.

Table 5-1: Summary of retained methods for method testing

Name of method	Citation	Identification	Page
Bayesian relative survival	Demiris & Sharples, (2006) ¹⁵¹	SLR (1 st iteration)	83
Relative survival	Lambert <i>et al.</i> , (2007) ¹⁵⁸	SLR (1 st iteration)	90
Combine model parameters	Hawe <i>et al.</i> , (2014) ¹⁷¹	SLR (1 st iteration)	94
Standard piecewise	Larkin <i>et al.</i> , (2015) ¹⁴⁶	SLR (1 st iteration)	103
Original MPES	Guyot <i>et al.</i> , (2017) ¹⁵³	SLR (1 st iteration)	109
Summary data piecewise	Pennington <i>et al.</i> , (2018) ¹⁷³	SLR (1 st iteration)	117
Expert opinion	Cope <i>et al.</i> , (2019) ⁹¹	SLR (1 st iteration)	120
Informative prior on shape	Soikkeli <i>et al.</i> , (2019) ¹⁹⁹	SLR (2 nd iteration)	135
Blended survival	Che <i>et al.</i> , (2023) ²¹⁴	SLR (2024 update)	151
M-spline MPES	Jackson (2023) ²¹⁶	SLR (2024 update)	157

Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

For the retained methods, the next step was to assess their feasibility for HTA through applying them to some example case studies. For this chapter, executable *R* code was prepared for each of the retained methods, which were then applied to three different case studies (please see Chapter 2, Section 2.3.3 for why *R* was chosen for the research presented in this thesis). At first, a ‘vanilla’ implementation was applied – here, ‘vanilla’ refers to implementations that are as standardised and hands-off as practically possible, relying solely on simulations with no human input in the form of editing or tweaking the methods to improve results. This was a necessary first step to determine a baseline from which to conduct further investigations.

Then, following the preparation of the vanilla implementation, exploratory analyses were conducted to better understand how the methods compare, and how different settings influence results, and critically, under what circumstances they might produce unusual results. It is important to note that there are potentially infinite model specifications that could have been plausibly explored by editing the specific settings and assumptions underpinning each method. For example, a piecewise model could be run using any cut-point greater than zero but less than the maximum follow-up time from the clinical trial. The goal of this chapter was not to explore all possible specifications of methods; however, exploratory analyses were conducted to gain a better understanding of how the methods work, and indeed how they might be best compared via the NCS. Therefore, specific investigations were determined according to a clear rationale for why a particular analysis may be of interest. In addition, it was not the goal of this chapter to demonstrate that any specific method was ‘bad’ – despite this, intentionally exploring combinations of methods and scenarios to establish when certain methods might produce unrealistic results was important to the overall goal of the thesis.

In NICE DSU TSD 21, the authors presented findings from a simulation study to compare different flexible survival analysis methods.⁵² In addition to the full simulation study results, the authors also considered a range of potential scenarios using the simulated data to investigate when biases may arise when applying methods using specific settings or assumptions. For example, the authors considered what might happen if incorrect background mortality rates were used by applying rates from Finland instead of England.⁵² While TSD 21 addressed some potentially interesting scenarios, the full range of scenarios were not specific to the range of methods explored as part of this thesis (i.e., those that aim to leverage external evidence to inform estimations of survival). Nevertheless, these types of investigations were considered important in order to demonstrate and understand how the methods appear to behave under specific conditions. Therefore, like TSD 21, results from a combination of a simulation study (i.e., the NCS) and results from a range of exploratory analyses were presented, to provide a holistic view of how the methods work under a range of circumstances.

A key difference in the approach taken in the thesis compared to TSD 21 is the ordering of the analyses. While TSD 21 presented results from the simulation study and then exploratory scenarios to investigate specific analyses of interest, the thesis adopts the reverse order: method testing is conducted first (in this chapter), followed by the NCS (in Chapter 6). This sequence was intentional. By beginning with exploratory analyses, it was possible to gain a clearer understanding of how the methods behave when applied using ‘real’ data, including how external evidence is incorporated and the potential implications of different modelling choices. These insights directly informed the design of the NCS, ensuring it was focused on the aspects of performance most relevant to real-world use.

In addition to the case studies, this chapter also includes further method investigations using hypothetical (simulated) data, in order to better understand how the methods behave in circumstances not covered by the three case studies. The advantage of using simulated data is that analyses can be performed by ‘stress testing’ specific features of the underlying data informing the models, which may not be fully captured using real-world examples. For example, with simulated data, one may infer how methods fare when the external data source is assumed to have a follow-up duration in excess of what may be considered typical for HTA. Such investigations were expected to be especially useful for the design and execution of the NCS (Chapter 6).

The goal of this chapter was to apply methods to a range of case studies and hypothetical evidence contexts, using the same software package (i.e., *R*). Ultimately, a combination of the findings from this chapter and the NCS (Chapter 6) were anticipated to be useful in determining the suitability of methods that leverage external evidence for survival extrapolation in HTA practice. For this chapter specifically, findings were also anticipated to be relevant for the design of the NCS (such as clarifying which methods could reasonably be compared in an NCS, and which scenarios may be of relevance). Importantly, the goal was *not* to assess method performance – it is not possible to assess the performance of a particular method via case studies alone, as this relies on knowing the true survival experience for the population of interest. Method performance was considered in the NCS (Chapter 6), which systematically compared the ‘vanilla’ implementation of each method, where the true survival experience was simulated. This was deemed necessary to maintain the neutrality of the comparison.

5.2 Methods

In this section, the steps taken to set up *R* code to use each of the different methods is described (Section 5.2.1), as well as the three case studies used (Section 5.2.2). In addition, once the code was prepared and the case studies were set up, specific requirements for each method were identified (Section 5.2.3), and then the final ‘vanilla’ model specifications were defined (Section 5.2.4).

5.2.1 Code preparation

Of the ten retained methods described in Table 5-1:

- Three were possible to execute with relative ease due to the publication of either statistical analysis code or a package (in *R*), or because it is possible to execute the original code via *R* using an interface for a different software package.^{151, 215, 216}
 - For the Bayesian relative survival approach by Demiris & Sharples, this was made possible through obtaining example code via personal communication

with the study author.¹⁵¹ Through this communication, the original study author explained that the Cox-like version of the method appeared to work best, and so this version was carried forward into the methods testing.

- Four were possible to execute by developing relatively simple *R* scripts from first principles.^{146, 171, 173, 199}
 - Statistical analysis code for the method by Soikkeli *et al.* was developed for an initial exploration of methods using data from a population with prostate cancer – see Bullement & Kearns, (2022).¹
- Two were possible to execute using code that was published using the same method that followed the original publication.^{91, 158}
 - The expert opinion approach by Cope *et al.*, (2019) was followed by a publication and corresponding *R* package by Cooney & White (2023).^{91, 253}
 - The relative survival approach identified first in the SLR via a study by Lambert *et al.*, (2007) was followed by a tutorial by Sweeting *et al.*, (2023), including worked examples.^{158, 221}

However, one (method 3) needed substantially more work to be executed in *R*, compared with the other nine methods: the original MPES approach by Guyot *et al.*, (2017).¹⁵³ This method has been cited a number of times in the literature, including by several of the other studies identified in the SLR. Until the publication of the more recent, alternative MPES approach ‘*survextrap*’ by Jackson (2023), the MPES approach by Guyot *et al.* was the only published method capable of leveraging multiple sources of external evidence within one overall model fitting process. In addition, both MPES approaches had not been formally compared. Therefore, as part of this chapter, the original MPES code by Guyot *et al.* was used to prepare an equivalent execution in *R*. Further details concerning the re-creation of the MPES approach of Guyot *et al.*, (2017) in *R* is described separately in a tutorial paper (Bullement *et al.*, [2024]).⁴

A summary of code availability is provided in Table 5-2.

Table 5-2: Method testing code availability

Name of method	Details
<i>Executable code</i>	
Bayesian relative survival	Provided through personal communication with the study authors.
Blended survival	Worked example <i>R</i> code provided in an appendix.
M-spline MPES	An <i>R</i> package was prepared by the authors.
<i>No executable code, but could be prepared from first principles with relative ease</i>	
Combine model parameters	-
Standard piecewise	-
Summary data piecewise	-
Informative prior on shape	-
<i>Latterly published code</i>	
Relative survival	A tutorial by Sweeting <i>et al.</i> , (2023) provides worked examples. ²²¹

Expert opinion	Cooney & White, (2023) present a general and easily implementable <i>R</i> package for incorporating expert opinion into survival models. ²⁵³
<i>No executable code, which could not be prepared from first principles with relative ease</i>	
Original MPES	Code available in an appendix, but could not be run (makes reference to an external website where a custom function was provided, but this is no longer available).

Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

A further requirement for the methods was to edit the tutorial code by Sweeting *et al.* to allow for the specification of UK-based life tables, instead of the default US-life tables. As this process was not straightforward, the code required to use UK life tables is provided in Appendix F.

5.2.2 Case studies

Once all ten of the retained methods had been successfully re-produced in *R*, they were applied to three different case studies to better understand how the methods compare in terms of potential performance, evidence requirements, and assumptions needed. The methods were applied to each of the three case studies described in Table 5-3. The case studies needed to provide sufficient evidence such that each of the methods could be applied and were focused on different cancer types per the overall aims of the thesis – to investigate different methods to incorporate external evidence into survival modelling in oncology.

Table 5-3: Case studies for method testing

Label	Description	Justification	External evidence
SCCHN	Patients with locoregionally advanced SCCHN were enrolled in the UAB-9901 clinical trial (NCT00004227), comparing cetuximab + radiation therapy with radiation therapy alone.	A population with SCCHN was used by Guyot <i>et al.</i> , (2017) to demonstrate the MPES approach. ¹⁵³ All inputs are publicly available, as reported by Guyot <i>et al.</i> in the supplementary appendix.	SEER registry data (as described by Guyot <i>et al.</i> , [2017]).
Melanoma	Patients with treatment-naïve advanced melanoma were enrolled in the CA184-024 clinical trial (NCT00324155), comparing ipilimumab + dacarbazine with dacarbazine alone.	Melanoma was the topic of my previous research ²⁵⁴ , and the method used by the company for estimating survival when submitting to NICE was published by Larkin <i>et al.</i> , (2015). ¹⁴⁶	AJCC registry data (as described by Larkin <i>et al.</i> , [2015]).
NSCLC	Patients with treatment-naïve, PD-L1 positive, advanced NSCLC were enrolled in the KeyNote-024 clinical trial (NCT02142738), comparing pembrolizumab with investigator's choice of chemotherapy.	A population with NSCLC was considered by Vickers (2019) in their comparison of different survival extrapolation methods. ¹⁷⁴ This included the MPES approach by Guyot <i>et al.</i> , (2017). ¹⁵³	KeyNote-010 (NCT01905657), an RCT of pembrolizumab versus docetaxel in patients with previously-treated NSCLC.

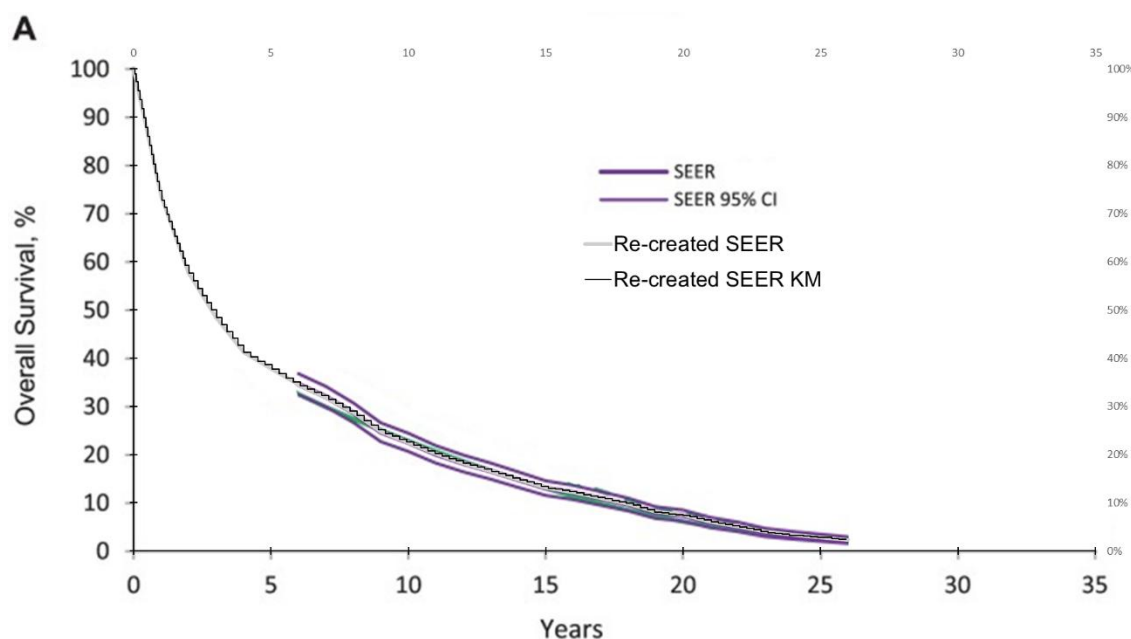
Key: AJCC, American Joint Committee on Cancer; NSCLC, non-small-cell lung cancer; SEER, The Surveillance, Epidemiology, and End Results Program; SCCHN, squamous cell carcinoma of the head and neck.

It should also be noted that some methods require transformation of input data to align with the requirements of a given method, which for the purpose of this chapter is defined as a method-specific setting. This included, for some methods, interpreting external data as if it were provided as clinical opinion (e.g., a point estimate of survival at a given landmark).

For the SCCHN case study specifically, the external data source (data collected in the SEER registry) did not take the form of either 'true' or 'pseudo' patient-level data. Some methods required that the external data are provided in patient-level format (as opposed to point estimates of survival, which was sufficient for the MPES approach of Guyot *et al.*). Therefore, an estimate of the patient-level data was produced via the following process:

- Figure 2A from Guyot *et al.*, (2017) provides estimated 1-year conditional survival probabilities for each year starting at Year 1 until Year 26. However, these data are provided via a diagram only. Using digitised estimates of the conditional survival probabilities, a re-created survival curve was produced. The survival curve was superimposed over Figure 4A from Guyot *et al.*, (2017) to sense check the accuracy of the estimated survival curve (see Figure 5-1, grey line).
- The largest number at risk reported by Guyot *et al.* for the matched SEER cohort was 358 at 6 years (see Table B1 in the supplementary appendix of Guyot *et al.*, [2017]). Therefore, based on the estimated survival probability at 6 years being close to one-third (approximately 34.5%), the starting cohort was assumed to be approximately n=1,000. For simplicity, a value of 1,000 was chosen, and as the cohort was expected to be approximately 1,000 assuming no missing data.
- Using the re-created SEER data, the plotted curve was re-digitised to avoid all events being estimated to occur at annual time points. Then, patient-level data for a hypothetical cohort of 1,000 patients was produced using a published algorithm (see Guyot *et al.*, [2014]), to produce pseudo patient-level data. From these data, a re-estimated KM can be plotted as a final sense check (see Figure 5-1, black line).

Figure 5-1: Overlay of SEER estimates from Guyot *et al.*, (2017) versus re-creation



Key: CI, confidence interval; KM, Kaplan-Meier; SEER, Surveillance Epidemiology and End Results (database).

Note: Original plot taken from Guyot *et al.*, (2017) – Figure 4 (edited to remove superfluous estimates). © The Author(s) 2016. Re-produced here in accordance with the Creative Commons Attribution 4.0 License (CC BY 4.0) (<https://www.creativecommons.org/licenses/by/4.0/>). Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information. *Med Decis Making*. 2017 May;37(4):353-366. doi: 10.1177/0272989X16670604.

The case studies were chosen based on the types of topics commonly seen in NICE STAs of cancer drugs, to ensure their relevance to real-world decision-making contexts (see Table 5-3 for details of the case studies). However, the case studies were not selected with the intention of favouring or disfavouring any method. Rather, it was merely necessary to ensure that the various methods under consideration could be applied to these case studies. As noted previously, the purpose of investigating different methods applied to the case studies was not to assess method performance – method performance is addressed separately in the NCS (see Chapter 6).

5.2.3 Determining method requirements

Outside the context of specific data needs (e.g., patient-level data versus point estimates of survival), some of the methods have requirements that are unique to their specification. For example, methods that adopt a piecewise approach require determining the time point(s) at which a different estimate of survival is used to extrapolate, often described as the ‘cut point(s)’. For each of the retained methods, the model specification was inspected to determine if any specific choices were required for the user to make in order to execute the model outside any default settings imposed by statistical software.

A summary of the method-specific requirements needed to execute each of the different methods is provided in Table 5-4.

Table 5-4: Summary of method-specific requirements

Name of method	Method-specific requirement(s)
Standard piecewise	Cut-point timing(s)
Bayesian relative survival	Baseline hazard specification
Original MPES	Knot locations General population specification HR specification
Expert opinion	Timing and point estimates for expert opinion
Relative survival	General population specification
Combine model parameters*	-
Summary data piecewise	Cut-point timing Long-term hazard interval
Informative prior on shape [†]	-
Blended survival	Timings for blending interval Beta distribution parameters for blending interval
M-spline MPES [§]	General population specification

Key: HR, hazard ratio; MPES, multi-parameter evidence synthesis.

Note: *While no specific settings were required to run this method, it should be noted that the data source used to inform this method must be a randomised controlled trial; [†]The strength of the informative prior on the shape parameter needs to be defined for this method, but it was arbitrarily assumed to take a value equal to 5% of the mean for simplicity; [§]The M-spline MPES method was run using default settings. While the method includes the ability to specify an informative prior for the HR, it does not allow the user to specify that this applies only for a specific time interval. This is different to the original MPES method, which specifically introduces an informative prior for the HR only for a defined time interval. As the default setting is to not enable HR adjustment, the M-spline MPES fitted in this chapter for the purpose of method comparison does not consider this functionality.

To run the Bayesian relative survival model, a custom ‘mu’ file is required, but no guidance was presented by the original study authors for how such a file should be created. For the purpose of this chapter, a custom mu file was created based on the following approach:

- A general population KM estimate was produced using the inputted cohort average age and sex, for annual time points up until the maximum plotted time horizon. This allowed for extraction of estimates of hazard rates.
- An external data KM estimate was produced using the case study specific external data source, for annual time points up until the maximum follow-up in the external data source. This allowed for extraction of estimates of hazard rates.
- The two sets of hazard estimates were combined by simply taking the external data source in preference to the general population estimate, for as long as the external data source provided data to inform the hazard estimates.

This approach was consistent with the worked example presented in the original paper by Demiris & Sharples.¹⁵¹ In principle, the Bayesian relative survival method (and indeed most of the methods included in this chapter) could be executed under a limitless number of specifications – for example, the two data sources could have been combined using any specific cut-off, potentially using different time increments and/or weighted average procedures. This particular means of combining the sources was used as a simple approach,

placing preference of the source expected to be more closely aligned with the trial population, in line with the approach taken by the original study authors.

In addition to the method-specific requirements, for any method that included specification of a parametric model, unless there was a clear rationale to choose a different model, a log-logistic model was selected. The log-logistic model was chosen as an example of a parametric model which has a non-monotonic hazard function, which is expected to more accurately represent the hazard function for the immunotherapy treatments in the melanoma and NSCLC case studies, as well as the SCCHN case study based on the SEER registry data presented in Figure 5-1. Given that the purpose of this study was to explore how the methods work, the choice of a specific parametric model was not deemed critical to the goal of this chapter (despite the importance of parametric model selection in HTA practice).

For general population mortality, the UK was selected as the base-case region, and therefore used to inform the included methods. However, one exception was made based on Guyot's MPES, as the background mortality requirement for this method requires a specific estimate of the number of survivors at one time point, and these data were already presented in the original study by Guyot *et al.*, (2017) based on US survival statistics.¹⁵³ For simplicity and alignment with the original model specification, the general population input parameters were left unchanged for Guyot's MPES.

5.2.4 Method-specific settings (vanilla implementations)

The method-specific inputs for the base-case analysis (i.e., the vanilla implementation) of each of the three case studies is provided in Table 5-5.

Table 5-5: Summary of method-specific inputs (vanilla implementations)

Method	Item	SCCHN	Melanoma	NSCLC
Standard piecewise	Timing (in months) for switching between data sources	From trial to external data: 48 From external to gen pop data: 240 <i>Rationale: 48 months is close to the end of trial follow-up. 240 months expected to be long enough in the future where survival would be per gen pop.</i>	From trial to external data: 48 From external to gen pop data: 240 <i>Rationale: Same rationale as per SCCHN.</i>	From trial to external data: 12 From external to gen pop data: 240 <i>Rationale: 12 months is close to the end of trial follow-up. 240 months chosen for same reason as per SCCHN.</i>
Original MPES	Timing (in log months) for knot locations*, gen pop specification, and HR specification. Numerical values for general population refer to numerator and denominator for survivors at a given time in years until the next year. [†]	Knots: 0, log(30), log(187), log(721) General population*: 34 years, 23,080 of 30,196 HR: $\sim Normal(1.0, 0.1^2)$, 6 to 35 years <i>Rationale: All settings specified as per original paper by Guyot et al., (2017).</i>	Knots: 0, log(36), log(72), log(240) General population*: 34 years, 23,080 of 30,196 HR: $\sim Normal(1.0, 0.1^2)$, 10 to 20 years <i>Rationale: Knots placed at 3, 6, and 20 years, aligned with rationale of original case study by Guyot et al. General population set same as SCCHN. HR assumption arbitrary but expect that hazards will be similar after 10 years.</i>	Knots: 0, log(12), log(48), log(360) General population*: 34 years, 23,080 of 30,196 HR: $\sim Normal(1.0, 0.1^2)$, 10 to 20 years <i>Rationale: Knots placed at 1, 4, and 30 years, aligned with rationale of original case study by Guyot et al. General population set same as SCCHN. HR assumption arbitrary but expect that hazards will be similar after 10 years.</i>
Expert opinion	Timing (in months) and mean point estimates (SE [†]) for expert opinion	60: 40% (0.05); 120: 30% (0.05); 312: 3% (0.05) Values from SEER data <i>Rationale: 60 months approximately end of trial follow-up. 120 months represents double maximum follow-up of the trial. 312 months is the maximum follow-up for the external SEER data.</i>	24: 30% (0.05); 60: 20% (0.05); 180: 7% (0.05) Values from AJCC data <i>Rationale: 24 months within trial follow-up. 60 months approximately end of trial follow-up. 180 months is the maximum follow-up for the external AJCC data.</i>	18: 60% (0.05); 36: 40% (0.05); 72: 20% (0.05) Values from KeyNote-010 <i>Rationale: 18 months within trial follow-up. 36 months represents double the maximum follow-up of the trial. 72 months is the maximum follow-up for the external KeyNote-010 data.</i>
Summary data piecewise	Timing (in months) for switching between data sources and the period over which to estimate the long-term constant hazard	From trial to external data: 48 Upper limit for estimating long-term hazard: 300 <i>Rationale: Trade-off between maximum follow-up of trial (60 months) and external data (312 months).</i>	From trial to external data: 36 Upper limit for estimating long-term hazard: 120 <i>Rationale: Trade-off between maximum follow-up of trial (60 months) and external data (180 months).</i>	From trial to external data: 12 Upper limit for estimating long-term hazard: 48 <i>Rationale: Trade-off between maximum follow-up of trial (18 months) and external data (72 months).</i>
Blended survival	Timing (in months) for blending interval, and Beta distribution parameters for blending interval [‡]	Blending interval: 48 to 600 Blending parameterisation: $\sim Beta(3,3)$ <i>Rationale: Lower interval equal to summary piecewise choice. Upper limit is maximum time horizon (50 years). Beta parameterisation per default settings of R package (given that authors did not identify this to have a large impact on results).</i>	Blending interval: 36 to 600 Blending parameterisation: $\sim Beta(3,3)$	Blending interval: 12 to 600 Blending parameterisation: $\sim Beta(3,3)$

Key: AJCC, American Joint Committee on Cancer; Gen pop, general population; HR, hazard ratio; NSCLC, non-small-cell lung cancer; SCCHN, squamous cell carcinoma of the head and neck; SE, standard error; SEER, Surveillance Epidemiology and End Results (database).

Note: *Knot locations are applied on the log scale; [†]SE's arbitrarily set at 0.05; [‡]Blended survival method does not use general population data in this context; ^{††}In Guyot's MPES, general population survival is applied based on the number of expected survivors for a specified one-year period. For example, in their analysis, Guyot *et al.* assumed that at 34 years from baseline, using data from published life tables there would be an estimated 30,196 people in an equivalent age- and sex-adjusted general population that would still be alive, but only 23,080 of these people would still be alive one year later. For further details, please refer to the original paper by Guyot *et al.*, (2017).¹⁵³

5.2.5 Additional simulated data

While case studies are a valuable tool for exploring and testing the different methods, they can only reveal so much about the methods themselves. By their nature, case studies are specific to the specific data and context they refer to, and these types of investigations may highlight strengths or limitations of a method in those settings, they cannot fully capture the diversity of real-world applications. Any insights gained from case studies may not generalise well beyond the scenarios they represent. This is particularly important to consider in the context of this thesis, as the three case studies were selected because they were considered good candidates for producing survival estimates with external evidence (see Section 5.2.2).

If methods were only ever investigated using examples with ‘ideal’ characteristics, this would risk overlooking important issues that might arise in evidence contexts with less ‘ideal’ characteristics. To address this, further explorations were conducted using simulated data, and were designed to reflect hypothetical contexts that might be encountered in practice, particularly those that deviate from the ideal. Specifically, four scenarios were considered:

1. Short-term follow-up trial combined with long-term follow-up 'perfect' external data.
2. Short-term follow-up trial combined with only slightly longer-term follow-up 'perfect' external data.
3. External evidence indicating substantially better survival than trial data.
4. External evidence suggesting substantially worse survival than trial data, but much longer-term follow-up.

The simulated data were generated using the *R* package *simsurv*²⁵⁵, using simple Weibull distributions. The specific parameters used to generate data are provided in Table 5-6.

Table 5-6: Parameters for simulated data using in method testing

Scenario		Trial population	External population
All scenarios		N = 1,000	N = 10,000
1	External population is 'perfect' and has much longer follow-up	Median survival = 5 years Maximum follow-up = 3 years Shape parameter = 0.8	Median survival = 5 years Maximum follow-up = 30 years Shape parameter = 0.8
2	External population is 'perfect' and has only slightly longer follow-up	Median survival = 5 years Maximum follow-up = 3 years Shape parameter = 0.8	Median survival = 5 years Maximum follow-up = 5 years Shape parameter = 0.8
3	External population has better survival	Median survival = 3 years Maximum follow-up = 5 years Shape parameter = 1.2	Median survival = 6 years Maximum follow-up = 10 years Shape parameter = 1.2
4	External population has poorer survival, but much longer follow-up	Median survival = 6 years Maximum follow-up = 5 years Shape parameter = 1.2	Median survival = 3 years Maximum follow-up = 30 years Shape parameter = 1.2

Key: N, number of patients (sample size).

For the piecewise and Bayesian relative survival methods, it was necessary to define a timepoint after which survival would be informed via general population data. This was chosen

to be the maximum follow-up time for the simulated external population (e.g., in scenario 1, this would be 10 years, unless all patients had died prior to this time, in which case the latest survival time would be used). All other specific settings for each of the models, including the specific functions used to generate the data and fit the models, can be found in the *R* code developed for the method testing, which is available from an online GitHub repository (available via this link: <https://github.com/ABullement/method-testing>).

As with the case studies, the objective of fitting the models to these simulated datasets was not to validate the methods *per se*, but rather to explore their behaviour and outputs under controlled but potentially extreme conditions. This approach cannot cover the full range of possible combinations of survival parameters for both the trial and external populations, but these investigations were intended to complement to the case studies by demonstrating potential issues that may only emerge outside of ‘ideal’ settings. These findings were expected to prove useful in general, as well as for the design of the NCS (Chapter 6).

It should be noted that for the investigations using simulated data, only methods that could be fitted to single-arm data were considered, to simplify the simulation process (i.e., simulate data for a single treatment group, for both the trial and external populations, but not a control arm for an RCT as well). Therefore, Guyot’s MPES and the ‘combine model parameters’ approaches were not considered as part of these investigations.

5.3 Results

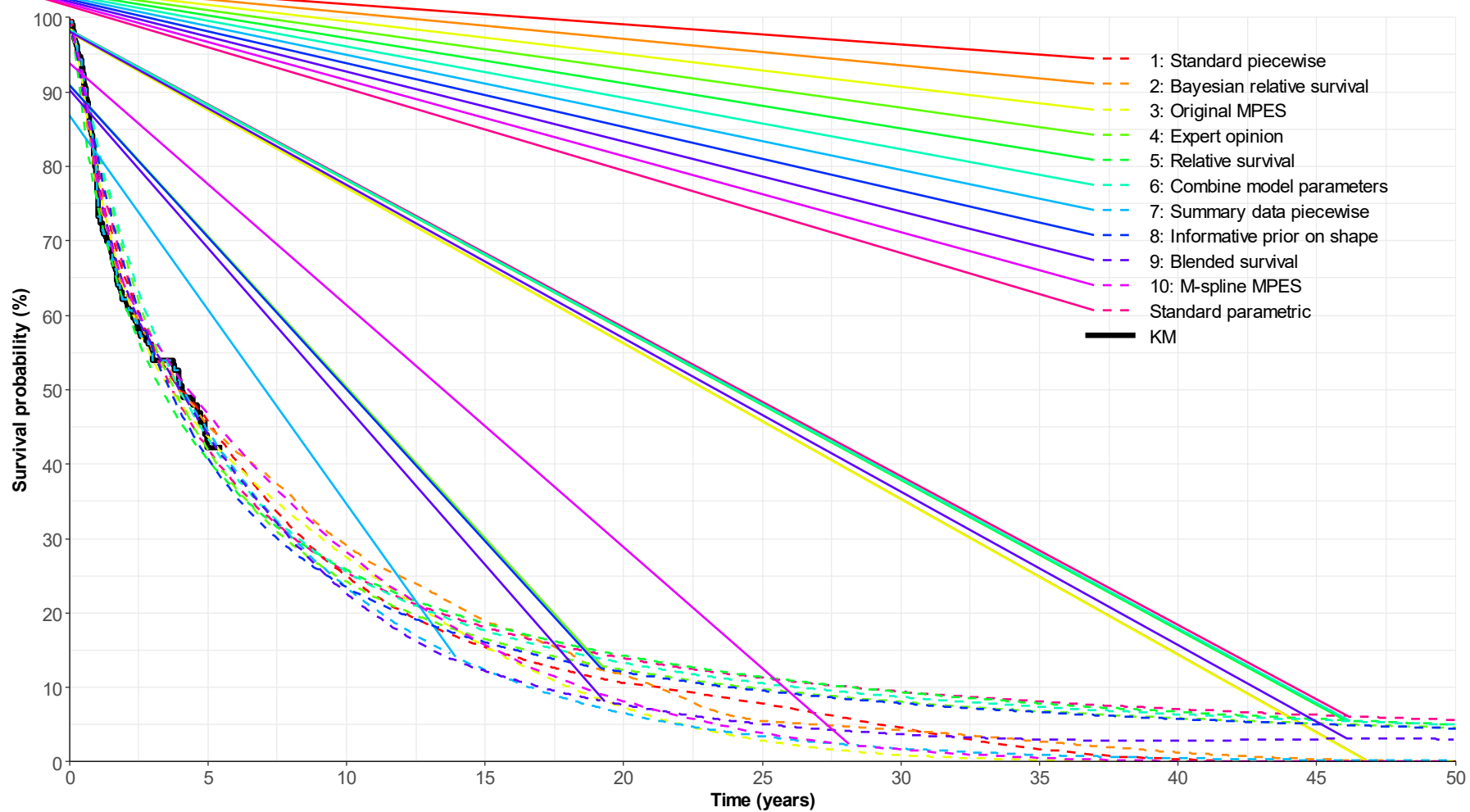
This section presents the results of applying the different methods to selected case studies. The initial analyses compared outputs generated using default settings (i.e., vanilla implementation), to establish a baseline and confirm that the methods were functioning as expected (Section 5.3.1). This is followed by a series of additional analyses, presented in Section 5.3.2, which were designed to explore the methods in greater depth and under a broader range of model specifications, using the same case studies. Investigations using simulated data are then provided in Section 5.3.3, which consider circumstances where methods may be used outside those captured by the case studies.

These results were not intended to compare the performance of the methods, as such comparisons are not feasible within the context of case studies. Instead, analyses were undertaken to illustrate how the methods operate in practice and to highlight potential challenges in their application. The specific results themselves are not the primary focus of this chapter, or of the thesis more broadly. Rather, the aim was to determine whether the methods produce results that can be meaningfully evaluated, and if these results revealed anything of interest for use of the methods in general. This exploration laid the groundwork for the development of the NCS, presented in Chapter 6.

5.3.1 Comparison of vanilla implementations applied to case studies

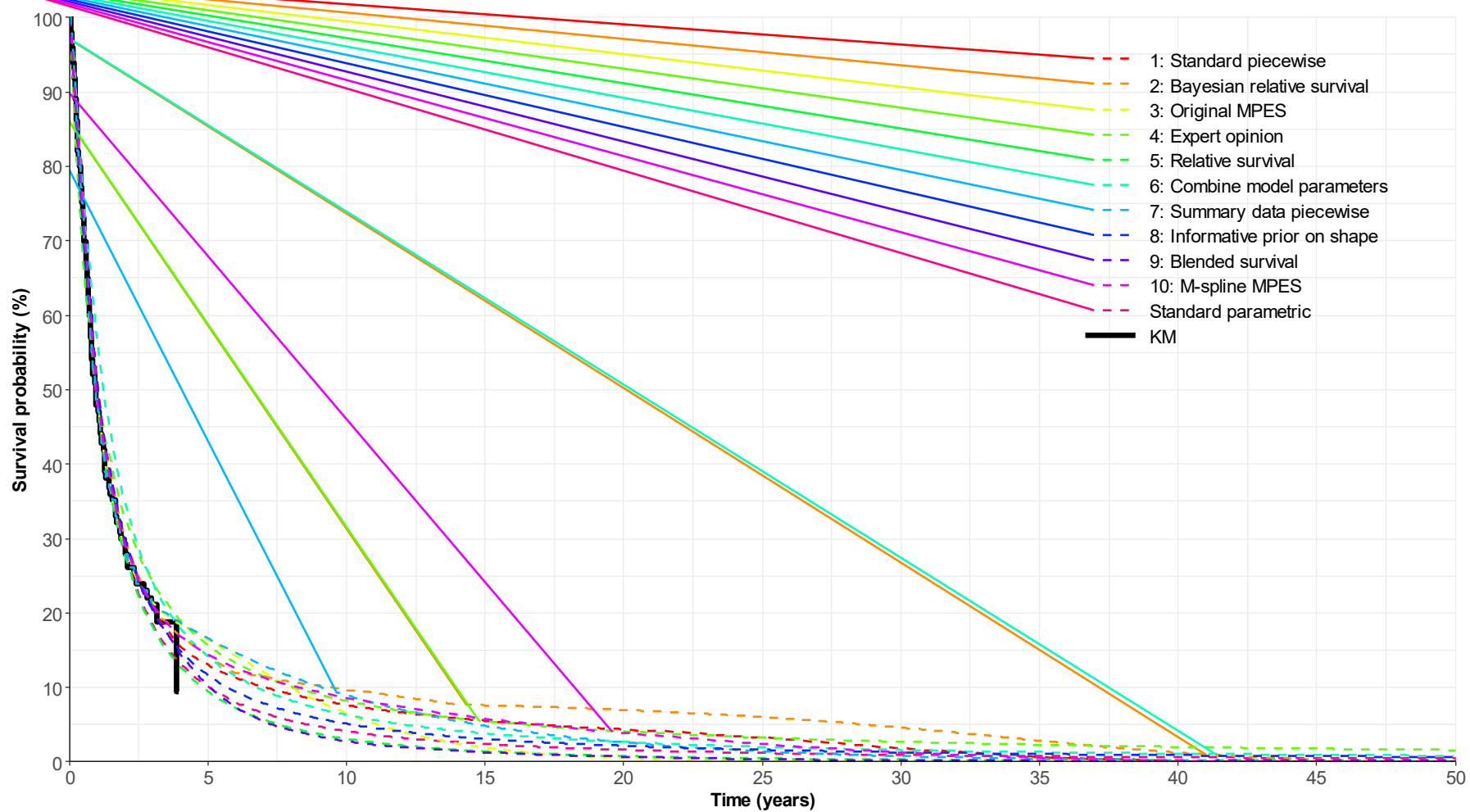
The base-case model fits for each of the three case studies are presented in Figure 5-2 (SCCHN), Figure 5-3 (melanoma), and Figure 5-4 (NSCLC).

Figure 5-2: Method testing – base-case results (SCCHN)



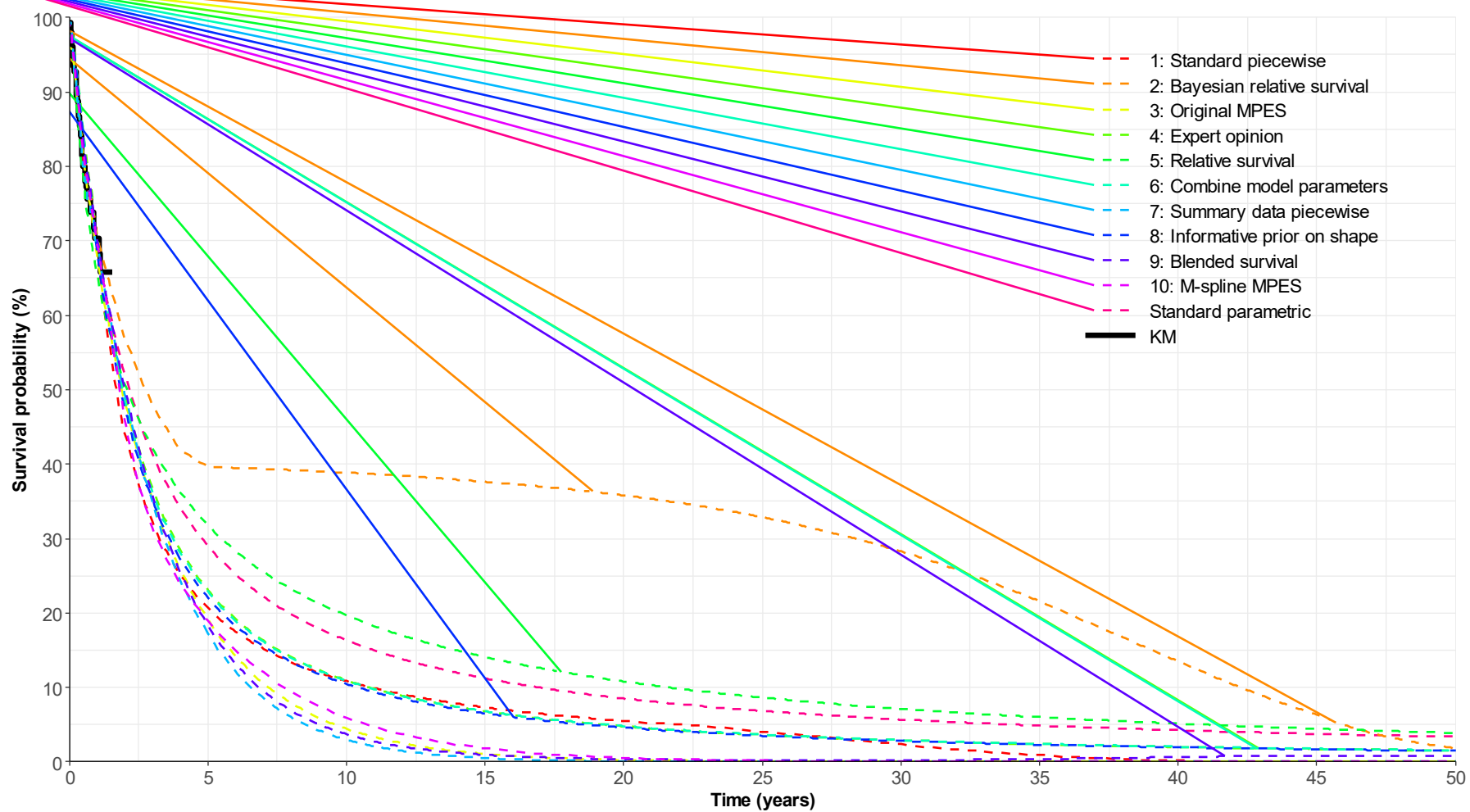
Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis; SCCHN, squamous cell carcinoma of the head and neck.

Figure 5-3: Method testing – base-case results (melanoma)



Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

Figure 5-4: Method testing – base-case results (NSCLC)



Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis; NSCLC, non-small-cell lung cancer.

For the SCCHN and melanoma case studies, each of the methods produced extrapolations that were broadly alike; but for the NSCLC case study, a greater spread was observed. This may be due to the relative immaturity of the trial data informing the analysis (e.g., median survival not reached for the NSCLC case study). The clear outlier of all the model fits across the three case studies is the Bayesian relative survival model for the NSCLC case study. This model produced some unexpected estimates compared to the other methods, with an ‘elbow’ in the curve when ~40% of patients are estimated to still be alive. To understand why this occurred, it is necessary to dig further into this method, and how the model is specified.

To run the Bayesian relative survival method, it was necessary to determine estimates of hazard rates to inform the baseline hazard of the model (labelled as ‘mu’ in the method code). As described in Section 5.2, no specific guidance is provided by the original study authors to determine the inputs for the mu file. Therefore, to fit this method in this chapter, the external data were used to elicit annual estimates of hazard rates until no longer feasible, and then the values of mu were informed by life tables. In the NSCLC case study, this translated to only having a total of five data points in mu informed by the external data, which provided estimated hazard rates that were much greater than the life tables. Moreover, the external hazard estimates did not provide much additional follow-up versus the KM estimate for the trial data. Combined, this means the baseline hazard for the Bayesian relative survival model reflects an abrupt change in hazard when switching from one data source (i.e., the external data) to another (i.e., life tables) and so the resultant survival curve is notably higher than the other methods once this switch occurs.

5.3.2 Further case study method investigations

Further investigations were conducted on specific combinations of methods and case studies. The goal of these further investigations was to investigate any potential issues that may be encountered when considering including each of these methods within the neutral comparison study (see Chapter 6). The further investigations were carried out using one of the three case studies for illustrative purposes. Each of these investigations were structured as follows:

Motivation: Why was this investigation undertaken (e.g., to compare methods that seem similar in terms of their design)?

Methods: How was the investigation performed?

Results: What was found, linking to relevant figures/tables as needed, and what does this mean for the NCS, or use of the method in general?

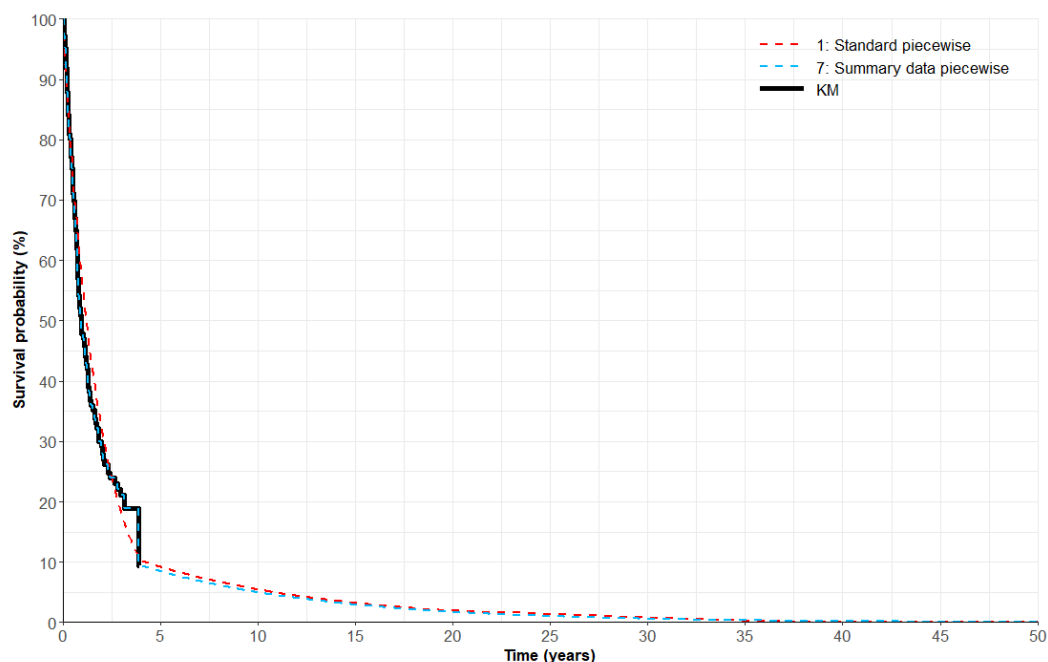
I: Standard versus summary data piecewise approaches

Motivation: Given that both approaches use piecewise methodology, further analyses were undertaken to understand how similar these approaches are.

Methods: In this exploratory analysis, the melanoma case study was used. For the standard piecewise model, an exponential model was selected. For the summary data model, the cut-point was changed to the same as the standard piecewise model (48 months).

Results: The results of this comparison are shown in Figure 5-5. As illustrated in this plot, it is plausible for both approaches to yield very similar extrapolations, assuming an exponential model is used (which is the only functional form that is possible for the summary data model). Selection of the timing of the cut-point(s) can have a marked impact on the results for both models, as documented in previous research.²⁵⁶ The summary data model is more restrictive compared to the standard piecewise model, which may hinder its ability to accurately capture long-term survival estimates and represent a realistic hazard function, but it can be used when there are insufficient data to produce a full piecewise model. In principle, the summary data model could be adapted further to include a general population cut-point. It was therefore determined that both methods were worthy of taking forward into the NCS. However, in general, it is likely that the standard piecewise model would be preferred if feasible to fit, since it can be fitted with a range of different specifications for the latter part of the extrapolation.

Figure 5-5: Melanoma results for ‘standard’ and ‘summary’ piecewise models



Key: KM, Kaplan-Meier.

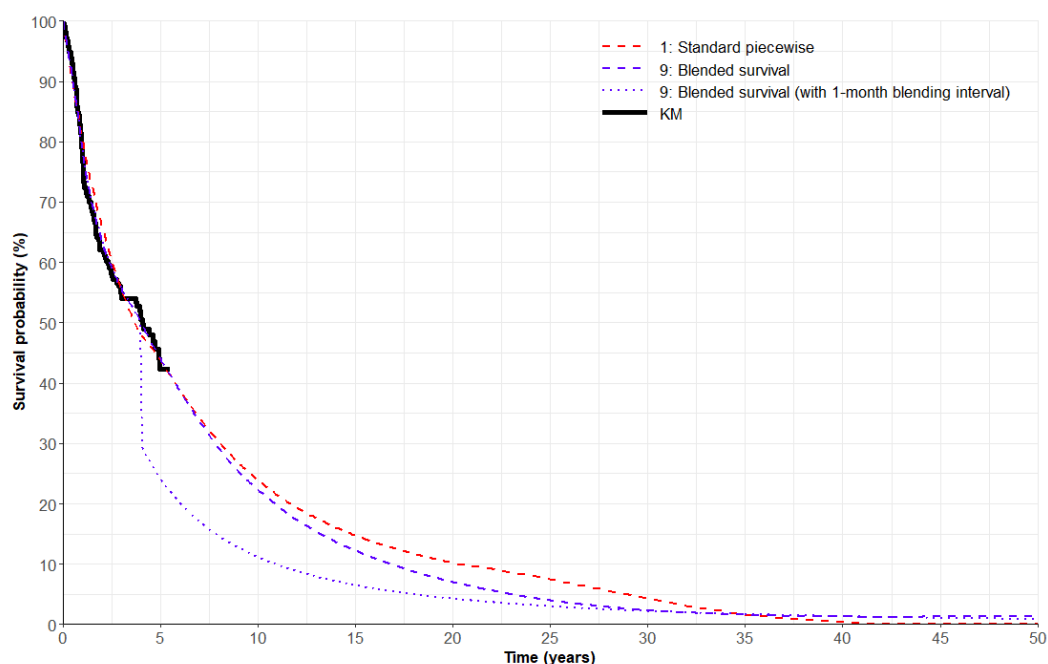
II: Standard piecewise versus blended survival approaches

Motivation: Conceptually, the blended survival model may appear similar to a standard piecewise model, except that there is a 'blending interval' over which the survival extrapolation gradually changes to be informed from one source to another, rather than an instantaneous change in survival. This analysis explores the difference in results between these approaches if the blending interval is reduced, such that it resembles an instant switch in survival.

Methods: The results for SCCHN were plotted to compare both approaches, except the blending interval was set to 1 month (rather than 552 months per the base-case analysis).

Results: The results of this comparison are shown in Figure 5-6. When the blending interval is set to a small value (in this case, 1 month), there is a 'jump' from one function to the other, rather than a smooth blending of sources when a larger interval is selected. This confirms that the blended survival method truly represents blended *survival*, rather than blended *hazards* (i.e., even with a small blending interval, the blended survival model is not the same as a piecewise model). This result illustrates that the blended survival model may produce implausible extrapolations when the blending interval is short, particularly if the reference source differs markedly in absolute survival. The model also does not account for background mortality, though this could be adjusted for *post hoc* (e.g., within a cost-effectiveness model in Excel). Despite these findings, both methods were considered worthwhile taking forward into the NCS.

Figure 5-6: SCCHN results for standard piecewise and blended survival models



Key: KM, Kaplan-Meier.

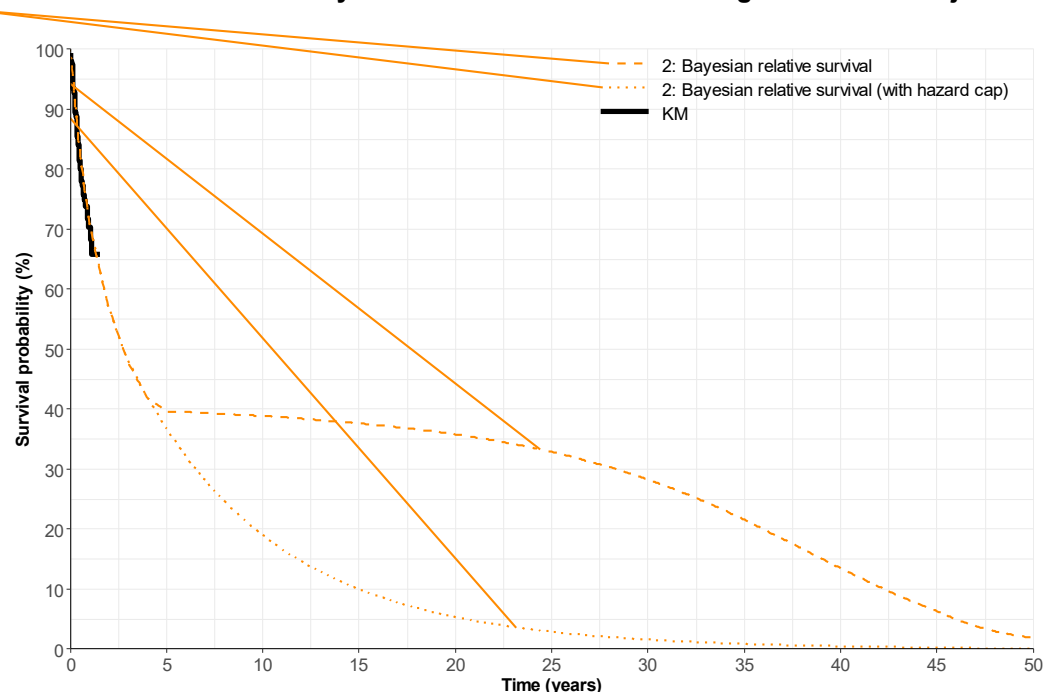
III: Bayesian relative survival model long-term hazards

Motivation: When running the Bayesian relative survival model for the NSCLC case study, an abrupt change in hazards was noted at the timepoint aligned with the maximum duration of follow-up for the external data source.

Methods: Given the unusual result produced for NSCLC (see Figure 5-2), an exploratory analysis was conducted where the estimated hazard in the long-term (i.e., after the maximum follow-up time contained within the external data source) cannot be any lower than the previous period (estimated in months using the code prepared for this analysis).

Results: The results of this comparison are shown in Figure 5-7. As expected, if the estimated hazard in the long-term is set to the hazard estimated in the previous period, the resultant extrapolation appears to provide a more reasonable extrapolation. However, this analysis illustrates that the method may not be suitable for estimating survival in populations where external data have limited follow-up, and where the overall risk of death is substantially greater than that of the age- and sex-adjusted general population, unless the user recognises this and places in a restriction as per this worked example. Based on this finding, and the other results for this method, it was determined that the Bayesian relative survival method was unlikely to be suitable for consideration in the NCS.

Figure 5-7: NSCLC results for Bayesian relative survival with long-term hazard adjustment



Key: KM, Kaplan-Meier.

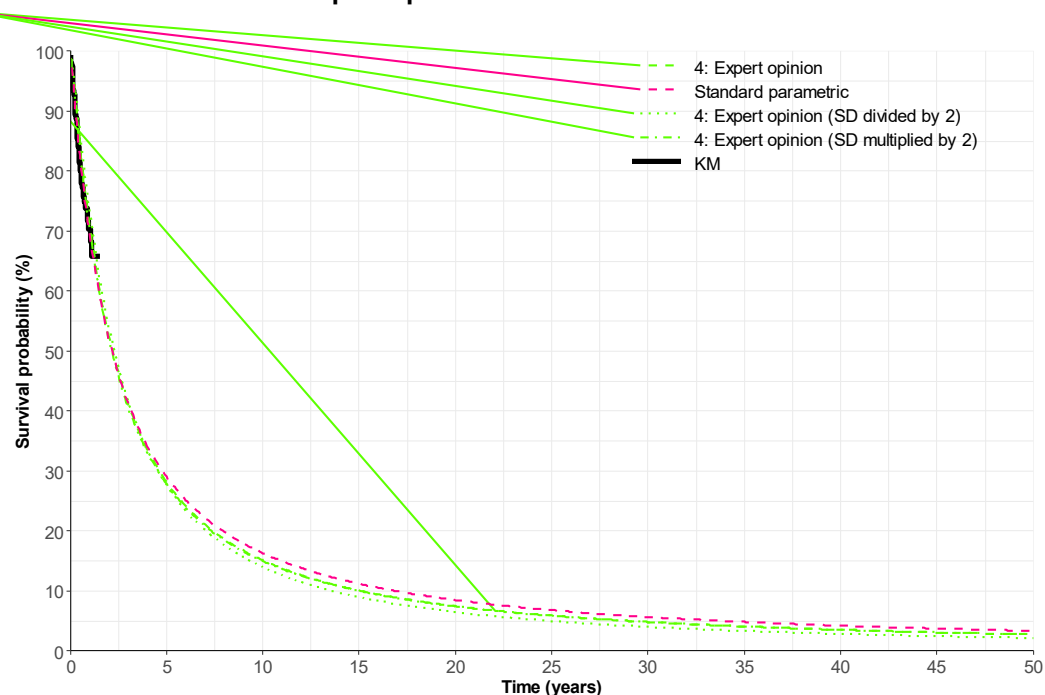
IV: Expert opinion standard deviation inputs

Motivation: The expert opinion method requires inputs related to expert opinion, which require three values: a time point (e.g., 10 years), an estimated survival proportion (e.g., 10%), and standard deviations (SDs) for these estimates (e.g., 0.01). While methods to elicit expert opinion are beyond the scope of this exercise, further analyses were explored to understand how the SD inputs may influence results using this survival extrapolation method.

Methods: Using the NSCLC case study as an example, exploratory analyses were conducted where the SDs for expert opinion timepoints were changed from the default value of 0.01 to reflect increased (SD divided by 2) or decreased (SD multiplied by 2) certainty in the opinion provided. Of note, for the purpose of the case studies, 'opinion' was determined based on the external evidence source, and so SD inputs are arbitrary.

Results: The results of this comparison are shown in Figure 5-8. In this case study, the expert opinion estimates were relatively close to a standard parametric model fitted without any external evidence. Therefore, the SD inputs did not have a great influence on results. However, estimates were still different to the standard model regardless of the strength of the prior, suggesting that even a vaguely informative prior can have a notable impact on extrapolations. For the purpose of the NCS, it was considered important to include the method, but acknowledging that determining suitable SD inputs is challenging.

Figure 5-8: NSCLC results for expert opinion survival with standard deviation edits



Key: KM, Kaplan-Meier; SD, standard deviation.

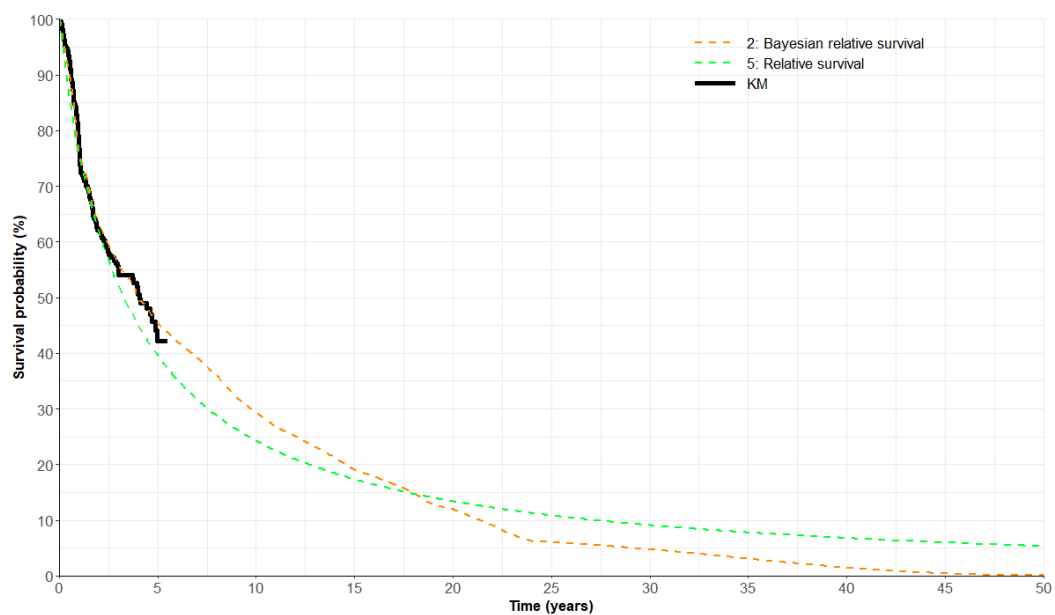
V: Bayesian versus frequentist relative survival approaches

Motivation: Based on their descriptions, the Bayesian and frequentist relative survival models may be expected to generate similar results. To determine whether this occurred, the models were compared.

Methods: The base-case models using either approach to capturing a relative survival model were presented on the same axes, in isolation of all other methods, using the SCCHN case study.

Results: The comparison of these models is provided in Figure 5-9. As evident from the plotted models, these approaches yield different results. These approaches, while sharing a relative survival framework, cannot be considered alternative specifications of the same model under different types of statistical inference. Critically, the specification of the Bayesian relative survival model relies on determining the baseline hazard, which is informed through a combination of external data and life tables; whereas the frequentist approach only uses life tables. In addition, the Bayesian approach considers a ‘Cox-like’ model; whereas the frequentist approach requires a parametric form (e.g., a Weibull model). There are also some further considerations to take into account when interpreting differences in results, such as the strength of priors in the Bayesian model. Overall, it was determined that these are two fundamentally different methods, with differences that go beyond distinguishing between Bayesian and frequentist inference.

Figure 5-9: SCCHN results for Bayesian versus frequentist relative survival



Key: KM, Kaplan-Meier.

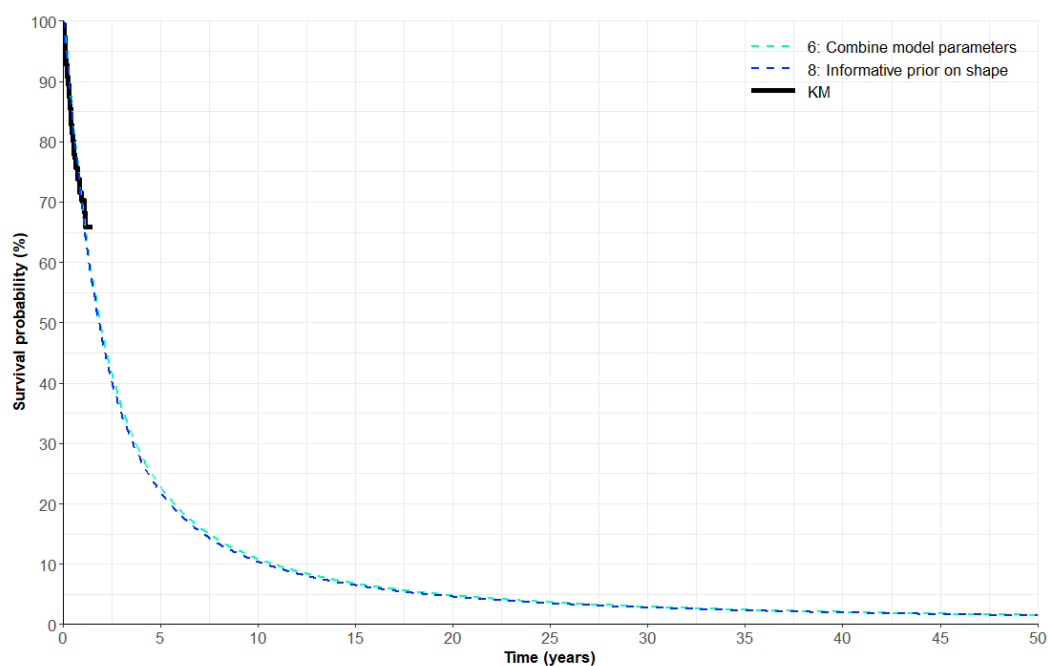
VI: Bayesian versus frequentist approaches for common shape parameter(s)

Motivation: The ‘combine model parameters’ (Hawe) should theoretically yield similar results to the ‘informative prior on shape’ (Soikkeli) method, since both approaches assume a common shape parameter. The former represents a frequentist approach, where the external data source informs the estimation of a shape parameter which is considered fixed; whereas the latter approach adopts a Bayesian approach where the shape parameters obtained from the external data source is used as an informative prior.

Methods: The base-case results for NSCLC were plotted to compare both approaches relying on a consistent shape parameter.

Results: The results of this comparison are shown in Figure 5-10. In this case study, the methods produced nearly identical results. Differences would be expected if the strength of the prior was reduced, which is explored in a separate analysis. Based on this analysis, it was considered unnecessary to include both methods in the NCS, since the approach of Soikkeli *et al.* may be considered a more flexible version of the approach of Hawe *et al.*

Figure 5-10: NSCLC results comparing Hawe and Soikkeli



Key: KM, Kaplan-Meier.

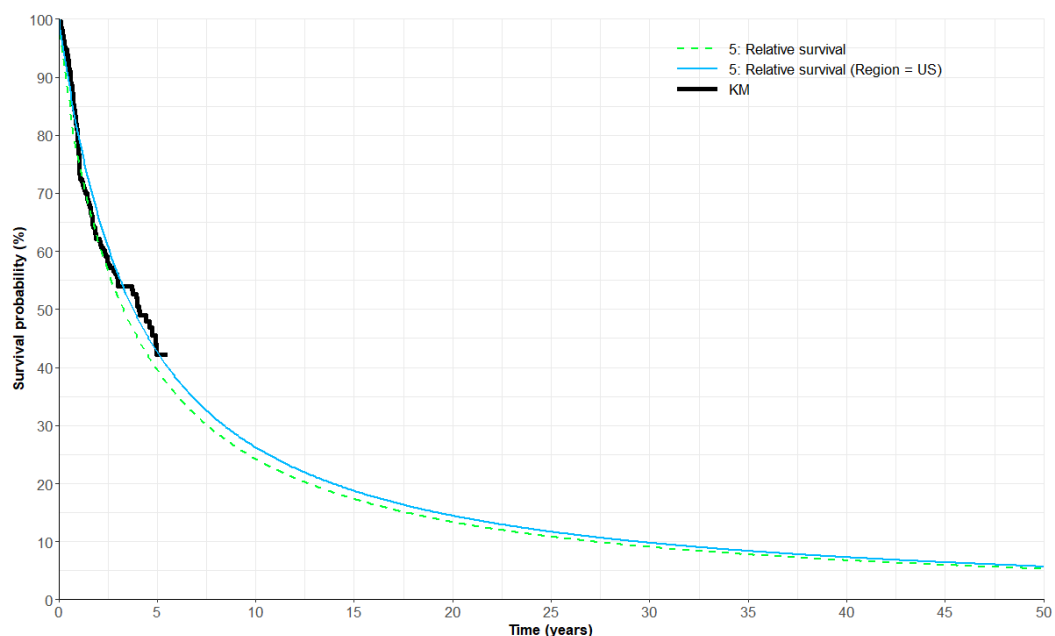
VII: Reference region for a relative survival model

Motivation: When fitting the relative survival model, it is necessary to define an external population from which estimates of expected survival are derived (e.g., the region where life table data are obtained for). In the tutorial by Sweeting *et al.*, (2023), US life tables are used by default.²²¹ However, for HTA in the UK, it is standard to use ONS life tables relevant to the UK population. The choice of region may have an impact on extrapolated survival estimates.

Methods: The relative survival model was fitted using the same settings and assumptions, except that the life table region was set to either the UK (base-case) or the US.

Results: The results of the relative survival models are shown in Figure 5-11. While the estimates are broadly similar, some differences can be seen. Use of UK life tables led to slightly lower estimates of survival until approximately 30 years. While not presented here, the same result was seen when changing the parametric model (e.g., Weibull instead of log-logistic). Also not presented here, the regions were changed for the other case studies, and the differences between the extrapolated estimates were smaller. For the NCS, it was considered unnecessary to explore the impact of different regions for use of life tables, given that any differences in general population mortality by region would be unlikely to materially influence extrapolations.

Figure 5-11: SCCHN results comparing alternative relative survival regions



Key: KM, Kaplan-Meier; US, United States.

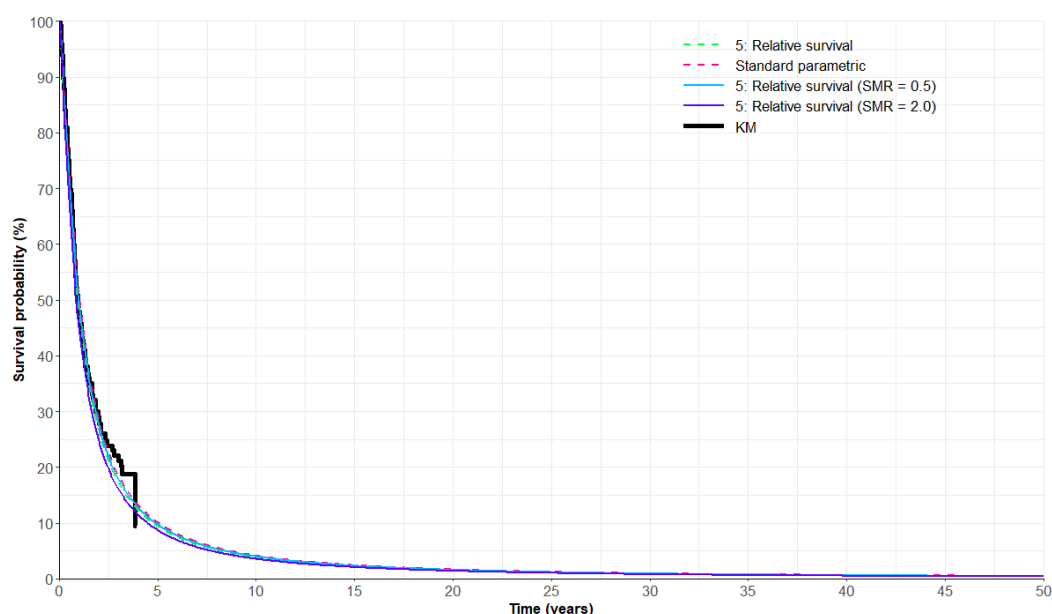
VIII: Magnitude of background hazards for a relative survival model

Motivation: The typical application of a relative survival model involves specifying a background hazard based on published population-level life tables. However, the background hazard experienced by a given population may not be fully represented by standard life tables. It may be the case, for example, that background mortality for a particular group of patients could be higher than what is expected for the general population.

Methods: In this analysis, an alternative relative survival model was specified for the melanoma case study where background hazards were adjusted using a standardised mortality ratio (SMR) applied to the life table estimates. Two different SMRs were specified – 0.5 (background mortality rates divided by 2) and 2.0 (background mortality rates multiplied by 2).

Results: The results of the relative survival models are shown in Figure 5-12. In this case study, including an SMR did not lead to major differences in results. It is likely difficult to establish what the difference in background mortality is between a patient population and the general population, particularly if an SMR does not accurately capture this difference (e.g., older people with a given cancer may be at a higher risk of death from other causes versus the general population, but this may not apply for younger people). Based on these results, for the NCS, it was considered that incorporating the concept of an SMR was unlikely to be worthwhile exploring.

Figure 5-12: Melanoma results comparing alternative relative survival models



Key: KM, Kaplan-Meier; SMR, standardised mortality ratio.

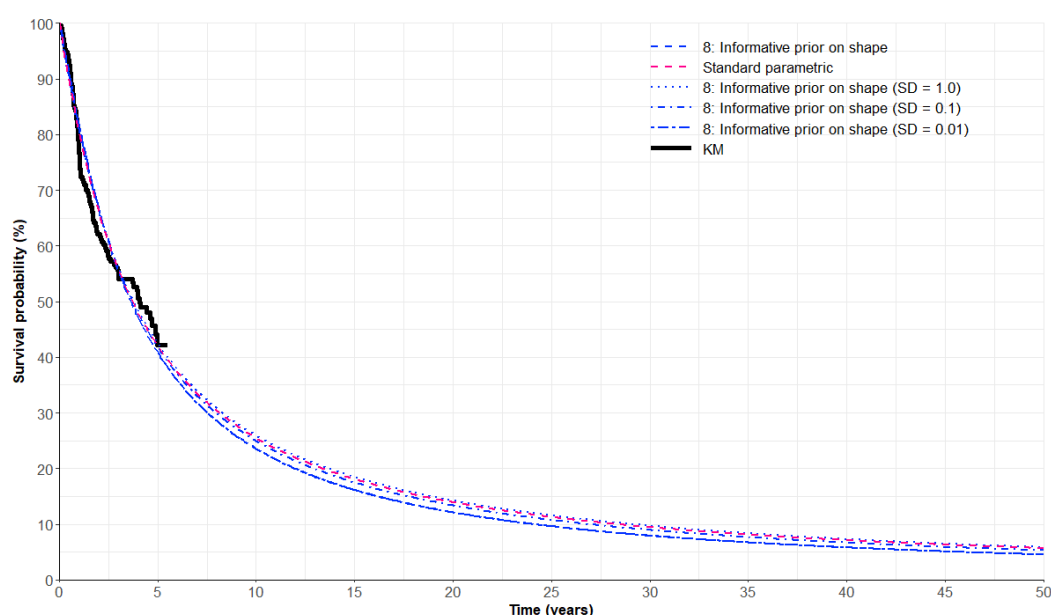
IX: Strength of prior for common shape parameter(s)

Motivation: The ‘common shape parameter’ approach requires the user to specify the strength of a prior for the common shape parameter(s). To understand how results are affected by the choice of prior, further analyses were explored.

Methods: In the base-case analysis, the SD parameter (which determines the strength of the prior) was set to 0.001. Three additional scenarios were conducted, changing this to 1.0, 0.1 and 0.01, to see how reducing the strength of the prior influences results, using the SCCHN case study. With an SD of 1.0, it would be reasonable to expect that the ‘common shape parameter’ approach would yield results similar to the corresponding standard parametric model fitted in a frequentist framework (as this means that the prior is weakly informative).

Results: The results of this analysis are shown in Figure 5-13. For this case study, changing the strength of prior had little influence on the overall model fit, and each of the models produced extrapolations that were similar to a standard parametric model fitted without external evidence. For context, the shape parameter obtained from the external evidence was 1.1635, compared to the shape parameter from the standard parametric model fitted to the trial data which was 1.0826 (around 7% lower, but still above 1). It is therefore unsurprising that even a strongly informative prior on the shape parameter would likely have a small impact on results given that the shape parameters themselves are reasonably similar. For the NCS, it was deemed necessary to specify a prior strong enough to have a tangible impact on the survival extrapolations relative to a standard parametric model.

Figure 5-13: SCCHN results with varying strength of prior for common shape parameter(s)



Key: KM, Kaplan-Meier; SD, standard deviation.

Alternative specifications of MPES

Several sensitivity analyses were undertaken to explore alternative specifications of MPES as part of the tutorial paper published in *Pharmacoeconomics* (Bullement *et al.*, 2024).⁴ In the sub-sections that follow, some notable sensitivity analyses are presented, which highlight key considerations for an MPES approach. For these analyses specifically, the original SCCHN case study presented by Guyot *et al.* is considered, and results are presented for both the intervention and the comparator.

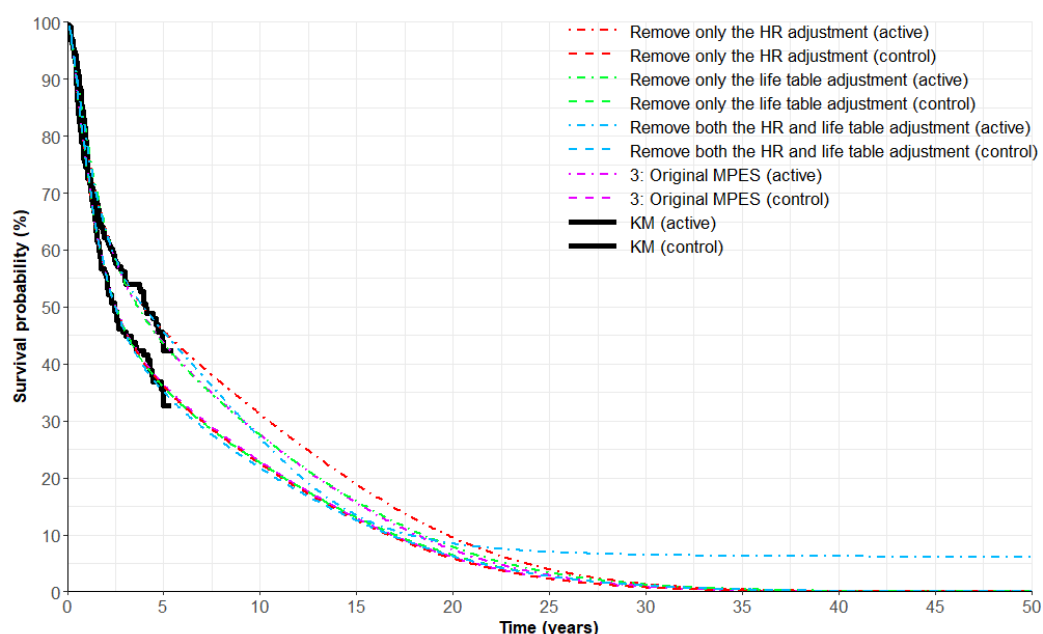
X: Hazard ratio and life table adjustments in Guyot's MPES

Motivation: Guyot's MPES involves synthesising evidence from a clinical trial, a registry, life tables, and clinical opinion concerning the treatment effect in the longer term (based on specification of an informative prior regarding the implied hazard ratio [HR] after a given time point). To determine how influential the latter two features are on the resultant model fits, and ultimately determine if this renders the models suitable for informing the NCS (which relies on estimating survival for a single group of patients), sensitivity analyses were explored where these features were disabled.

Methods: Three scenarios were considered: (i) remove only the HR adjustment; (ii) remove only the life table adjustment; and (iii) remove both. These were explored to understand how the model behaves when general population survival is removed from consideration (life table adjustment), and an explicit link between the treatment arms (HR adjustment).

Results: The results of this comparison are shown in Figure 5-14. There was little impact on the control arm based on removing the life table or HR adjustment (the latter of which would be expected to have no impact, since this parameter only impacts estimates of survival for the active arm). However, for the intervention arm, removing both constraints led to implausible long-term extrapolations. Removing just one of these also led to potentially important differences in the long-term estimates of survival (e.g., the red dot dashed line shows the impact of only removing the HR constraint, which provides the most optimistic estimates of survival until ~20 years when the model removing both constraints plateaus indefinitely).

Figure 5-14: SCCHN results disabling hazard ratio and life table adjustments



Key: HR, hazard ratio; KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

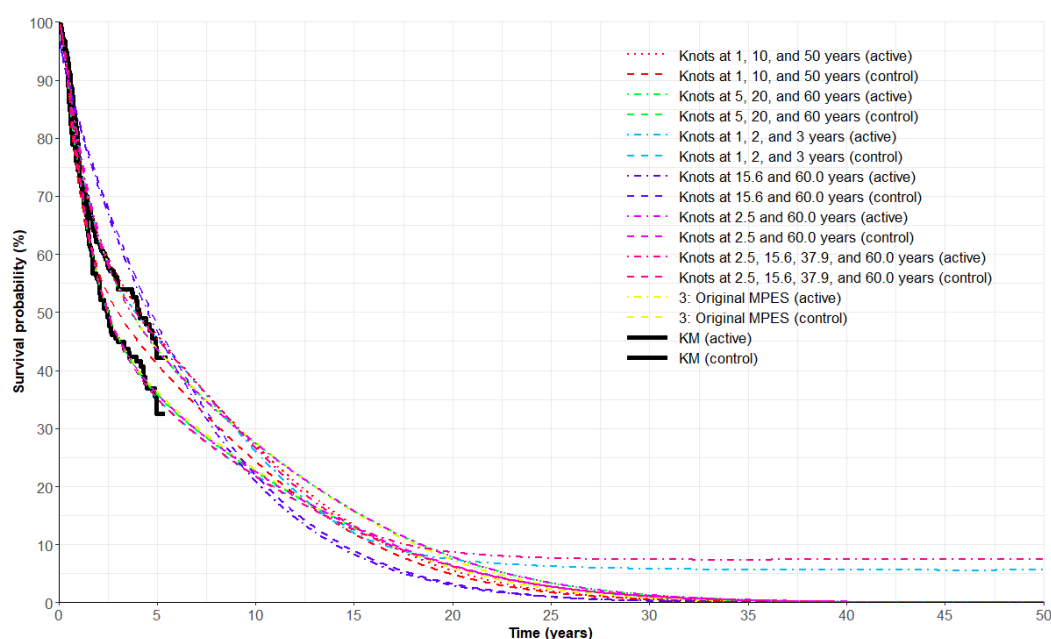
XI: Number and location of knots for Guyot's MPES

Motivation: Guyot's MPES requires the user to determine where to place knots, and the original model specification included three knots placed at 2.5, 15.6, and 60.0 years. To establish how important the placement of knots is for this method, the location and number of knots were varied.

Methods: Six scenarios were considered with knots placed at: (i) 1, 10, and 50 years; (ii) 5, 20, and 60 years; (iii) 1, 2, and 3 years; (iv) 15.6 and 60.0 years; (v) 2.5 and 60.0 years; and (vi) 2.5, 15.6, 37.9, and 60.0 years.

Results: The results of this comparison are shown in Figure 5-15. While this plot is crowded, it reveals some notable findings. The largest impact appears to be for the active arm when only two knots are used (dot dashed pink, blue, and purple lines). For the purple lines, the short-term estimates are notably poor, given that the first internal knot is specified at 15.6 years – a time point substantially later than the trial follow-up period. Consequently, it appears that visual fit is compromised greatly when knots are not spread over a suitable period, and a simple solution to address this in the context of a model with only two knots is to add a third knot (in a sensible location).

Figure 5-15: SCCHN results changing number and location of knots



Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

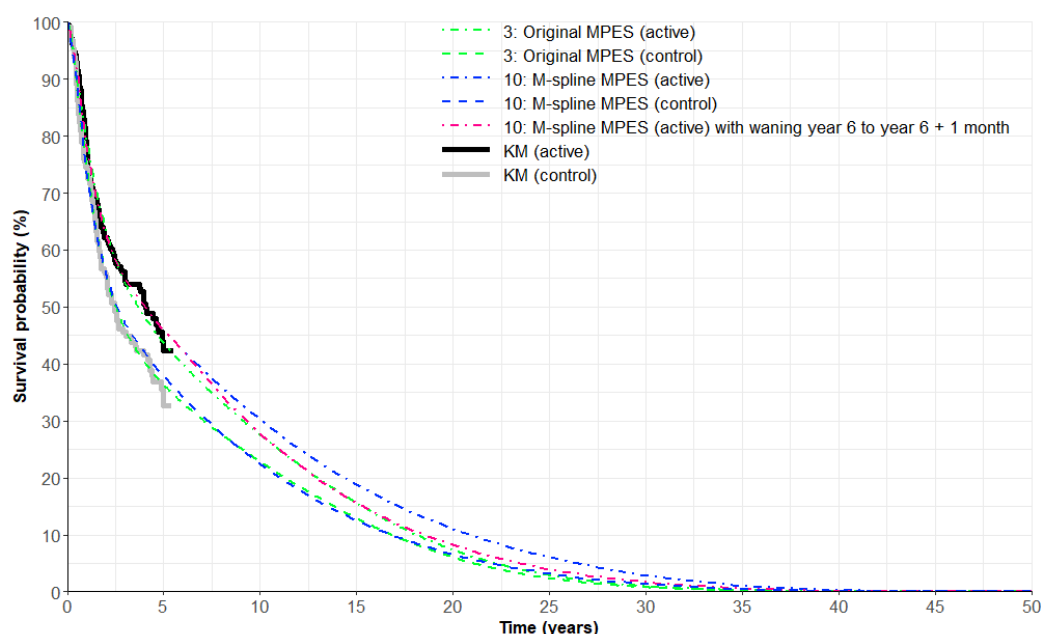
XII: Comparing Guyot's MPES with Jackson's MPES

Motivation: Following publication of Guyot's MPES, Jackson published an alternative specification of an MPES. Using the same inputs, both approaches were compared to determine how similar the results appear to be. In this analysis, Jackson's MPES was re-run fitting to both arms (whereas the base-case model was fitted only to the intervention arm).

Methods: Jackson's MPES was run using the same inputs as per Guyot's MPES, except for the absence of the HR adjustment since this is not a feature of Jackson's MPES. In addition, to emulate the HR adjustment of Guyot's MPES, the *post-hoc* waning functionality in Jackson's MPES was used to capture a waning effect between year 6 and year 6 + 1 month, such that the hazards for the intervention arm were equal to those of the control arm after 6 years and 1 month (one month selected as it is not possible to apply the waning effect instantly).

Results: The results of this comparison are shown in Figure 5-16. The control arms are near-identical, since the input data is identical for this group. However, the active arms are different, which is expected to be due to the HR specification in Guyot's MPES. This prior specification essentially assumes that hazards are equal after 6 years, which in this case causes the extrapolated curve to be lower than that of Jackson's MPES. When including the *post-hoc* waning functionality was included in Jackson's MPES, the resultant extrapolations for the intervention arm were notably similar. Given that the NCS focuses on estimates for a single group only, it was considered sufficient to only include one MPES formulation (Jackson's MPES – rationale presented later in this chapter).

Figure 5-16: SCCHN results comparing both MPES approaches fitted to two arms



Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

5.3.3 Additional investigations using simulated data

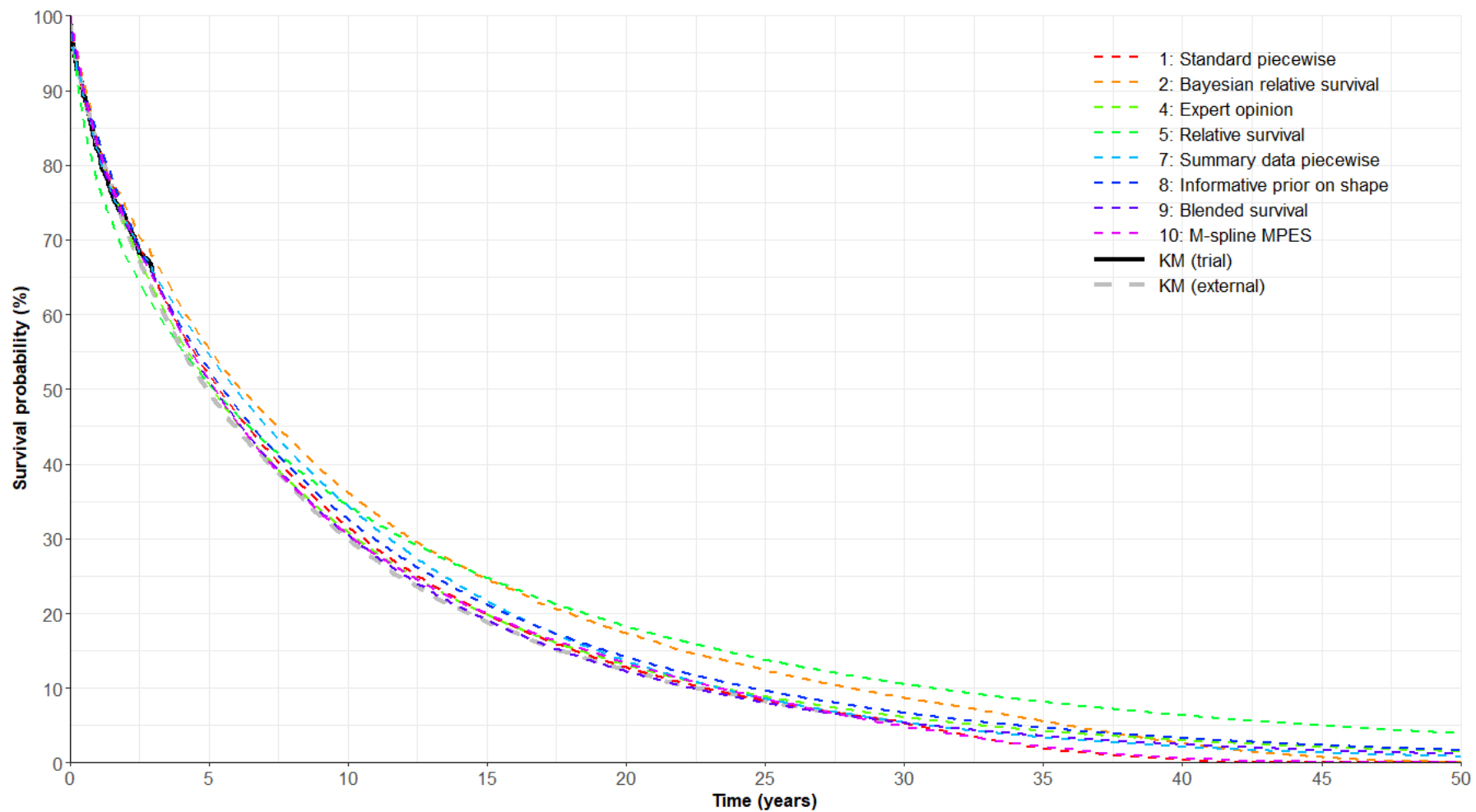
The base-case model results under the four different scenarios are presented in Figure 5-17 to Figure 5-20, respectively. When observing the results, it is important to note that while the external KM estimate may resemble a survival curve, this is not a fitted model.

In the first and second scenarios, the external data was assumed to be a perfect match to the trial, but with different periods of follow-up. These scenarios demonstrated how most of the methods produced accurate estimates of long-term survival, provided the external data yields sufficiently long follow-up. The relative survival method produced the least accurate estimates of long-term survival, though it should be noted that this method only considers general population data for the external data source. For the second scenario specifically, it can be observed that methods relying on a switch to general population mortality require careful consideration of the timepoint – a switch to general population mortality when excess mortality is expected to be non-zero could lead to an unusual ‘elbow’ or ‘kink’ in the survival curve.

In the third scenario, methods were fitted to a combination of trial and external data where the external data exhibited notably improved survival. This approach showed that some of the methods (notably, blended survival and expert opinion) produced estimates of survival that differed substantially from the trial data. For the blended survival model specifically, survival estimates could increase over time, which is clearly not appropriate. In addition, the same background mortality ‘elbow’ or ‘kink’ can be seen for the standard piecewise and Bayesian relative survival methods, though this was not as ‘obvious’ as per the second scenario.

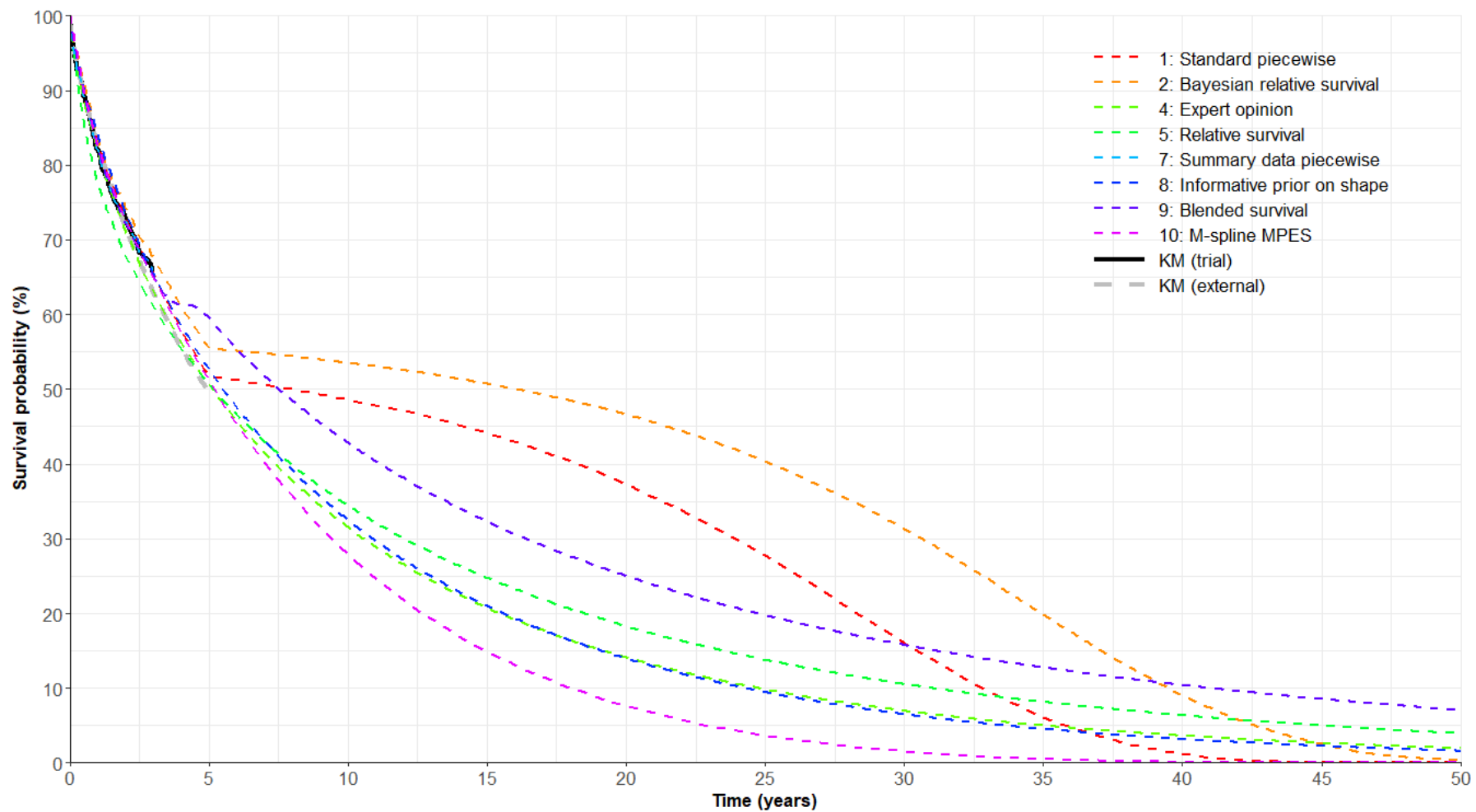
In the fourth scenario, it can be seen clearly that again, the expert opinion method can differ substantially from the trial survival estimate (based on the strength of prior), despite the external population exhibiting notably lower survival than the trial population. This scenario also reveals how the external evidence can “pull down” the survival curve for most of the methods, but not all methods. Notably, the Bayesian relative survival, relative survival, and informative prior on shape methods did not appear to be overly influenced by the magnitude of survival reflected by the external evidence.

Figure 5-17: Method testing – simulated data results (scenario 1)



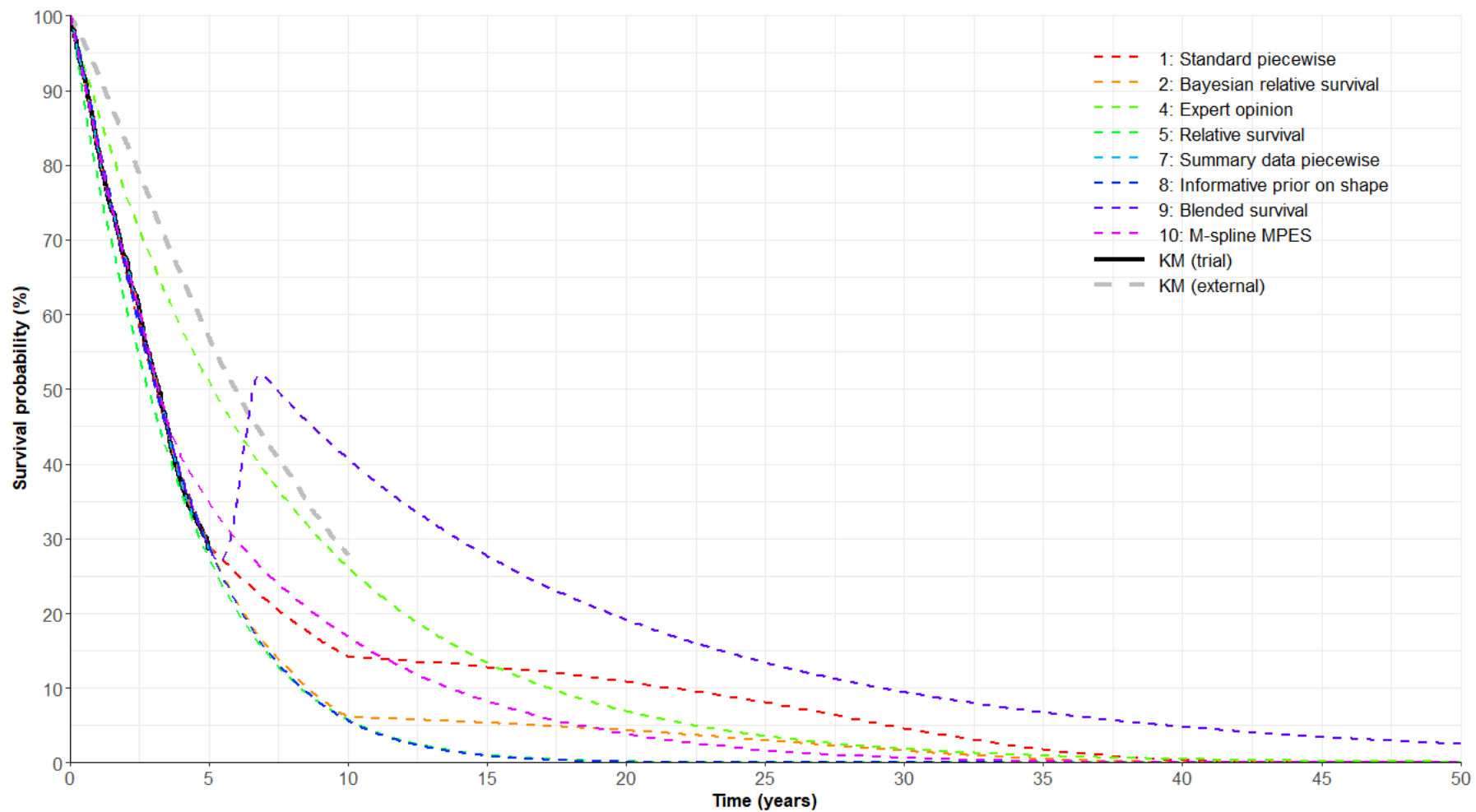
Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

Figure 5-18: Method testing – simulated data results (scenario 2)



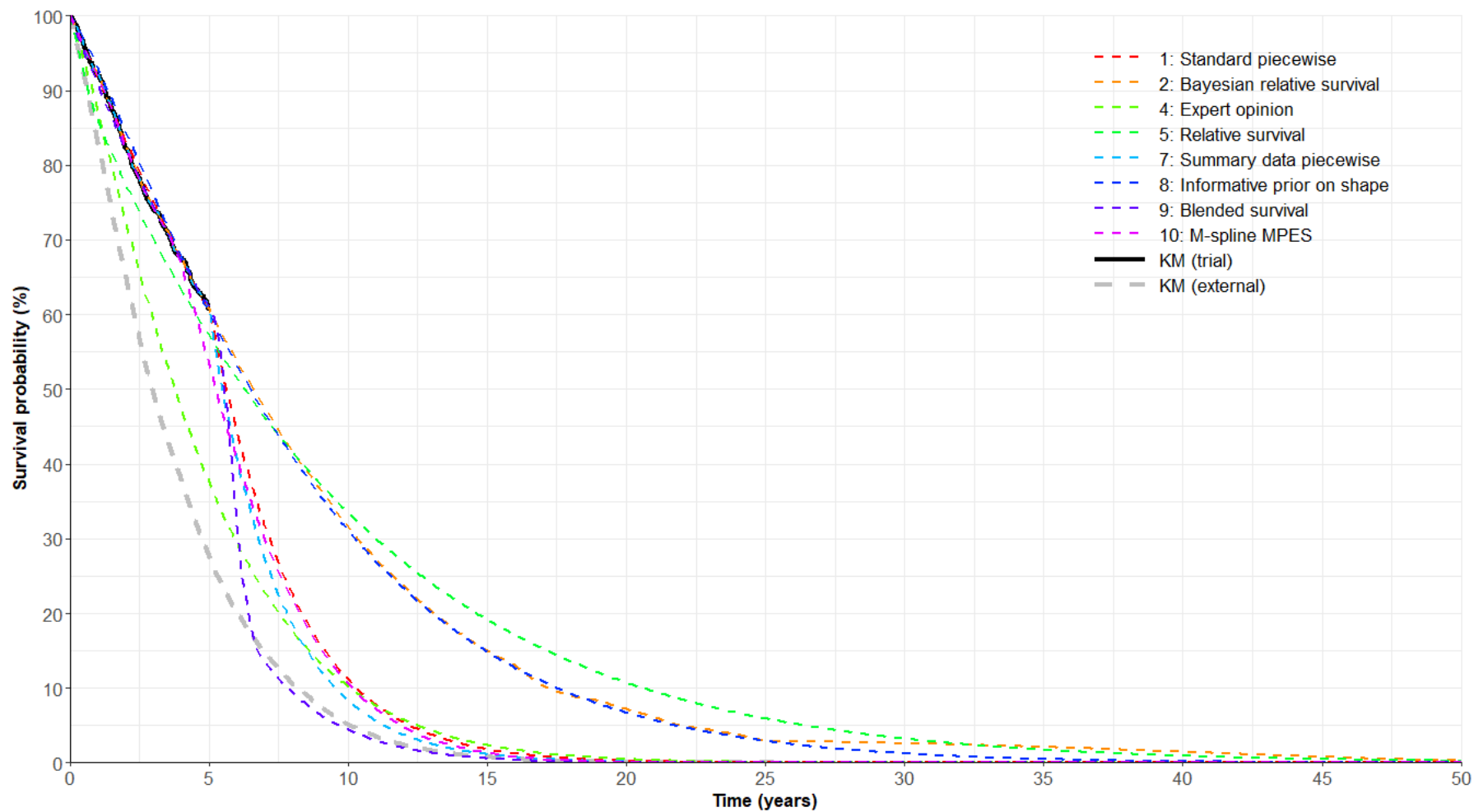
Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

Figure 5-19: Method testing – simulated data results (scenario 3)



Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

Figure 5-20: Method testing – simulated data results (scenario 4)



Key: KM, Kaplan-Meier; MPES, multi-parameter evidence synthesis.

Beyond the specific findings arising from each individual scenario, the testing of methods using simulated data highlighted several broader considerations, both in relation to the application of the methods in general, and the design of the NCS. One practical challenge encountered was the difficulty in automating model fitting procedures when using simulated data. In some cases, by chance, the simulated data did not extend sufficiently far into the future to allow for model estimation. For example, the simulated KM estimate of survival for the external population may simply not reach the required time point specified by a given method per its specification (e.g., switch to general population survival at 10 years based on maximum follow-up, but the external population may have all died prior to this point in time). This has implications for the NCS design, where it may be necessary to define time-based cut-points earlier than the end of follow-up to minimise the risk of model fitting errors.

A further issue arose in the specification of priors for Bayesian methods, particularly the expert opinion method. If priors are too strongly informative, the resulting models may fail to converge or produce errors, suggesting that the prior is incompatible with the trial data. On the other hand, if priors are specified as too weakly informative, they may have a negligible influence on the posterior distribution, effectively rendering the external information redundant. This highlights the challenge of calibrating prior strength in a meaningful and balanced way, which is particularly challenging in the context of simulated data, where the 'true' level of compatibility between datasets needs to be considered. These findings suggest that Bayesian methods of this nature may appear to have variable performance when applied in the context of the NCS.

An additional issue emerged through running the third scenario, where the blended survival model produced estimates of survival that increased over time. This is clearly implausible in the context of modelling OS for use in HTA. This observation was therefore considered a noteworthy finding, prompting the need for further investigation into how frequently such 'logic errors' may occur (for the blended model and indeed any other method) across a wider range of scenarios. Incorporating this into the NCS design was therefore deemed important for assessing the robustness and plausibility of survival estimates generated using different methods.

5.4 Discussion

5.4.1 Summary

The method testing using three case studies, in addition for further testing of method using simulated data, revealed several important learnings for how specific methods function, how their results compare under certain conditions, and when they might not function as intended.

The case studies explored in this chapter reflect typical clinical trials that may be used to inform HTAs of cancer treatments – indeed all three were the topics of NICE appraisals: TA145

(SCCHN), TA319 (melanoma), and TA531 (NSCLC).^{55, 257, 258} The first two case studies (in SCCHN and melanoma) had similar durations of follow-up and maturity of survival data, and sample size (n=211 and n=250 on the intervention arm, for SCCHN and melanoma, respectively); whereas the third case study (in NSCLC) was notably less mature with a smaller sample size (n=154 on the intervention arm). There was a greater spread in estimates produced for the NSCLC case study, compared to the SCCHN and melanoma case studies. Therefore, as may be expected, a combination of sample size and data maturity are both likely to be influential in terms of how the different methods perform.

The primary external evidence sources used in each case study also exhibited different levels of data maturity and similarity to the trial population. The trial data for the SCCHN and melanoma case studies appeared to be more similar to their respective external data sources, which comprised of long-term estimates of survival that were broadly aligned with the clinical trial estimates. In contrast, the NSCLC case study revealed challenges with attempting to use historical trial data as a proxy for long-term external evidence. This was especially clear from some of the analyses explored in the MPES tutorial paper,⁴ as well as the Bayesian relative survival method. Consequently, the overall compatibility of external evidence with the trial data is likely to also be an important consideration for method performance.

Some of the exploratory method investigations were designed to test extreme situations, where methods would be expected to yield somewhat unrealistic results. For example, in investigation II, the blending interval was set to one month – here, the blended survival model would be expected to generate somewhat unusual results because typically a much longer blending interval would be considered more plausible. Therefore, it bears repeating that the purpose of this chapter, and indeed the thesis as a whole, was not to demonstrate that any particular method is ‘bad’ – rather, it was to explore these methods more broadly, including certain extreme scenarios, in order to better understand how they work, and when they might be most useful to consider.

The investigations using simulated data aimed to investigate the different methods in circumstances not captured by the three case studies. The scenarios with perfectly aligned external data (the first and second scenarios) showed that most methods could produce accurate long-term estimates, provided that the external data had sufficiently long follow-up, though the relative survival method performed less well due to it using solely general population data. In contrast, the scenarios with substantial mismatch between trial and external data (the third and fourth scenarios), the expert opinion and blended survival methods produced survival estimates that diverged notably from the trial data. These investigations also revealed broader challenges relevant to the design of the NCS, such as considering issues with random sampling for follow-up, and specifying suitably strong priors.

It should be noted that the case study investigations presented in this chapter were not exhaustive, in the context of the full range of diagnostic plots and model specifications that could have been explored to inform interpretation. For example, a comparison of smoothed hazard plots was not generated for each of the investigations, alongside the modelled hazard functions, though these may prove useful to consider in practice when consider the overall appropriateness of a given survival extrapolation, as per guidance set out in TSD 21.⁵²

5.4.2 Neutral comparison study design considerations

Specifically in relation to the design of the NCS, findings from the method investigations are summarised in Table 5-7. It is important to note that the analyses described in this chapter are based on specific applications of the methods, with ‘vanilla’ implementations executed in keeping with the descriptions of the methods in the papers in which they were originally reported in, and specific analyses undertaken to further understand the methods. While alternative implementations or adaptations of these methods may be undertaken (and may well prove more effective in certain contexts), exploring all possible variations is beyond the scope of the thesis, and would not be practical for the NCS. Further research would be needed to investigate how these methods perform under a broad range of different specifications or in other settings. Also, it is unlikely to be practical to simultaneously compare a broad range of different method applications using simulation techniques, owing to the sheer volume of simulations required to evaluate various different model specifications. The goal of this chapter was not to exhaustively optimise or tailor each method to show it in its best light, but rather to assess their performance as they were originally applied and described, and then perform specific further investigations to better understand how they work in general, and how they could be compared in the NCS.

Table 5-7: Top-level findings from method investigations

#	Finding(s)
I	Necessary to include both standard and summary data piecewise methods in the NCS.
II	Necessary to include both the standard piecewise and blended survival methods in the NCS.
III	Bayesian relative survival method (Demiris & Sharples, 2006) unlikely to be suitable for consideration in the NCS.
IV	Necessary to include expert opinion method in the NCS, noting challenges with SD inputs.
V	Frequentist and Bayesian specifications of relative survival models are two fundamentally different methods, beyond distinguishing between Bayesian and frequentist inference.
VI	Unnecessary to include the ‘combine model parameters’ approach in the NCS.
VII	Unnecessary to explore the impact of different regions for use of life tables in the NCS.
VIII	Unnecessary to explore the concept of an SMR in the NCS for the relative survival model.
IX	Necessary to specify a prior strong enough for the ‘informative prior on shape’ approach, to have a tangible impact on the survival extrapolations.
X	Hazard ratio and life table adjustments have a large impact on Guyot’s MPES output.
XI	Number and location of knots both have a large impact on estimates using Guyot’s MPES.
XII	Sufficient to include only one MPES formulation.

Key: HR, hazard ratio; MPES, multi-parameter evidence synthesis; NCS, neutral comparison study.

5.4.3 Considerations for use of methods in practice

Outside of considerations related to the design of the NCS, the findings from this chapter revealed other important findings for use of the methods in general, which are discussed in this sub-section.

Some of the methods required substantially more effort to produce survival extrapolations compared to others. For example, one of the methods (the Bayesian relative survival approach, by Demiris & Sharples) was re-produced following contact with the original study authors to provide analysis code, which required a .txt file to inform the model.¹⁵¹ Part of this process also led to the identification of an error in reporting in the original study publication, which was confirmed via personal communication with the original study author. The issues with the Bayesian relative survival approach identified as part of this chapter may also have been found when using a different type of approach that considers some kind of change in either survival (e.g., the blended survival model) or hazards (e.g., a piecewise model) beyond a specific cut-point.

In practice, to avoid the issues seen with the Bayesian relative survival approach when applied in the scenarios explored in this thesis, an additional model fitting step could be introduced into the survival modelling process to first provide a more reasonable basis to inform the 'mu' file (e.g., a blended survival model could be fitted to a combination of the external data and life tables, and then this model result could be used to inform the 'mu' file). This was not explored further in this chapter, as running models to inform other models was considered beyond the scope of the thesis.

Contrastingly, some of the other methods had a published package, and could be run in a matter of seconds, particularly those methods which do not require bespoke statistical analysis code. Moreover, some methods could be performed using standard survival analysis outputs (e.g., standard parametric models fitted to the trial data), paired with simple calculations performed in an Excel cost-effectiveness model file (e.g., the summary data piecewise method). This finding may partially explain why some methods appear to have received little attention in contemporary HTA (as evidenced by the review of NICE TAs discussed in Chapter 4), whereas others would be considered somewhat standard and feature heavily both in previous NICE TAs and DSU guidance (e.g., piecewise methods).

Several of the methods investigated could generate survival curves that were not 'smooth', and the final model would exhibit a sharp 'elbow' or 'kink' – such as the piecewise methods, and the blended survival method. This may be deemed clinically implausible, and this finding could highlight possible incompatibility of evidence sources (e.g., an external data source which projects an unexpectedly high hazard). Some methods have been developed to address

this issue – for example, by specifying a sufficiently long blending interval, the blended survival method may produce smooth estimates. However, it could be argued that this approach ‘masks’ an issue of incompatibility of data sources (which is essentially obfuscated by the blending interval).

Some of the methods include use of general population data. The methods that do not include a general population modification could, in principle, be adapted to include an adjustment in some way. van Oostrum *et al.*, (2021) previously compared different approaches for incorporating general population mortality within survival extrapolation, showing different ways this could be achieved and theoretically combined with a broad range of extrapolation methodologies.²⁰² While the goal of the method exploration was not to change the methods, it became apparent through the investigations that some of the shortcomings of some methods could (at least partially) be addressed through a general population specification (e.g., to prevent implausibly long tails in estimated survival curves). Nevertheless, linked to the point made above, a general population adjustment (particularly if imposed *post-hoc*) could theoretically mask an underlying issue with the method itself.

For the relative survival method, it is possible that specifying a different reference region could have an impact on results, though this does not appear to be large based on the explorations undertaken hitherto and reported elsewhere. For example, Stroup *et al.*, (2014) previously explored relative survival estimates for specific US states versus the US as a whole, finding some small differences depending on the sources used.²⁵⁹ In addition, this was also explored as part of TSD 21, wherein the authors intentionally mis specify background mortality for England by using data for a Finnish population instead, finding that the resultant model is less accurate than the correctly specified model, but still performs better than a model which does not adjust for background mortality at all.⁵²

The extent of impact for different specification of a relative survival model is expected to vary depending on the specific properties of the survival data available (e.g., how long patients were followed up for, on average), and could also be influenced by the choice of parametric model to capture disease-specific survival. The choice of region in general can be challenging to consider when attempting to align the target population for decision-making (e.g., the UK, in the case of a NICE technology appraisal) versus the trial population (e.g., an RCT which recruited participants from the US, UK, and [potentially many] other countries). For the NCS, a simplified approach is likely to be appropriate owing to the limited impact on results, but in practical applications this may warrant careful consideration.

On occasion, some methods led to error messages (e.g., failure to converge, failure to find a solution, computation error). For example, in some model runs not reported in this chapter,

the expert opinion model occasionally failed to initialise after more than 100 attempts. This seemed to occur when the expert opinion was defined in such a way which suggested small uncertainty (e.g., a small standard error [SE] such that the expert opinion point was incompatible with the specified model parameterisation and the available trial data). Other approaches, such as the informative prior method, sometimes failed to initialise depending on how informative the prior was specified to be (e.g., if 'too strong' a prior was specified, the model may not initialise correctly). These programming errors will have important implications for the NCS, as based on the method testing explored as part of this chapter, there are highly likely to be some simulations where an error is produced, and it is important to understand how often this occurs (and ideally also, why errors could be generated [and ideally prevented]).

The Bayesian methods investigated require establishing a connection between *R* and a Bayesian software package. For the methods explored in this chapter, three different Bayesian software packages were used: WinBUGS, JAGS, and Stan; but all were executed from the same base *R* script (for ease of combining results, generating plots, etc.). While not an issue for this thesis in general, nor the NCS specifically, expertise in several different programming languages could represent a barrier to entry for the use of some of these methods in HTA more broadly; and, as previously noted, could explain their limited uptake.

5.4.4 Methods for inclusion in the neutral comparison study

With respect to the design of the NCS, the method testing identified a critical consideration in terms of the evidence base required to inform the different methods. Of the ten methods, two explicitly require data from at least one RCT, whereas the other eight methods can be fitted to data for a single group of patients (most of which can only be fitted to data for a single group of patients). Comparing all ten methods in a truly neutral comparison is therefore not feasible, since two of the methods require some additional data that are incompatible with most of the other methods (i.e., data from a control arm).

Taking into consideration the full range of findings from the method investigations, three methods were not carried forward into the NCS, summarised in Table 5-8.

Table 5-8: Summary of discarded methods following method testing

Name of method	Citation	Identification	Page
Bayesian relative survival	Demiris & Sharples, (2006) ¹⁵¹	SLR (1 st iteration)	83
Combine model parameters	Hawe <i>et al.</i> , (2014) ¹⁷¹	SLR (1 st iteration)	94
Original MPES	Guyot <i>et al.</i> , (2017) ¹⁵³	SLR (1 st iteration)	109

Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

The reasons for excluding these methods from the NCS are presented below:

Bayesian relative survival

Presented by Demiris & Sharples, (2006).¹⁵¹ The derivation of inputs for this method (per its original specification) is complex, requiring specification of a .txt file to read into the WinBUGS code, as well as use of the R2WinBUGS package to facilitate the analysis, both of which pose practical challenges for the NCS (e.g., needing to generate potentially 1,000 different versions of a .txt file to run this model). However, more importantly, the method was originally proposed for a population with cardiovascular disease, wherein it may be reasonable to assume similar hazards to the general population by the end of trial (or external data source) follow-up. As shown in the original paper, the method worked well in the context of a population with cardiac arrhythmia, as it is expected that long-term excess mortality tends to zero. However, this is seldom the case in the context of oncology (except in the context of a curative intervention), which provides context to why the method appeared to perform poorly especially in the context of the NSCLC case study. This method was therefore excluded on the basis that it is unlikely to be a relevant method for consideration in HTAs of cancer treatments, which is the focus on the thesis.

It would theoretically be possible to include this method in the NCS if a clear decision-making process was determined to edit the .txt 'mu' file required for specifying hazards for the reference population. However, such an approach goes against the ethos of the NCS, which was to evaluate 'vanilla' specifications of the identified approaches, without attempting to improve upon these by editing the approach itself.

Guyot's MPES

Presented by Guyot *et al.*, (2017).¹⁵³ While substantial efforts were made to facilitate a comparison between the original MPES method and the other methods identified, several issues remain affecting this method and its suitability for inclusion within the NCS:

- The original specification by Guyot *et al.* explicitly requires an RCT. Including a method that requires an RCT poses a large challenge for the NCS, as this means generating control arm data that are ignored by the methods that do not require an RCT (i.e., are incompatible with RCT data).
- As noted by Jackson²¹⁶, there are some mathematical pitfalls of Guyot's MPES which can cause problems with model fitting. In theory, Guyot's MPES can model negative hazards. Personal communication with one of the original study authors of Guyot's MPES supported the idea that the M-spline approach represents an improvement upon the original restricted cubic spline model.
- Including Guyot's MPES would require eliciting two different sources of external evidence – one for conditional survival, and one for clinical opinion. This unique feature

of Guyot's MPES presents a further complexity, wherein it is necessary to simulate both survival estimates (which is common across all methods) and an expectation of a treatment effect over time. Evidence synthesis methods, including indirect treatment comparisons, are beyond the scope of this thesis.

- The original method itself includes an element of human checking of an original model fit, before then making changes and re-running. It is impractical for human checking to be carried out for each iteration of the analysis used in the NCS (as this would require a person individually checking results for 1,000 different model fits). While some of these processes could be automated (e.g., checking for background mortality violations), this raises a question of 'fairness' for the NCS (e.g., is it appropriate to allow only one method to be fitted multiple times if the first fit was deemed sub-optimal or implausible?).
- Finally, it is anticipated that in the future, Jackson's MPES will effectively supersede the use of the original MPES for HTA, given its user-friendly implementation and the use of M-splines which do not suffer from the same issues that can be encountered with using restricted cubic spline models, and that it can be fitted in the single-arm study setting as well as the RCT setting.

Therefore, taken together, the original MPES was excluded due to a combination of practicalities, relevance, comparability, and theoretical considerations. However, the principle reason was that it was designed to address an RCT context, whereas in this thesis the focus is on methods that can be used to extrapolate survival from any data source – not only RCTs.

Combine model parameters

Presented by Hawe *et al.*, (2014).¹⁷¹ This method is conceptually similar to the informative prior on shape parameter(s) method by Soikkeli *et al.*, (2019)¹⁹⁹, except for two differences:

- The combine model parameters method explicitly requires data from an RCT, which is how external evidence is brought into the model estimation.
- The informative prior on shape parameter(s) method allows for a degree of control for specifying the strength of prior on the shape parameter taken from the external study.

Therefore, given that the informative prior on shape parameter(s) method is more flexible and does not require data from an RCT, the combine model parameters method was excluded.

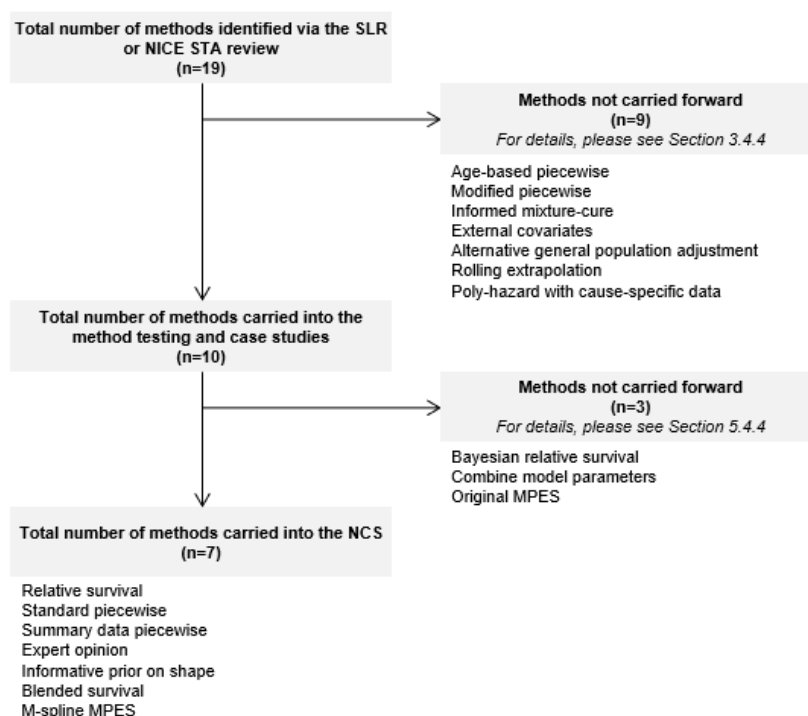
A total of seven methods were ultimately carried forward to the NCS in Chapter 6:

- Standard piecewise by Larkin *et al.*, (2015)¹⁴⁶ and others.^{147, 148}
- Expert opinion by Cope *et al.*, (2019)⁹¹ and others.²⁵³
- Relative survival by Lambert *et al.*, (2007)¹⁵⁸ and others.^{156, 160-163, 182, 202, 221}

- Summary data piecewise by Pennington *et al.*, (2018)¹⁷³
- Informative prior on shape by Soikkeli *et al.*, (2019)¹⁹⁹
- Blended survival by Che *et al.*, (2023)²¹⁴
- M-spline MPES by Jackson (2023).²¹⁶

An updated summary of method identification and determination of which methods are carried forward in the remainder of the thesis is presented in Figure 5-21.

Figure 5-21: Methods summary following completion of method testing and case studies



Key: MPES, multi-parameter evidence synthesis; NCS, neutral comparison study; SLR, systematic literature review; STA, single technology appraisal.

5.4.5 Conclusion

In this chapter, the methods identified in earlier chapters were tested through three case studies to assess their practical application in HTAs of cancer drugs. By applying each method to real-world scenarios, the testing process revealed key data requirements, potential implementation challenges, and methodological limitations. The findings informed the selection of suitable approaches for the NCS in Chapter 6, ensuring that comparable methods suitable for HTA of cancer treatments were carried forward.

The three case studies considered different cancer types, with specific combinations of clinical trial data, external evidence, and assumptions (as mandated by some of the specific methods explored). While most methods could be implemented using existing or custom-developed code, the MPES approach proposed by Guyot *et al.*, (2017)¹⁵³ required redevelopment in R and Stan due to issues with publicly available code. The testing process also highlighted

practical challenges such as software setup complexities, unrealistic survival estimates in some cases, and the potential for error messages. In conclusion, this chapter provided critical insights for the structured comparison of methods in the NCS, ensuring that methodological limitations are well understood before formal evaluation.

6. Neutral comparison study (NCS)

Chapter synopsis

In this chapter, a neutral comparison study (NCS) of methods to incorporate external evidence into survival extrapolation for health technology assessment (HTA) is presented. An NCS adopts simulation study techniques to provide a comparison of different methods, from the perspective of a neutral party (i.e., a researcher with no involvement in the development of the methods being compared). By using simulation techniques, methods can be evaluated under controlled conditions, and further insights into how these work, as well as where they seem to perform well or less well can be determined.

The NCS was developed based on the core structure of a previous simulation study used to inform TSD 21⁵², in keeping with guidance by Morris *et al.*, (2019).²⁶⁰ A range of data generating mechanisms were explored to determine how the methods performed, covering different hazard profiles, sample sizes, magnitudes of survival, similarity of the trial and external populations, and the duration of trial follow-up. In total, 32 scenarios were considered. Mean survival time over a lifetime horizon was the key estimand of relevance. Eight different methods were compared, comprising of seven methods incorporating external evidence, and a standard parametric model serving as a baseline from which to determine how the performance for the methods that introduce external evidence compared. The main performance measure was the mean bias – that is, the difference in mean survival between the fitted model and the ‘truth’. Defining the truth as either the trial population or the external population was explicitly factored into the design of the NCS.

No single method consistently outperformed others for the key estimand, under either truth context; instead, performance varied by scenario and method. Among approaches using external evidence, M-spline MPES and piecewise methods generally produced lower bias than the reference model. Other methods, such the expert opinion approach, showed inconsistent improvements and occasionally greater bias. Key drivers of bias included survival magnitude, hazard profile, and the similarity of the trial and external populations. Additional findings showed that some methods took a long time to execute, and the blended survival method may generate logic errors (survival increasing between two time points).

Future research could expand on the NCS by simulating RCT data to evaluate methods for multiple treatment arms, exploring complex data-generating mechanisms such as treatment switching and subsequent therapies, and addressing distinctions between external data and external information in model fitting. Ultimately, based on the findings of the NCS, it is clear that a detailed assessment of the theoretical strengths and weaknesses of possible candidate methods is critical if external evidence is intended to be leveraged for HTA decision-making.

6.1 Introduction

In Chapter 5, the identified methodological approaches to leverage external evidence for survival extrapolations were applied to case studies, to better understand how they can be used, and how they might be compared. While helpful to understand some practical implications for how the methods could (or should) be used, including where methods may result in inappropriate extrapolations of survival, applying methods only in the context of case studies does not provide a holistic view of how some methods appear to perform in general, including how they might compare to alternative approaches (be those that leverage external evidence in another way, or compared to methods that do not consider external evidence as part of the model fitting process).

Traditionally, when authors propose a new statistical method (related to survival analysis or some other context), it is common to see the performance of the method demonstrated via a simulation study. Morris *et al.* describe simulation studies as computer experiments that involve creating data by pseudo-random sampling from known probability distributions.²⁶⁰ Hoaglin & Andrews highlight the parallels of simulation studies for statistics with experiments in other sciences, noting the need to carefully consider the experimental setting, measuring equipment, and how to analyse results.²⁶¹ Ultimately, the goal of executing a simulation study is to quantitatively assess the performance of a given statistical model under controlled conditions (i.e., by making use of synthetic data), as it is not feasible to assess performance using ‘real’ data.

By assessing the performance of methods to extrapolate estimates of survival for HTA decision making, evidence can be generated to help determine the case for or against using specific methods under certain circumstances. Without an assessment of the performance of a given method, analysts would otherwise need to rely on *expected* performance based on real-world case studies. Real-world case studies will naturally be limited by real-world problems, such as relying on data from studies with limited long-term follow-up and restricted sample sizes. By undertaking a simulation study, the performance of methods can be objectively assessed since the ‘truth’ is known (because it is generated as part of the study).

Despite their value, some simulation studies have been criticised due to issues in their design, execution, and interpretation.^{260, 262-264} Boulesteix *et al.* have previously described issues with method developers being the same researchers that design and execute a comparison study (using simulation techniques) to assess the performance of their new method versus established alternatives.²⁶² Based on a survey by Boulesteix and colleagues, most comparison study papers which were included as part of proposing a new method identified the new method as a ‘winner’. A ‘winner’ means that the new method provides some kind of advantage over existing methods, be that related to performance, efficiency, or some other metric defined

by the authors. However, in pure comparison studies where no new method was being simultaneously proposed, Boulesteix *et al.* found that a winner (or winners) may not always be identified.

For a comparison study to be considered ‘neutral’, Boulesteix *et al.* define three criteria:

1 The focus of the study is the comparison itself

- This means no new method should be proposed as part of the study (i.e., a method is presented with a corresponding simulation study to demonstrate its performance). Such simulation studies may either implicitly or explicitly be biased in favour of the new method (e.g., by only considering scenarios where the new method would be expected to have some type of advantage over conventional methods).

2 The author(s) should be ‘reasonably neutral’

- While there is no definitive way of determining ‘reasonable neutrality’, As an example, Boulesteix *et al.* suggest that an author that published a new method six months prior is unlikely to be ‘neutral’.

3 The evaluation criteria, methods, and data sets should be chosen in a rational way

- **Evaluation criteria:** In addition to simple evaluation criteria (e.g., related to error), it may also be helpful to consider computational aspects (e.g., computation time) when considering the overall goal of the comparison to assess the usefulness of a method in practice.
- **Methods:** Methods should be representative of current practice, relevant to the research context, and selected based on a transparent, well-justified rationale.
- **Data sets:** Boulesteix *et al.* note that in their survey, a median of five data sets were considered in each comparison study – this relatively small number of data sets may not be sufficient to cover (to the extent that is feasible) the full range of datasets where the method could be applied.

The above criteria were deemed to be met, and so an NCS was considered feasible:

- The focus of my comparison is specifically to compare methods, not to present a given method and then demonstrate its performance via a simulation study.
- I have not developed a new method, and so I can be considered ‘reasonably neutral’.
- The evaluation criteria and number of data sets* were defined or chosen in a rational way (see Section 6.2.1).
- The methods were identified via an SLR (see Chapter 3), and then some practical reasons for excluding some methods were presented (see Section 6.2), and so the selection of methods should be reasonably ‘neutral’. While it could be argued that some of the included methods do not appear to be frequently used in current practice (see Chapter 4), it was still considered important to include them to determine if they could be better utilised in HTA.

* Boulesteix *et al.* use the term ‘data sets’ to describe the data used to inform comparison studies performed using either real data or simulated data.

In this chapter, an NCS is described and undertaken, in keeping with the criteria above. The third criterion is covered in detail as part of the Methods section, which describes the design of the NCS. Ultimately, the goal of the NCS was to objectively compare methods to leverage external evidence via a range of HTA-relevant scenarios.

6.2 Methods

In TSD 21, the authors report findings from a simulation study which was conducted to assess the performance of different survival extrapolation methods that could be used to inform technology appraisals.⁵² The TSD 21 simulation study considers the same target audience as the intended audience of the NCS – that is, analysts and researchers involved in producing survival estimates for HTA. Therefore, this previous simulation study serves as a useful foundation from which to develop the NCS.

Separate to the TSD 21 simulation study, guidance for the development of simulation studies in general was proposed by Morris *et al.*, (2019). One of the co-authors of TSD 21 is also a co-author of the simulation study tutorial by Morris *et al.*, (2019). It is therefore anticipated that the principles outlined by Morris *et al.* were adhered to as part of the design of the TSD 21 simulation study, further reinforcing its suitability as a foundation from which to develop the NCS. More broadly, the NCS was conducted in keeping with the principles set out by Boulesteix and colleagues, both in their original commentary paper published in 2013²⁶², and in a 2024 editorial for the special collection “Towards neutral comparison studies in methodological research”.²⁶⁵

6.2.1 ADEMP

Following guidance by Morris *et al.*, the ADEMP structure was followed in the design of the NCS.²⁶⁰ The ADEMP structure aims to clearly describe the design of a simulation study, by drawing attention to five key elements, which are described in turn below.

A: Aim(s)

To investigate the performance of different survival extrapolation methods using the same evidence base for a single treatment group, including at least one source of external evidence.

D: Data generating mechanisms (DGMs)

The data generating mechanisms (DGMs) are based on a combination of some of those used to inform the TSD 21 simulation study, and some that are relevant to consider given that the NCS leverages external evidence. In TSD 21, the following DGMs were used, giving a total of 32 scenarios:

- Hazard profile: (i) Weibull decreasing hazards, (ii) Weibull increasing hazards, (iii) mixture-Weibull (high initial hazard that decreases and then slowly decreases over time), and (iv) cure (with Weibull distribution for uncured people).
- Sample size: (i) small (n=100), and (ii) medium (n=500).
- Survival magnitude: (i) low, and (ii) medium.
- Unmeasured heterogeneity: (i) low, and (ii) medium.

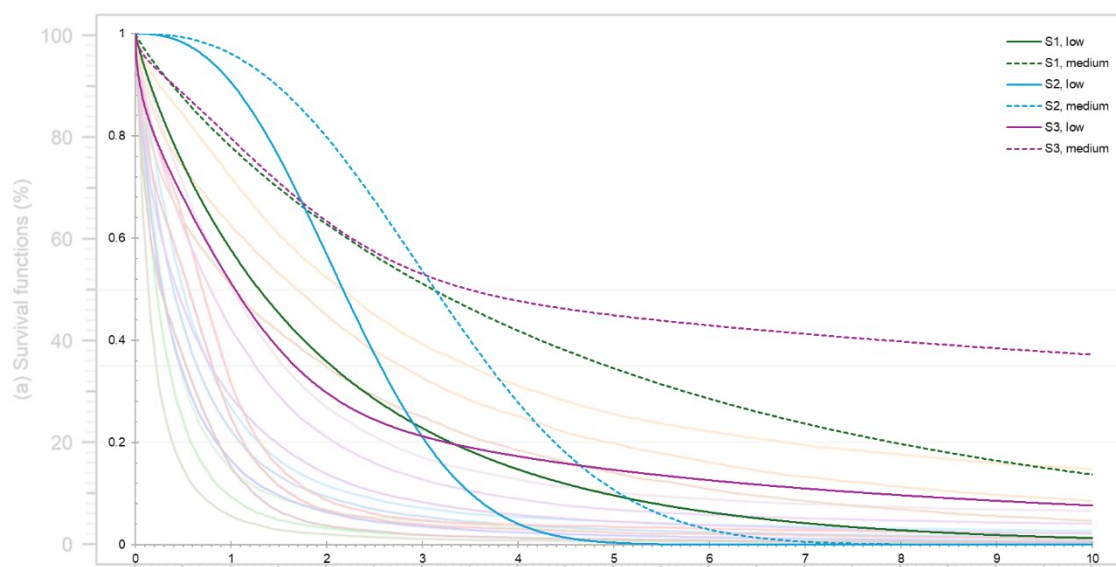
While not explicitly stated by the authors, it is expected that the hazard profiles used in the TSD 21 simulation study were intended to represent a range of plausible hazard profiles seen in typical HTAs for disease-specific mortality. The fourth hazard profile considered in TSD 21 included a 'cured' proportion of patients. This hazard profile was not considered in the NCS because the included methods were not explicitly designed to assess this type of hazard profile. For example, the M-spline MPES method by Jackson (2023)²¹⁶ includes the ability to specify a cure fraction, but the other methods included do not allow for this to be enabled or disabled. Evaluation of methods to estimate survival within the context of a potentially curative intervention was therefore deemed outside the scope of this study.

Before the final set of DGMs was determined, the suitability of the remaining three hazard profiles used in TSD 21 were compared to long-term registry survival data from five different cancers[†], based on a study by Gray *et al.*, (2021).²⁶⁶ This study presented survival data for a range of different cancer types that have each been the topic of previous NICE appraisals (see Chapter 4[‡]). Comparing estimates of hazards from real-world data to parametric models can be challenging, as hazards are not observed. In the study by Gray *et al.*, hazard estimates were generated using the 'muhaz' R package.²⁶⁷ Therefore, as a preliminary first step, the estimates of survival were first compared, as shown in Figure 6-1.

[†] Gray *et al.* considered five different cancer types: breast cancer, colorectal cancer, small-cell lung cancer (SCLC), non-small-cell lung cancer (NSCLC), and pancreatic cancer.

[‡] The STA review covers each of these cancer types, except pancreatic cancer for which there has not been a recent STA, but there have been past appraisals of treatments for pancreatic cancer, such as TA440 of pegylated liposomal irinotecan for treating pancreatic cancer after gemcitabine (<https://www.nice.org.uk/guidance/ta440>).

Figure 6-1: Overlay of TSD 21 simulation study and Gray *et al.*, (2021) survival estimates



Key: TSD, technical support document; S, scenario.

Note: Original plot taken from Gray *et al.*, (2021) – Figure 1, panel A. © The Author(s) 2021. For ease of comparisons, the original figure was adjusted for transparency, and a new plot was overlayed. The figure is reproduced here in accordance with Sage journals pre-approved permission requests, details of which can be found here: <https://uk.sagepub.com/en-gb/eur/pre-approved-permission-requests-journals>. In the original plot from Gray *et al.*, a total of 15 different survival estimates were presented, corresponding to five different cancer types (breast cancer, colorectal cancer, small-cell lung cancer (SCLC), non-small-cell lung cancer (NSCLC), and pancreatic cancer); with each cancer type split into three different age groupings (18-59 years, 60-69 years, and 70+ years).

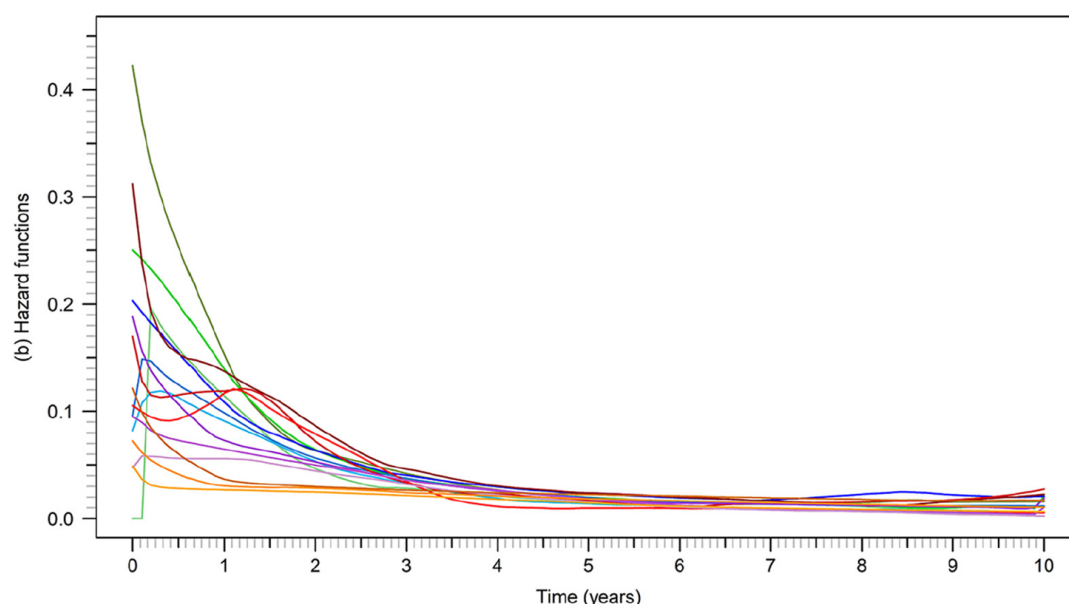
Based on Figure 6-1, compared to the real-world populations from Gray *et al.*, (2021), the TSD 21 simulation study profiles appeared to generally estimate larger estimates of median survival (consistently greater than 1 year), whereas all estimates from Gray *et al.* except for the estimates for breast cancer (orange lines) show median survival less than 1 year. It should be noted that Gray *et al.* report data from the SEER program, which is a population-based registry of cancer patients managed by the National Cancer Institute in the US.^{86, 266}

In NICE STAs of cancer treatments, median survival may not have been reached for the pivotal clinical trial at the time of submission. Through plotting the TSD 21 hazard profiles (or ‘scenarios’) on the same axes, it was noted that differences in survival magnitude were not consistent across the different profiles. For example, the ‘medium’ survival magnitude for hazard profile/scenario 2 (Weibull increasing hazards) yielded an estimated restricted mean survival time (RMST) at 10 years of approximately 3.2 years, whereas for hazard profile/scenario 3 (mixture-Weibull [high initial hazard that decreases and then stabilises]) the corresponding estimate was approximately 5.2 years – a difference of approximately 2 years.[§]

[§] RMST is defined as the area under the survival curve evaluated from baseline (i.e., time zero) until a defined cut-point (in this case, 10 years).

It can also be seen from Figure 6-1 that for hazard profile/scenario 2 (Weibull increasing hazards), survival estimates were effectively zero by around 5 or 8 years (depending on the magnitude of survival used). This pattern of survival was not observed in any of the real-world populations considered by Gray *et al.*, and hazards appeared to reduce in the long term, regardless of the type of cancer. This can be seen in the plot of hazards produced by Gray *et al.*, which is re-produced in Figure 6-2.

Figure 6-2: Gray *et al.*, (2021) hazard estimates



Note: Plot taken from Gray *et al.*, (2021) – Figure 1, panel B. © The Author(s) 2021. The figure is re-produced here in accordance with Sage journals pre-approved permission requests, details of which can be found here: <https://uk.sagepub.com/en-gb/eur/pre-approved-permission-requests-journals>.

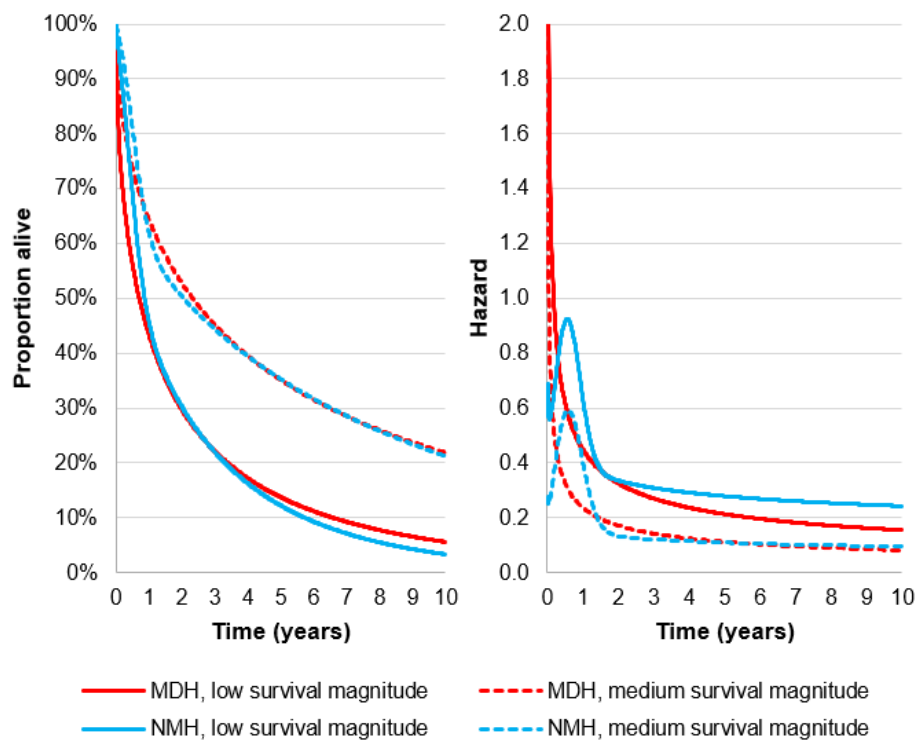
Based on Figure 6-1 and Figure 6-2, it was determined that the hazard profiles used in the TSD 21 simulation study may not be reflective of the expected hazard profiles for real-world populations with cancer, and in particular the types of hazard profiles where long-term survival extrapolations are required for HTA. Despite this, it should be acknowledged that clinical trial populations may (by design) reflect populations that have a better prognosis than an average patient treated in real-world clinical practice, owing to inclusion/exclusion criteria applied in the context of a clinical trial. For the NCS, the following changes were made to the DGMs related to survival profiles:

- The monotonically increasing hazard profile was omitted, since it is unlikely to represent the hazard profile of cancer populations that are the topic of NICE STAs (and especially those where external evidence would be expected to play a role in estimating long-term survival). The remaining hazard profiles were defined as follows:
 - Monotonically decreasing hazards (based on a Weibull model).

- Non-monotonic hazards (based on a mixture-Weibull model).
- To aid with interpretation, the parameters for the remaining survival profiles were re-calibrated such that 10-year RMST was identical for scenarios that shared the same survival magnitude DGM. By re-calibrating the parameters in this way, the concepts of survival magnitude and hazard profile can be more easily distinguished. Magnitudes of survival were quantified as follows:
 - Low: 10-year RMST = 2 years.
 - Medium: 10-year RMST = 4 years.

The resultant survival profiles considered in the NCS are presented in Figure 6-3 (left panel: survival; right panel: hazards). The specific parameters used for these survival profiles are presented in Table 6-1.

Figure 6-3: Hazard profiles used in the NCS (up to 10 years) for disease-specific mortality



Key: MDH, monotonically decreasing hazards; NCS, neutral comparison study; NMH, non-monotonic hazards.
Note: The survival proportions presented in the left panel do not take into consideration background mortality.

Table 6-1: Parameters for hazard profiles used in NCS

Profile label	λ_1	γ_1	λ_2	γ_2	ρ
MDH, low survival magnitude	0.835*	0.538	-	-	1.000
MDH, medium survival magnitude	0.440*		-	-	1.000
NMH, low survival magnitude	2.467*	2.000	0.481*	0.800	0.300
NMH, medium survival magnitude	1.981*		0.189*		0.300

Key: MDH, monotonically decreasing hazards; NCS, neutral comparison study; NMH, non-monotonic hazards.
Note: Greek letters refer to the scale (λ), shape (γ), and mixture (ρ) parameters of the mixture-Weibull model.

When considering the hazard profiles used in the NCS, it is important to note that the ‘true’ values of RMST evaluated at 10 years are slightly lower than 2 years and 4 years, as the NCS accounts for other-cause mortality. The hazard of all-cause mortality comprises the sum of disease-specific and other-cause hazards. Incorporation of other-cause mortality was consistent with the approach taken in the TSD 21 simulation study – namely, using a Gompertz distribution with shape $\lambda = 0.000028$, and scale $\gamma = 0.0936$. As described by the authors of TSD 21:

“These shape and scale parameters were obtained from fitting a Gompertz distribution to English mortality rates in 2009 in females.” (Rutherford *et al.*, 2020, p.56).⁵²

It should be noted that in HTA practice, there are a multitude of real-world and clinical trial design considerations that can influence how reasonable a given survival extrapolation appears for decision-making purposes. These include, but are not limited to, treatment switching (and associated adjustment methods), the use of subsequent therapies (which may or may not reflect routine clinical practice), and the need for population adjustment (e.g., using a MAIC or another indirect comparison method). While these factors are clearly relevant in real-world HTAs, they are not explicitly represented in the NCS.

Attempting to emulate these kinds of complexities in the NCS would not only be practically challenging but would also introduce additional assumptions for data generation and model fitting. These, in turn, would also obscure the key aim of the NCS, which was to assess the performance of different survival extrapolation approaches that leverage external evidence under controlled conditions. Including complex features such as treatment switching or subsequent treatment pathways would risk “muddying the waters”, making it harder to attribute differences in the performance of each method to the modelling approach itself rather than to differences in how these external factors were specified. It is also worth noting that the TSD 21 simulation study also did not attempt to emulate these features.

The aim of the NCS was not to reproduce all features of applied HTA decision problems, but to provide a clear and interpretable assessment of the performance of survival models that leverage external evidence. Consequently, it was considered necessary for the NCS to adopt a simplified and transparent approach to data generation that does not address all possible issues that may influence the determination of a most suitable survival curve for HTA decision-making processes.

The TSD 21 simulation study also included the sample size of the study, by considering sample sizes of $n=100$ (a ‘small’ sized study), or $n=500$ (a ‘medium’ sized study). This DGM was retained as the sample size may be relevant for consideration when determining the potential pros and cons of using methods that leverage external evidence.

Separate to hazard profiles, survival magnitude, and sample size, the TSD 21 simulation study also included the concept of unmeasured heterogeneity in individual survival functions, or 'frailty'. In TSD 21, scenarios were run incorporating small and large amounts of unmeasured heterogeneity. In the interest of restricting the number of DGMs involved in the NCS, and acknowledging that the concept of external evidence similarity is something that needs to be captured separately, this additional frailty DGM was excluded from the NCS.

Further DGMs were also considered, based on the aim of the NCS. Firstly, as noted above in relation to the frailty DGM, the NCS needed to consider how similar the external evidence was to the trial data, and how this might impact the performance of specific methods. To do this, two different DGMs were explored. The first DGM considered the situation where the external evidence comes from a population that is indistinguishable from the clinical trial, such that prognosis is identical. This could represent, for example, data collected in another clinical trial. The other DGM considered a likely more realistic situation: the external evidence comes from a population with a generally poorer prognosis compared to the clinical trial population.

It should be acknowledged that in previous NICE STAs, some companies have opted to adjust background mortality data using an SMR to reflect differences in expected long-term survival for people who may be considered 'cured' versus the age- and sex-adjusted general population. Considering the specification of an SMR within the design of the NCS introduces a number of issues:

- When producing estimates of survival involving an SMR, this would typically be treated as an input to the survival model. For instance, in TA677 the company applied an SMR of 1.09 to adjust the trajectory of long-term survivors within its cost-effectiveness analysis (i.e., the value of 1.09 was an input used directly in the calculations that determined the resultant survival curve).²⁴⁸ Including an SMR in the data generation for the NCS may, to an extent, be seen as somewhat redundant, given that this would require specifying an SMR for both the data generation and model fitting steps.
- Incorporating an SMR into the design of the NCS would add a further layer of complexity in the interpretation of the results. In principle, there could be scenarios undertaken with some kind of imbalance between the background mortality data used to inform the model (e.g., adjusted using an SMR), and the background mortality data assumed by the model (e.g., a relative survival model using unadjusted life table data). While potentially of interest, this highlights a broader issue of external evidence compatibility, which is already addressed in the NCS in a simpler way (i.e., considering similar versus poorer prognosis for the external population, versus the trial population).

Based on the above, adjustment of background mortality via an SMR was not considered in the design of the NCS.

In addition to external evidence similarity, scenarios were also considered that varied the duration of clinical trial follow-up. Two DGMs were considered: 'short' follow-up (a maximum follow-up of 2 years), and 'long' follow-up (a maximum follow-up of 5 years). The concept of trial follow-up was considered explicitly in order to better understand possible differences in method performance depending on when the methods are applied in relation to the overall extent of maturity of clinical trial data. This was acknowledged to likely be an important consideration of method performance as part of the method testing (see Chapter 5). It should be noted, however, that the maximum follow-up for the external population was considered fixed at 10 years, across all scenarios. This was a pragmatic choice to reflect a follow-up duration that was notably longer than the maximum trial follow-up (set to either 2 or 5 years).

While the DGM related to the duration of clinical trial follow-up covers the concept of administrative censoring, it is also worth noting that aligned with the TSD 21 simulation study, the data generating process also accounted for some random censoring. Random censoring times were drawn from an exponential distribution with rate $\lambda = 0.1$, across all scenarios.

In total, 32 unique scenarios were considered in the NCS, adopting a full factorial design:

- Sample size: small ($n=100$) or medium ($n=500$).
- Clinical trial data follow-up: short (2 years) or long (5 years).
- Hazard profile: monotonically decreasing hazards (Weibull model) or non-monotonic (increasing then decreasing) hazards (mixture-Weibull model).
- Survival magnitude: low (10-year RMST: 2 years) or medium (10-year RMST: 4 years).
- External evidence similarity: identical prognosis or poorer prognosis.

This design is presented in tabular format in Table 6-2 for ease of reference.

Table 6-2: Tabulated summary of scenarios in the neutral comparison study

#	Sample size	Maximum follow-up	Hazard profile*	Survival magnitude	External evidence similarity	Description (for reference)
1	100	2 years	Monotonically decreasing	Low	Identical prognosis	n=100; fol=2yrs; haz=mon-dec; mag=low; ext=same
2					Poorer prognosis	n=100; fol=2yrs; haz=mon-dec; mag=low; ext=poorer
3				Medium	Identical prognosis	n=100; fol=2yrs; haz=mon-dec; mag=med; ext=same
4					Poorer prognosis	n=100; fol=2yrs; haz=mon-dec; mag=med; ext=poorer
5			Non-monotonic (increasing then decreasing)	Low	Identical prognosis	n=100; fol=2yrs; haz=non-mon; mag=low; ext=same
6					Poorer prognosis	n=100; fol=2yrs; haz=non-mon; mag=low; ext=poorer
7				Medium	Identical prognosis	n=100; fol=2yrs; haz=non-mon; mag=med; ext=same
8					Poorer prognosis	n=100; fol=2yrs; haz=non-mon; mag=med; ext=poorer
9		5 years	Monotonically decreasing	Low	Identical prognosis	n=100; fol=5yrs; haz=mon-dec; mag=low; ext=same
10					Poorer prognosis	n=100; fol=5yrs; haz=mon-dec; mag=low; ext=poorer
11				Medium	Identical prognosis	n=100; fol=5yrs; haz=mon-dec; mag=med; ext=same
12					Poorer prognosis	n=100; fol=5yrs; haz=mon-dec; mag=med; ext=poorer
13			Non-monotonic (increasing then decreasing)	Low	Identical prognosis	n=100; fol=5yrs; haz=non-mon; mag=low; ext=same
14					Poorer prognosis	n=100; fol=5yrs; haz=non-mon; mag=low; ext=poorer
15				Medium	Identical prognosis	n=100; fol=5yrs; haz=non-mon; mag=med; ext=same
16					Poorer prognosis	n=100; fol=5yrs; haz=non-mon; mag=med; ext=poorer
17	500	2 years	Monotonically decreasing	Low	Identical prognosis	n=500; fol=2yrs; haz=mon-dec; mag=low; ext=same
18					Poorer prognosis	n=500; fol=2yrs; haz=mon-dec; mag=low; ext=poorer
19				Medium	Identical prognosis	n=500; fol=2yrs; haz=mon-dec; mag=med; ext=same
20					Poorer prognosis	n=500; fol=2yrs; haz=mon-dec; mag=med; ext=poorer
21			Non-monotonic (increasing then decreasing)	Low	Identical prognosis	n=500; fol=2yrs; haz=non-mon; mag=low; ext=same
22					Poorer prognosis	n=500; fol=2yrs; haz=non-mon; mag=low; ext=poorer
23				Medium	Identical prognosis	n=500; fol=2yrs; haz=non-mon; mag=med; ext=same
24					Poorer prognosis	n=500; fol=2yrs; haz=non-mon; mag=med; ext=poorer
25		5 years	Monotonically decreasing	Low	Identical prognosis	n=500; fol=5yrs; haz=mon-dec; mag=low; ext=same
26					Poorer prognosis	n=500; fol=5yrs; haz=mon-dec; mag=low; ext=poorer
27				Medium	Identical prognosis	n=500; fol=5yrs; haz=mon-dec; mag=med; ext=same
28					Poorer prognosis	n=500; fol=5yrs; haz=mon-dec; mag=med; ext=poorer
29			Non-monotonic (increasing then decreasing)	Low	Identical prognosis	n=500; fol=5yrs; haz=non-mon; mag=low; ext=same
30					Poorer prognosis	n=500; fol=5yrs; haz=non-mon; mag=low; ext=poorer
31				Medium	Identical prognosis	n=500; fol=5yrs; haz=non-mon; mag=med; ext=same
32					Poorer prognosis	n=500; fol=5yrs; haz=non-mon; mag=med; ext=poorer

Key: ext, external evidence similarity; fol, follow-up; haz; hazard profile; mag, survival magnitude; med, medium; mod-dec, monotonically decreasing hazards; n, sample size; non-mon, non-monotonic (increasing then decreasing hazards).

Note: *Parameters for the hazard profiles are presented in Table 6-1.

The sample size of the external data set was always assumed to be ten times the size of the trial, based on the expectation that larger external data sets would likely be available for clinical trials that were able to recruit larger numbers of participants, versus smaller clinical trials of potentially less common cancers. This was not varied as a separate DGM for pragmatic reasons (since incorporating this additional DGM would double the number of scenarios). This essentially assumes that the external data comes from a disease registry, or a large clinical trial, and not (for example) a small, earlier phase study.

To summarise the process for the simulations described so far:

- First, sample three groups of survival times per scenario and per simulation: (i) the trial data sample, (ii) the external data sample, and (iii) the truth (noting the latter is based on a very large sample size).
- Next, fit each of the included models to the trial $\pm$ external data, and record the necessary metrics to assess performance (discussed later in this sub-section).
- Then, compare these metrics to the 'truth'.
- Finally, average the results over the number of simulations and present other metrics as per the performance measures of interest (again, discussed later in this sub-section).

Assuming it was appropriate to only have 1 simulation, the above would mean a total of:

- 32 trial data sets
- 32 external data sets
- 8 truth data sets (as there is some duplication across the 32 scenarios, which is clarified further in Table 6-3).

However, this total number of simulated data sets ($32 + 32 + 8 = 72$) increases markedly when taking into account the fact that a sufficiently large number of simulations is required in order for the results of the NCS to be meaningful. To determine the total number of simulations, it is useful to consider the role of Monte Carlo Standard Error (MCSE). This can be extracted alongside other performance measures to check that the total number of simulations is sufficient to yield broadly reliable results. However, there is no formal 'rule' to determine if enough simulations have been completed. Determining the optimal number of simulations is especially challenging in the context of this NCS because it is not possible to output a confidence interval for all methods explored – some methods are frequentist, others are Bayesian, and some are a combination of approaches that ultimately do not yield a confidence interval or other formalised measure of uncertainty.

Therefore, for the purpose of this NCS, MCSE was extracted and used in the following way:

- MCSE for the main estimand of interest (see next sub-section for details) was compared to the true value.
- If the ratio was $\leq 5\%$, this would suggest that the total number of simulations ($n=1,000$) is sufficient; however, if the ratio was $>5\%$, this would suggest that the total number of simulations ($n=1,000$) is not sufficient, and that further simulations would be required to produce stable results.
- It was expected that if the base number of simulations (1,000) was insufficient, then doubling this (to 2,000) would likely address the issue. However, a decision would need to be made based on how close the ratio was to 5% (for example, if the ratio was 5.1%, it may not be necessary to run an additional 1,000 simulations, whereas if the ratio was much larger than 5%, then a further 1,000 may still not be sufficient).

The purpose of this ‘rule of thumb’ was to ensure that the MCSE (which measures the uncertainty due to the use of sampling in the NCS) is reasonably small versus the true value of the quantity being estimated. Notably, the choice of 5% is arbitrary, but provides a pragmatic balance between efficiency and precision – any further reductions in MCSE (relative to the true value) would be expected to yield diminishing returns in terms of overall accuracy, and it is unlikely that any additional stability in results would affect any key conclusions reached by the NCS whilst running times would be increased. Nevertheless, it should be noted that if several methods exhibit bias differences smaller than MCSE, it may not be possible to reliably rank their performance, as differences in estimated bias will reflect a combination of true differences and random error. This could occur even when the MCSE is less than 5% of the true value, underscoring the importance of reporting the MCSE alongside performance metrics.

E: Estimand(s)

The primary estimand is difference in mean LYs between the survival model and the ‘true’ data (derived using a KM estimate). This was chosen to align with the main outcome of interest, related to survival, in the context of an HTA (wherein estimates of mean survival are used to determine total QALYs, and by extension, the ICER). In the TSD 21 simulation study, estimates of RMST were taken based on two timepoints:

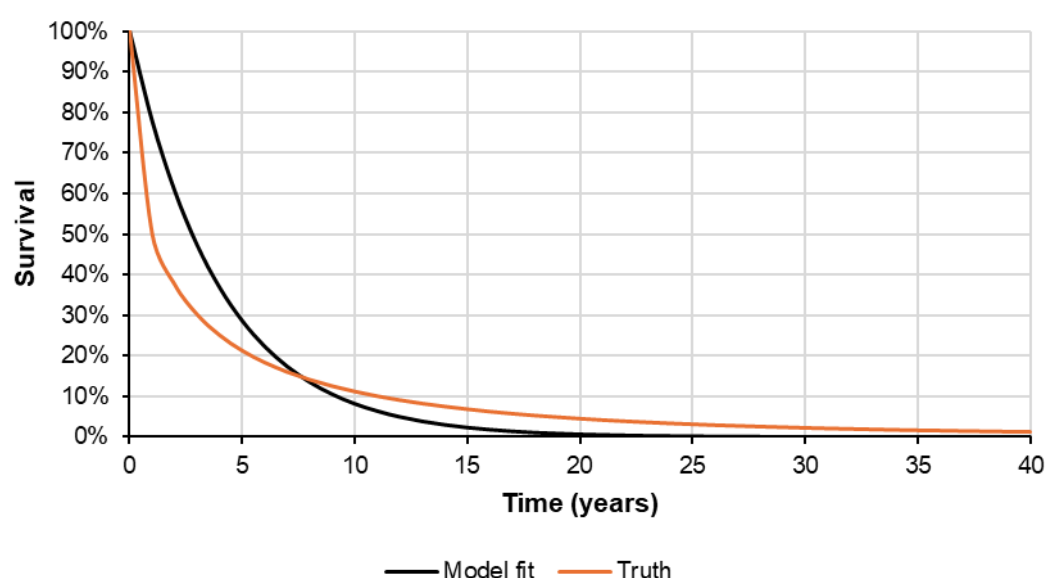
- A short-term extrapolation (assumed to be 3 years in TSD 21).
- A lifetime extrapolation (assumed to be 99 years in TSD 21).

For the NCS, this set of timepoints was expanded to cover a broader range of potentially relevant time points: 2, 3, 5, 10, and 40 years (lifetime). ‘Lifetime’ was determined as 100 years minus the average age of the simulated cohort at baseline (set to 60 years, consistent with TSD 21), hence this was 40 years. The earlier milestones of 2, 3, 5, and 10 years were chosen

as a combination of time points aligned to the DGMs (2, 5, and 10 years) and the original estimand in TSD 21 (3 years). However, the main estimand of interest was the difference in mean LYs between the ‘truth’ and the fitted model over a lifetime horizon (or in other words, RMST evaluated over a lifetime horizon).

In addition to RMST, point estimates of survival were also extracted at the same time points. A comparison of point estimate accuracy, particularly at later time points, was considered to possibly help further understand the overall accuracy of some the methods which RMST could mask. As an example, consider Figure 6-4 – these two survival curves yield identical RMST values when evaluated over a 40-year (lifetime) horizon, but evidently the model does not provide a good visual fit to the ‘truth’.

Figure 6-4: Illustration of equal 40-year RMST but poor fit



Key: RMST, restricted mean survival time.

Defining ‘the truth’

An important concept relevant to this NCS specifically is how to define the ‘truth’, as there are theoretically multiple options that could be considered. The first option is to consider the trial population as the ‘truth’, based on convention in HTA that survival models are fitted to the trial data and this population is assumed to be generalisable to the real-world population. Another option is to instead consider that the external evidence population represents the ‘truth’, as this population may better reflect real-world outcomes experienced by patients outside the context of the clinical trial.

There is also a third option – something else entirely. For example, in the context of an HTA, it may be the case that the trial population is considered to likely have better survival outcomes

versus the population that would be treated in practice, due to trial inclusion/exclusion criteria. It may also be considered that any external data could reflect poorer outcomes than those that would be expected for the population that would be treated in practice – this could be because of standard of care improving over time, despite the same active treatment(s) being used. Consequently, one could argue that the ‘truth’ is neither the trial nor the external population, but something different (though likely ‘similar’ to both populations).

In the context of the NCS, it is not feasible to consider this third option for the ‘truth’ because this would require simulating data which is in no way connected to the data used in the model fitting process, and so any performance measures would be directly linked to decisions made regarding how the ‘truth’ is simulated. As such, only the first two options (trial as truth and external as truth) were deemed suitable for consideration in the NCS.

Given that both options may be of interest, parallel sets of results were prepared considering each definition of what comprises the ‘truth’. In scenarios where the external population and the trial population are identical, results are expected to be indistinguishable (i.e., no differences, other than small differences due to random number generation and sampling). For both versions of the truth, the truth was determined based on simulating 10 million patients with no restriction on follow-up (except to not predict any patient to live longer than the maximum time horizon of 40 years). When the trial population was considered the ‘truth’, 10 million patient survival times were estimated using the same parameters as used to sample the trial population survival; and all relevant summary estimates were prepared (RMST and point estimates of survival, at all the previously highlighted time points). When the external population was considered the ‘truth’, the same process was followed, but using the parameters used to sample the external population survival, instead of trial population survival.

To aid with interpretation of findings, a ‘truth lookup’ was created, presented in Table 6-3. Using this table, it can be seen which scenarios are intended to be evaluated against the same ‘truth’. For example, scenario 1 always shares the same ‘truth’ as scenario 17 – the only thing that differs between these scenarios is the sample size of the clinical trial.

Table 6-3: Truth lookup for scenarios in the neutral comparison study

#	Description (for reference)	Truth (as trial)*				Truth (as external)†							
		1	2	3	4	1	2	3	4	5	6	7	8
1	n=100; fol=2yrs; haz=mon-dec; mag=low; ext=same	✓				✓							
2	n=100; fol=2yrs; haz=mon-dec; mag=low; ext=poorer						✓						
3	n=100; fol=2yrs; haz=mon-dec; mag=med; ext=same		✓					✓					
4	n=100; fol=2yrs; haz=mon-dec; mag=med; ext=poorer								✓				
5	n=100; fol=2yrs; haz=non-mon; mag=low; ext=same			✓						✓			
6	n=100; fol=2yrs; haz=non-mon; mag=low; ext=poorer										✓		
7	n=100; fol=2yrs; haz=non-mon; mag=med; ext=same				✓							✓	
8	n=100; fol=2yrs; haz=non-mon; mag=med; ext=poorer												✓
9	n=100; fol=5yrs; haz=mon-dec; mag=low; ext=same	✓				✓							
10	n=100; fol=5yrs; haz=mon-dec; mag=low; ext=poorer						✓						
11	n=100; fol=5yrs; haz=mon-dec; mag=med; ext=same		✓					✓					
12	n=100; fol=5yrs; haz=mon-dec; mag=med; ext=poorer								✓				
13	n=100; fol=5yrs; haz=non-mon; mag=low; ext=same			✓						✓			
14	n=100; fol=5yrs; haz=non-mon; mag=low; ext=poorer										✓		
15	n=100; fol=5yrs; haz=non-mon; mag=med; ext=same				✓							✓	
16	n=100; fol=5yrs; haz=non-mon; mag=med; ext=poorer												✓
17	n=500; fol=2yrs; haz=mon-dec; mag=low; ext=same	✓				✓							
18	n=500; fol=2yrs; haz=mon-dec; mag=low; ext=poorer						✓						
19	n=500; fol=2yrs; haz=mon-dec; mag=med; ext=same		✓					✓					
20	n=500; fol=2yrs; haz=mon-dec; mag=med; ext=poorer								✓				
21	n=500; fol=2yrs; haz=non-mon; mag=low; ext=same			✓						✓			
22	n=500; fol=2yrs; haz=non-mon; mag=low; ext=poorer										✓		
23	n=500; fol=2yrs; haz=non-mon; mag=med; ext=same				✓							✓	
24	n=500; fol=2yrs; haz=non-mon; mag=med; ext=poorer												✓
25	n=500; fol=5yrs; haz=mon-dec; mag=low; ext=same	✓				✓							
26	n=500; fol=5yrs; haz=mon-dec; mag=low; ext=poorer						✓						
27	n=500; fol=5yrs; haz=mon-dec; mag=med; ext=same		✓					✓					
28	n=500; fol=5yrs; haz=mon-dec; mag=med; ext=poorer								✓				
29	n=500; fol=5yrs; haz=non-mon; mag=low; ext=same			✓						✓			
30	n=500; fol=5yrs; haz=non-mon; mag=low; ext=poorer										✓		
31	n=500; fol=5yrs; haz=non-mon; mag=med; ext=same				✓							✓	
32	n=500; fol=5yrs; haz=non-mon; mag=med; ext=poorer												✓

Key: ext, external evidence similarity; fol, follow-up; haz; hazard profile; mag, survival magnitude; med, medium; mod-dec, monotonically decreasing hazards; n, sample size; non-mon, non-monotonic (increasing then decreasing hazards).

Note: *When considering the trial data as the truth, there are four different versions of the truth, based on the hazard profile and survival magnitude (i.e., $2 \times 2 = 4$); †When considering the external data as the truth, there are eight different versions of the truth, based on the hazard profile, survival magnitude, and the similarity of the external and trial data (i.e., $2 \times 2 \times 2 = 8$).

M: Method(s)

Seven methods that leverage external evidence were included in the NCS, shown in Table 6-4. In addition to the seven methods presented in Table 6-4, a standard parametric model without any use of external evidence was also included. This allowed for a comparison of all seven methods incorporating external evidence with a standard approach that does not leverage external evidence in any way. Therefore, in total, eight methods were considered in the NCS.

Table 6-4: Summary of retained methods for neutral comparison study

Name of method	Citation	Identification	Page
Relative survival	Lambert <i>et al.</i> , (2007) ¹⁵⁸	SLR (1 st iteration)	90
Standard piecewise	Larkin <i>et al.</i> , (2015) ¹⁴⁶	SLR (1 st iteration)	103
Summary data piecewise	Pennington <i>et al.</i> , (2018) ¹⁷³	SLR (1 st iteration)	117
Expert opinion	Cope <i>et al.</i> , (2019) ⁹¹	SLR (1 st iteration)	120
Informative prior on shape	Soikkeli <i>et al.</i> , (2019) ¹⁹⁹	SLR (2 nd iteration)	135
Blended survival	Che <i>et al.</i> , (2023) ²¹⁴	SLR (2024 update)	151
M-spline MPES	Jackson (2023) ²¹⁶	SLR (2024 update)	157

Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

Importantly, methods were not adjusted in any way before inclusion within the NCS, as doing so would contradict the impartiality of the NCS itself. In other words, the neutrality of the comparison would be compromised if one or more methods were edited to better fit into the NCS. For example, the original MPES by Guyot *et al.* could theoretically be adapted to the single-arm setting (as evidenced in the tutorial paper, see Chapter 5), but this was not the original intention of the method – it was designed in the context of an RCT. The full reasons for discarding the original MPES by Guyot *et al.*, are provided in 5.4.4.

While the relative survival method was taken forward for inclusion in the NCS, it should be noted that this method may not be considered strictly comparable to the other methods because the only external source used is life tables. It could be argued that it is unfair to include this method in the NCS per its original intention (i.e., using background mortality only), because none of the other methods were designed to only use life tables. In addition, van Oostrum and colleagues previously published a paper called: “*Comparison of Parametric Survival Extrapolation Approaches Incorporating General Population Mortality for Adequate Health Technology Assessment of New Oncology Drugs*”.²⁰² In this paper, the authors set out to explicitly compare different methods to incorporate general population mortality into survival models for HTA, including a relative survival approach (but did not attempt to compare these to other methods that could leverage different types of external evidence).

Given that the goal of the NCS was to objectively assess the performance of methods that *can* leverage external evidence, the relative survival method was retained since it could be of

interest to ascertain how well the method performs despite it not making use of any external data that could be considered more relevant to the disease population.

Based on the availability of *R* code for each method, one of three approaches was taken to implement the method used in the NCS. In some cases, the original study authors had already provided *R* code, either through a dedicated package or as supplementary material published alongside the paper. For other methods, while the original publication did not supply code, later work by other researchers had led to the publication of either a tutorial including code or preparation of an *R* package, which were used in lieu of *R* code which had not been provided by the original method authors. Where neither of these options existed, bespoke *R* code was developed from first principles to implement the method. Table 6-5 summarises the approach taken for each method.

Table 6-5: Implementation of retained methods for neutral comparison study

Name of method	Application
Relative survival	Sweeting <i>et al.</i> , (2023) ²²¹ tutorial code
Standard piecewise	Bespoke <i>R</i> code
Summary data piecewise	Bespoke <i>R</i> code
Expert opinion	Cooney & White, (2023) ²⁵³ <i>R</i> package
Informative prior on shape	Bespoke <i>R</i> code
Blended survival	<i>R</i> code provided by original authors
M-spline MPES	<i>R</i> code provided by original authors

Key: MPES, multi-parameter evidence synthesis.

As previously highlighted in Chapter 5, each of the methods require specific model settings (see Table 5-5) and/or translation of external evidence into a format that the method requires (e.g., point estimates of survival at specified time points). Each of these decisions could have an impact on the resultant model fit. In the interest of maintaining neutrality, settings were determined via the following principles:

- If time points were required for piecewise-type approaches, these were selected as the end of trial follow-up, the end of external evidence follow-up (if a second point was required), and the midpoint between these times (if a third point was required). If more than three time points were required, time was divided into annual intervals. It could be argued that more granular inputs could be provided, but annual intervals were considered the most likely HTA-relevant landmarks.
- If a distribution needs to be specified, this was chosen to be Weibull in all cases. A Weibull model formed the basis of the DGMs related to hazard profile for disease-specific mortality. For the mixture-Weibull DGM, it could be argued that a different model should be chosen. Conversely, for piecewise methods, it could also be argued that specifying two (or more) non-monotonic distributions (e.g., log-logistic) would

introduce bias against this method. For both simplicity and consistency across all model fits, the same distribution (Weibull) was selected, including a standard parametric model fitted without using any external evidence.

- If a user-specified SD or SE was required, this was arbitrarily set at a fixed value of 0.1. These inputs would typically be derived using some form of expert elicitation. In the study by Soikkeli *et al.*, (2019), based on information provided in the supplementary appendix, values for the SD were estimated to be in the region of 0.07 for the informative prior, and 0.007 for the strongly informative prior.¹⁹⁹ A value of 0.1 was chosen for pragmatic reasons as a trade-off between two extreme approaches: a much larger value may render the priors being essentially noninformative, whereas a much smaller value may lead to issues in model convergence. In reality, the choice of these values would be based on an assessment of the available evidence, but this is not practically achievable within the context of an NCS, hence an arbitrary choice was made here, in keeping with the values used in the study by Soikkeli *et al.*, (2019).¹⁹⁹

For each method, the resultant survival curve x- and y-coordinates (i.e., a table of survival times and proportions alive) were extracted, and then used to determine model performance.

P: Performance measure(s)

Consistent with the TSD 21 simulation study, the bias of the estimates is of primary interest, and is therefore the key performance measure for the NCS. It should be noted that some of the methods included in the NCS are not single statistical models produced from standard software packages. As such, additional metrics such as standard deviations in estimates of mean survival are not readily producible, which limits the number of performance measures that could be considered.

Bias is defined as the amount by which the estimate ($\hat{\theta}$) is either greater than or less than the true value (θ_{True}):

$$Bias = \hat{\theta} - \theta_{True}$$

Bias can take a positive or negative value – positive values indicate an over-estimate, whereas negative values indicate an under-estimate (on average). A potential limitation in only considering bias as the measure of model performance is that this does not differentiate between over- or under-estimation of survival. In practice, it may be helpful to understand if a given method tends to produce several estimates which are over- and under-estimates, which effectively cancel each other out when considering the mean bias over a number of scenarios. Mean absolute bias (or mean absolute error) was therefore also calculated:

$$Absolute\ bias = |\hat{\theta} - \theta_{True}|$$

A summary of the mean absolute bias was calculated for each combination of method and scenario.

In addition, for ease of interpretation, percentage bias results were also estimated. It can sometimes be difficult to determine whether the mean bias estimates are meaningful (e.g., if the difference between the mean estimated RMST and the true RMST seems important or not). By expressing bias as a percentage, the relative size of the bias can be more easily inferred. Percentage bias (or sometimes termed ‘relative bias’) is calculated as follows:

$$\text{Percentage bias} = \frac{\hat{\theta} - \theta_{True}}{\theta_{True}}$$

Another important consideration when assessing the performance of each method is the variability of estimates across simulations. Even if a method performs well on average (i.e., has low bias), it may still produce widely varying results. This variation can be quantified using the empirical standard error (empSE), which reflects the degree of precision in the model's estimates across simulations. The formula for empSE is provided below, which requires in addition to the measures defined previously, the mean result across all simulations ($\bar{\theta}$), and the total number of simulations (n_{sim}):

$$\text{empSE} = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{\frac{\sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2}{n_{sim} - 1}}$$

A high value of empSE indicates that the method produces estimates that are more variable across repeated samples, suggesting lower precision. Conversely, a low empSE suggests that the method yields more consistent results. A summary of empSE was calculated for each combination of method and scenario. It should be noted that empSE does not vary according to which version of the ‘truth’ is considered.

When considering these performance measures, it may be helpful to consider which version of the ‘truth’ is most closely aligned with the method itself, as this may aid with interpretation of the findings of the NCS (or, in other words, what the most relevant *estimand* is that the method is suited to address). For example, a standard parametric model using no external data would be expected to show lower bias when considering the trial population as the truth, versus the external population as the truth (since this method does not use any external data). For the methods that leverage external evidence in some way, it is less straightforward to anticipate which version of the truth may be more closely aligned.

To aid with interpretation of results, Table 6-6 provides some commentary on which version of the ‘truth’ may be expected to show lower bias than the other for the primary estimand (bias in the difference in mean LYs between the model and the truth), based on rationale linked to

the method itself. However, this commentary should be interpreted with caution – in practice, there are many factors that influence how accurate survival estimates are likely to be, compared to the objective truth (or in this case, truths inferred from different data sources). For example, a standard parametric model may, in principle, be well suited to describing survival for a trial population, but if the trial data are immature, extrapolated estimates of survival may be overly optimistic. If such an overestimate is then compared against an external data source with a poorer prognosis (like in the NCS), it might appear that the standard parametric model provides a better estimate for a different population than for the one from which it was fitted. This apparent paradox does not imply that a given model truly generalises better to a different population. Rather, it reflects the combined influence of multiple factors (such as data maturity or hazard profile) which may, by chance, make the model appear to perform better or worse when compared against an alternative data source.

Table 6-6: Truth considerations for each method

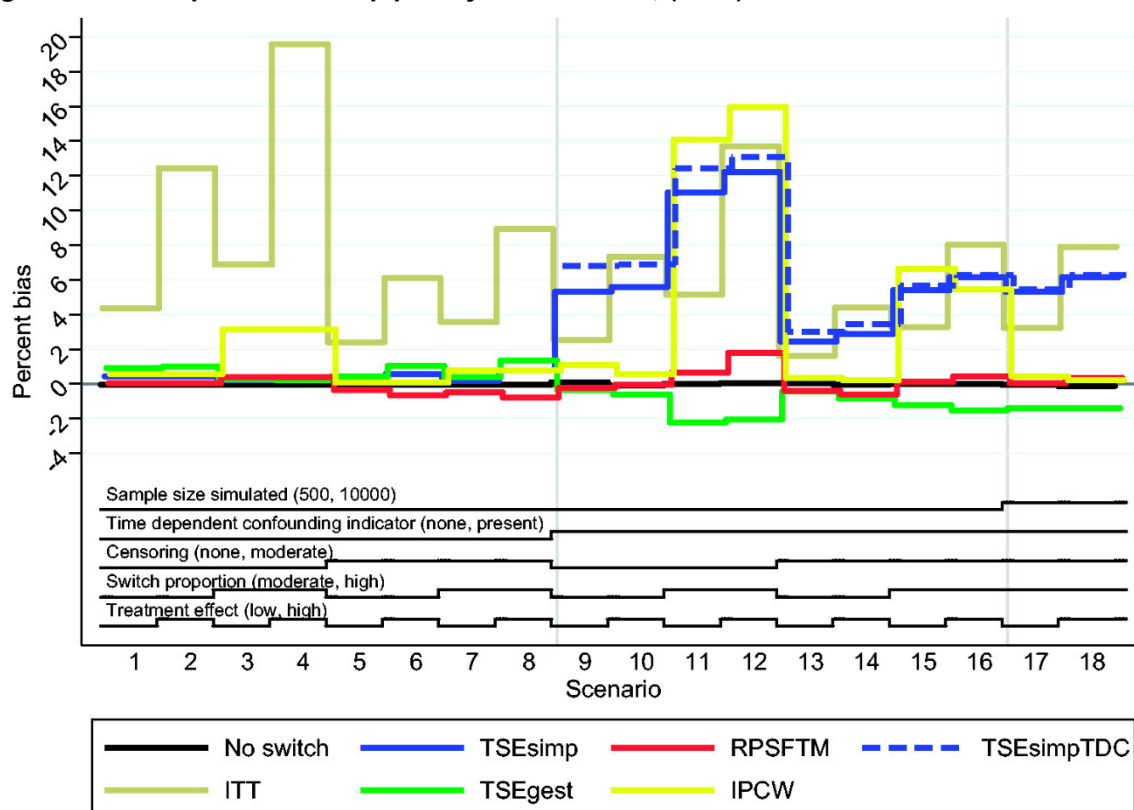
Name of method	Which truth is expected to show lower bias and associated rationale*	
Standard parametric	Trial	No external evidence used, so this method would be expected to show lower bias when the trial population is assumed to represent the truth, versus the external population.
Relative survival	Trial	The only external evidence used to inform this method is life tables (i.e., not the external population data that is generated as part of the NCS). Because of this, it would be expected that this method would show lower bias when the trial population is assumed to represent the truth, versus the external population. However, if the addition of the data from life tables caused survival estimates to decrease, <u>and</u> the external population exhibited lower long-term survival, it may be the case that by chance, assuming the external population represents the truth would lead to lower bias.
Standard piecewise	Unclear	A piecewise model will reflect the trial population exactly up until the specified cut-point. After this, depending on how influential the data for the external population are on the extrapolated tail, it may be the case that bias is lower using either population to represent the truth.
Summary data piecewise		
Expert opinion	External	Subject to the strength of prior specified, the expert opinion model would be expected to closely align with the external population survival at a given landmark. Therefore, for this model, lower bias would be expected when the external population survival is assumed to represent the truth (provided the prior is specified to be strong enough).
Informative prior on shape	Trial	For this model, the <i>shape</i> parameter is informed via the external population. This should mean that the trial population lies closer to the truth, given that the <i>scale</i> parameter is more likely to dictate large changes in the magnitude of survival over time (and this is not directly informed by the external population).
Blended survival	Unclear	Depending on a mixture of survival magnitude, study follow-up, and the specification of the blending interval, the blended survival model may be expected to show lower bias when considering the external population as the truth, versus the trial population. If the blending interval is very short, most of the modelled survival will be based on the external population, and so this population would likely show lower bias when assumed to represent the truth. However, if the blending interval is long (and does not apply until most of the survival curve has been observed), then the trial population may show low bias when this is assumed to represent the truth.
M-spline MPES	Unclear	The underlying specification of this model leads to the production of one overall hazard estimate based on all the inputs provided. It is therefore not possible to infer from the outset which version of the truth is more likely to represent the truth, other than to acknowledge that the model relies on individual-level data from the trial population, and aggregate conditional survival data from the external population. It may therefore be the case that subject to how the model is specified, this method may produce lower estimates of bias when the trial population is assumed to represent the truth (as the model is provided with “more data” from the trial population).

Key: MPES, multi-parameter evidence synthesis; NCS, neutral comparison study.

Note: *In this table, rationale is provided focussing on the primary estimand (bias in mean survival).

The bias-related and empSE performance measures were presented in the form of a nested loop plot (NLP). An NLP is a way to visually present the results of the NCS where multiple DGMs are varied in combination (in this case, five different DGMs which make up the total $n=32$ [2⁵] scenarios). In an NLP, the different levels of each DGM are represented via lines either above or below the main plot, where the lower level represents the first entry in the list for that particular DGM. The x-axis of the main plot shows each individual scenario, and the corresponding levels above the main plot show which combination of DGMs comprise that scenario. The y-axis shows the result of interest (in this case, bias). The term "nested loop" refers to how the simulations are organised: one loop runs through all values of one DGM (e.g., trial sample size), and within these levels, another loop runs through values of the second DGM (e.g., maximum trial follow-up), akin to how loops are used in statistical programming. The plot helps to quickly compare the performance of the different methods across the full range of scenarios. An example of an NLP is presented in Figure 6-5.

Figure 6-5: Example nested loop plot by Latimer *et al.*, (2020)



Key: ITT, intention to treat; TSEsimp, two-stage estimation with simple Weibull model; TSEgest, two-stage estimation with g-estimation; RPSFTM, rank preserving structural failure time model; IPCW, inverse probability of censoring weights; TSEsimpTDC, two-stage estimation with simple Weibull model and time-dependent covariates.

Note: Original plot taken from Latimer *et al.*, (2020) – Figure 5. © The Author(s) 2020. Re-produced here in accordance with the Creative Commons Attribution 4.0 License (CC BY 4.0) (<https://www.creativecommons.org/licenses/by/4.0/>). Latimer N, White I, Tilling K, Siebert U. Improved two-stage estimation to adjust for treatment switching in randomised trials: g-estimation to address time-dependent confounding. *Statistical Methods in Medical Research*. 2020;29(10):2900-2918.

Coverage was not included as a performance metric in the NCS because several of the methods do not output readily interpretable or consistently defined 95% CIs for the primary estimand of interest (mean survival). Methods such as piecewise models consisting of a KM estimate followed by a parametric model present a challenge in terms of how uncertainty should be reflected across the full model. Ultimately, as the main goal of the NCS was not to fully capture uncertainty in survival estimates (as the NCS was focused mostly on bias in mean survival), coverage was not considered further (along with any other performance measures that would require further work to extract additional metrics from the included methods).

Separate from the traditional performance measures described above, three further performance measures were also considered for practical considerations. The first of these was for each method, the proportion of simulations within each scenario in which the method produced an error. This was not captured in the TSD 21 simulation study, most likely due to the consideration of only frequentist methods that were fitted to a single set of data. This is analogous to ‘% convergence’ which has been used in previous simulation studies – for example Alshreef *et al.*, (2025).²⁶⁸ The second was the proportion of model fits for each method that reflected an increase in survival at any point in time, given that this is not biologically plausible (i.e., an increase in survival would mean that people could die more than once). This additional performance measure was considered for completeness, based on one of the findings of Chapter 5 related to the blended survival method, and the MPES tutorial paper also discussed as part of Chapter 5.⁴ Finally, the overall computation time required to execute each method was recorded, to rank the methods in terms of how long they took to execute (on average). This latter performance measure does not capture how well the method estimates survival, but could nevertheless help assess the usefulness of a method were it to be used in practice.

To account for differences in run time between different computers (due to differences in computer processing power based on different hardware), an additional simulation was carried out running only the first 10 iterations (i.e., the first 1%) all using the same computer. For both the full simulation runs (on different computers) and the 1% simulation runs (on the same computer), the rank of the method in terms of run time was determined (i.e., the fastest method was assigned rank 1, and the slowest method was assigned rank 8).

Therefore, in total, the NCS considers the following performance measures:

- In either prose text or tabular format:
 - Frequency of model estimation errors.
 - Frequency of logical errors in model outputs.
 - Time elapsed to generate model estimates.

- As NLPs:
 - Bias.
 - Absolute bias.
 - Percentage bias.
 - EmpSE.

6.2.2 Code and script generation

Given that the TSD 21 simulation study provides analysis code in the public domain, this study formed the basis of the code for the NCS. However, despite being available in the public domain, the script was developed in Stata.¹⁰⁷ As discussed in Chapter 5, all methods that could be considered in the NCS were implemented in one consistent programming language: *R*. Therefore, the first step in the development of the code for the NCS was to translate the core functionality of the TSD 21 simulation study code into an executable *R* script (but not the whole TSD 21 simulation study, given the differences in the aims of this study and the NCS).

For this thesis, it was not necessary to re-produce the model fitting aspects of the TSD 21 simulation study. The models included in the TSD 21 simulation study were fitted using the following Stata packages:

- `stmix`²⁶⁹ (for the mixture models).
- `stpm2`²⁷⁰ (for flexible parametric models).
- `streg`²⁷¹ (for parametric models).
- `strsmix`²⁷² (for a Weibull cure model).

Some of these packages do not have a counterpart in *R* – for example, while mixture models can be fitted via `flexsurv`¹⁰⁸, there is no direct counterpart to `stmix`. Conversely, there is an *R* package called `rstpm2`²⁷³ which was designed as a replication of `stpm2` in Stata. It should be noted, however, that Stata and *R* packages may be maintained by different people, so corresponding packages are not guaranteed to have identical functionality indefinitely. Even for standard parametric models, there are some important differences between `flexsurv` and `streg` – for example, `streg` does not allow fitting of the two-parameter gamma model, whereas this can be executed via `flexsurv`. Therefore, for the reasons highlighted above, the model fitting aspect of the TSD 21 simulation study was not replicated, given that this was not required to address the aims of the NCS. Instead, focus was placed on re-producing the aspects of the simulation code that were intended to be used as a foundation for the NCS (i.e., primarily the data generation processes).

In total, the NCS uses 13 different *R* packages:

- `dplyr`²⁷⁴ for data handling.

- `expertsurv`²⁵³ for fitting the ‘expert opinion’ model.
- `flexsurv`¹⁰⁸ for fitting parametric survival models.
- `ggplot2`²⁷⁵ for plotting results.
- `progress`²⁷⁶ for outputting and updating a progress bar in the *R* console.
- `rsimsum`²⁷⁷ for summarising the outputs of the NCS.
- `simsurv`²⁵⁵ for simulating survival data.
- `survextrap`²¹⁶ for fitting the ‘M-spline MPES’ model.
- `survHE`¹¹⁵ for fitting Bayesian survival models.
- `survival`³⁴ for producing survival estimates (e.g., KM estimates).
- `survtools`²⁷⁸ for extracting KM estimate coordinates for plotting.
- `rstudioapi`²⁷⁹ for setting the working directory (not necessarily a requirement, but used here for ease of programming).
- `purrr`²⁸⁰ for logic checks of the model fits.

The full simulation code used to execute the NCS is provided through an online GitHub repository (available via this link: <https://github.com/ABullement/external-evidence-ncs>). It should be noted that in the `simsurv` *R* package, the function `simsurv()` requires specification of the ‘true’ parameters, which are called using the argument ‘betas’. In the code linked above, the argument ‘beta’ is specified (i.e., the singular ‘beta’, rather than plural ‘betas’). This does not affect the results of the analysis, because the package allows for partial matching in argument names (i.e., *R* reads beta as if it were betas because these are deemed to be similar enough). In practice, it would be preferred to use the full argument names when using this (or indeed any) *R* package. However, this discrepancy was only noted after executing the full NCS, and so this note serves as a clarification on this difference between the NCS code and the package documentation for `simsurv`. An alternative version of the code using ‘betas’ instead of ‘beta’ has also been added to the GitHub repository for completeness.

The overall code script generation process can be described in three main parts, mirroring the approach for TSD 21, which is summarised in Figure 6-6.

Figure 6-6: Description of script generation in relation to TSD 21

1 Generate the data

- To generate the sample data to mimic a clinical trial, each line of code was translated into R, with adjustments made where necessary to account for differences in syntax. The Stata code makes use of a command called 'survsim' to simulate survival data. For the NCS code, the R package *simsurv* was used. In the help file for *simsurv*, it is noted that this R package was modelled on *survsim* for Stata. It should be noted that there is also an R package called *survsim*, but this was not designed to mirror *survsim* in Stata.
 - The code echoes the same process for the external data as the trial sample data, except using different input parameters where applicable (e.g., the size of the external data set relative to the trial sample size).
 - The same process was also followed to produce the code for determining the truth. Compared to the trial and external data, a much larger sample size was used for the truth.

2 Fit the models

- For fitting the models, the code developed as part of Chapter 5 was edited such that each model fit can be called as a function, and generates results as required for the NCS.

3 Analyse the simulations

- To analyse the simulations, the R package *rsimsum* was used. This package was designed to mirror the Stata package *simsum*. The TSD 21 simulation study does not specifically note the use of *simsum*, but given the overlap in authorship, it was assumed that *simsum* was used to consolidate results.

Key: NCS, neutral comparison study; TSD, Technical Support Document.

During script development, a potential issue with estimating the truth was identified, when using a mixture of two Weibull models (mirroring the approach taken in the TSD 21 simulation study). In brief, when sampling many different survival times, running the code using the default settings of the *simsurv* R package led to simulation errors. Following personal communication with the package developer (which can be found here: <https://github.com/sambrilleman/simsurv/issues/21>), the 'interval' argument was found to be the cause of the problem. Below is a brief step-by-step explanation for why this error occurred, and how changing the 'interval' argument solved the problem:

- When simulating survival times, *simsurv* draws a uniform random number, U , and determines the associated time, T , such that $S(T) = U$.
- To establish the corresponding value of T , the *simsurv* function requires inverting the survival function, such that $T = S^{-1}(U)$.
- For mixtures of parametric distributions (e.g., in this case, a mixture-Weibull model), the inverse survival function does not have a closed form solution (i.e., cannot be expressed as a precise formula). Therefore, *simsurv* searches for the solution numerically, using an R function called 'uniroot'.

- The 'uniroot' function requires specification of an interval (or sometimes described as a bracket) over which the solution is expected to be found. The default interval is $c(1 \times 10^{-8}, 500)$.
- As my code sampled 10 million values for U , it was possible that at least one of these values of U caused the solution to $T = S^{-1}(U)$ to fall outside the default interval, resulting in the error message.
- If the interval is widened (for example, to $c(1 \times 10^{-20}, 1 \times 10^{20})$), this means a solution can be found, but the code takes longer to run. Importantly, this change does not alter the process of data generation; rather, it simply ensures that a numerical solution can be found.

Other than this edit, the code for generating the truth was essentially the same as the code for generating sample data.

6.2.3 Error handling

As highlighted in Chapter 5, it is possible that some methods may, on occasion, produce errors as part of the model fitting process. In the context of running a script which generates thousands of simulations, it was therefore necessary to implement error handling within the simulation code. For all lines of code where an error was plausible, 'tryCatch' syntax was used (from base *R*). If a given line of code generated an error, this finding was recorded, and the output was recorded as 'NA'. Of note, a resultant model was only deemed an error if it produced an error message meaning that no numerical output was obtained. This means that any resultant model fit, regardless of any logical checks, was **not** considered a model error. This distinction is important, as if a method runs without error but produces an output which lacks face validity (e.g., survival increasing over time), this needs to be considered separately (even though it may be described as an 'error' from a lay perspective).

6.2.4 Random number generation

To ensure that the findings from the NCS are reproducible, it is important to consider the role of random number generation. In *R*, a random number seed can be set using `set.seed()` to ensure replicability of simulations. However, for Bayesian methods, there are potentially multiple layers of random number generation that occur, and it can be challenging to account for all of these when considering reproducibility of code. Multiple layers of random number generation introduces challenges when considering reproducibility across different computers, as differences in computing environments may lead to differences in results.

Of the eight methods considered in the NCS, four may be considered Bayesian: 'expert opinion', 'informative prior on shape', 'blended survival', and 'M-spline MPES'. The 'expert opinion' and 'informative prior on shape' methods both use a model fitting process based on

the *fit.models()* command from the *survHE* package.¹¹⁵ This command allows the user to specify a random number seed to ‘pass’ to *rstan*, using the argument ‘*seed* =’. For the ‘M-spline MPES’ method, a random number seed can also be passed to *rstan*, using the argument ‘*seed* =’. This leaves the ‘blended survival’ method. To ensure reproducibility for this method, a random number seed is required to be passed through to *rstan* on three different occasions (all using an argument ‘*seed* =’).

Outside of setting random number seeds, if the code is intended to be run over different computers, it is also necessary to consider multithreading and the number of cores being used. Multithreading allows software packages to perform more than one task simultaneously, which normally would be advantageous, but this can cause issues with reproducibility if different computers handle simultaneous tasks differently. Similarly, restricting Stan to using only one core means that only one chain is run at a time (using a single core). Combined, these settings mean that while the code runs more slowly (versus if these options are not used), it should be reproducible across any computer, regardless of technical specifications or user settings. In addition, it is critical that all software and package versions are aligned across computers to obtain identical results.

6.2.5 Script execution

Due to the large number of simulations required for the NCS, the analysis script was prepared such that it can be run across multiple computers. In total, four different computers were used to run the analysis script, details of which are summarised in Table 6-7.

Table 6-7: Computers used in the NCS

#	Name	Processor	Memory
1	A Lenovo Yoga Pro 7i.	13th Gen Intel Core i7-13700H	16GB
2	Asus Vivobook Pro M3401QA.	AMD Ryzen 7 5800H	16GB
3	HP Envy x360 15-ee000na.	AMD Ryzen 5 4500U	8GB
4	Lenovo IdeaPad Flex 5 16ABR8.	AMD Ryzen 5 7530U	16GB

Key: GB, gigabytes; NCS, neutral comparison study.

First, Computer 1 was used to generate sample data, generating the truth as if it were aligned with the trial population. Computer 2 was also used to generate sample data, but generated the truth as if it were aligned with the external population. Then, eight different versions of the model fitting script were used to fit the various methods to the sample data (details provided in Section 6.3). It should be noted that Computer 1 had a different processor (Intel) versus Computers 2, 3, and 4 (AMD) – differences in processor may introduce small inconsistencies between Bayesian analyses, and so Computer 1 was not used for any of the Bayesian analyses. The outputs from each code script were combined and analysed on Computer 4.

In principle, a super computer could have been used to increase the speed of the code execution – for example, the University of Sheffield’s High-Performance Computing (HPC)

clusters. However, given that it was possible to execute the analyses using consumer level equipment, this avoided the need to add further code, schedule time on the supercomputers, and most importantly, ensure the results were comparable to what a 'typical' user of the approaches would obtain. Furthermore, this means that the performance measure related to computation time could be assessed on readily-accessible computing hardware.

6.3 Results

In this sub-section, the results of the NCS are presented, grouped based on the different performance measures considered. Where it was possible to 'rank' the different methods based on performance, a rank was assigned. Any ties were treated as such (i.e., it was possible for multiple methods to be assigned first place), with the next-best method being assigned the appropriate rank below (e.g., if two methods were tied for first place, then the next-best method was noted as third place).

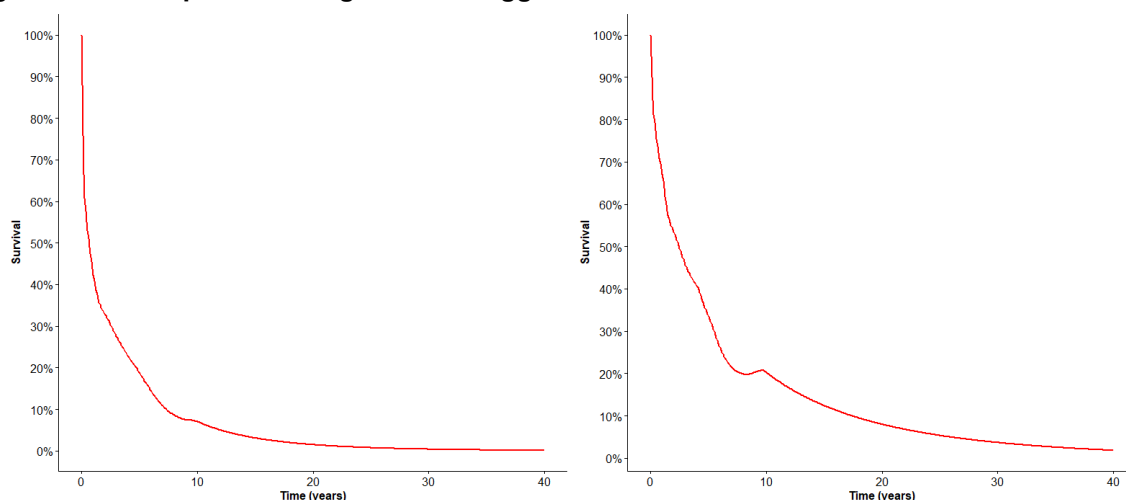
6.3.1 Error and logic checking

Prior to the inspection of the model fits, some initial checks were performed on the data sets produced. First, the frequency of errors was determined by establishing how often a given method did not produce a valid result – i.e., the result was recorded as 'NA' due to an error message (see Section 6.2.3 for further details). Across all of the simulations, no errors were produced for any of the model fits.

Based on a total of 1,000 simulations for each scenario, the MCSE for bias in mean survival was consistently <5% of the true value (range: 0.1% to 1.6%). This means that the MCSE associated with the estimated bias was small relative to the magnitude of the true value, indicating that the number of simulations was sufficient to obtain suitably stable estimates of bias across all scenarios. Therefore, results based on 1,000 simulations were deemed sufficient for all analyses presented henceforth. See Section 6.3.3 for more results related to bias specifically.

An additional check was performed on the resultant model fits, to determine how often (if at all) any of the model fits generated estimates of survival that increased over time. Across all the simulations, only the blended survival model produced logic errors. This occurred for 9,926 of the 32,000 simulations (31.0% of all simulations). Therefore, in just under one-third of the model fits, the blended survival method produced estimates of survival which increased at some point in time, even if only once between two specific time points. Some of these logic errors were far less consequential than others from a face validity perspective. Figure 6-7 demonstrates model fits produced on two separate iterations where a logic error was flagged.

Figure 6-7: Example of two logic errors flagged for the blended survival method



Note: The left-hand panel denotes a relatively minor logic error where survival increases by 0.003 from 9.9 years to 10.0 years – this does not occur at any other time point and is not possible to infer purely by looking at this survival plot (scenario 5, iteration 83). Conversely, the right-hand panel denotes a clear logic error of greater consequence, where between 8.3 to 9.7 years, survival is increasing (scenario 27, iteration 3).

A breakdown of the logic errors by scenario is also provided in Appendix G (see Appendix Table 17), which reveals that this occurred consistently across all of the scenarios explored.

Notably, if the external population was considered to have a poorer prognosis than the trial population, this consistently led to a smaller number of model fits with a logic error (e.g., compare scenarios 1 and 2, through to scenarios 31 and 32). This is broadly expected, given that a poorer prognosis for the external population would (on average) lead to a lower estimate of survival at a given landmark versus the trial population (and therefore, a smaller chance of survival increasing when using a blended survival approach). Despite this, there is still a possibility that survival beyond the end of follow-up in the trial may be greater for the external population versus the trial population based on random sampling. This is evidenced when comparing the scenarios paired according to sample size – the scenario with the smaller sample size consistently shows a greater number of logic errors compared to the larger sample size scenario (e.g., compare scenarios 1 and 17, through to scenarios 16 and 32).

Outside of the external population similarity and the sample size, there appeared to be some specific combinations of the remaining DGMs that inform each scenario which generally (though not always) led to more logic errors than others. Below, these DGMs are discussed alongside potential reasons for why there appeared to be less/more logic errors:

- When the maximum trial follow-up was set to 2 years instead of 5 years, more logic errors were recorded across all but one of the paired scenarios (i.e., all scenario pairings based on trial follow-up, except for scenarios 6 and 14). It is important to note that the blended survival model was executed assuming a blending interval from the end of trial follow-up (i.e., either 2 or 5 years), and the end of follow-up for the external

population (fixed at 10 years). These results suggest that when trial follow-up is longer (i.e., 5 years instead of 2 years), there appears to be a lower chance that survival for the external population will be greater than the trial population.

- When the hazard profile was defined as monotonically decreasing, there appeared to be a greater number of logic errors, versus a non-monotonic hazard profile in all but two scenario pairs (scenarios 3 and 7, as well as scenarios 9 and 13). No clear reason was identified for why the hazard profile may have led to a difference in logic errors.
- Logic errors appeared more frequently when survival magnitude was set to 'medium' versus 'low' (except for one scenario pair: scenario 13 and 15). This is likely because scenarios with lower survival magnitude were generally more complete (i.e., a lower proportion of censored observations), thus relied less heavily on extrapolation and therefore being less prone to errors.

While the points above are somewhat speculative, they may partially explain why some specific scenarios exhibited a greater number of logic errors than others. It was not feasible to inspect all 9,926 blended survival model outputs to determine how often these logic errors appeared to be 'important' (i.e., a situation like the right-hand panel of Figure 6-7, as opposed to the left-hand panel).

Overall, this finding was considered important to highlight, but no specific action was taken to exclude these potential logical errors in the NCS, as any attempts to *post-hoc* adjust the blended survival model fitting approach and/or represent only a sample of the outputted model estimates may inadvertently influence the neutrality of the methods comparison. Nevertheless, any results related to bias for the blended survival model should be interpreted with this important finding in mind and the requirement for interpretation noted. The logic issues noted with the blended survival model are important to consider in the context of the NCS – the blended survival model would likely not yield anywhere near the same volume of errors when applied in real-life situations, because the method would generally be only applied when it 'makes sense' to apply it.

6.3.2 Run time

The total time elapsed when fitting a given model to each of the 32,000 simulated data sets (i.e., 1,000 x 32 scenarios) is presented in Table 6-8, allowing for the determination of which methods were quickest to execute.

Table 6-8: Run time – model fitting (full run)

Name of method	Computer used	Run time (minutes)	Rank
Standard parametric	Computer 1	65.9	2
Standard piecewise	Computer 1	72.2	3
Expert opinion	Computer 2	1,476.2	5
Relative survival	Computer 3	6,652.0	7
Summary data piecewise	Computer 1	5.8	1
Informative prior on shape	Computer 2	1,165.4	4
Blended survival	Computer 3	2,313.9	6
M-spline MPES	Computer 2	10,346.0	8

Key: MPES, multi-parameter evidence synthesis.

As noted in Section 6.2.1, to account for differences in computing environments, a second set of model fits were produced using 1% of the total data set all using the same computer, with results presented in Table 6-9.

Table 6-9: Run time – model fitting (sample run)

Name of method	Computer used	Run time (minutes)	Rank
Standard parametric	Computer 4	1.2	2
Standard piecewise	Computer 4	1.5	3
Expert opinion	Computer 4	13.9	5
Relative survival	Computer 4	50.0	7
Summary data piecewise	Computer 4	0.5	1
Informative prior on shape	Computer 4	11.7	4
Blended survival	Computer 4	18.4	6
M-spline MPES	Computer 4	112.0	8

Key: MPES, multi-parameter evidence synthesis; SLR, systematic literature review.

Ranks were consistent between the full and sample runs, though due to differences in computer specifications that magnitudes of difference in time for the full run across multiple computers differed versus the sample run results. The fastest method to execute was the summary data piecewise method. This is perhaps unsurprising, since it is the only method that does not involve some form of regression or other complex model estimation process – instead, the model is estimated through a combination of producing a KM estimate and then via simple algebra, a constant probability of death for the remainder of the modelled time horizon. The slowest method to execute was the M-spline MPES method. This was expected since this method is arguably the most complex of all the methods included in the NCS.

As expected, the Bayesian methods generally took longer to run compared to the frequentist methods. It should also be noted that Bayesian methods will take different lengths of time to execute in practice depending on the final model specification, including the number of iterations for sampling, chains, etc. However, an unexpected finding was that the frequentist relative survival method was one of the slowest methods to execute. This method was implemented based on the published tutorial paper by Sweeting *et al.*, which includes estimation of a relative survival model, and then a post-estimation function to determine the

estimation of OS which may have contributed to the overall slow speed of computation, versus the other methods.

6.3.3 Bias, absolute bias, percentage bias, and empSE

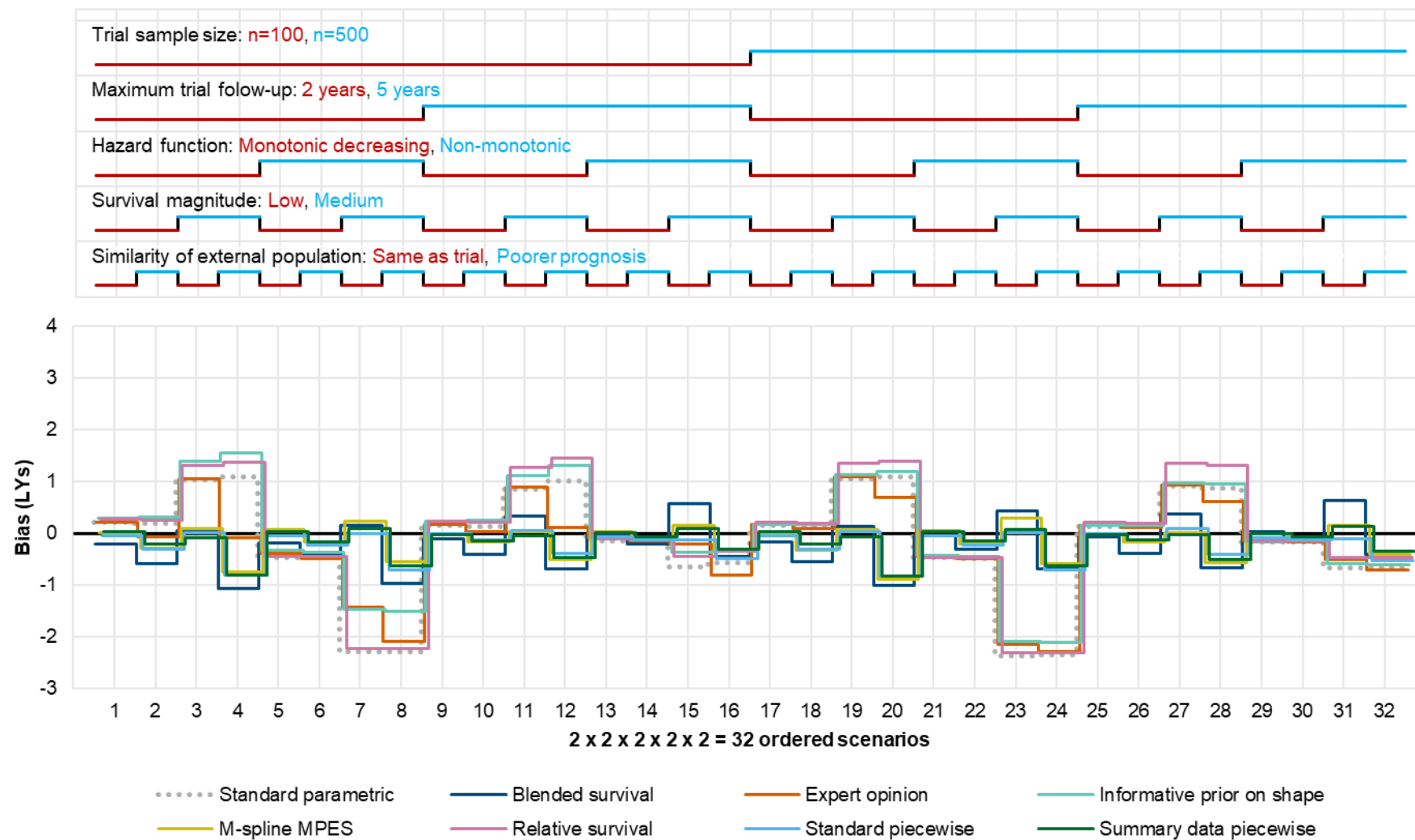
In this sub-section, results are presented and discussed for the bias-related performance measures, as well as empSE. First, results are presented for the main estimand of interest (mean survival over a lifetime horizon). Following the results for bias in mean survival, summary results are presented for absolute bias, percentage bias, and empSE; also for mean survival. Finally, additional results are discussed for the other estimands (RMST and point estimates of survival evaluated at 2, 3, 5, or 10 years).

Main estimand: Bias in mean survival

NLPs for all scenarios

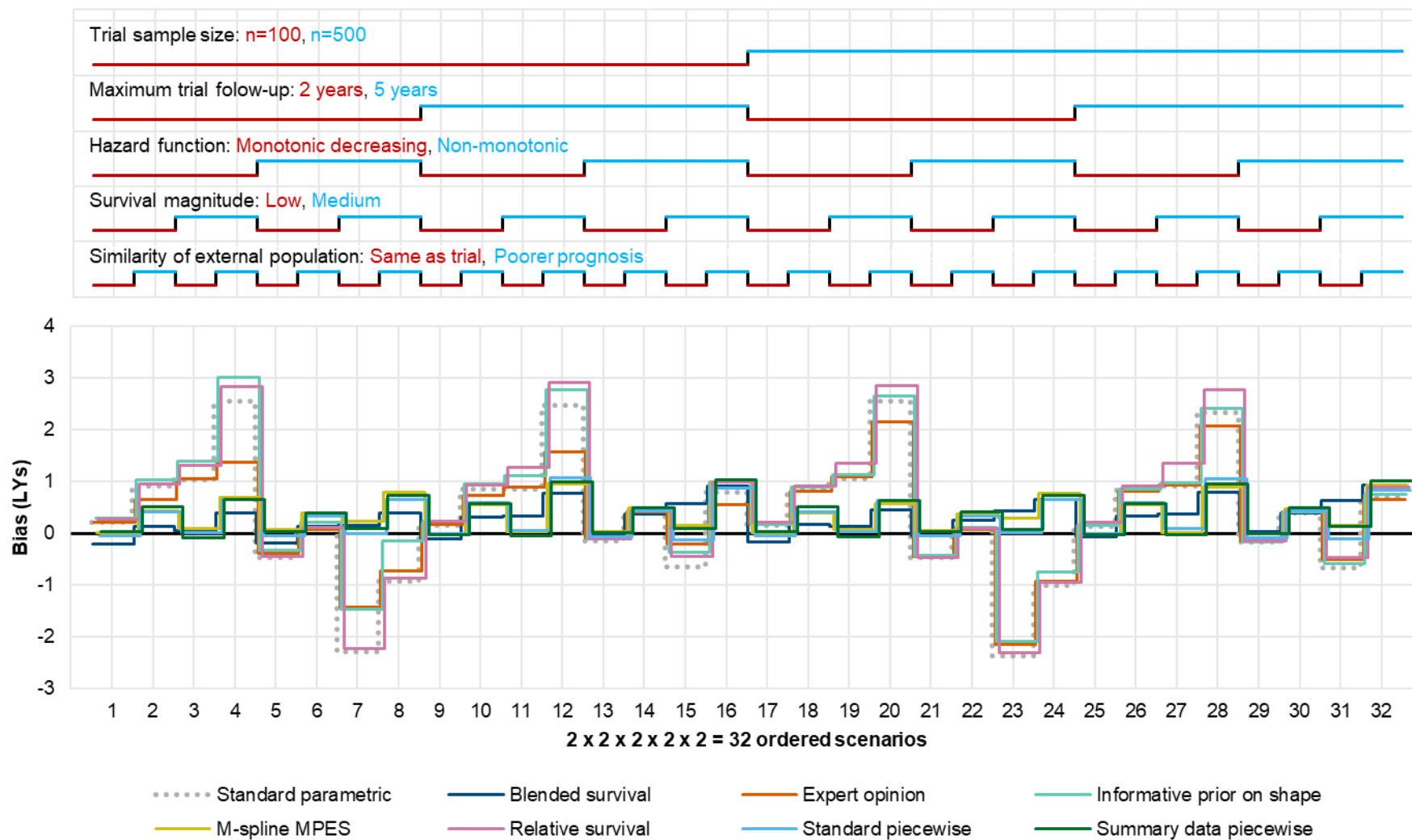
Firstly, results are presented for the main estimand of interest – the difference in mean survival between the fitted models and the ‘truth’ over a lifetime horizon (note: a description of the scenarios is provided in Table 6-2). NLPs showing results for bias using both versions of the ‘truth’ are provided in Figure 6-8 (for trial as truth) and Figure 6-9 (for external as truth). Full tabulated results for this estimand, including MCSE, are provided as tables in Appendix G.

Figure 6-8: NLP: Bias in mean survival, all scenarios, trial as truth



Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Figure 6-9: NLP: Bias in mean survival, all scenarios, external as truth



Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Commentary on NLPs for all scenarios

In comparison to the reference method (a standard parametric model – the dashed grey line in the NLPs), no method was consistently worse from a bias perspective – either the method that incorporated external evidence performed better (on average) when the truth was assumed to be the trial population, or when the truth was assumed to be the external population. In addition, the best-scoring method across each scenario was never the reference method. This suggests that across all scenarios, there was always some benefit of including the external evidence.

The following methods appeared to generally produce lower estimates of bias versus the reference method (a standard parametric model), across both truth contexts:

- M-spline MPES.
- Standard piecewise.
- Summary data piecewise.

For the remaining four methods, the following results were noted:

- Blended survival: This method appeared to generally perform better than the reference method, but due to the volume of logic errors, this does not necessarily mean that the resultant survival estimates were more valid.
- Expert opinion: This method appeared to generally provide similar estimates of bias versus the reference method, or a modest improvement. This method consistently produced larger estimates of bias compared to at least one other method that leveraged the external population data.
- Informative prior on shape: This method generally performed similarly to the reference method, but often led to greater estimates of bias (e.g., consider scenarios 4 and 12 across both truth contexts).
- Relative survival: Similar to the informative prior on shape method, the relative survival method generally performed similarly to, or possibly worse than, the reference method.

Averaging mean bias across all of the scenarios is generally ill-advised, as this can mask important differences in results within each scenario. Despite this, the methods with the lowest mean bias were identified for each of the truth options:

- When the *trial* population represented the truth, the model associated with the smallest mean absolute bias was the informative prior method, with a mean absolute bias across all 32 scenarios of 0.002. However, as is discussed in the remainder of the results section, this may essentially be because of positive biases being ‘cancelled out’ by negative biases. The model with the largest bias was the blended survival method,

with a mean bias of -0.222 (meaning that this method under-estimated survival, on average). Only the informative prior method produced a positive mean bias, meaning the remaining seven methods produced a negative mean bias.

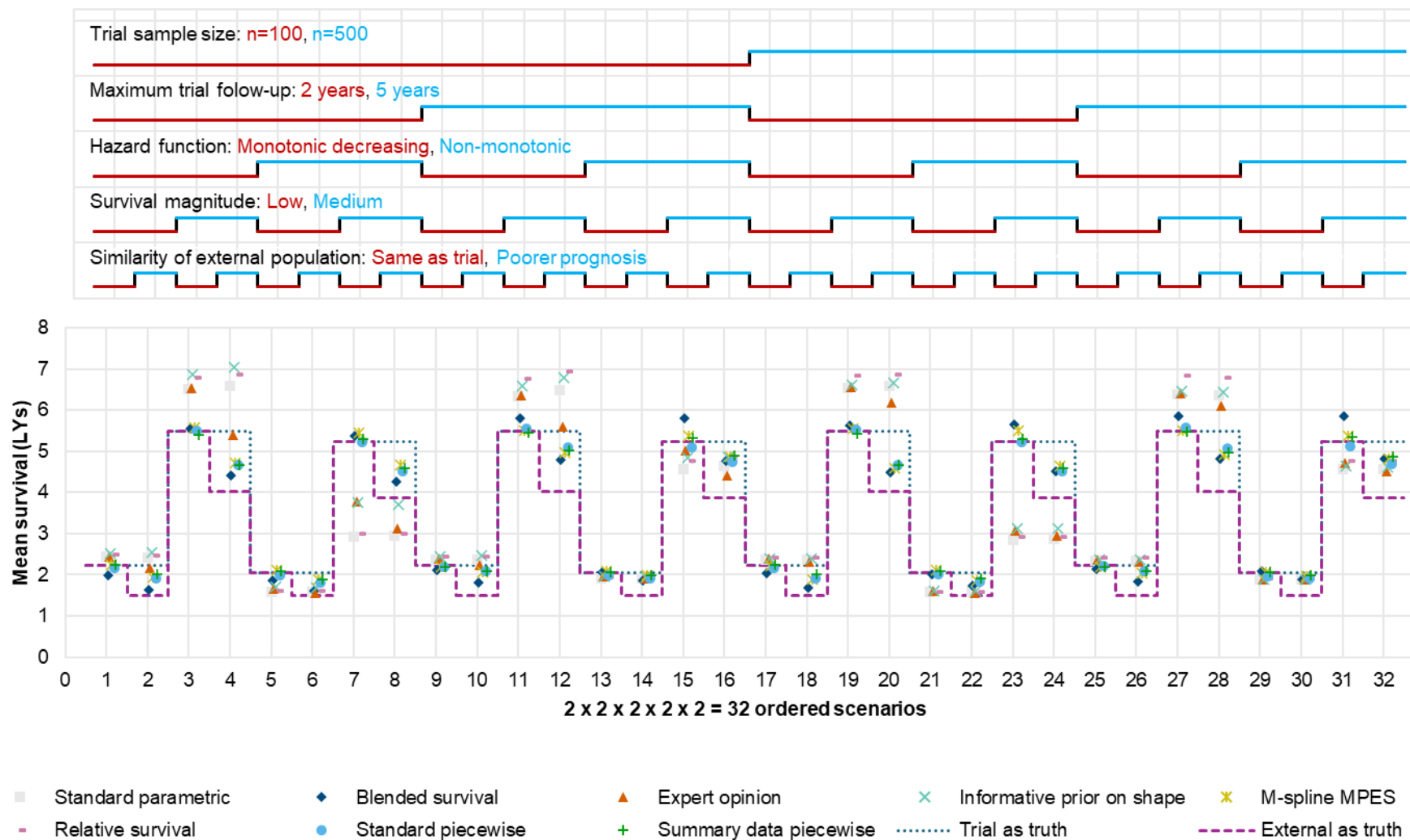
- When the *external* population represented the truth, the model associated with the smallest absolute bias was the blended survival method, with a mean bias across all 32 scenarios of 0.289. However, it should be noted that the blended survival method suffered from a number of logic errors (see Section 6.3.1). The next-best method was the standard piecewise model. The model with the largest bias was the informative prior method, with a mean bias of 0.514 (meaning that this method generally over-estimated survival). All methods produced a positive mean bias.

As an alternative way of considering the results for the main estimand of interest (bias in mean survival), a different type of NLP was prepared showing estimates of mean survival versus both truths. This is presented in Figure 6-10. From this plot, it can be seen that some methods appear to over-estimate survival in some scenarios regardless of which version of the truth is considered (e.g., consider scenarios 3, 4, 11, 12, 19, 20, 27, and 28 – these show consistent over-estimation of survival for the standard parametric, relative survival, and informative prior methods). Of interest, when the external population has a lower prognosis, this plot can be used to see how some estimates of survival sit between the two truths. More broadly, these findings are helpful to consider alongside the ‘truth considerations’ outlined in the Methods section (see Table 6-6). Here, for example, it was noted that for many of the methods, the estimand of relevance could be argued to be some third truth that falls between the trial and external populations (which is not possible to consider directly in the NCS).

- For the standard parametric and relative survival estimands, it was considered likely that the methods would typically align closer to when the trial population was considered the truth. This generally appears to be the case, except for some scenarios (e.g., scenario 8) where all methods appear to under-estimate survival. This may also be due to the specification of models based on a monotonic hazard function (i.e., a Weibull model) versus an underlying hazard function for disease-specific mortality which is simulated to be non-monotonic.
- For the expert opinion method, it was considered likely that estimates would fall closer to the external population. This can be seen clearly for some scenarios (e.g., see scenarios 4 and 12), but the opposite was seen for others (e.g., see scenario 8). As per the point above, this may be due to the specification of a Weibull-based model to estimate survival for a population simulated to have non-monotonic disease-specific hazards.

- For the other methods, it was expected that estimates of survival (if reasonably accurate) may fall between the two truths. This was generally seen for most scenarios. For example, consider scenarios 3 and 4 – when the external population is assumed to have a poorer prognosis, estimates of mean survival for blended survival, expert opinion, M-spline MPES, standard piecewise and summary data piecewise, all sit between the trial and external ‘truths’. However, when the methods produced estimates of survival that were

Figure 6-10: NLP: Mean survival, all scenarios, both trial and external as truth

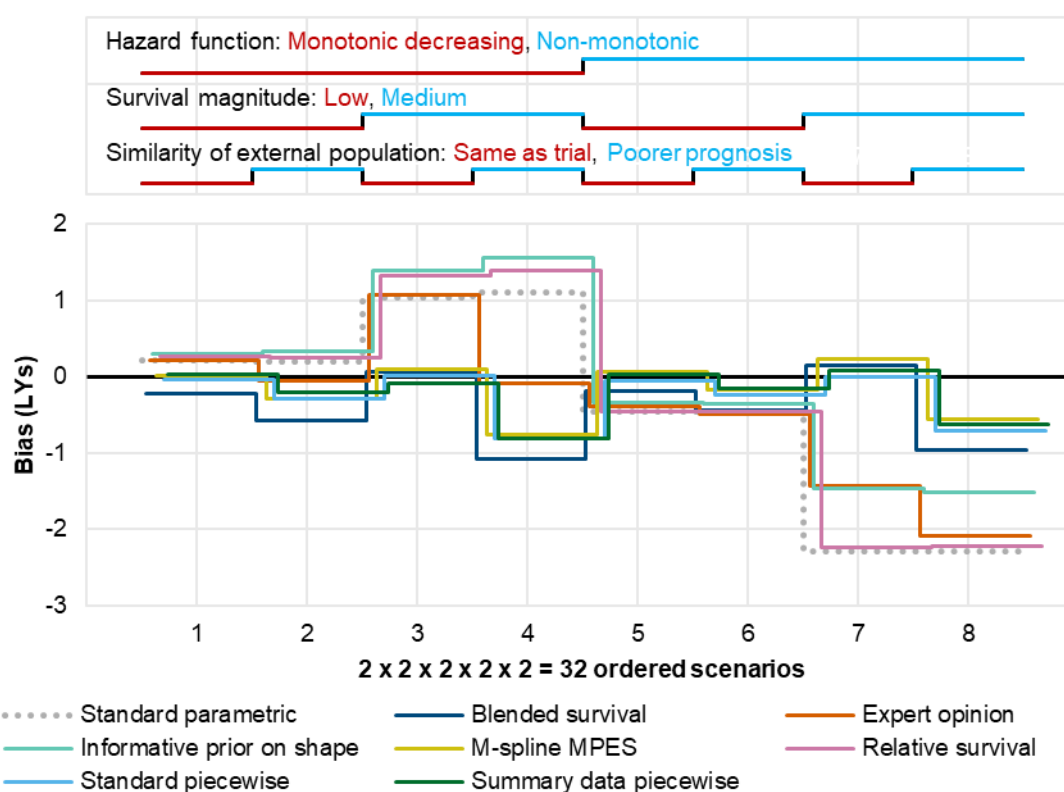


Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

NLPs for restricted set of scenarios

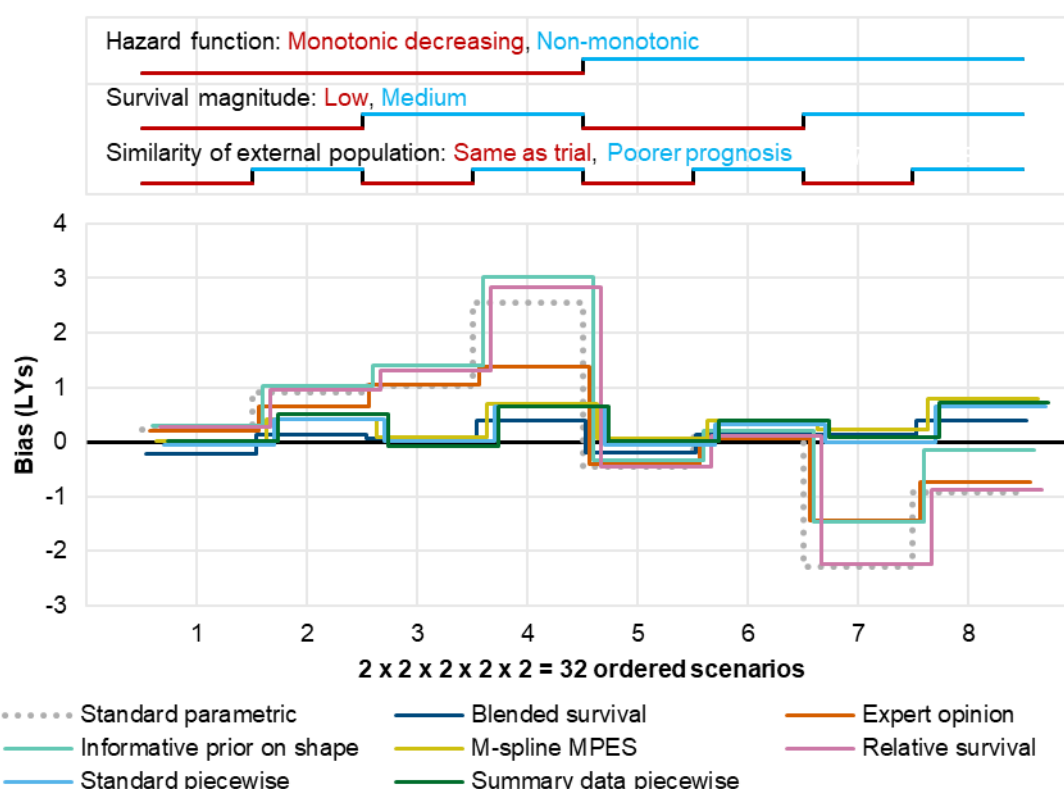
Based on the NLPs presented in Figure 6-8 and Figure 6-9, the sample size of the clinical trial, and the maximum duration of follow-up in the clinical trial did not appear to have a major impact on results (i.e., a [mostly] 'repeat pattern' can be seen in the NLPs for scenarios 1-16 versus 17-32; and for scenarios 1-8 versus 9-16 versus 17-24 versus 25-32). These DGMs are therefore not discussed in further detail. Instead, focus is placed on the remaining three DGMs. To help with interpretation, alternative versions of the NLPs focusing on the first eight scenarios are presented in Figure 6-11 (trial as truth) and Figure 6-12 (external as truth). It should be noted that bias for scenarios 1, 3, 5, and 7 are identical across truth contexts (since these scenarios consider the external population as having the same prognosis as per the trial population); however, due to scaling of the plots, results may look slightly different in Figure 6-11 (trial as truth) and Figure 6-12 (external as truth).

Figure 6-11: NLP: Bias in mean survival, first eight scenarios, trial as truth



Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Figure 6-12: NLP: Bias in mean survival, first eight scenarios, external as truth



Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Commentary on NLPs for restricted set of scenarios

Prior to the execution of the NCS, there was no general expectation for how the different methods would perform based on the underlying hazard profile, but it was of interest to understand how the methods appear to fare when fitted to different patterns of survival data. Ultimately, there was no obvious pattern observed in the bias estimates produced by the NCS according to hazard profile alone. However, it was noted that the non-monotonic hazard function was (on average) more likely to predict a greater hazard of death in the longer-term, versus the monotonic hazard function, as shown in Figure 6-3.

For survival magnitude, it was expected that a lower survival magnitude would mean that survival estimates are likely to exhibit smaller bias, when measured in absolute terms, given that a lower survival magnitude means that the extrapolated proportion of the survival curve is relatively less influential on estimates of lifetime mean survival. For survival magnitude, as expected, a low survival magnitude generally resulted in model estimates with lower absolute bias compared to those scenarios with a medium survival magnitude.

When considering survival magnitude and hazard profile together, three methods in particular stood out as producing particularly large estimates of bias under specific conditions. The informative prior on shape, relative survival, and expert opinion methods all performed poorly

with a medium survival magnitude, with substantial bias in opposite directions depending on whether hazards were monotonically decreasing or non-monotonic. In these cases, the methods tended to over-estimate survival with monotonically decreasing hazards (scenarios 3 and 4), and under-estimate survival with non-monotonic hazards (scenarios 7 and 8), a pattern that was also evident for the standard parametric models. Each of these approaches relies on the specification of an overall parametric form to capture all-cause mortality (informative prior on shape, expert opinion, and standard parametric) or disease-specific mortality (relative survival), which likely underpins the direction and scale of the observed biases. By contrast, other methods either do not require specification of a parametric form (e.g., M-spline MPES) or only partially rely on it (e.g., standard piecewise) and consequently appeared less sensitive to these scenario characteristics.

For the external evidence similarity scenario characteristic, prior to the execution of the NCS, it was expected that when the truth is assumed to be the *trial* population, scenarios where the external population is identical to the trial population should yield *lower* bias than those where the external population has a poorer prognosis than the trial population, for all methods that leverage external evidence from the external population. In these scenarios, there is a ‘perfect’ match between the evidence sources informing the model, and so it is expected that these scenarios would yield the lowest bias.

Based on the above, the results in Figure 6-11 (trial as truth) may be viewed moving from left to right, and the magnitude of bias should broadly be similar or increase when moving between pairs (scenario 2 showing greater bias than scenario 1, scenario 4 showing greater bias than scenario 3, etc.). As expected, this pattern was generally seen in the results – for some methods in particular (e.g., informative prior on shape) there was a notable ‘jump’ in bias when there was a misalignment between the trial and external populations incorporated into the data-generating process. For the expert opinion method, the opposite result was seen when moving from scenario 3 to scenario 4. This may be due to the combination of these scenarios considering a medium survival magnitude (hence a greater need to rely on extrapolations), and a decreasing hazard function, and so it may be the case that the expert opinion method prevented unrealistic long-term hazards, especially when prognosis was poorer for the external population (and so incorporating this forced the model extrapolation to be ‘pulled down’). However, this possible interpretation is speculative.

Importantly, the opposite of this pattern in results is not expected when the *external* population represents the truth, and indeed was not reflected in the results shown in Figure 6-12 (external as truth). Depending on the method, when the external population is assumed to represent the truth, estimates of bias could increase, decrease, or be approximately the same when comparing across scenarios, depending on the simulated prognosis of the external population.

In the monotonically decreasing hazards scenarios, for both the standard parametric and relative survival methods, bias is greater for scenarios 2 and 4 versus 1 and 3 (monotonic decreasing hazards). However, this is not the case for scenarios 5-8 (non-monotonic hazards), which is likely due to these methods not adequately capturing non-monotonic hazards, and the magnitude of bias being counteracted somewhat when compared survival estimates to a lower value (external population survival).

For the standard parametric and relative survival methods, the only differences in mean survival estimates between scenarios that had an external population that was similar to the trial, and those that had an external population with worse prognosis than the trial, would be related to sampling variation, and therefore, as expected, there was no consistent pattern in the difference in bias related to the similarity of the external population. However, when the external population was considered the truth and the external population had a poorer prognosis versus the trial population, estimates of bias generally increased, as would be expected. Despite this, the direction of travel for bias (in this case, generally increasing) should be considered based on what the estimate of bias was when the trial population was assumed to represent the truth. For example, when a method had under-estimated survival when the trial population represented the truth, the specification of the truth being the external population may 'cancel out' some of the under-estimation of the model fits. While true for any method, this finding is particularly noteworthy for the standard parametric and relative survival methods, since neither of these are informed directly by the simulated external data.

For the remaining six methods, the external evidence influences model fits directly. For the informative prior method, the impact of the prognosis of the external population was occasionally negligible, which may be expected given that the informative prior was placed on the *shape* parameter, not the *scale* parameter.

For some of the methods, when the prognosis of the external population was the same as the trial population, this led to an over-estimation of survival. This is unsurprising, since by chance, a given method may over-estimate survival despite being informed by data from the same sampling distribution. However, when considering paired scenarios where the prognosis of the external population was assumed to be poorer than the trial population, then the resultant bias was typically lower. In some cases, where there was a positive bias when prognosis was assumed equal, there could be a negative bias when prognosis was assumed poorer for the external population (since here, the trial population is assumed to represent the truth). This links back to commentary made in Table 6-6, wherein it was noted that some methods may essentially attempting to reflect a 'third truth' which sits between the trial and external populations.

If the *external* population represents the truth, then there is a difference in the expected survival of the population over the period of follow-up in the clinical trial and the external evidence. Consequently, it can be seen across all six methods that leverage external evidence from the external population, that the average bias is greater when the prognosis of the external population is poorer. This result suggests that across all methods, the trial data remains the main data source which influences survival extrapolations. Ultimately, this means that the average bias is greater when there is a notable discrepancy between the trial and external populations, and the *external* population is assumed to represent the truth.

In general, the results suggest that all methods performed reasonably well when survival magnitudes were low, although this represents a situation where external evidence may add relatively little value for extrapolation. Conversely, performance differed across the methods when considering scenarios with a medium survival magnitude. The piecewise, MPES, and blended survival methods seemed to perform relatively better, notwithstanding the logic issued previously described for the latter method, particularly when the trial population was assumed to represent the truth. Importantly, across all methods, substantial bias arose when the prognoses of the trial and external populations differed, regardless of which population was taken to represent the truth. This limitation is especially relevant for practical use in HTA, as differences between trial and external populations are generally the norm rather than the exception.

Other performance measures

For the remainder of this section, results are presented assuming that the *trial* population represents the truth, for ease of interpretation, except for a brief paragraph on the *external* population as truth results provided at the end. NLPs for absolute and percentage bias when the *trial* population is assumed to represent the truth are presented in Figure 6-13 and Figure 6-14, respectively. Corresponding NLPs for when the *external* population is assumed to represent the truth are shown in Appendix G. An NLP for empSE is presented in Figure 6-15.

Absolute bias shows the average magnitude of deviation between estimated and true values, irrespective of direction (i.e., under- or over-estimation of mean survival). These results highlight that survival magnitude plays a particularly important role: scenarios with medium survival magnitude tend to exhibit notably higher absolute bias, especially for the following methods: standard parametric model, informative prior on shape, expert opinion, and relative survival. Interestingly, sample size does not appear to be strongly associated with the absolute magnitude of bias. For instance, scenarios 7 and 8 (with smaller sample sizes) exhibit similar or lower levels of absolute bias compared to scenarios 23 and 24 (with larger samples),

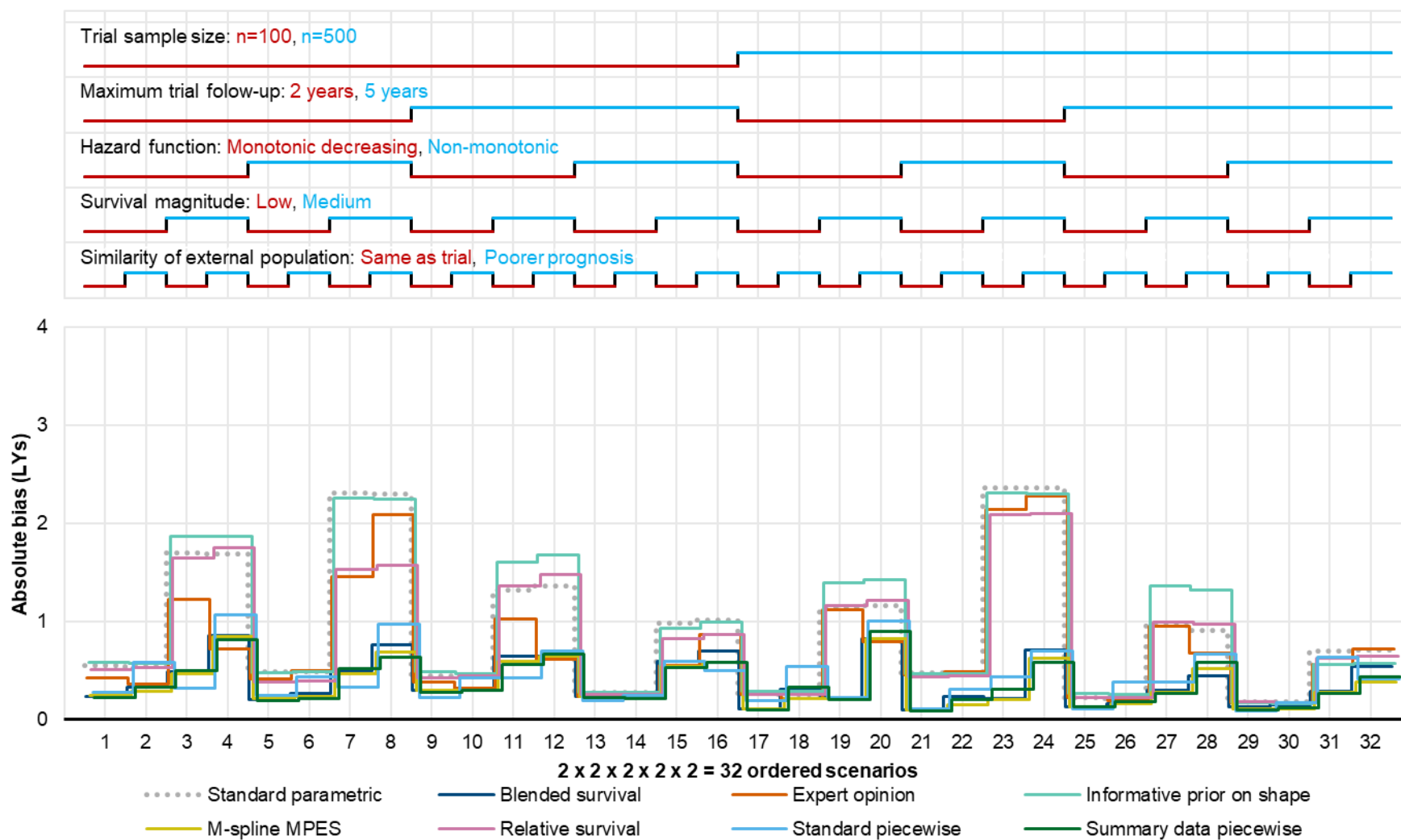
suggesting that the other scenario characteristics are more influential in terms of absolute bias.

Percentage bias, by contrast, reflects bias relative to the true value of the estimand, making it a scale-sensitive measure. It captures how large the errors are in proportional terms, enabling comparisons across estimands of different magnitudes. Across all methods and scenarios, percentage bias ranges from -45.3% up to +28.3%. This demonstrates that, in some cases, methods can deviate substantially from the true value, particularly in long-term extrapolations. Nonetheless, the overall pattern of results remains broadly consistent with those observed for both bias and absolute bias. The overall findings for specific methods based on absolute and percentage bias were generally consistent with those for bias.

Finally, empSE quantifies the variability of estimates across repeated simulations and behaves largely as expected. As sample size increases, empSE tends to decrease, reflecting greater precision. In addition, empSE is clearly influenced by the survival magnitude across all methods: longer survival times imply a longer extrapolation period, which in turn increases uncertainty. Notably, a specific combination of low survival magnitude and non-monotonic hazard functions results in consistent empSE values across all methods, suggesting that under certain circumstances (e.g., when there is a lesser need to rely on extrapolated estimates of survival), different methods may produce similarly precise estimates (regardless of the specifics of the methodological approaches).

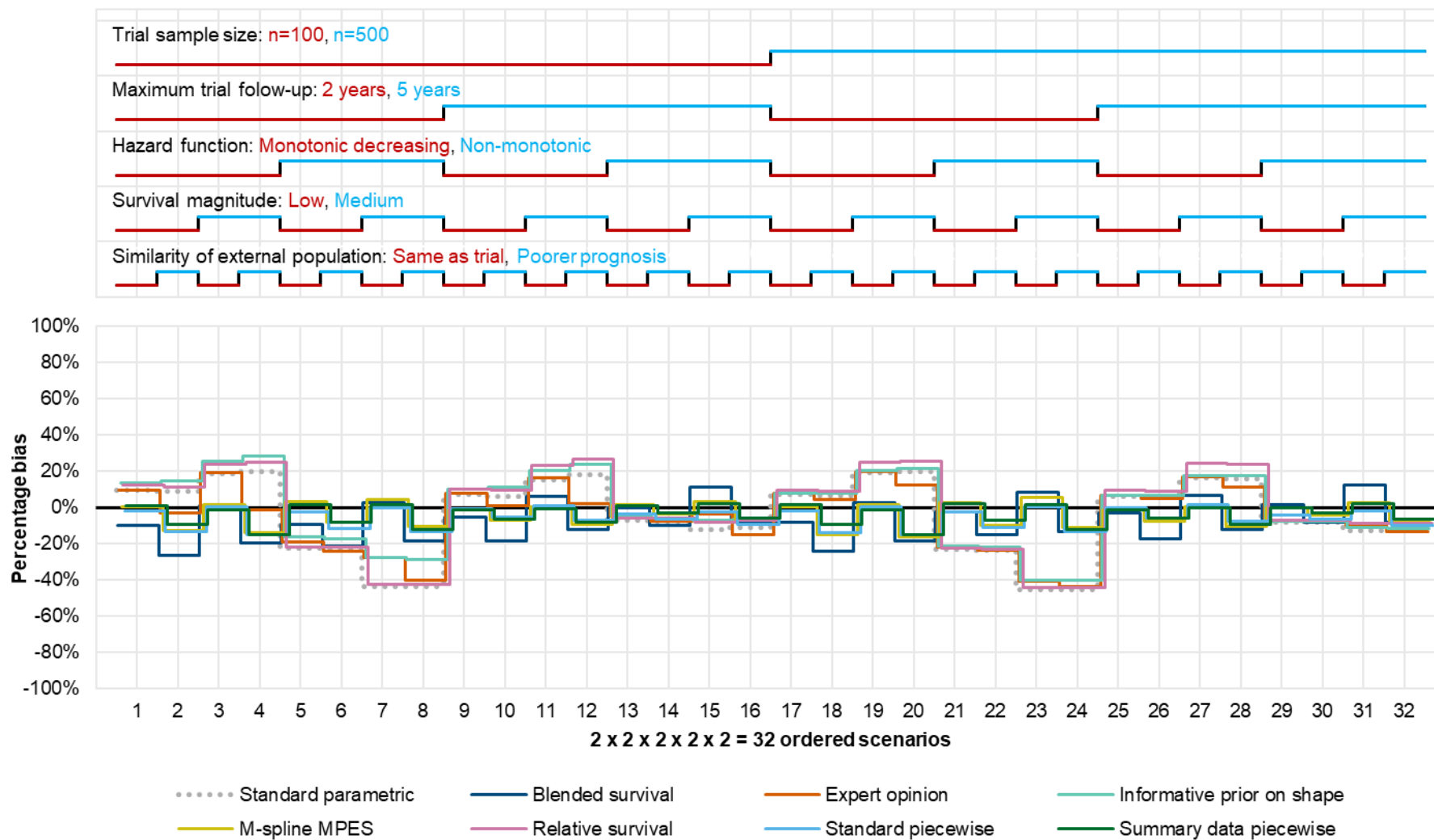
For the methods specifically, empSE was generally high for the relative survival, standard parametric, expert opinion, and informative prior methods. Overall, the pattern of empSE was broadly consistent with the pattern observed for bias. This suggests there was no clear 'bias versus precision' trade-off across methods; rather, the methods associated with larger bias also tended to exhibit greater empirical variability. Consequently, there do not appear to be any specific methods that gain improved precision in exchange for higher bias, indicating that poorer method performance reflects both reduced accuracy and reduced precision.

Figure 6-13: NLP: Absolute bias in mean survival, all scenarios, trial as truth



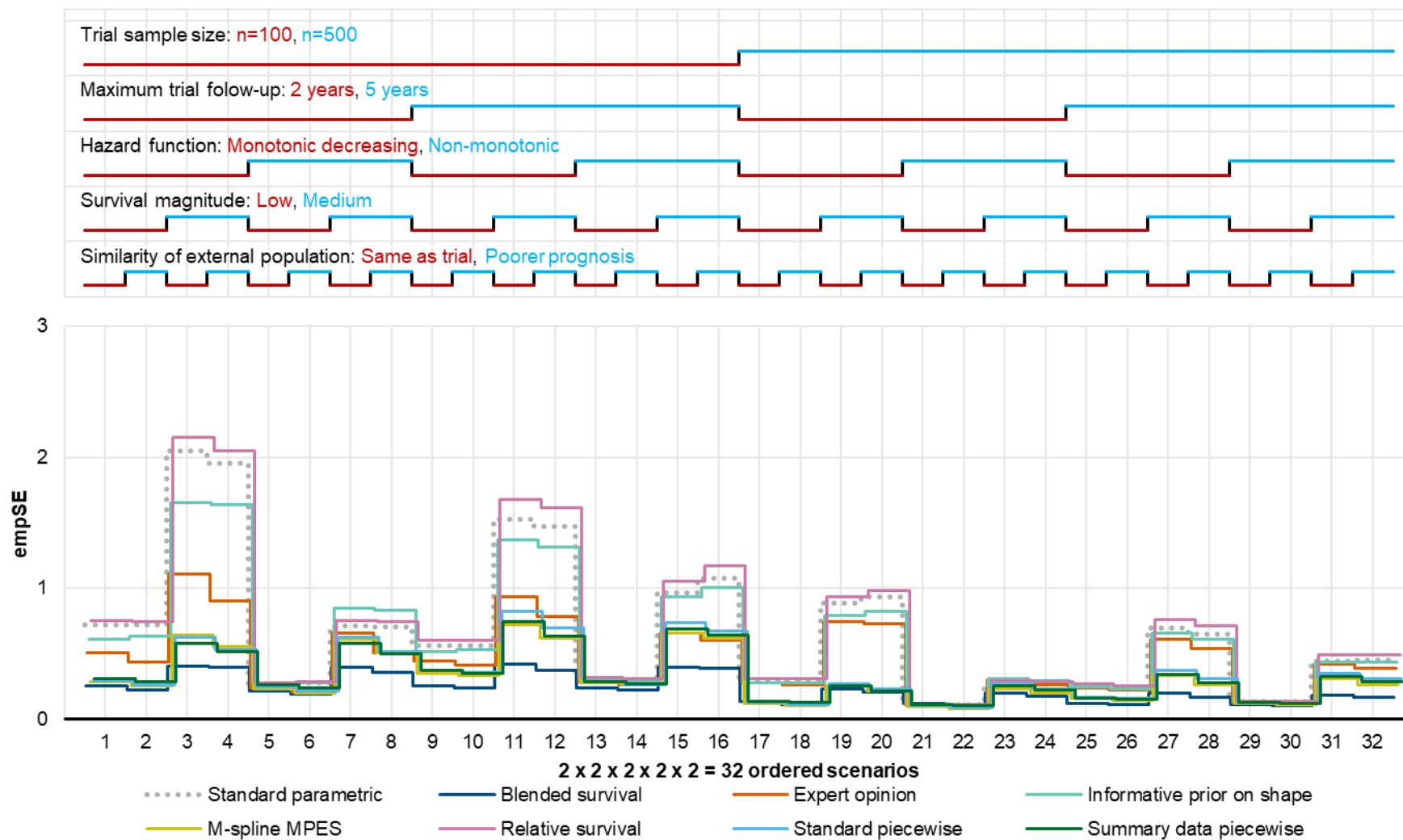
Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Figure 6-14: NLP: Percentage bias in mean survival, all scenarios, trial as truth



Key: MPES, multi-parameter evidence synthesis.

Figure 6-15: NLP: empSE in mean survival, all scenarios



Key: MPES, multi-parameter evidence synthesis.

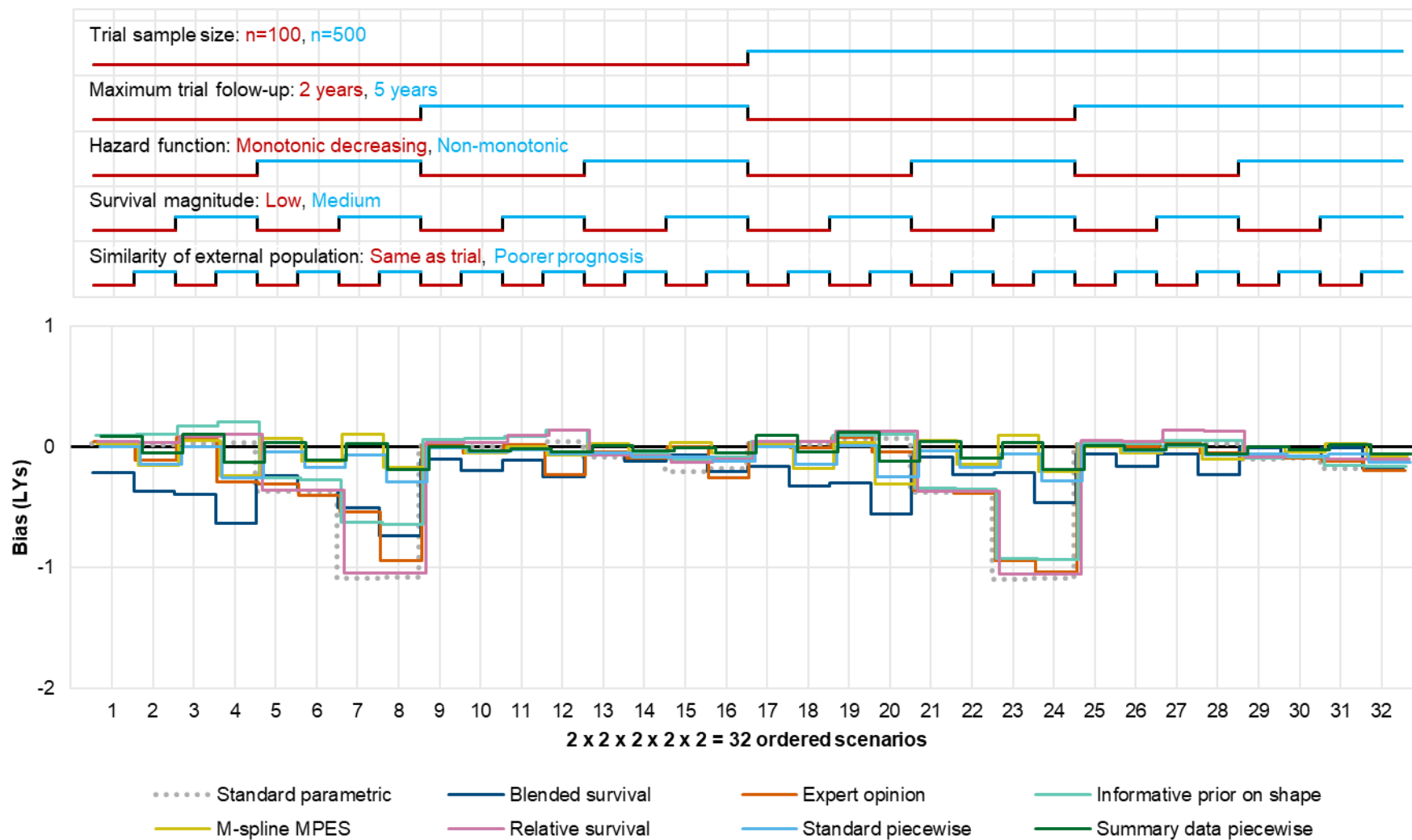
Additional estimands

The results discussed until now refer to the primary estimand of interest: bias in mean survival over a lifetime horizon. However, additional survival measures were also generated to provide a more comprehensive view of method performance. These were RMST and point estimate survival probabilities evaluated at 2, 3, 5, and 10 years. While these were not the primary focus of the NCS, results for these estimands were considered to provide additional context, notably covering potential differences in method performance over the short-term. To avoid the inclusion of a further 16 NLPs, a subset of results is presented in this thesis, focusing on selected NLPs that illustrated notable findings.

Among the supplementary estimands, RMST and point estimate survival at 10 years are arguably the most informative, as these scenarios reflect a time point after the period of trial follow-up, but not the entire modelled time horizon. NLPs for these two estimands are provided in Figure 6-16 (RMST) and Figure 6-17 (point estimate), assuming that the *trial* population represents the truth. Across the full range of scenarios, methods tended to under-estimate, rather than over-estimate, RMST or point estimates of survival evaluated at 10 years. This pattern suggests that the incorporation of external evidence may “pull down” survival projections, particularly when the external evidence is misaligned with the trial data (or, more specifically, when the external population has a poorer prognosis compared to the trial population).

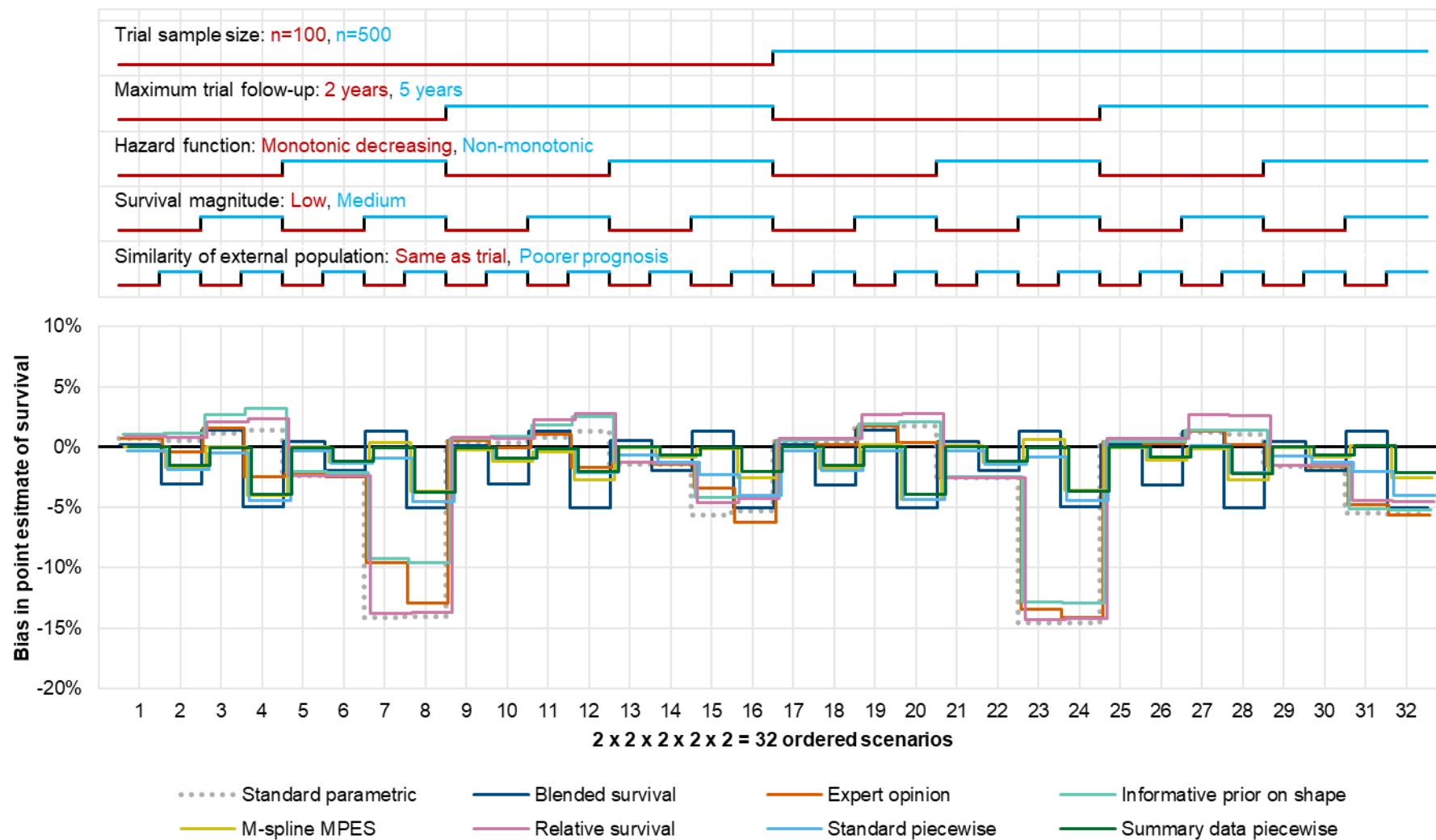
To illustrate the differences in RMST evaluated over a lifetime horizon versus an earlier time point (10 years) further, Figure 6-18 shows the NLP for RMST at 10 years superimposed over the NLP for bias in mean survival. This shows that while over a lifetime horizon some methods appeared to greatly over- or under-estimate survival, this materialised *after* 10 years, illustrating how influential assumptions concerning long-term survival may be on lifetime estimates used to inform HTA decision making. The largest differences in RMST between 10 years and over a lifetime horizon were seen for the standard parametric, relative survival, informative prior on shape, and expert opinion methods.

Figure 6-16: NLP: Bias in RMST at 10 years, all scenarios, trial as truth



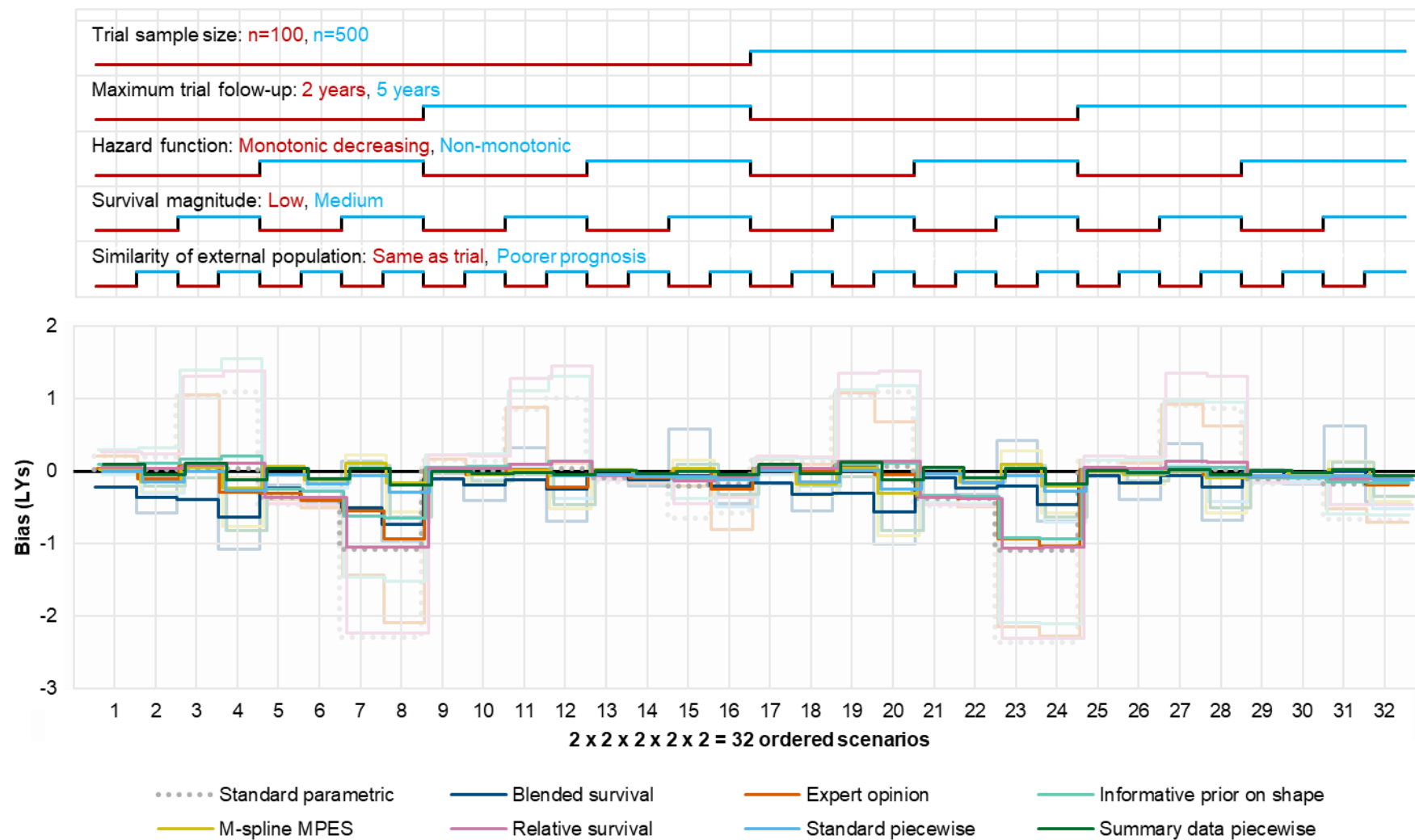
Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Figure 6-17: NLP: Bias in point estimate of survival at 10 years, all scenarios, trial as truth



Key: MPES, multi-parameter evidence synthesis.

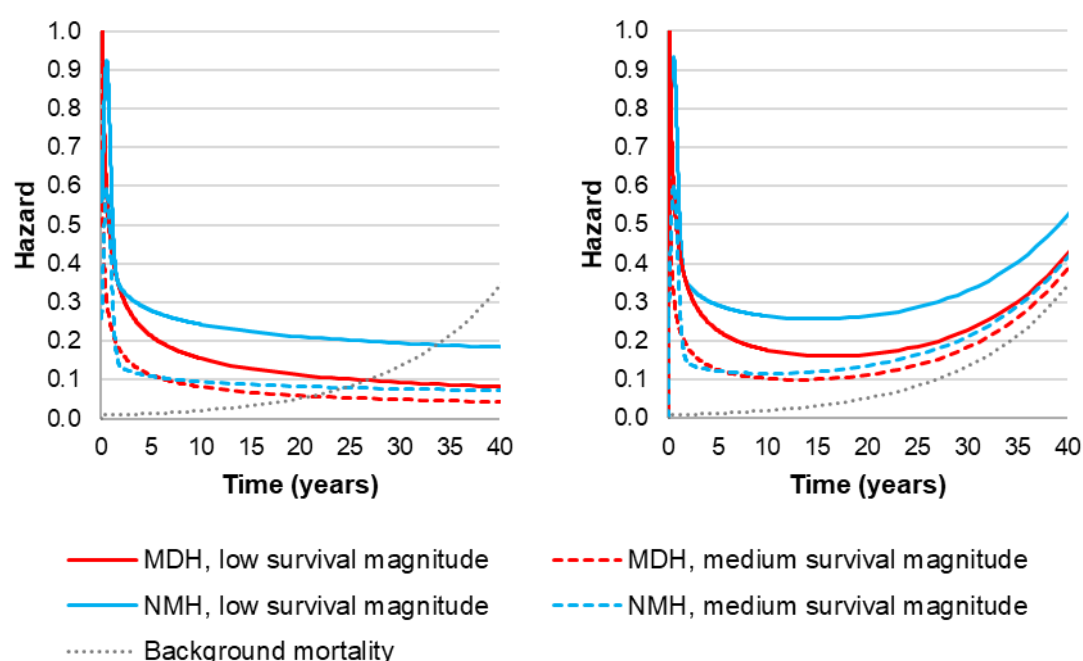
Figure 6-18: NLP: Bias in mean survival (faint) versus RMST at 10 years (bold), all scenarios, trial as truth



Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

For the relative survival method in particular, this finding may reveal a somewhat limited role of background mortality in the scenarios explored within the NCS. Across the range of results for each of the scenarios explored, it can be inferred that the estimated hazards rarely declined to a level comparable with background mortality (i.e., there was always a non-zero excess mortality). This is consistent with the hazard profiles that informed each of the scenarios, as shown in Figure 6-3. Another version of this plot is shown in Figure 6-19, focusing on only hazards, up to a 40-year time horizon, and with background mortality rates superimposed (based on the Gompertz model used in TSD 21).

Figure 6-19: Hazard profiles used in the NCS (up to 40 years)



Key: MDH, monotonically decreasing hazards; NCS, neutral comparison study; NMH, non-monotonic hazards.
Note: Left panel = cause-specific hazards, right panel = all-cause hazards. Background mortality shown for reference on both panels.

Based on Figure 6-19, it can be seen that the earliest timepoint at which background mortality rates exceed the simulated disease-specific mortality is at 20 years. At this point in time, across all four scenarios, there are <3.0% of patients expected to be alive (range: 0.1% [non-monotonic hazards, low survival magnitude, solid blue line] to 3.0% [monotonically decreasing hazards, medium survival magnitude, red-dashed line]). Since background mortality would therefore have a relatively small impact on results, this shows why the relative survival approach produced results that were similar to the standard parametric model.

External population as truth

Findings for absolute and percentage bias when the external population was considered the truth were consistent with the findings presented for bias. For absolute bias, there was a notable increase in bias for some scenarios where there was a combination of medium survival magnitude and poorer prognosis for the general population for the methods that did not leverage the external data directly as part of the model fitting process (standard parametric, relative survival, and informative prior on shape). For percentage bias, some combinations of scenarios and methods showed large estimates of bias (close to a positive percentage bias of 80%), especially for the specific combinations of methods and scenarios described previously for bias (e.g., relative survival when external population assumed to have a poorer prognosis than the trial population). NLPs for absolute and percentage bias are provided in Appendix G.

6.4 Discussion

6.4.1 Summary and overall findings

When choosing between methods, analysts would ideally want to understand how each performs to make an informed decision. To support this, an NCS was conducted to evaluate methods objectively across a common set of scenarios. The NCS compared seven approaches for incorporating external evidence into survival extrapolations against a standard parametric model, with the aim of assessing their relative performance.

None of the model fits produced error messages, but the blended survival method produced a large number of model fits where survival at some point in time increased relative to an earlier time point. This was an important finding from the NCS, as without this context, analysts may assume that results generating increasing estimates of survival are due to user error, rather than this simply being a possible output from this method. Some logic errors may be considered to be of greater consequence than others, but nevertheless are important for awareness. In practice, analysts should ensure to sense check model fits using the blended survival method to ensure survival does not increase at any point in time – if this occurs, it may be suitable to consider an alternative model specification, or possibly use a different survival extrapolation method.

From a practical perspective, some of the methods took substantially longer to run compared to others. The slowest method was the M-spline MPES approach, which took over a full week to generate all the model fits required for the NCS, compared to the fastest method, which took under 10 minutes. In practice, it is unlikely that such long run times would be encountered, based on the smaller number of estimates of survival that are generally required for HTA. However, run time may be an important consideration for analysts looking to generate many survival estimates, potentially for combinations of treatment arms, subgroups, and endpoints.

Overall, when considering the key estimand of mean lifetime survival, the results of the NCS showed that no single method was universally preferred, but some methods appeared to perform better than others under specific circumstances. In summary, the following findings were observed when considering the full range of scenarios explored:

- Among the methods that leveraged external evidence, M-spline MPES, standard piecewise, and summary data piecewise approaches generally produced lower bias than the standard parametric reference model.
- The remaining methods (blended survival, expert opinion, informative prior on shape, and relative survival) typically performed similarly to, or modestly better than the reference model (but with inconsistent improvements over the standard parametric reference model). In some cases, these methods that leverage external evidence yielded greater bias than the reference method (which does not leverage any external evidence).

Looking more closely into the features of specific scenarios, the following results were noted:

- When considering the impact of survival magnitude and hazard profile, three methods (informative prior on shape, relative survival, and expert opinion) yielded large estimates of bias for scenarios with a medium survival magnitude (with bias in opposite directions depending on the hazard profile specified in the scenario being evaluated). These methods rely on the specification of a parametric model which has a notable impact on results. While other methods (such as a piecewise model) also involve selecting a parametric model (or models), these generally yielded lower bias, as these models play less of a direct role in estimating survival over a full lifetime horizon.
- Focusing on the similarity of the external population compared to the trial population, bias generally increased as the external population diverged from the trial population, particularly for methods like the informative prior on shape approach. An exception was observed for the expert opinion method, which showed reduced bias in one scenario pair related to external evidence similarity. This was expected to be caused by the specific characteristics of the method, which relies on choosing a point estimate for external evidence. This means that possibly by chance, the external evidence reflected poorer survival than expected at a given time point, but caused the survival extrapolation to be affected in such a manner that the average bias was still lower than if the external evidence aligned correctly at this specific time point.
- The informative prior method showed limited sensitivity to the external population's prognosis, likely because the prior was placed on the *shape* rather than the *scale* parameter; in some cases, bias was reduced or reversed when a poorer prognosis

was assumed. This particular method may therefore be of value when there is a known difference in prognosis between data sources, but the overall 'shape' of the survival estimate for the external population is expected to be similar.

- The relative survival method did not outperform the standard parametric model in most scenarios when the external population was considered the truth. This suggests that background mortality was not highly influential in the scenarios explored in the NCS. However, as previously noted, the NCS was expected to show relatively poor performance for the relative survival method versus the other approaches, owing to the external data not being utilised by the relative survival method (i.e., only life tables are used).

When considering estimates of mean survival (rather than bias in mean survival), the methods can be simultaneously compared to both truths (trial and external). This proved useful in illustrating that most of the methods are theoretically intended to provide estimates of survival that sit somewhere between the two truths. This is also shown when comparing some of the NLPs for the different truths – for some combinations of scenarios and methods, the NLPs show 'mirror images' (i.e., positive bias for one truth, and negative bias for the other). This is also expected, since some methods would be expected to produce estimates of survival that fall between the two truths (which is evident when the external population is assumed to have a poorer prognosis versus the trial population).

The NCS also considered additional performance measures and estimands. As may be expected, absolute bias was notably higher in scenarios with medium versus low survival magnitude, particularly for methods like standard parametric, informative prior on shape, expert opinion, and relative survival. Sample size appeared to have little impact on absolute bias, suggesting other DGMs (such as trial follow-up) played a more influential role. Percentage bias, which importantly reflects bias relative to the actual size or value of the estimand, showed wide variation across methods and scenarios, though the pattern was consistent with the primary estimand (bias). In addition, empSE decreased with larger sample sizes and increased with longer survival times. In scenarios with limited 'need' to extrapolate, empSE was consistent across methods (i.e., similar levels of precision) despite methodological differences in how external evidence was used. Where there was a greater value for external evidence (e.g., scenarios with a medium survival magnitude), empSE was notably higher for the relative survival, informative prior, and expert opinion methods.

Results for RMST and point estimates of 10-year survival were the most informative of the additional estimands explored. Across scenarios, methods tended to under-estimate these values, likely due to external evidence 'pulling down' survival projections when misaligned with the trial data. These results, considered alongside the primary estimand (bias in mean survival

over a lifetime horizon), clearly demonstrate how assumptions or features of methods related to long-term projections can substantially impact survival estimates.

6.4.2 Strengths and weaknesses

The neutrality of the methods comparison represents one of the key advantages of this study. Unlike previous studies, which may have been influenced by specific methodological preferences, assumptions, or stakeholder interests, this NCS was designed to provide an impartial, balanced assessment of alternative methods. This neutrality ensures, to the extent that is possible, that the findings are not inadvertently biased in favour of any particular method. In the context of methods to incorporate external evidence, this NCS, to the best of my knowledge, represents the first documented and formalised comparison of alternative approaches. By systematically evaluating multiple methods under consistent conditions, the NCS offers a comprehensive evidence base that can inform analysts and decision-makers in selecting the most appropriate methods for their specific needs, without undue influence from methodological biases.

The range of scenarios explored in this NCS was designed to capture the key considerations that may influence survival extrapolations in HTA practice. By incorporating a set of scenarios intended to mirror situations often encountered in HTA, such as varying levels of external evidence relevance and different data availability conditions (e.g., duration of follow-up), the NCS provides a holistic view of how different methods perform under real-world situations. Ultimately, the structured, neutral nature of this study means that it can serve as a reliable reference point for anyone seeking objective guidance on survival extrapolation methods that leverage external evidence. Despite this, the results of the NCS cannot be used to determine which method (or methods) is (or are) objectively 'better' than another in all circumstances – each method has its own pros and cons, and there are circumstances wherein each method could plausibly produce accurate estimates for decision-making.

Like any study, the NCS was not without its limitations. While the scenarios explored were carefully selected to capture key considerations relevant to survival extrapolations in HTA, additional scenarios could have provided further insights. For example, varying the relationship between the sample size of the trial and external populations might have offered a deeper understanding of how population size impacts the integration of external evidence. However, the NCS was purposefully designed with a relatively restricted number of scenarios to ensure that the results remained interpretable. By focusing on a manageable set of conditions, the study aimed to strike a balance between comprehensiveness and clarity, providing meaningful insights without overwhelming complexity.

Furthermore, while the scenarios included in the NCS addressed a range of practical considerations, they could theoretically have been expanded to explore additional dimensions, such as varying definitions of ‘similarity’ between trial and external populations (e.g., ‘identical’, ‘poorer’, and ‘much poorer’ prognoses of the external versus trial populations). However, a simple 2^5 fully factorial design was chosen as the most pragmatic and transparent approach for this first formalised NCS of its kind. This design allowed for a structured and systematic exploration of key factors, providing a foundation upon which future studies could build, or new methods also be investigated.

In the NCS, survival data were simulated using hypothetical distributions for both disease-specific and other-cause mortality. The Weibull distribution was selected for simulating disease-specific mortality, and the same distribution was selected for any of the methods that required selection of a given parametric model. In reality, there is no specific distribution which informs disease-specific mortality, and a range of possible parametric models may be explored. Ultimately, it would not have been practical to explore all possible combinations of model settings and assumptions (such as exploring different parametric model fits) within the context of this NCS, without needing to evaluate hundreds of different model specifications. Considering different parametric models would also mean choosing which other settings or assumptions should be evaluated (e.g., the choice of cut-point in a piecewise modelling approach). Nevertheless, it should be recognised that a key limitation of the NCS is the need to choose a given distribution for simulating disease-specific mortality, as well as specific model settings, including (but not limited to) a parametric model distribution.

Another potential limitation of the NCS is the mismatch between the data-generating mechanism and the model fitting process in terms of other-cause mortality (i.e., background mortality). Specifically, the simulated datasets were generated using the same Gompertz model as used in the TSD 21 simulation study, whereas the models were fitted using general population life tables to account for background mortality (where applicable for a given method). This introduces a degree of misalignment between the simulated and modelled mortality risks. However, this issue is unlikely to have materially affected the conclusions. Findings from the TSD 21 simulation study suggest that mismatches between the external population and the general population lifetables used in modelling have a limited impact on the resulting estimates. Consistent with this, the simulation scenarios in Chapter 5 did not reveal major differences in results attributable to the use of general population lifetables.

It should also be noted that the design of the NCS may be considered to disadvantage some of the methods included. The clearest example of this is the relative survival approach – this approach does not use the simulated external data, and so it is perhaps to be expected that this method would perform less well versus the other methods that make use of these data as

part of the model estimation process. The bias in mean survival for this method, particularly acknowledging the lifetime horizon, is likely to be greater especially in the scenarios where ‘the truth’ was assumed to be the external population. Despite this, it was helpful to include the relative survival method, to understand how it performs in general, even if this NCS design does not fully explore how this one specific method may perform in situations most relevant to the method. A different simulation study designed to evaluate this method specifically may be helpful to better understand its performance.

6.4.3 Considerations for use of methods in the NCS versus HTA practice

In HTA practice, there are several stages that an analyst would go through to determine the most suitable model specification to produce survival extrapolations. For example, an analyst would consider the most suitable parametric model (potentially using statistical goodness-of-fit scores to guide their decision), and would also likely seek clinical input to ascertain if a given survival extrapolation was plausible. Analysts would also plausibly consider any warning messages that might affect use of the resultant model (e.g., if a Bayesian model would ideally be executed for a larger number of simulations). These considerations were not feasible to factor into the NCS – for example, while statistical goodness-of-fit scores (such as AIC and BIC, see Chapter 2) could have been used to select a given parametric model, it is relatively common practice in HTA for models to be chosen which are not the best fitting according to these measures. Ultimately, the findings of the NCS should be interpreted while acknowledging how a given survival model would likely be specified, run, and justified in contemporary HTA practice.

As alluded to previously in the context of the design of the NCS (see Section 6.2.1), an important decision with respect to interpreting the findings was how to define ‘the truth’. In the context of the NCS, it was important to consider multiple “truths” when evaluating method performance, as this acknowledges a key challenge in determining method performance, as ultimately there is uncertainty in what constitutes the ‘true’ survival experience of the population that would be treated in current real-world practice. For example, there will never be a ‘perfect’ data source which could be used to estimate survival for people treated with a given intervention from the date that the HTA decision is made, under all circumstances relevant to the decision maker (including, for example, which treatments are available in subsequent lines of therapy). By exploring performance across two versions of the ‘truth’, the NCS provides a broader evidence base for method users and decision-makers, who can then assess how different approaches perform under varying assumptions. This was considered preferable to asserting a single ground truth, as it reflects the complexity and uncertainty faced in real-world HTA decision-making.

All of the methods considered in this study face challenges related to the generalisability of the data sources they rely on. In practice, decision-making committees would carefully deliberate over which method (or methods) is (or are) most appropriate for a given context. For example, if standard of care has changed substantially over time, or at least survival for people managed with (approximately) the same standard of care, this has clear implications for the suitability of methods that produce survival estimates that are in some way related to older external data. In the case of approaches like blended survival, where survival beyond a certain time point is assumed to follow that of an external population *exactly*, outdated or misaligned external data could lead to implausible projections. In such scenarios, methods that use external data in a less specific way, such as piecewise models that use the longer-term hazard estimates or those incorporating informative shape priors, may offer more accurate (and, importantly, more plausible from a decision-maker perspective) extrapolations.

As noted previously, across all of the simulations, no errors were produced for any of the model fits. Importantly, this does not mean that the methods are incapable of producing errors. Rather, the NCS was designed such that the inputs would reasonably be expected to run without error. It is entirely possible to specify some methods to occasionally produce an error (e.g., by specifying a very small number for the SD inputs to expert opinion method, to force the model produce an error). It was considered most suitable for the NCS to specify each method in such a way that would be expected to work across most simulations. A potential criticism of this approach, however, is that the results of the NCS may be considered inherently less favourable to methods which require more thoughtful or nuanced specification; particularly those that when run with less influential parameters may suffer in terms of their performance (e.g., a weaker prior may ensure the model runs, but may mean that the external evidence is less 'helpful' than it would be if a stronger prior was specified).

While none of the methods had individual runtimes that would render them unsuitable for HTA purposes, from a practical perspective, some of the methods took a relatively long time to execute (most notably, relative survival and M-spline MPES). While in practice analysts may only need to produce a handful of model fits, this was considered an important output of the NCS for several reasons. The ability to fit models efficiently is crucial for iterative model development, and as evidenced in the MPES tutorial paper (discussed in Chapter 5), analysts may need to explore different model specifications, adjust covariates, and experiment with prior distributions to identify the most appropriate approach for the data. Faster runtimes enable this process to be conducted more efficiently. Additionally, sensitivity analysis plays a key role in ensuring the robustness of survival models, and long runtimes can be prohibitive when needing to conduct sensitivity analyses.

6.4.4 Areas for further research

Future research could build upon the findings of this NCS by expanding the scope of evaluation to address additional complexities encountered in real-world applications. One possible area for future research could consider simulating RCT data, so that methods that consider multiple treatment arms (e.g., Guyot's MPES) could be compared to methods that can consider more complex extrapolations for more than one treatment arm. Additionally, future studies could explore more complex data-generating mechanisms that incorporate treatment switching or the impact of subsequent therapies on survival, which are commonly raised as key issues in contemporary HTAs of cancer drugs.

There is also a possible need to specifically address the difference in how external *data* (i.e., facts or figures about something) and external *information* (i.e., communication about how data have been understood or interpreted) are considered as part of the model fitting process. In this NCS, this distinction was not focused on, which may mean that the NCS does not represent a fully comprehensive assessment of the expert opinion method in particular. Relatedly, it may be the case that external *information* would practically be used to help assess how plausible survival estimates appear to be which leverage external *data*. In other words, it may be the case that if external data are available, expert opinion would be reserved for validation purposes.

Outside of the NCS specifically, it should be acknowledged that none of the methods included were designed to adjust the characteristics of the trial population to reflect those of the real-world or decision-making population. Instead, the focus of the included methods was typically on extrapolating survival outcomes for the trial population (i.e., using the external evidence to supplement the trial data, but still with a view that this is the population relevant for decision making). This distinction is important: while such extrapolations may provide internally valid estimates, they may not always be aligned with the estimand of primary interest in HTA, which is survival for the real-world population who would be treated with the intervention undergoing assessment. Consequently, there is a clear need for further research to explore how emergent methods might incorporate population adjustment or other strategies to better reflect the decision-making context. As new approaches continue to be developed, careful attention should be paid to clarifying what each method is truly estimating. An explicit focus on the estimand (i.e., what quantity is being targeted and for whom) is critical to ensure methodological advances can be applied in contemporary HTA decision-making contexts.

6.4.5 Conclusion

In this chapter, an NCS was conducted to evaluate different methods for incorporating external evidence into survival extrapolation for HTA. By using simulation study techniques, the NCS provided a structured comparison of seven methods (plus a reference method [standard

parametric model] that did not use external evidence), assessing their performance across 32 scenarios with varying survival magnitudes, hazard profiles, trial follow-up, sample sizes, and population similarities. Mean bias was used as the primary performance measure, allowing for an objective assessment of how each method compared to a standard parametric model.

The findings demonstrated that different methods may produce wide-ranging estimates of survival subject to the underlying data and assumptions underpinning the model. As reiterated throughout the discussion, and as is perhaps expected, there appears to be no single gold standard method that should always be used – ultimately, performance varies depending on the specific context, and in some cases, methods incorporating external evidence may even be less accurate than methods that do not use external evidence to inform model estimates. This underscores the need for a rigorous assessment of the theoretical strengths and weaknesses of candidate methods, and evidence sources, before their application in HTA. Future research could extend the NCS to more complex scenarios, including multiple treatment arms and treatment switching, further refining our understanding of how external evidence influences survival modelling. Ultimately, these results highlight the necessity of thoughtful model specification when leveraging external evidence for HTA decision-making.

7. Overview, discussion, recommendations, and conclusions

7.1 Overview

In Chapters 1 and 2, I set up the motivation behind the thesis, and discussed key background information related to terminology, evidence sources, and how survival modelling is typically carried out for HTA. These chapters aimed at setting out the necessary background information for the remainder of the thesis, including the specification of the three research questions that guided the development of the thesis, which are discussed in turn within the context of the later chapters.

In Chapters 3 and 4, I detailed the findings from two complimentary reviews aimed at identifying methodology for the use of external evidence in survival analysis for HTA (Chapter 3), and understanding which methods are typically used in contemporary HTAs (Chapter 4). These chapters addressed my first research question: which methods have been previously used to combine external evidence with trial data to make long-term projections of OS? The two reviews allowed an exploration of methods that had been published versus those that had been used in ‘real life’ decision-making. Following identification of methods, the next stage of the thesis concerned the application of the methods themselves.

In Chapter 5, I presented findings from testing the identified methods within the context of three case studies. The goal of the method testing was to apply each method to different ‘real-world’ scenarios where they may be used, to identify exactly which methods require specific input data and settings to be determined, as well as to reveal any potential issues that may be encountered with attempting to use the methods for the purpose of the thesis – to incorporate external evidence into survival modelling in oncology for HTA. This chapter partially addressed my second research question: how do the different methods compare; and provided the necessary groundwork to undertake additional method comparisons in Chapter 6.

In Chapter 6, I provided the findings from an NCS of different methods to incorporate external evidence into survival extrapolation for HTA. The NCS used simulation study techniques to provide a comparison of different methods, from the perspective of a neutral party. By using simulation study techniques, the methods were evaluated under controlled conditions, to better understand where they seem to perform well or less well. This chapter further addressed my second research question (how the different methods compared), and critically provided some evidence to help address my third research question: how should external evidence be incorporated into estimates of survival?

In summary, the first six chapters of the thesis addressed two of the three research questions: (i) which methods have been previously used to combine external evidence with trial data to make long-term projections of OS?, and (ii) how do these methods compare?

7.2 Discussion

The findings from the previous chapters of the thesis are considered together to inform this discussion section of the thesis, which aims to address (to the extent that is possible) the third research question of the thesis: how should external evidence be incorporated into estimates of survival? This discussion considers both the circumstances in which external evidence could be used, as well as the specific methodological approaches which appear to be most appropriate to consider under different conditions, based on the findings of the thesis.

What do we *really* mean by ‘external evidence’?

Throughout the thesis, external evidence is used to describe data and/or information obtained from outside of the pivotal trial(s) of primary interest. When considering a method that leverages external evidence, it is apparent that most of the methods explored in this thesis make the explicit or implicit assumption that the external evidence tells us about the ‘true’ survival experience of a given population. None of the methods explored in this thesis ultimately aimed to address this well-known limitation of using trial-derived estimates of survival for HTA decision-making – i.e., that trial populations typically have a better prognosis than real-world populations. More specifically, the methods explored in this thesis do not allow the user to ‘tell’ the model what they think about the external evidence directly, though some methods can be used to indirectly impose upper and lower limits (e.g., through specification of strong informative priors in a Bayesian framework). The NCS addressed this by showing how the methods performed differently when the external evidence is assumed to represent lower survival versus the clinical trial (see Chapter 6).

Relatedly, it is important to consider precisely what an external evidence source is likely to tell us (and, consequently, what it could be used for). Unsurprisingly, most external evidence sources used in contemporary NICE appraisals were from either older clinical trials or disease registries (see Chapter 4). Given that an external evidence source may only reveal a particular bound for plausible outcomes (most likely, a lower bound based on the expectation that a new intervention would be expected to improve survival), it could be argued that most of the methods identified would seem ill-placed to inform extrapolations of survival for new interventions, and therefore may provide under-estimates of longer-term survival. On the other hand, longer-term survivors that received standard of care may be relatively fitter than the average longer-term survivor treated with a new intervention, since the new intervention may have helped people live longer than they would have done if they were managed with standard of care, and so long-term survival could be over-estimated.

How should external *information* be collected?

One of the most unique external evidence sources considered by some of the methods explored in this thesis is clinical expert opinion. Some of the methods, such as Jackson's MPES, can (hypothetically) leverage combined input from multiple clinical experts, with relatively recent methodology proposed to explore how SEE can be done while being cognisant of potential biases.²⁸¹ Also, TSD 26 was recently published offering clear guidance for conducting SEE to inform long-term survival estimates specifically.⁶⁵ Despite this, in contemporary HTA decision-making by NICE (and likely other agencies internationally), committees frequently rely on the input from a small number of clinical experts who are invited to attend meetings (usually one or two at a 'typical' meeting). This could mean that in practice, a methodologically robust approach to elicit expert opinion is discarded in favour of the opinions expressed by the invited experts (given that these opinions are heard first-hand by the committee).

To an extent, this established practice in HTA decision making calls into question the need for any more advanced methodology that synthesises expert opinion as part of a model fitting process (despite guidance from the NICE DSU demonstrating how expert elicitation could be undertaken such that outputs are suitably robust for HTA decision making).⁶⁵ Instead, it could be argued that processes should evolve to consider a more timely and collaborative approach to elicit and apply expert opinion into survival modelling – for example for NICE STAs, representatives from the company, the EAG, and/or NICE could attend a meeting with clinical experts to see first-hand the responses given to questions raised, ideally with a clearly specified protocol for SEE, and then the outputs could be applied within the company's submission (e.g., to inform model fitting). Unless there is a clear change in how the expert opinion is obtained in the first instance, it would seem unnecessary to dedicate further research efforts to making use of such outputs, if they are effectively cast aside in favour of real-time input from other experts present at an appraisal committee meeting.

What happens if we have multiple sources of 'external evidence'?

Most methods explored in this thesis use only one source of external evidence (see Chapter 3), yet there are typically multiple sources of evidence that could be used to help guide survival extrapolations – though these might be considered useful in different ways. For a given case, there may be (for example) multiple historical trials, disease registry data, life tables, and clinical expert opinion – each of these evidence sources might reveal something about the survival experience for a given population. However, there remains an unanswered question as to how much evidence should be used to inform survival model estimation specifically. In other fields (such as artificial intelligence and machine learning), datasets are usually divided

into ‘training’ and ‘validation’ subsets – here, the ‘validation’ set represents a sample of the full data that is reserved for validation purposes (and therefore does not impact model estimation). As data collection efforts look to provide increasing volumes of evidence for consideration in HTA, this raises the question of if a similar approach should be considered (or at least for cases where there are sufficient data to consider ‘leaving some out’ for validation).

Where there are multiple sources of external data that could be used (e.g., two historical clinical trials), there remains a lack of established guidance on the best approach for integrating such data. Options include naïve pooling, some form of meta-analysis, using each source separately and comparing results, or selecting a ‘preferred’ dataset based on perceived quality or relevance. The principles of evidence-based medicine emphasise the importance of using all available and relevant evidence to inform decision-making and reduce bias through selective exclusion.²⁸² Despite this, each source of data is likely to have its own strengths and limitations (such as differences in baseline characteristics, sample size, and duration of follow-up) which may impact survival estimates if they were to be leveraged as part of a model-fitting process. Further research is needed to develop clearer frameworks for deciding how to incorporate multiple external sources in survival modelling for HTA.

Should we be seeking the ‘best’ model?

Based on the findings of the thesis, one could argue that perhaps focus should be shifted away from identifying the ‘best’ estimate of survival over a lifetime horizon (as is generally considered standard practice, through companies and EAGs specifying a ‘base-case’ analysis), and instead be shifted towards identifying a range of ‘most plausible’ estimates. The full set of methods explored in this thesis ought not to be seen as a list of methods to ‘try’ (echoing sentiments raised by the authors of TSD 21⁵² when describing possible flexible methods for survival analysis), but there is perhaps some merit in presenting alternatives from a number of methods that ‘make sense’ in the context of a given evaluation, rather than historical convention being to fit the ‘standard’ six (or seven) parametric models, and then choosing and defending a ‘most plausible’ extrapolation.

In the context of a NICE STA, different stakeholders naturally bring their own perspectives and motivations to the appraisal process. For a company, there is a strong incentive to submit a base-case analysis that presents its product in the most favourable light, often through possibly optimistic (or at least, “non-pessimistic”) survival extrapolations compared with standard of care. By contrast, the EAG, while independent, may view the company’s modelling with caution, recognising the potential for optimism in its assumptions. Invited clinical experts, even in the absence of any declared conflicts of interest, may be inclined to support access to a promising new treatment, given their focus on improving patient outcomes. Meanwhile, the

appraisal committee must ensure that decisions represent a responsible use of finite public resources. These differing perspectives highlight the inherent challenges in selecting, justifying, and defending a single survival model, further suggesting a range of models may be more suitable for informing decision making.

Despite the potential value in considering a range of plausible methods using different extrapolation approaches (with or without external evidence), more complex methods for HTA can mean new challenges are introduced that are not routinely encountered when using conventional survival analysis methods. These could include, for example, elements of subjectivity in terms of model specification and translation of input data (which is not an issue in conventional regression models fitted to trial data). There is also a practical point when considering transparency in decision-making – some methods rely on using ‘true’ patient-level data (e.g., when considering age and sex for relative survival), and therefore some methods may simply not be possible to re-produce from first principles without access to the full range of inputs. In addition, software package updates may also hinder re-reproducibility efforts – for example, updates to *stanheaders*^{*} can affect Bayesian methods that use Stan, and during the development of this thesis, a new package was launched, removed, and then re-added to The Comprehensive R Archive Network (CRAN) – *expertsurv*.²⁵³

How neutral is the NCS?

In the NCS, efforts were made to provide as neutral a comparison of different methods as possible, in order to determine which specific methods appear to work well or less well under certain circumstances. Despite this, it is near-impossible to remain definitively neutral while also considering a manageable set of DGMs, and the mere fact that some methods were designed to apply under specific scenarios, whereas others were designed for others. The simplest case of this applies to the relative survival approach, which attempts to only account for hazards related to expected mortality (i.e., non-disease-related), and therefore (by design) this method does not attempt in any way to leverage external evidence about the survival experience of the specific population under investigation.

A relative survival approach models the difference in hazards between two data sets (index and reference populations). If these hazard estimates converge, this means that the excess hazard (i.e., the additional hazard experienced by the index population, compared to the reference population) is assumed to be zero. Beyond the point the hazards converge, hazards are dictated entirely by the reference population. Consequently, in principle, one could design a relative survival model that considers use of long-term data from a disease-specific source (e.g., a disease registry), rather than life tables. However, there are complexities involved in

^{*} Note: ‘*stanheaders*’ is an R package which provides the C++ header files (i.e., files that store pre-defined functions) for Stan.

specifying such an approach (related to, for example, limitations in sample size, follow-up, etc.), no such approaches were identified in the literature (see Chapter 3), and importantly, any edits to the methods to change how they work (i.e., improve upon them) would have directly impacted the neutrality of the NCS itself.

One aspect of maintaining neutrality in the NCS was the way methods were implemented. The focus was on using 'vanilla' implementations (i.e., standard, minimally customised versions of each method) to evaluate performance. This avoided the introduction of bias through selective tailoring, such as fitting certain methods multiple times and choosing the best-fitting result, while restricting others to a single parametric form. While this means the results may not reflect each method's optimal performance, the ethos behind comparing vanilla implementations was intentional, as it aimed to minimise the extent to which user input or subjective choices could influence performance. Further research is needed to explore each method in more detail under conditions where optimal model specification and refinement are prioritised over automation and a 'hands-off' approach. However, such work falls outside the scope of this thesis, and most likely falls outside the context of a simulation study or separate NCS.

What about estimating survival for treatments compared in an RCT?

This thesis considers the case of needing to produce extrapolations of OS for a single group of patients (more specifically, a group of people treated with a particular intervention). The reasons for this focus of the thesis are described in detail in Chapter 1, but in brief, modelling multiple treatment groups simultaneously introduces substantial complexities (such as challenges related to treatment switching and non-proportional hazards). For HTA, estimates are required for at least two groups to assess the impact of the proposed (new) health technology against some established standard of care, but there could be even more groups to consider in the context of different subpopulations and/or comparators.

Two methods[†] were excluded from the NCS described in Chapter 6, specifically because they relied upon the existence of RCT data to generate survival estimates (i.e., these methods could not be applied in the context of a single treatment group). The exclusion of these methods from the NCS does not mean that the methods are inappropriate for consideration in HTA – rather, it means that they were developed specifically with an RCT providing the relevant data for the decision problem in mind, which precludes a fair comparison to other methods designed with a single patient group in mind.

In principle, some of the methods that were designed with a single treatment group in mind could be adapted to cover a joint modelling approach. For example, taking a simple case, a

[†] These were Guyot's MPES and 'combine model parameters' by Hawe *et al.* – both described in detail in Chapter 3.

piecewise model could be fitted with treatment arm as a covariate, or the KM estimate could be arm-specific until the specified cut-point. However, determining the most suitable approach to capture the treatment effect requires answering a number of questions to determine a suitable model specification (for example, should the same cut-point and/or distributions be used across both arms?). Perhaps more importantly, considering multiple treatment groups calls into question the relevance of the external evidence – it cannot be the case that the evidence is simultaneously aligned with both the intervention and comparator arms, unless there is no expected difference in survival.

Given that some methods have been specifically designed in the context of an RCT (e.g., Guyot's MPES), this raises the question of what could or should be done in this situation (i.e., where the RCT is fully aligned with the decision problem). Taking an evidence hierarchy perspective, it could be argued that the potential merits of introducing external evidence are outweighed by the benefits of maintaining the comparison of the RCT, and so external evidence ought not to be used when an RCT otherwise provides the best possible evidence of differences in survival between two treatment groups. However, it was precisely the RCT context where some methods were developed, and so while external evidence could be useful, the following questions should be considered:

- How representative is the RCT of real-world practice, including (for example) use of subsequent therapies (which may be more commonly used in the control group within the period of follow-up available at the time of HTA submission)?
- When choosing a particular method, how does it account for the comparator? Does the method estimate treatment effect as part of the extrapolation, or does this need to be applied separately?
- Could the same method be applied for both arms independently? What should be done if the resultant estimates appear contradictory (e.g., crossing survival curves)? If the same method cannot (or should not) be used, is there a clear justification for this?

Even where an RCT is present, there may be issues in using a standard intention-to-treat analysis of OS for decision-making – for example, for RCTs where the control arm could 'switch' to receive the active treatment following disease progression (a fairly common occurrence for cancer trials used to inform HTA, described in TSD 24²⁸³). For these contexts, there are established methods for adjusting for treatment switching, though it is less straightforward how these could (in principle) be combined with methods to incorporate external evidence into survival extrapolation. For example, methods such as two-stage estimation and the rank-preserving structural failure time model both generate counterfactual survival times as part of their estimation process, which can then be used to model long-term survival incorporating external evidence; whereas the inverse probability of censoring

weighting method produces patient-level weights which can be used to fit survival models (i.e., not counterfactual survival times).

When faced with needing to select suitable survival models in an RCT context, standard guidance should be followed regardless of the availability and/or suitability of external evidence. For example, TSD 14 and TSD 21 highlight good practice principles, including (but not limited to) carefully considering the suitability of assuming that PH holds, not overly relying upon statistical goodness-of-fit scores, and assessing the implied treatment effect over time by plotting estimated hazards.^{10, 52}

What about cure models?

As noted previously and throughout this thesis, several methods have been developed specifically to model survival outcomes in the context of treatments that may result in a proportion of patients being 'cured'. In addition, some methods considered the possibility that people may not die of their cancer (i.e., no excess mortality related to the cancer), but they may be at some elevated risk of death owing to the development of comorbidities and/or complications as a consequence of their previous cancer diagnosis and treatment. In practice, this concept has been integrated into cost-effectiveness models using an SMR to increase background mortality rates, relative to the age- and sex-adjusted general population.

Cure-based models, with and without an SMR adjustment, were not explored in detail in this work for several reasons. First, comparing survival modelling approaches across curative and non-curative treatment contexts introduces considerable complexity, particularly due to differences in underlying assumptions and data requirements. Second, many cure-based models are closely aligned with the structure of cost-effectiveness models that explicitly include a 'cured' health state, whereas the focus of this thesis is on OS extrapolation methods rather than full economic model design. Finally, the definition and identification of 'cure' in a practical sense is itself subject to uncertainty and variation across disease areas (including the potential for additional mortality risk in the long-term, even if this is not directly related to the cancer).

In addition to cure models, any modelling approach using an SMR was not considered in detail as part of the thesis. In practice, SMRs are highly context-specific, as the specific value used will depend on the disease area, and they are typically derived from long-term registry data (or possibly based on clinical expert input). In addition, as SMRs are typically applied within cure-based or cure-like modelling frameworks, this raises questions concerning what is truly disease-specific versus expected survival (for example, should a death due to a comorbidity that may or may not be linked to the disease be counted as disease-related or not?). In the context of the NCS conducted in this thesis, SMRs would have been required as both an input

to the data-generating process and an input to the fitted models, and so because of this, the inclusion of an SMR would be unlikely to have had a material impact on the overall findings of the NCS.

Importantly, when considering situations where external evidence implicitly reflects a cure fraction (with or without some kind of adjustment using an SMR or other approach), there is no reason to expect the methods evaluated in this thesis to be completely unsuitable. In such cases, the approaches explored in this thesis should naturally incorporate the evidence of cure that is present in the external evidence; though based on the features of specific methods, some may be expected to perform less well compared to a cure-based method. For example, the expert opinion explored in the NCS requires an assumption of one over-arching parametric model, which may not be sufficiently flexible to take into account a cure-like plateau in survival.

Taking into consideration all of the above, the omission of cure models represents a limitation of this work. Further research would be valuable to compare the performance of the methods included in this thesis with models that specifically attempt to include some kind of cure effect, in scenarios where cure is present, which is a distinct research question separate to the goal of this thesis. There are multiple published sources of guidance on the topic of cure models^{69, 104, 201}, but there is an absence of specific investigations into comparisons of models that explicitly versus implicitly capture some kind of cure effect.

What is the estimand of relevance?

A central issue highlighted throughout the thesis, and in particular when designing and evaluating the results from the NCS, is the question of what constitutes the “truth” that extrapolation methods aim to estimate. In principle, several definitions could be justified. The trial population may be considered the truth, consistent with standard HTA practice where survival models are fitted directly to trial data and assumed to be generalisable. Alternatively, one might argue that an external evidence population may better represent the truth, on the grounds that it more closely reflects real-world outcomes experienced by patients outside the trial setting, though this might not be available for all treatments that are relevant to a given HTA. A further possibility is that the truth is neither of these, but rather some intermediate construct that accounts for differences between the trial and external populations, such as selection effects in trial enrolment or improvements in standard of care over time. In the NCS, only the first two definitions could feasibly be implemented, but even within this restricted framing, it is clear that the choice of estimand has a direct influence on how model performance is assessed and interpreted.

This observation underscores the importance of explicitly defining the estimand of interest when selecting and applying extrapolation methods. Different methods carry implicit

assumptions about which population they are targeting (i.e., trial, external, or some blend of the two). Importantly, these assumptions are not necessarily apparent based on method-based publications – of note, the SLR described in Chapter 3 did not identify any clear discussion from authors with respect to estimands. Without clarity about the estimand, there is a risk of drawing misleading conclusions about how well the model is estimating survival. Looking forward, there is scope for methodological development to create approaches that can better align the estimand with the decision problem, ideally also paired with guidance for method users (e.g., methods that include processes to match trial and external data sources).

What about outcome measures other than OS?

As discussed throughout the thesis, focus was placed on the outcome of OS, and so any methods which evaluated OS-adjacent outcomes and/or relative effects were considered beyond the scope of the thesis, including where OS was estimated indirectly (e.g., as was the case in TA1017²⁸⁴). This restriction on the scope of the thesis was made to ensure that sufficiently similar methods could be identified and compared. That there may be a range of other possible methods to make use of all relevant evidence to inform accurate estimates of OS (directly or otherwise) should be acknowledged, however. For example, a limitation of the standard PartSA is that estimates of OS and PFS are determined by fitting independent survival models. It would therefore only seem fitting to acknowledge that any methods that incorporate external evidence for estimating OS may mean estimates of PFS are no longer appropriate (e.g., if the resultant extrapolated estimates 'cross').

The example of OS and PFS crossing is only one of likely many cases where methods to incorporate external evidence into extrapolations of OS could have unintended consequences for the remainder of the cost-effectiveness analysis. Hypothetically, if one were to produce a real-world adjusted estimate of OS by leveraging external evidence, this could mean other trial-based cost-effectiveness analysis inputs would also require adjustment for consistency, such as: PFS, TTD, dosing information, adverse event occurrence, and HRQL. It may not be feasible or even justifiable to adjust these other aspects of the cost-effectiveness analysis (for example, based on the availability of evidence), and there could be further challenges with facilitating an ITC (e.g., if the within-trial analysis is deemed inappropriate for decision-making, is it then reasonable to apply a within-trial analysis of a different trial to facilitate an ITC?).

Should we perhaps look to estimate survival indirectly?

As discussed in Chapter 4, there are several cases where the use of external evidence is considered 'routine' for decision-making, with respect to estimating OS, even if OS is not estimated directly to inform the economic model. This includes use of life tables to adjust long-term estimates and/or consider survival experience to be comparable to background mortality,

the use of ITCs to facilitate comparisons to non-trial comparators, or where models that rely on indirect estimation of OS through intermediate health states or events is required (e.g., interventions used in a (peri-/neo-) adjuvant setting [i.e., around the time of surgery]). These approaches to incorporating external evidence into survival estimates may not be deemed as such in the context of HTA, since these may be deemed modelling assumptions rather than statistical methodology (for example, including a 'cured' health state may be deemed a modelling assumption, versus a parametric mixture-cure model which may be considered a statistical method). This reveals a potential 'grey area' between the role of the statistician and the health economist in the design and conduct of a cost-effectiveness analysis for HTA.

Noting the above considerations, some NICE appraisals (such as TA746²³⁰ of nivolumab for adjuvant treatment of resected oesophageal or gastro-oesophageal junction cancer, which was identified in Chapter 4), involved multiple elements of external evidence as part of the submitted cost-effectiveness model. The company used published data to capture transitions from 'post-recurrence' to 'dead'. A more recent example (published after the NICE appraisals review) is TA1017²⁸⁴ of pembrolizumab with chemotherapy before surgery (neoadjuvant) then alone after surgery (adjuvant) for treating resectable NSCLC; which included the use of life tables for transitions from 'event-free' to 'dead' for cured patients (noting that the plausibility of a 'cured' proportion of patients was based on clinical expert opinion provided to the company) and registry data for capturing later transitions from 'loco-regional recurrence or progression' to 'dead'.

In these appraisals, it was determined by the submitting companies that the best (or possibly only) way of obtaining reasonable estimates of cost-effectiveness for decision-making was through using multiple sources of evidence to indirectly estimate OS through intermediate health states. This situation is likely to become more common in STAs of interventions for earlier-stage cancers, as increasingly complex cost-effectiveness models may be required to capture the full implications of introducing a new health technology into NHS practice.

How should the estimates be incorporated into an economic model?

Throughout the thesis, focus has been placed primarily on obtaining an estimate of OS over a lifetime horizon; rather than how this could be integrated within an economic model. In HTA practice, most models that are submitted to HTA agencies, such as NICE, are developed in Microsoft Excel, and it is therefore required for the analyst to extract the outputs of the survival curve and integrate these appropriately within a health economic model file.²⁸⁵ For a standard parametric survival model, this is straightforward – typically, analysts will extract the survival model parameters, along with a variance-covariance matrix, and include these as economic model inputs, varying coefficients within probabilistic sensitivity analysis (PSA) through

random sampling under an assumed multivariate Normal distribution. Unlike standard parametric models, there are some additional considerations that apply for most of the methods explored throughout the thesis.

For Bayesian methods, analysts must extract samples from the posterior distribution to appropriately implement PSA. This reflects that uncertainty in Bayesian outputs is represented by the posterior distribution rather than a closed-form parametric distribution. While this approach is commonplace for Bayesian NMA outputs, it remains less typical in survival modelling for HTA, where standard parametric models are usually estimated using frequentist methods. So, while this is not a 'new' consideration for health economic models in general, it would require some additional consideration for implementation (and may contribute to large file sizes, when accounting for the number of samples required for a stable PSA result).

For piecewise methods, further challenges arise when representing uncertainty within PSA. A key issue concerns how to appropriately represent uncertainty across the full model when separate parametric functions are applied to different time periods. Independent sampling of segment-specific parameters could lead to misleading results, as parameter dependence across each segment would not be explicitly represented (i.e., there is no linking parameter estimated in standard implementations of a piecewise model). Where part of the model relies directly on the KM estimate, further issues apply regarding how uncertainty in this segment should be accounted for. Beyond parameter uncertainty, cut-point selection introduces structural uncertainty that is not reflected within PSA, though exploring alternative cut-points via deterministic scenario analyses has been considered in the past (e.g., in TA736, which was captured within the NICE STA review, but this was not a piecewise model that made use of external evidence as part of model estimation). No specific guidance on uncertainty propagation for piecewise models was identified in the NICE STA review (see Chapter 4), suggesting that further research is needed to understand the extent to which uncertainty may be misrepresented in piecewise modelling approaches.

For models incorporating general population mortality, these inputs are typically treated as fixed within economic models, as they are derived from large population datasets and uncertainty is assumed negligible. The TSD 21 simulation study demonstrated that the choice of reference population appeared to have only a modest impact on overall results, suggesting limited consequences from omitting parameter uncertainty in life table-based estimates.⁵² However, this assumption may not hold when applying SMR adjustments, where the magnitude of the SMR may introduce substantial additional uncertainty (though it should be noted that owing to the exploratory nature of models that include an SMR adjustment, it may be challenging to truly capture the uncertainty associated with the choice of SMR in the context

of the economic model, as the specification of an SMR is itself an assumption about additional excess mortality).

In addition to how survival estimates are incorporated while accounting for sensitivity analysis, it is also important to consider how OS inputs interact with other model inputs. These include, but are not limited to, other time-to-event endpoints (such as PFS and TTD), utility values, and adverse event rates. Adjustments to OS may mean that there could or should be meaningful changes in other parameters that are implicitly linked to survival outcomes.

For example, PartSA models commonly estimate OS and PFS independently, with a logical constraint ensuring that PFS does not exceed OS at any time point. If external evidence reduces long-term OS estimates, it is unclear whether PFS (and/or TTD) should also be adjusted for consistency. This relates closely to the estimand of interest for decision making. If the objective is to estimate OS in a real-world treated population, then using an OS model that leverages external evidence as part of its estimation with purely trial-based estimates of PFS or TTD may introduce inconsistency (as the OS reflects both trial and external evidence). Extending this further, utility values may differ across populations (e.g., poorer prognosis may be associated with lower baseline utility), and adverse event incidence may also vary with prognosis.

These additional model inputs are likely to be intrinsically linked to OS estimation, yet it remains unclear whether and how they should be adjusted to reflect the incorporation of external evidence. Obtaining external evidence for these other model inputs is often infeasible due to data limitations, particularly for inputs like utility values, given that EQ-5D data are rarely captured in real-world settings in a manner that would allow for the re-estimation of health state utility values. Consequently, in HTA practice it is important to consider how OS adjustments may influence other key drivers of the economic model and the resulting cost-effectiveness conclusions, though for now this would be limited to commentary rather than further statistical analysis.

Can we combine external evidence survival analysis methods with other HTA methods?

A key practical question arising from this thesis is whether survival models incorporating external evidence can be combined with other HTA methodologies that may be relevant based on appraisal-specific context. In principle, provided these methods do not directly interfere with the estimation of OS for a single group of patients, then they should be relatively straightforward to incorporate within the same economic model.

A relatively situation is consideration of an NMA. From a mechanical perspective, there is no reason why HRs derived from an NMA could not be combined with a base curve produced via fitting a survival model that leverage external evidence. However, it is important to consider

the appropriateness of taking NMA estimates that are typically derived using RCT evidence, and ‘transporting’ the treatment effect to apply to the base curve obtained from a combination of trial and external evidence. For example, if the external evidence suggests that there are important differences in treatment effect modifiers between the trial and external populations, then adjusting the trial-based survival estimate using external evidence would have knock-on implications for the applicability of the trial-based NMA. Similar considerations would also apply when using a MAIC, although implementation may be more complex due to the reweighting of IPD – in a survival model that uses both trial and external data, it is no longer straightforward to consider which patients are still ‘at risk’ over time. Similar challenges would likely also apply for other PAIC methods (for example, for STC, the regression model would need to be based on one specific data set, rather than a combination of trial and external evidence).

More complex problems apply when considering methods that adjust for treatment switching or subsequent therapies. These approaches aim to estimate counterfactual treatment effects or survival times specifically within the trial population to account for later use of treatments, whereas the external evidence methods explored in this thesis typically seek to improve long-term survival estimation or generalisability. These methods may be viewed as trying to address similar problems, but combining them is difficult for similar reasons as per PAIC methods. It would therefore be most suitable to determine whether it is more important to adjust for subsequent use of treatments (e.g., switching), or to consider external evidence – doing both at the same time may produce misleading results.

Treatment effect waning presents a further area of interaction. Some external evidence approaches, such as those implemented within the M-spline MPES approach by Jackson (2023)²¹⁶, allow waning to be applied within the survival modelling framework (albeit post estimation, as described in the MPES tutorial paper [see Bullement *et al.*, 2024]⁴). Any of the methods could be adjusted *post-hoc* to account for treatment effect waning, but doing so creates additional challenges when representing uncertainty, as waning assumptions are often implemented deterministically or explored only through scenario analyses.

Finally, some methods may be fundamentally incompatible with most of the methods explored in this thesis. For example, FP-NMA generates a treatment-specific baseline survival function as part of the modelling process, and so it is not possible to then subsequently incorporate external evidence into survival estimates without changing the overall interpretation of the FP-NMA. In such situations, attempts to combine approaches are likely not possible, and so analysts may need to choose which issue seems the most pertinent to decision making, or consider presenting a range of scenarios where simpler approaches are taken to explore

uncertainty more holistically (e.g., adopting a simpler ITC approach so that survival estimates leveraging external evidence could be considered).

Overall, while many external evidence survival modelling approaches could be combined with established HTA methods, doing so may introduce additional complexity in implementation and interpretation. Importantly, it is necessary for the analyst to determine what are the main drivers of the cost-effectiveness results, and based on this determine which method (or methods) is (or are) most likely to provide valuable insights for the decision maker. The appropriateness of combining HTA methods depends not only on practical feasibility but also on the primary estimand of interest relevant to the economic model.

What can be done about SACT data collected via the CDF?

This thesis discusses the potential role of external evidence to help aid survival extrapolation for use in HTA, noting that some of the external evidence sources may be best described as RWE (see Chapter 2 for definitions). In recent years, several reports have explored how RWE might be used more extensively (or perhaps more effectively) to inform HTA decision making. Particular attention has been given to the use of SACT data, which are typically collected when interventions are recommended for use in the CDF.

In 2024, a pilot study was conducted by the NICE DSU exploring how SACT data could be used in cost-effectiveness analyses during re-evaluations of four technologies exiting the CDF, though this initial pilot study was not published. The existence of this report was only apparent when a subsequent report was published in 2025 by the DSU, which explored the feasibility and value of using SACT data across a larger sample of eight case studies.²⁸⁶ The case studies involved digitising KM estimates of OS and TTD (collected in SACT), to produce parametric survival models. Data for PFS are not collected in SACT, so PFS curves were estimated based on assumed relationships between data sources and endpoints.

The DSU report found that in most cases, analyses based on the SACT data resulted lower incremental QALYs (through typically lower OS estimates), and the impact on the ICER was mixed, but could lead to different decisions being reached (at least hypothetically). The authors highlighted that while trial-based analyses are likely to remain the primary analysis for NICE decision-making, analyses based on SACT data could provide useful scenarios, particularly if improvement to SACT data access could be achieved allowing for more detailed analysis (e.g., if data on disease progression could be added).

While this thesis focuses on situations in which trial data remain the primary source for estimating OS, it would be remiss not to acknowledge that there may be circumstances in which the pivotal trial data are no longer the most informative source for HTA decision making. This may occur, for example, where a managed access agreement generates a substantial

body of RWE. In practice, however, such situations may be relatively uncommon. Given the typical timelines between initial appraisal and CDF re-evaluation, trial data will usually still provide the longest follow-up available for the new intervention, and it would therefore seem inappropriate to disregard this data source entirely.

Following the DSU report, an HTA Lab report was published (latest version September 2025), which considered the DSU's findings alongside broader implications for NICE methods, processes, and decision making.²⁸⁷ This report included several recommendations regarding the collection and use of SACT data. Among these was a specific recommendation to: *“Develop technical guidance on how to incorporate SACT-based survival insights into trial-based survival curves by applying a factor to reduce absolute survival, while retaining the original curve shape and the benefit of longer follow-up in the trial data”* (HTA Lab project working group, 2025, p. 16).²⁸⁷

This recommendation aligns closely with the broader research question explored in this thesis, namely how external evidence might be used to inform survival extrapolation when trial data alone are insufficient. However, the aim of this thesis is not specifically to generate lower survival estimates relative to models based solely on trial data, but rather to explore principled ways in which external evidence can be incorporated into survival modelling more generally. It will be interesting to observe, in the coming years, whether the HTA Lab recommendations are implemented and how they may ultimately influence HTA decision making for technologies that enter the CDF.

7.3 Method-specific recommendations

Based on the findings presented hitherto, some recommendations can be made concerning the suitability of the methods explored as part of this thesis in practice for HTA, which are described in turn below.

Background mortality and relative survival

Background mortality data are publicly available, and so in principle it makes sense to endeavour to account for this when producing survival extrapolations for HTA, even if only for face validity purposes. Of note, background mortality data should not be conflated with external evidence about the population of interest, and so its merits cannot necessarily be compared to those associated with incorporating other sources of evidence (e.g., data from a disease registry). Methods that incorporate background mortality are generally expected to outperform a standard parametric model in the long-term, and so would generally be encouraged to consider in favour of an approach that does not leverage background mortality data (in keeping with recommendations reached by van Oostrum *et al.*, [2021]).²⁰² However,

only adjusting for background mortality may have relatively little effect on model estimates as these data do not reflect the population of primary interest.

Despite the potential merits in using a relative survival approach, there may be some issues that could be encountered. As part of the NCS (see Chapter 6), it was noted that the published tutorial by Sweeting *et al.* can take a long time to run, and does not automatically allow for incorporation of different population life tables – to address this, code is provided in an appendix to show how UK life tables could be substituted for the default US life tables (see Appendix F). In addition, the findings from the NCS suggested that the mechanism by which background mortality information is applied does not always lead to improved accuracy of extrapolations (for example, using a relative survival parametric model, versus a standard parametric model without external evidence). Therefore, a relative survival approach should be considered carefully, and ideally compared to a standard parametric model approach to ensure the results exhibit face validity.

Relative survival modelling for HTA purposes specifically may be a useful avenue for further research – the simulation study reported in TSD 21⁵² is, to the best of the author’s knowledge, the only simulation study in addition to the study reported in this thesis that investigates this approach in an HTA context. In addition, Chen *et al.*, (2024) compared models for survival extrapolation under both all-cause and relative survival frameworks, using long-term registry data from the Swedish Cancer Registry.²⁸⁸ The results of this study suggested that relative survival frameworks, particularly when combined with spline models, provided closer alignment with observed long-term outcomes. However, the NCS undertaken to inform this thesis showed that under the scenarios explored, relative survival approaches may yield results that are less accurate compared to standard models fitted in an all-cause mortality framework. While this is likely due to the specification of excess hazard functions that generate long-term estimates that sit consistently above zero, this is nevertheless an area for further research to better understand when these approaches may be beneficial for use in HTA.

Piecewise approaches

Piecewise methods are arguably a good option for consideration owing to their relative simplicity, particularly where data are limited, though resultant hazard functions are unlikely to be fully realistic. The summary data method could be especially useful in cases where very few data exist to help guide extrapolations, but this comes at the cost of needing to assume a simplistic extrapolation (e.g., an exponential ‘tail’). Piecewise methods fitted to patient-level data from multiple sources also rely on selecting at least two different parametric models (or specifying a KM estimate followed by an extrapolated tail) – as such, selecting a best fitting model may be challenging. Despite the limitations, the piecewise methods explored in the

NCS (see Chapter 6), generally performed well, and may therefore represent a good option for a simple, transparent approach to incorporating external evidence.

Blended survival

Some methods do not make sense when applied under specific circumstances which should be recognised. For example, the blended survival method only produces sensible outputs when the estimated survival for the external population is never higher than that of the trial population, in the time period covered by the blending interval. It is therefore 'obvious' that applying the blended survival method in cases where this is unlikely to apply could generate increasing estimates of survival over time. The ability for this method to exhibit increasing survival proportions over time, as seen in the NCS (see Chapter 6), is a clear indication that there will be some cases where the blended survival method is unsuitable (at the very least, for some specifications of this method). Despite this, it should be noted that the blended survival method may be useful under specific circumstances, but caution should be exercised if using this method in practice, given that the NCS highlights the consequences of applying this method without deliberately determining suitable data and model settings.

Multi-parameter evidence synthesis approaches

With respect to MPES methods (covering both Guyot's and Jackson's MPES), as part of the MPES tutorial (Bullement *et al.*, [2024]), the B-A-K-E recommendations were developed, which are re-produced in Figure 7-1.⁴

Figure 7-1: B-A-K-E: Recommendations for analysts (MPES methods)

Basis: A clear rationale

- Due to its complexity, it is recommended that analysts carefully consider the expected benefits of specifying a more complex model like an MPES approach, and how any losses (e.g., the risk of 'over-fitting') are warranted, and obtain clinical input of the plausibility of the assumptions underpinning the model. Analysts should provide results alongside a detailed justification for why the MPES method was deemed suitable. In addition, when relying on any parametric model for extrapolation, the rationale for the pattern of long-term hazards should be made clear (e.g., do we expect to see constant/increasing/decreasing/complex hazards after the end of trial follow-up?).

Alternative: Consider other specifications

- The approach of Guyot *et al.* is one possible specification of an MPES approach, and *survextrap* by Jackson is another. Both of these should be considered, as MPES is not a single method - it is a bespoke approach to estimating survival using multiple sources of evidence. Other modifications, such as accounting for a treatment effect over time, relative survival and/or cure proportions, may also be relevant. Be aware that alternative specifications can yield drastically different results, including for Guyot's MPES potentially implausible extrapolations.

Knots: Avoid getting tied up

- Seemingly small changes in the number of knots and the knot locations can have a profound impact on survival estimates (particularly for Guyot's MPES). Guyot's MPES requires user-specification of the number of knots, unlike Jackson's M-spline implementation where the number of knots can be determined automatically. Analysts should ideally explore alternative knot positionings before settling on a final model. Inspection of hazard plots may help determine optimal knot locations.

Evidence: Garbage In, Garbage Out

- An MPES approach is only capable of generating plausible long-term estimates of survival if the inputs are sensible. It is eminently possible to generate unrealistic extrapolations using an MPES approach, but this does not mean an MPES approach is itself flawed. Analysts are therefore urged to carefully consider the inputs used to inform an MPES model (e.g., using established data quality and/or risk of bias checklists).

Key: MPES, multi-parameter evidence synthesis.

Source: Excerpted and reprinted with permission from Pharmacoeconomics. (Copyright © 2024, Springer Nature Switzerland AG). Bullement A, Edmondson-Jones M, Guyot P, Welton NJ, Baio G, Stevenson M, Latimer NR. MPES-R: Multi-Parameter Evidence Synthesis in R for Survival Extrapolation-A Tutorial. *Pharmacoeconomics*. 2024 Dec;42(12):1317-1327. doi: 10.1007/s40273-024-01425-4.

In addition, as shown in the MPES tutorial paper, Guyot's MPES is capable of producing increasing proportions of survivors over time when the active arm is not linked to the control arm.⁴ It can be challenging to anticipate when and why these types of logic errors could occur, owing to the underlying complexity of the MPES model specified by Guyot *et al.* relative to other methods (e.g., the blended survival model). It is therefore recommended that researchers carefully investigate the outputs of the blended survival and Guyot's MPES methods to sense check that survival estimates are constantly decreasing – within the group of methods explored in Chapters 5 and 6, these appear to be the only methods capable of generating negative hazards.

Jackson's MPES does not suffer from the same issues as per Guyot's MPES (described above) and based on the NCS (see Chapter 6) generally produced low estimates of bias. Since Jackson's MPES is a relatively recent method there are limited applications in the

literature. Recently, a pre-print of a simulation study by Timmins *et al.* (2025) was made available online, which assessed the performance of Jackson's MPES specifically.²⁸⁹ While this simulation study was not considered to provide a fully comprehensive assessment of all possible modelling choices, the authors concluded that Jackson's MPES demonstrated “*good performance across a range of realistic scenarios*” (Timmins *et al.*, 2025, p.28).²⁸⁹ Of note, the NCS demonstrated relatively little difference in performance between this method and a standard piecewise method, though the NCS compared relatively simplistic method specifications and data generation scenarios.

Expert opinion

Methods that rely on eliciting expert opinion are likely to be subject to increased scrutiny versus methods that leverage external data, since expert opinion-based methods rely on an additional layer of methodology to obtain the expert opinion in the first instance. Elicitation techniques were not explored as part of this thesis, but it should be noted that expert opinion should be termed ‘external information’ as opposed to ‘external data’, and therefore needs to be handled as such (e.g., most of the methods explored in this thesis would not be compatible with expert opinion if this was the only source of external evidence). The NCS did not make this distinction, and so investigating the implications of using external *data* versus external *evidence*, as well as a combination of both, remains an outstanding area for future research.

Informative prior

The informative prior method relies on an assumption of the same ‘shape’ parameter across sources. This may be entirely reasonable, or could prove implausible (especially in the context of using data for groups of patients with different survival trajectories – e.g., patients that receive interventions with different mechanistic properties or that could be cured). The strength of prior specified is, to an extent, at the discretion of the analyst, and for the method to generate estimates that are different to a standard parametric model fitted to the trial data only, the prior typically needs to be strong (see Bullement & Kearns, [2022]).¹ In this situation, the informative prior methods may be seen as essentially setting the shape parameter to a fixed value, which may call into question the overall face validity of the method (i.e., if the method only produces reasonable estimates when it is ‘told’ which parameters to use, is this appropriate for decision making?).

7.4 General recommendations

No matter the method (or methods) used, it is important to consider the balance of methodological complexity and transparency. In other words, what good is a very complex method if decision makers cannot ‘trust’ its outputs for decision making? Conversely, the complexity of methodologies should not be restricted without good cause. Decision-makers

do not necessarily need to be able to explain the intricacies of a given method, but need to have confidence that the method generates reliable results. Given that true survival data will never be known, this places further importance on the role of simulation studies and NCSs to provide evidence concerning model performance under a range of different circumstances.

As described in Chapter 6, the NCS aimed to investigate the performance of different methods under a range of HTA-relevant scenarios. However, the range of scenarios explored was by no means exhaustive. For example, the NCS only considered two measures of external evidence similarity (by design, for ease of interpretation). Given this, and the current state of play for methods applied in current HTA practice (Chapter 4), there needs to ideally be even more research using and comparing *multiple* different methods, to better understand how the methods work, and critically how they might be used in real-life HTA decision making.

A recurring theme throughout this thesis is the absence of any methodological approach which is capable of explicitly addressing the difference between real-world and trial populations in terms of survival experience. There is perhaps cause to suggest that as a field, efforts should be moved away from methods to estimate OS for a population receiving a new intervention, and instead focus on estimating ‘current care’ survival and then apply a treatment effect through whatever means are suitable and/or necessary (e.g., via an ITC). Avoiding fitting a survival model to the intervention group from the pivotal clinical trial would represent a paradigm shift from current convention in HTA, but nonetheless may lead to more representative estimates of survival outcomes for real-world populations. It may be argued, however, that there may be conflicting rationale to approach survival estimation for HTA in this way – for example, the act of not fitting a survival model to the intervention group from the pivotal clinical trial could be seen as a statement on the non-generalisability of the clinical trial population. Such an approach could introduce a plethora of issues for decision-making (e.g., if OS is not generalisable, then what about the other outcomes or data collected?).

Another complicating factor for NICE decision-making is the application of the severity modifier, which was introduced in January 2022 as part of NICE’s updated methods for technology appraisal (i.e., adoption of ‘the manual’).³¹ In brief, the severity modifier is intended to reflect societal preferences for allocating greater weight to health gains in more severe conditions, which is put into practice by applying multipliers of either x1.2 or x1.7 to incremental QALYs, based on the estimated total QALYs for the comparator group (i.e., current standard of care). Thus, in principle, if external evidence can be leveraged to demonstrate poorer outcomes under standard of care versus relying on extrapolations based on trial data alone, a higher weighting of incremental benefits may be achieved. Consequently, there may be a strategic motivation for companies to use a method that leverages external evidence if this could lead to qualifying for a greater severity modifier, even if this comes at the cost of the

magnitude of the incremental QALY gain projected for the intervention undergoing assessment.

Outside of recommendations for currently-available methods, and though perhaps somewhat obvious, methods that come ‘pre-packaged’ in easy-to-use software (e.g., the M-spline MPES approach by Jackson, using *survextrap* in R²¹⁶), are much more likely to be used versus older methods that are not made available in the same way (e.g., Guyot’s MPES¹⁵³). In this thesis, any bias introduced by ‘ease of use’ was addressed head-on by implementing all relevant methods into the same R-based analysis (with the publication of the MPES tutorial intending to make both MPES approaches feasible for use in practical applications). However, this does not address the underlying differences in difficulty for users to apply different methods, and it is likely the case that users will naturally gravitate towards methods that can be easily produced; and some methods, despite the potential advantages, are unlikely to be used much if they are seen to be prohibitively difficult to apply in practice.

Throughout the development of this thesis, several challenges were identified related to the availability of scripts to execute specific models, which included:

- A website that was no longer active, which re-directed to a plumbing company based in North Carolina, instead of a website related to WinBUGS development.
- Code that existed based on personal communication with study authors, which upon investigation led to clarification of an error that was in the original publication.
- An R package which was withdrawn and later re-instated on CRAN.

If any new methods are proposed in the literature, it would seem especially important to ensure that they can be readily applied – this means dedicating ample time towards preparing worked examples, dummy data, clearly commented code, and ideally publishing approaches as specific software packages (e.g., as per *survextrap*).²¹⁶ Sharing code via online repositories such as GitHub, would also likely help avoid issues with privately-hosted websites (though there is no guarantee that GitHub will be hosted indefinitely).

In addition to usability, any new methods should also ideally identify their use case as part of the publication – i.e., what data context(s) is(are) the new method attempting to address? What has been done to address this historically, and why is the new method expected to be ‘better’ than what has been done previously? Acknowledging that simulation studies conducted alongside proposing new methods are somewhat problematic (as described in detail by Boulesteix *et al.*²⁶²), such comparisons would allow for a clear documentation of what the new method is at least trying to achieve. Consequently, researchers aiming to choose between both established and emerging methods would stand a better chance in terms of fully comprehending which methods would seem to be relevant for their work.

It is also worth noting that it can take a long time from when any ‘new’ method is first proposed through to it being considered one of the standard tools for use in HTA. A study by Jiu *et al.*, (2024) describes a three-stage process for new method adoption in HTA: (i) identification, (ii) development, and (iii) implementation.²⁹⁰ The implementation phase in particular comprises various activities that would all help with getting any new method used in HTA practice, covering activities such as conference talks, workshops, tutorials, training materials, and published case studies. These activities consider an audience that extends beyond method users, as all relevant stakeholders (including, for example, decision makers and reviewers) need to also be comfortable with new methods for them to be truly accepted for use in HTA.

A further consideration for any new methodological developments is to consider how the method could ultimately be used. Convention in HTA is for a health economist to take survival analysis results from a statistician’s statistical analysis, and put these into a (Microsoft Excel) cost-effectiveness model. Outputs from any new statistical software packages should therefore be mindful of this workflow, and ideally be as simple as possible to apply within models used for decision-making – including considerations for sensitivity analyses (e.g., outputs for sampling in probabilistic sensitivity analysis).

The focus of this thesis was on methods to estimate lifetime OS for use in HTA. As noted throughout the thesis (and in Chapter 4 especially), there are plenty of cases where OS is not directly estimated for use in cost-effectiveness modelling, though OS is instead estimated indirectly through intermediate health states. There is a distinct lack of guidance for modelling in earlier-stage cancers, where (as highlighted in the context of TA1017²⁸⁴ as a recent example) multiple evidence types have been used in decision-making. Further research into the use of external evidence from an economic modelling perspective (rather than a statistical methodology perspective) may be a useful avenue for further research, though falls outside the scope of this thesis.

In the context of HTA, survival modelling is often based on data from an RCT. My thesis focuses on estimating survival for a single treatment group only. In principle, as described previously, the same method could be applied independently to multiple treatment arms in an RCT – this is an approach frequently used in HTAs where separate extrapolations are performed for each arm. However, alternative strategies could be considered when both treatment and control arm(s) data are available, and are relevant to the HTA decision problem.

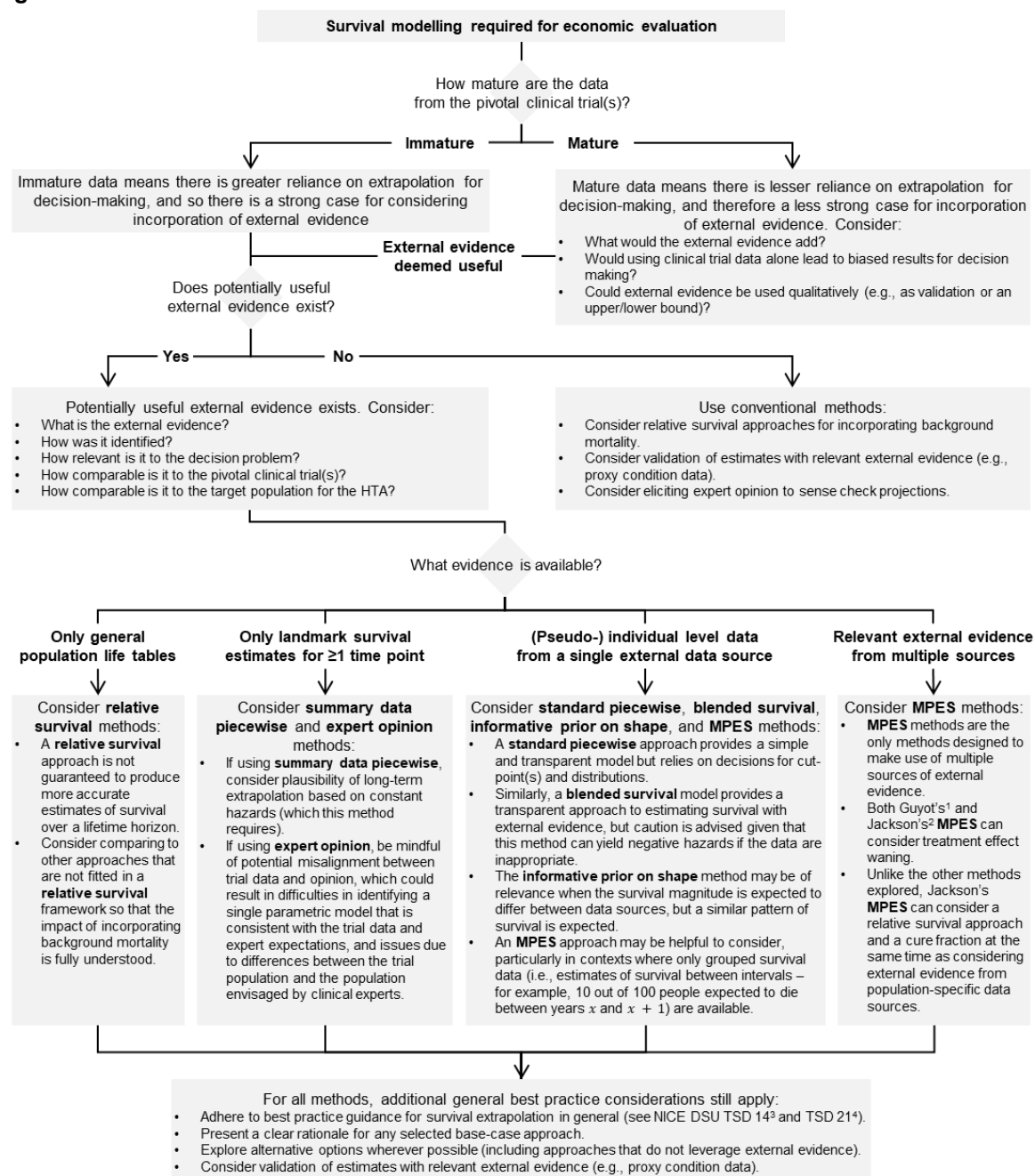
As an example, one approach could be to use the relative treatment effect estimated from the RCT (e.g., an HR) and apply this to a survival extrapolation based on external data. For example, in NICE’s ongoing assessment of isatuximab + pomalidomide + dexamethasone for treating relapsed and refractory MM (NICE ID4067²⁹¹ – a CDF review of TA658⁵⁸), the

committee preferred a survival extrapolation based on data from SACT for the intervention group (isatuximab + pomalidomide + dexamethasone), but an HR obtained from the pivotal clinical trial (ICARIA-MM) to determine survival for one of the comparator regimens (pomalidomide + dexamethasone).²⁹¹ To an extent, this approach addresses the issue of efficacy versus effectiveness, which, as described previously, is often overlooked in HTAs. However, this is most often due to a distinct lack of data that could practically be used in lieu of the clinical trial data to directly estimate OS. This particular appraisal (ID4067) is somewhat unique in that it is a CDF review, part of which involved prior collection of data via SACT as opposed to a standard STA where such data would be unavailable.

While methods that allow for the estimation of survival for multiple treatment groups simultaneously could be extremely valuable for HTA, exploring and evaluating such methods in the context of my research would be extremely challenging. Firstly, in HTA decision-making, we often have little more than the RCT itself to inform both survival extrapolation *and* treatment effect estimation. Furthermore, we may only have external data that are relevant for one treatment group (most likely [though not always], the control arm). In addition, capturing any kind of non-PH relationship between treatment arms likely necessitates a somewhat complex model specification, which may essentially be equivalent to independent models fitted separately to each arm (albeit with extra steps). Ultimately, further research is needed to explore the intersection between survival extrapolation and modelling treatment effects over time, but such research is likely to be fraught with practical difficulties with disentangling absolute survival and survival benefit.

An overall summary of recommendations based on the findings of the thesis is provided in Figure 7-2, in the style of the survival model selection process algorithm provided in TSD 14.¹⁰

Figure 7-2: Flowchart of recommendations for use of external evidence for survival in HTA



Key: DSU, Decision Support Unit; HTA, health technology assessment; MPES, multi-parameter evidence synthesis; NICE, National Institute for Health and Care Excellence; TSD, Technical Support Document.

Sources: ¹Guyot P, Ades AE, Beasley M, *et al.* Extrapolation of Survival Curves from Cancer Trials Using External Information. *Med Decis Making* 2017; 37: 353-366; ²Jackson CH. survexptrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology* 2023; 23: 282; ³Latimer NL. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data, https://www.ncbi.nlm.nih.gov/books/NBK395885/pdf/Bookshelf_NBK395885.pdf (2011, accessed 2025/08/29).; ⁴Rutherford MJ, Lambert PC, Sweeting MJ, *et al.* NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis, <https://sheffield.ac.uk/media/34188/download?attachment> (2020, accessed 29-Aug-2025).

As echoed throughout the thesis, each of the different methods explored may be suitable under certain circumstances. To help illustrate this, below are some example situations, and how the findings of the thesis (as summarised in Figure 7-2) could help guide analysts to select suitable methodology for survival extrapolation:

Hypothetical scenario 1: Late-stage cancer with mature trial data (observed >75% of events), with some older clinical trial data in similar populations.

- Unclear value of external evidence, given that the trial data are considered mature.
- Consider using a relative survival approach, to incorporate background mortality.
- Any other external evidence may be best reserved for validation or context setting.

Hypothetical scenario 2: Advanced cancer with maturing data (observed around 50% of events), with some long-term registry or trial data for current care.

- Assuming long-term hazards are expected to be similar, these could be leveraged via a piecewise or MPES modelling approach.
- If survival experience is expected to be near-identical in the long-term (in terms of percentage alive), then a blended survival approach may be useful to consider.
- If the shape parameters are expected to be similar across data sources, consider the informative prior method.

Hypothetical scenario 3: Rare cancer with immature data (observed around 30% of events), with some clinical expert opinion concerning likely long-term survival at specific landmarks.

- Consider the expert opinion method, noting that this introduces challenges concerning validation (since expert opinion directs informs model estimation).
- Consider a simplistic piecewise approach using summary data to determine an exponential tail (acknowledging the assumptions made in terms of long-term hazards).

Of course, each real-life situation will be different, but these hypothetical examples illustrate the key trade-offs that analysts will need to make when considering different survival extrapolation approaches, and critically how their choices can be defended. The choice of a specific method should ideally ‘make sense’ in terms of what the external evidence tells us, and how this could be leveraged. For example, a piecewise model may be viewed as somewhat basic, but it may be preferred over a complex MPES approach if the underlying assumptions used to generate the model estimates are simple to communicate.

These hypothetical situations make the important assumption that the trial data are reasonably representative of the population for whom the HTA recommendation would apply. If this is not the case, other approaches such as using propensity score weighting or using other data sources as a reference curve may be relevant (though are beyond the scope of this thesis). It is also worth noting that these hypothetical situations assume that directly estimating OS is necessary for the economic model – some situations may realistically require a different type of economic model structure to appropriately estimate long-term costs and outcomes.

7.5 Conclusions

Estimates of OS over a lifetime horizon (i.e., 'survival extrapolations') are often influential on cost effectiveness results which inform HTA decision making. External evidence (that is, data or information collected outside the main clinical trial) can help with producing realistic extrapolations, but methods to incorporate external evidence are not well-established, and there are no standard guidelines for method selection. This thesis aimed to explore how external evidence could be incorporated into survival extrapolations, focusing on cancer interventions. The structure of the thesis, described in Chapter 1, is centred around three key research questions, which considered which methods had been previously used, how they compared, and how external evidence should be incorporated into estimates of survival for HTA decision-making. Important background context to ensure understanding of key terminology used throughout the thesis is provided in Chapter 2.

In 2017, Jackson *et al.* published a review of methods for extrapolating survival from randomised trials using external data.²⁶ From this review, a range of methods were highlighted, yet it was unclear which methods had been developed since this review was undertaken, if any methods had been 'missed', or if any unpublished methods had been used in recent NICE appraisals. Chapters 3 and 4 provide comprehensive reviews of methods that could be used to combine external evidence with trial data to make long-term projections of OS in HTA, highlighting limited use of published methodology in contemporary HTAs.

Prior to this thesis, there has been no formal comparison of different methods to leverage external evidence for survival extrapolation. To compare the different methods, three different case studies were considered in Chapter 5 to fit a variety of different model specifications, with the goal of fully understanding key assumptions and data requirements and how these influence results. Part of this work involved developing a tutorial to execute Guyot's MPES using R software.⁴ Chapter 5 ultimately set up the necessary groundwork for the NCS (Chapter 6). The NCS then aimed to formally evaluate the performance of comparable methods that leverage external evidence, described in Chapter 6.

The findings from the method comparison and NCS were used to inform recommendations detailed in this concluding chapter of the thesis. The conditions under which external evidence may be beneficial are described in this chapter, as well as a summary of which methods appear to be most appropriate in specific situations, drawing on the findings presented throughout the thesis. An important conclusion of the work is also that the use of external evidence does not guarantee accurate extrapolations, and therefore careful consideration and clear justification is required if such methods are to be used in contemporary HTA practice. In addition, the importance of defining the estimand of interest was demonstrated clearly through

the NCS, which showed performance varies markedly dependent on what exactly is the goal of producing survival extrapolations.

In conclusion, this thesis has explored a range of methods for incorporating external evidence into survival extrapolation for HTA, in line with the original research aims. As part of the method exploration, it was possible to document which methods exist, whether they have been used, where they could be applied, how they compare, how they perform, and, ultimately, when they could be considered in practice. The findings of this research could empower analysts to more appropriately determine survival extrapolation methods for use in HTA. Critically, this thesis sets out expectations for best practice when developing new methods for survival extrapolation, which could (if followed) lead to an improvement in how new methods are perceived and used in HTA decision-making. Ultimately, if methods are better understood and their relative strengths and weaknesses are clearly communicated, this should lead to increased procedural efficiency in determining if a new treatment comprises cost-effective use of NHS resources.

References

1. Bullement A and Kearns B. Incorporating external trial data to improve survival extrapolations: a pilot study of the COU-AA-301 trial. *Health Services and Outcomes Research Methodology* 2022. DOI: 10.1007/s10742-021-00264-6.
2. Vadgama S, Mann J, Bashir Z, et al. Predicting Survival for Chimeric Antigen Receptor T-Cell Therapy: A Validation of Survival Models Using Follow-Up Data From ZUMA-1. *Value Health* 2022; 25: 1010-1017. 2022/06/07. DOI: 10.1016/j.jval.2021.10.015.
3. Bullement A, Stevenson MD, Baio G, et al. A Systematic Review of Methods to Incorporate External Evidence into Trial-Based Survival Extrapolations for Health Technology Assessment. *Med Decis Making* 2023; 272989x231168618. 2023/05/01. DOI: 10.1177/0272989x231168618.
4. Bullement A, Edmondson-Jones M, Guyot P, et al. MPES-R: Multi-Parameter Evidence Synthesis in R for Survival Extrapolation—A Tutorial. *PharmacoEconomics* 2024. DOI: 10.1007/s40273-024-01425-4.
5. Petitjean A, Shang H, Bullement A, et al. Use of external data to inform overall survival extrapolation in NICE technology appraisals for oncology drugs. *J Med Econ* 2025; 28: 803-813. 2025/05/20. DOI: 10.1080/13696998.2025.2506968.
6. Bullement A, Critchlow S, Stevenson M, et al. A Review of External Evidence Used to Support Survival Extrapolation in Assessments of Cancer Drugs By the National Institute for Health and Care Excellence (NICE). In: *ISPOR Europe 2022* Vienna, Austria, 2022.
7. McQueen RB, Bullement A and Heeg B. 303: Incorporating External Evidence Into Extrapolation: Making Sense of Seemingly Endless Options. In: *ISPOR 2023 Annual Congress* Boston, MA, USA, 2023.
8. McQueen RB, Bullement A, Heeg B, et al. 229: Incorporating External Evidence into Extrapolations: A Debate on the Trade-Offs between Complexity and Simplicity. In: *ISPOR 2023 Annual European Congress* Copenhagen, Denmark, 2023.
9. Pedder H, Allignol A, Jen M-H, et al. 030: Survival at the Margins: Navigating HTA Challenges in Modern Oncology Trials. In: *ISPOR 2025 Annual European Congress* Glasgow, Scotland, 2025.
10. Latimer N. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data, https://www.ncbi.nlm.nih.gov/books/NBK395885/pdf/Bookshelf_NBK395885.pdf (2011, accessed 2025/08/29).
11. World Health Organisation (WHO). Health technology assessment, <https://www.who.int/publications/i/item/9789240110878> (2020, accessed 2025/08/29).
12. International Society for Pharmacoeconomics and Outcomes Research (ISPOR). What Is HTA?, <https://www.ispor.org/strategic-initiatives/health-technology-assessment> (2020, accessed 2025/08/29).
13. O'Rourke B, Oortwijn W and Schuller T. The new definition of health technology assessment: A milestone in international collaboration. *Int J Technol Assess Health Care* 2020; 36: 189-190. DOI: 10.1017/s0266462320000215.
14. Canada's Drug Agency (CDA). Canada's Drug Agency website, <https://www.cda-amc.ca/> (2025, accessed 2025/08/29).
15. Pharmaceutical Benefits Advisory Committee (PBAC). Pharmaceutical Benefits Advisory Committee (PBAC) Membership, <http://www.pbs.gov.au/pbs/industry/listing/participants/pbac> (2020, accessed 2025/08/29).
16. Agency for Care Effectiveness (ACE). ACE website, <https://www.ace-hta.gov.sg/> (2020, accessed 2025/08/29).
17. Institute for Clinical and Economic Review (ICER). ICER website, <https://icer.org/> (2020, accessed 2025/08/29).

18. Chamova J. Mapping of HTA national organisations, programmes and processes in EU and Norway, https://health.ec.europa.eu/system/files/2018-02/2018_mapping_npc_en_0.pdf (2017, accessed 2025/08/29).
19. European Commission. Joint Clinical Assessments - European Commission, https://health.ec.europa.eu/health-technology-assessment/implementation-regulation-health-technology-assessment/joint-clinical-assessments_en (2025, accessed 2025/08/29).
20. Bounds A. EU rules threaten to cut new treatments for rare diseases. *Financial Times*, 2024-08-13 2024.
21. Henderson N, Brassel S, O'Neill P, et al. *Do NICE's Decision Outcomes Impact International HTA Decision-making? OHE Contract Research Report, London: Office of Health Economics.* 2023.
22. National Institute for Health and Care Excellence (NICE). Draft manual consultation: NICE HealthTech programme manual, <https://www.nice.org.uk/guidance/gid-pmg10010/documents/consultation-document> (2025, accessed 2025/08/29).
23. Scottish Medicines Consortium (SMC). A Guide to the Scottish Medicines Consortium, <https://scottishmedicines.org.uk/media/7605/20231818-a-guide-to-the-scottish-medicines-consortium.pdf> (2018, accessed 2025/08/29).
24. UK Government. The National Institute for Health and Care Excellence (Constitution and Functions) and the Health and Social Care Information Centre (Functions) Regulations 2013. Part 2, Regulation 7, <https://www.legislation.gov.uk/uksi/2013/259/regulation/7/made> (2013, accessed 2025/08/29).
25. National Institute for Health and Care Excellence (NICE). NICE technology appraisal guidance, <https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance> (2020, accessed 2025/08/29).
26. Jackson C, Stevens J, Ren S, et al. Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods. *Med Decis Making* 2017; 37: 377-390. 2016/03/24. DOI: 10.1177/0272989x16639900.
27. Bell Gorrod H, Kearns B, Stevens J, et al. A Review of Survival Analysis Methods Used in NICE Technology Appraisals of Cancer Treatments: Consistency, Limitations, and Areas for Improvement. *Medical Decision Making* 2019; 39: 899-909. DOI: 10.1177/0272989x19881967 L3 - 10.1177/0272989x19881967.
28. National Institute for Health and Care Excellence (NICE). Proposals for increasing capacity within NICE's technology appraisal programme, <https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/technology-appraisals/increasing-ta-capacity-consultation.pdf> (2017, accessed 2025/08/29).
29. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of health economics* 1999; 18: 341-364. 1999/10/28. DOI: 10.1016/s0167-6296(98)00039-3.
30. National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal 2013, <https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781> (2013, accessed 2025/08/29).
31. National Institute for Health and Care Excellence (NICE). NICE health technology evaluations: the manual [PMG36], <https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation> (2022, accessed 2025/08/29).
32. Collett D. *Modelling Survival Data in Medical Research, Second Edition*. CRC Press, 2003.
33. Kaplan EL and Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 1958; 53: 457-481. DOI: 10.2307/2281868.
34. Therneau TM, Lumley TA, E. and Crowson C. R package 'survival'. 3-2-7 ed. CRAN2020, p. Contains the core survival analysis routines, including definition of Surv objects, Kaplan-Meier and Aalen-Johansen (multi-state) curves, Cox models, and parametric accelerated failure time models.

35. National Institute for Health and Care Excellence (NICE). NICE TA621: Osimertinib for untreated EGFR mutation-positive non-small-cell lung cancer, <https://www.nice.org.uk/guidance/ta621/> (2020, accessed 2025/08/29).
36. National Institute for Health and Care Excellence (NICE). NICE TA654: Osimertinib for untreated EGFR mutation-positive non-small-cell lung cancer, <https://www.nice.org.uk/guidance/ta654/> (2020, accessed 2025/08/29).
37. Soria J-C, Ohe Y, Vansteenkiste J, et al. Osimertinib in Untreated EGFR-Mutated Advanced Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 2018; 378: 113-125. DOI: 10.1056/NEJMoa1713137.
38. Ranganathan P and Pramesh CS. Censoring in survival analysis: Potential for bias. *Perspect Clin Res* 2012; 3: 40. 2012/02/22. DOI: 10.4103/2229-3485.92307.
39. Ramalingam SS, Vansteenkiste J, Planchard D, et al. Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC. *N Engl J Med* 2020; 382: 41-50. 2019/11/22. DOI: 10.1056/NEJMoa1913662.
40. Hernán MA. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass)* 2010; 21: 13-15. 2009/12/17. DOI: 10.1097/EDE.0b013e3181c1ea43.
41. Morris TP, Jarvis CI, Cragg W, et al. Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. *BMJ Open* 2019; 9: e030215. DOI: 10.1136/bmjopen-2019-030215.
42. Rohatgi A. WebPlotDigitizer Version 4.4. 2020, p. This program is distributed under the GNU Affero General Public License Version 3.
43. Guyot P, Ades AE, Ouwens MJNM, et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology* 2012; 12: 9. DOI: 10.1186/1471-2288-12-9.
44. European Society for Medical Oncology (ESMO). Final FLAURA Results Demonstrate Overall Survival Benefit with Osimertinib Over TKIs in Advanced NSCLC. 2019.
45. Cox DR. Regression Models and Life-Tables. In: Kotz S and Johnson NL (eds) *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY: Springer New York, 1992, pp.527-541.
46. Prasad V and Mailankody S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA internal medicine* 2017; 177: 1569-1575. 2017/09/12. DOI: 10.1001/jamainternmed.2017.3601.
47. Henderson RH, French D, Stewart E, et al. Delivering the precision oncology paradigm: reduced R&D costs and greater return on investment through a companion diagnostic informed precision oncology medicines approach. *Journal of Pharmaceutical Policy and Practice* 2023; 16: 84. DOI: 10.1186/s40545-023-00590-9.
48. Siddiqui M and Rajkumar SV. The high cost of cancer drugs and what we can do about it. *Mayo Clinic proceedings* 2012; 87: 935-943. 2012/10/06. DOI: 10.1016/j.mayocp.2012.07.007.
49. NHS England. Cancer Drugs Fund, <https://www.england.nhs.uk/cancer/cdf/> (2020, accessed 2025/08/29).
50. Scottish Medicines Consortium (SMC). Interim acceptance decision option, <https://www.scottishmedicines.org.uk/how-we-decide/interim-acceptance-decision-option/> (2020, accessed 2025/08/29).
51. Cancer Research UK (CRUK). Cancer incidence by age, <https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age> (2020, accessed 2025/08/29).
52. Rutherford MJ, Lambert PC, Sweeting MJ, et al. NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis, <https://sheffield.ac.uk/media/34188/download?attachment> (2020, accessed 29-Aug-2025).
53. Latimer NR and Adler AI. Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ Medicine* 2022; 1: e000094. DOI: 10.1136/bmjmed-2021-000094.

54. Carrigan G, Whipple S, Capra WB, et al. Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. *Clin Pharmacol Ther* 2020; 107: 369-377. 2019/07/28. DOI: 10.1002/cpt.1586.
55. National Institute for Health and Care Excellence (NICE). NICE TA319: Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma, <https://www.nice.org.uk/guidance/ta319> (2014, accessed 2025/08/29).
56. National Institute for Health and Care Excellence (NICE). NICE TA517: Avelumab for treating metastatic Merkel cell carcinoma, <https://www.nice.org.uk/guidance/ta517> (2018, accessed 2025/08/29).
57. National Institute for Health and Care Excellence (NICE). NICE TA559: Axicabtagene ciloleucel for treating diffuse large B-cell lymphoma and primary mediastinal large B-cell lymphoma after 2 or more systemic therapies, <https://www.nice.org.uk/guidance/ta559> (2019, accessed 2025/08/29).
58. National Institute for Health and Care Excellence (NICE). NICE TA658: Isatuximab with pomalidomide and dexamethasone for treating relapsed and refractory multiple myeloma, <https://www.nice.org.uk/guidance/ta658/> (2020, accessed 2025/08/29).
59. Nordon C, Karcher H, Groenwold RH, et al. The "Efficacy-Effectiveness Gap": Historical Background and Current Conceptualization. *Value Health* 2016; 19: 75-81. 2016/01/23. DOI: 10.1016/j.jval.2015.09.2938.
60. Monti S, Grosso V, Todoerti M, et al. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology* 2018; 57: vii54-vii58. DOI: 10.1093/rheumatology/key109 %J Rheumatology.
61. Averitt AJ, Weng C, Ryan P, et al. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *npj Digital Medicine* 2020; 3: 67. DOI: 10.1038/s41746-020-0277-8.
62. National Institute for Health and Care Excellence (NICE). Technology appraisal data: cancer appraisal recommendations, <https://www.nice.org.uk/what-nice-does/our-guidance/about-technology-appraisal-guidance/technology-appraisal-data-cancer-appraisal-recommendations> (2020, accessed 2025/08/29).
63. NHS England Cancer Drugs Fund Team. Appraisal and Funding of Cancer Drugs from July 2016 (including the new Cancer Drugs Fund) - A new deal for patients, taxpayers and industry, <https://www.england.nhs.uk/wp-content/uploads/2013/04/cdf-sop.pdf> (2016, accessed 2025/08/29).
64. Kang J and Cairns J. "Don't Think Twice, It's All Right": Using Additional Data to Reduce Uncertainty Regarding Oncologic Drugs Provided Through Managed Access Agreements in England. *PharmacoEconomics - Open* 2023; 7: 77-91. DOI: 10.1007/s41669-022-00369-9.
65. Oakley JE, Ren S, Forsyth JE, et al. NICE DSU Technical Support Document 26: Expert elicitation for long-term survival outcomes, <https://sheffield.ac.uk/nice-dsu/tsds/expert-elicitation-tds> (2025, accessed 2025/08/29).
66. Cancer Research UK (CRUK). Cancer Statistics for the UK, <https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk#heading-One> (2022, accessed 2025/08/29).
67. Woods B, Sideris E, Palmer S, et al. NICE DSU Technical Support Document 19. Partitioned Survival Analysis for Decision Modelling in Health Care: A Critical Review, <https://sheffield.ac.uk/media/34205/download?attachment> (2017, accessed 2025/08/29).
68. Woods BS, Sideris E, Palmer S, et al. Partitioned Survival and State Transition Models for Healthcare Decision Making in Oncology: Where Are We Now? *Value in Health* 2020; 23: 1613-1621. DOI: <https://doi.org/10.1016/j.jval.2020.08.2094>.
69. Latimer NR and Rutherford MJ. Mixture and Non-mixture Cure Models for Health Technology Assessment: What You Need to Know. *PharmacoEconomics* 2024; 42: 1073-1090. DOI: 10.1007/s40273-024-01406-7.
70. Kleinbaum DG. *Survival Analysis*. New York, NY: Springer, 1996.

71. Klein JP, Houwelingen HCv, Ibrahim JG, et al. *Handbook of Survival Analysis*. New York: Chapman and Hall/CRC, 2016, p.656.
72. The Britannica Dictionary. Definition of 'Survival', <https://www.britannica.com/dictionary/survival> (2025, accessed 2025/08/29).
73. Tableman M and Kim JS. *Survival Analysis Using S: Analysis of Time-to-Event Data*. CRC Press, 2003.
74. Pascariu MD and Torres C. Force of Mortality. In: Gu D and Dupre ME (eds) *Encyclopedia of Gerontology and Population Aging*. Cham: Springer International Publishing, 2020, pp.1-6.
75. Rutherford E. I. A radio-active substance emitted from thorium compounds. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1900; 49: 1-14. DOI: 10.1080/14786440009463821.
76. Dickman PW and Coviello E. Estimating and Modeling Relative Survival. *The Stata Journal* 2015; 15: 186-215. DOI: 10.1177/1536867X1501500112.
77. Sarfati D, Blakely T and Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International Journal of Epidemiology* 2010; 39: 598-610. DOI: 10.1093/ije/dyp392.
78. McGivern L, Shulman L, Carney JK, et al. Death Certification Errors and the Effect on Mortality Statistics. *Public Health Rep* 2017; 132: 669-675. 2017/11/02. DOI: 10.1177/0033354917736514.
79. Aalen OO. A linear regression model for the analysis of life times. *Statistics in Medicine* 1989; 8: 907-925. DOI: <https://doi.org/10.1002/sim.4780080803>.
80. Office for National Statistics (ONS). Life expectancy calculator, <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/articles/lifeexpectancycalculator/2019-06-07> (2025, accessed 2025/08/29).
81. National Institute for Health and Care Excellence (NICE). NICE TA893: Brexucabtagene autoleucel for treating relapsed or refractory B-cell acute lymphoblastic leukaemia in people 26 years and over, <https://alpha.nice.org.uk/guidance/ta893> (2023, accessed 2025/08/29).
82. Maurer MJ, Ghesquières H, Jais JP, et al. Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large B-cell lymphoma treated with immunochemotherapy. *J Clin Oncol* 2014; 32: 1066-1073. 2014/02/20. DOI: 10.1200/jco.2013.51.5866.
83. National Institutes of Health (NIH). NIH's Definition of a Clinical Trial, <https://grants.nih.gov/policy-and-compliance/policy-topics/clinical-trials/definition> (2017, accessed 2025/08/29).
84. Suresh K. An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *J Hum Reprod Sci* 2011; 4: 8-11. 2011/07/21. DOI: 10.4103/0974-1208.82352.
85. Monaghan TF, Agudelo CW, Rahman SN, et al. Blinding in Clinical Trials: Seeing the Big Picture. *Medicina (Kaunas)* 2021; 57 2021/07/03. DOI: 10.3390/medicina57070647.
86. National Cancer Institute (NCI). About the SEER Registries, <https://seer.cancer.gov/registries/> (2021, accessed 2025/08/29).
87. Food and Drug Administration (FDA). Expanded Access, <https://www.fda.gov/news-events/public-health-focus/expanded-access> (2022, accessed 2025/08/29).
88. Human Mortality Database (HMD). Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France), <https://www.mortality.org/> (2022, accessed 2025/08/29).
89. Office for National Statistics (ONS). National life tables: UK, <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables> (2025, accessed 2025/08/29).
90. O'Hagan T and Oakley J. SHELF: the Sheffield Elicitation Framework (version 3.0). School of Mathematics and Statistics, University of Sheffield, UK.2016.

91. Cope S, Ayers D, Zhang J, et al. Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia. *BMC Med Res Methodol* 2019; 19: 182. 2019/09/04. DOI: 10.1186/s12874-019-0823-8.
92. Stevenson M, Archer R, Tosh J, et al. Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for the treatment of rheumatoid arthritis not previously treated with disease-modifying antirheumatic drugs and after the failure of conventional disease-modifying antirheumatic drugs only: systematic review and economic evaluation. *Health Technol Assess* 2016; 20: 1-610. 2016/05/04. DOI: 10.3310/hta20350.
93. Stephens WP. *Canoe and Boat Building: A Complete Manual for Amateurs*. Forest and Stream Publishing Company, 1889.
94. Royston P and Parmar MKB. Flexible parametric 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. DOI: 10.1002/sim.1203.
95. Croxford R. *Restricted Cubic Spline Regression: A Brief Introduction*. 2016.
96. Ramsay JO. Monotone Regression Splines in Action. *Statistical Science* 1988; 3: 425-441, 417.
97. Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Medical Research Methodology* 2011; 11: 61. DOI: 10.1186/1471-2288-11-61.
98. Petersohn S, McGregor B, Klijn SL, et al. Challenges in conducting fractional polynomial and standard parametric network meta-analyses of immune checkpoint inhibitors for first-line advanced renal cell carcinoma. *Journal of comparative effectiveness research* 2023; 12: e230004. 2023/07/11. DOI: 10.57264/cer-2023-0004.
99. Heeg B, Garcia A, Beekhuizen SV, et al. Novel and existing flexible survival methods for network meta-analyses. *Journal of comparative effectiveness research* 2022 2022/09/13. DOI: 10.2217/cer-2022-0044.
100. Royston P and Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1994; 43: 429-453. DOI: <https://doi.org/10.2307/2986270>.
101. Yamaguchi K. Accelerated Failure-Time Regression Models with a Regression Model of Surviving Fraction: An Application to the Analysis of "Permanent Employment" in Japan. *Journal of the American Statistical Association* 1992; 87: 284-292. DOI: 10.1080/01621459.1992.10475207.
102. Ray RA, Perry RW, Som NA, et al. Using Cure Models for Analyzing the Influence of Pathogens on Salmon Survival. *Transactions of the American Fisheries Society* 2014; 143: 387-398. <https://doi.org/10.1080/00028487.2013.862183>. DOI: <https://doi.org/10.1080/00028487.2013.862183>.
103. Jiang C, Wang Z and Zhao H. A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research* 2019; 277: 20-31. DOI: <https://doi.org/10.1016/j.ejor.2019.01.072>.
104. Peng Y and Yu B. *Cure Models: Methods, Applications, and Implementation*. CRC Press, 2021.
105. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., <https://www.R-project.org/> (1995, accessed 2025/08/29).
106. SAS Institute Inc. SAS, Cary, NC: SAS Institute Inc. 1972.
107. StataCorp. Stata Statistical Software. College Station, TX: StataCorp LP. 1985.
108. Jackson C. flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software* 2016; 70: 1 - 33. DOI: 10.18637/jss.v070.i08.
109. R for HTA. R for Health Technology Assessment (HTA), <https://r-hta.org/> (2023, accessed 2025/08/29).

110. (FDA) FaDA. Study Data Technical Conformance Guide Technical Specifications Document, <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document> (2018, accessed 2025/08/29).
111. MRC Biostatistics Unit. BUGS, <https://www.mrc-bsu.cam.ac.uk/software/bugs-project> (1989, accessed 2025/08/29).
112. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003)*; Vienna, Austria 2003; 124.
113. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. 2012.
114. Guo J, Gabry J, Goodrich B, et al. rstan: R Interface to Stan. 2023.
115. Baio G. survHE: Survival Analysis for Health Economic Evaluation and Cost-Effectiveness Modeling. *Journal of Statistical Software* 2020; 95: 1 - 47. DOI: 10.18637/jss.v095.i14.
116. Bullement A, Cranmer HL and Shields GE. A Review of Recent Decision-Analytic Models Used to Evaluate the Economic Value of Cancer Treatments. *Appl Health Econ Health Policy* 2019; 17: 771-780. 2019/09/06. DOI: 10.1007/s40258-019-00513-3.
117. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997; 50: 683-691. 1997/06/01. DOI: 10.1016/s0895-4356(97)00049-8.
118. Dias S, Welton NJ, Sutton AJ, et al. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials, https://www.ncbi.nlm.nih.gov/books/NBK310366/pdf/Bookshelf_NBK310366.pdf (2011, accessed 2025/08/29).
119. Phillippo DM, Ades AE, Dias S, et al. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE, <https://www.sheffield.ac.uk/nice-dsu/tsds/full-list> (2016, accessed 2025/08/29).
120. Truong B, Tran L-AT, Le TA, et al. Population adjusted-indirect comparisons in health technology assessment: A methodological systematic review. *Research Synthesis Methods* 2023; 14: 660-670. DOI: <https://doi.org/10.1002/jrsm.1653>.
121. Signorovitch JE, Sikirica V, Erder MH, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health* 2012; 15: 940-947. 2012/09/25. DOI: 10.1016/j.jval.2012.05.004.
122. Caro JJ and Ishak KJ. No Head-to-Head Trial? Simulate the Missing Arms. *PharmacoEconomics* 2010; 28: 957-967. DOI: 10.2165/11537420-000000000-00000.
123. Ishak KJ, Proskorovsky I and Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics* 2015; 33: 537-549. 2015/03/22. DOI: 10.1007/s40273-015-0271-1.
124. Phillippo DM, Dias S, Ades AE, et al. Multilevel network meta-regression for population-adjusted comparisons. *Journal of the Royal Statistical Society Series A, (Statistics in Society)* 2020; 183: 1189-1210. 2020/07/21. DOI: 10.1111/rssa.12579.
125. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.1, www.training.cochrane.org/handbook (2019, accessed 2025/08/29).
126. Dodd M, Fielding K, Carpenter JR, et al. Statistical methods for non-adherence in non-inferiority trials: useful and used? A systematic review. *BMJ Open* 2022; 12: e052656. 2022/01/14. DOI: 10.1136/bmjopen-2021-052656.
127. Grutters JP, Sculpher M, Briggs AH, et al. Acknowledging patient heterogeneity in economic evaluation : a systematic literature review. *Pharmacoeconomics* 2013; 31: 111-123. 2013/01/19. DOI: 10.1007/s40273-012-0015-4.
128. Teerawattananon Y, Anothaisintawee T, Pheerapanyawaranun C, et al. A systematic review of methodological approaches for evaluating real-world effectiveness of COVID-19 vaccines: Advising

resource-constrained settings. *PLoS One* 2022; 17: e0261930. 2022/01/12. DOI: 10.1371/journal.pone.0261930.

129. Alshreef A, Latimer N, Tappenden P, et al. Statistical Methods for Adjusting Estimates of Treatment Effectiveness for Patient Nonadherence in the Context of Time-to-Event Outcomes and Health Technology Assessment: A Systematic Review of Methodological Papers. *Medical Decision Making* 2019; 39: 910-925. DOI: 10.1177/0272989x19881654.

130. Johnson SR, Tomlinson GA, Hawker GA, et al. Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of Clinical Epidemiology* 2010; 63: 355-369. DOI: 10.1016/j.jclinepi.2009.06.003.

131. Henrys PA, Thompson J, Smets G, et al. Review of statistical methods and data requirements to support post market environmental monitoring of agro ecosystems. *EFSA Supporting Publications* 2014; 11. DOI: 10.2903/sp.efsa.2014.en-582.

132. Latimer N. *The role of treatment crossover adjustment methods in the context of economic evaluation*. University of Sheffield, Sheffield, United Kingdom, 2012.

133. Schlosser RW, Wendt O, Bhavnani S, et al. 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: 567-582. DOI: 10.1080/13682820600742190.

134. Zwakman M, Jabbarian LJ, van Delden J, et al. Advance care planning: A systematic review about experiences of patients with a life-threatening or life-limiting illness. *Palliat Med* 2018; 32: 1305-1321. 2018/06/30. DOI: 10.1177/0269216318784474.

135. Derges J, Kidger J, Fox F, et al. Alcohol screening and brief interventions for adults and young people in health and community-based settings: a qualitative systematic literature review. *BMC Public Health* 2017; 17: 562. 2017/06/11. DOI: 10.1186/s12889-017-4476-4.

136. Salway SM, Payne N, Rimmer M, et al. Identifying inequitable healthcare in older people: systematic review of current research practice. *Int J Equity Health* 2017; 16: 123. 2017/07/13. DOI: 10.1186/s12939-017-0605-z.

137. Adamovic M. Text Analyzer, <https://www.online-utility.org/text/analyzer.jsp> (2020, accessed 2025/08/29).

138. Petticrew M and Roberts H. How to Find the Studies: The Literature Search. *Systematic Reviews in the Social Sciences*. United Kingdom: Blackwell Publishing Ltd, 2006, pp.79-124.

139. Chilcott J, Brennan A, Booth A, et al. The role of modelling in prioritising and planning clinical trials. 2003; 7. DOI: 10.3310/hta7230.

140. International Society for Pharmacoeconomics and Outcomes Research (ISPOR). ISPOR Presentations Database, <https://www.ispor.org/heor-resources/presentations-database/search> (2020, accessed 2025/08/29).

141. Guyot P, Welton NJ, Beasley M, et al. Extrapolation of Trial-Based Survival Curves Using External Information. *Value Health* 2014; 17: A326. 2014/11/01. DOI: 10.1016/j.jval.2014.08.587.

142. National Institute for Health Research (NIHR). PROSPERO: International prospective register of systematic reviews website, <https://www.crd.york.ac.uk/prospERO/> (2020, accessed 2025/08/29).

143. Bullement A, Shields G, Latimer N, et al. Systematic review of methods to incorporate external information into estimates of survival from a clinical trial. PROSPERO CRD42020202034, https://www.crd.york.ac.uk/prospERO/display_record.php?ID=CRD42020202034 (2020, accessed 2025/08/29).

144. Messori A and Trippoli S. A new method for expressing survival and life expectancy in lifetime cost-effectiveness studies that evaluate cancer patients (review). *Oncol Rep* 1999; 6: 1135-1141. 1999/07/30. DOI: 10.3892/or.6.5.1135.

145. Viscomi S, Pastore G, Dama E, et al. Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukemia. *Ann Oncol* 2006; 17: 167-171. 2005/10/27. DOI: 10.1093/annonc/mdj050.

146. Larkin J, Hatswell AJ, Nathan P, et al. The Predicted Impact of Ipilimumab Usage on Survival in Previously Treated Advanced or Metastatic Melanoma in the UK. *PLoS One* 2015; 10: e0145524. 2015/12/25. DOI: 10.1371/journal.pone.0145524.
147. Barlev A, Lin VW, Katz A, et al. Estimating Long-Term Survival of Adults with Philadelphia Chromosome-Negative Relapsed/Refractory B-Precursor Acute Lymphoblastic Leukemia Treated with Blinatumomab Using Historical Data. *Adv Ther* 2017; 34: 148-155. 2016/11/23. DOI: 10.1007/s12325-016-0447-x.
148. Fisher M, Nathan SD, Hill C, et al. Predicting Life Expectancy for Pirfenidone in Idiopathic Pulmonary Fibrosis. *J Manag Care Spec Pharm* 2017; 23: S17-s24. 2017/03/14. DOI: 10.18553/jmcp.2017.23.3-b.s17.
149. Thornton Snider J, Batt K, Wu Y, et al. The option value of innovative treatments for non-small cell lung cancer and renal cell carcinoma. *Am J Manag Care* 2017; 23: e340-e346. 2017/11/01.
150. Thornton Snider J, Seabury S, Tebeka MG, et al. The Option Value of Innovative Treatments for Metastatic Melanoma. *Forum Health Econ Policy* 2018; 21 2018/09/14. DOI: 10.1515/fhep-2016-0014.
151. Demiris N and Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Stat Med* 2006; 25: 1960-1975. 2005/10/13. DOI: 10.1002/sim.2366.
152. Benaglia T, Jackson CH and Sharples LD. Survival extrapolation in the presence of cause specific hazards. *Stat Med* 2015; 34: 796-811. 2014/11/22. DOI: 10.1002/sim.6375.
153. Guyot P, Ades AE, Beasley M, et al. Extrapolation of Survival Curves from Cancer Trials Using External Information. *Med Decis Making* 2017; 37: 353-366. 2016/09/30. DOI: 10.1177/0272989x16670604.
154. Heinävaara S and Hakulinen T. Parametric mixture model for analysing relative survival of patients with multiple cancers. *J Cancer Epidemiol Prev* 2002; 7: 147-153. 2003/04/01.
155. Maetani S, Nakajima T and Nishikawa T. Parametric Mean Survival Time Analysis in Gastric Cancer Patients. *Medical Decision Making* 2004; 24: 131-141. DOI: 10.1177/0272989X04263253.
156. Lambert PC, Smith LK, Jones DR, et al. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Stat Med* 2005; 24: 3871-3885. 2005/12/02. DOI: 10.1002/sim.2399.
157. Stare J, Henderson R and Pohar M. An individual measure of relative survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2005; 54: 115-126. DOI: <https://doi.org/10.1111/j.1467-9876.2005.00473.x>.
158. Lambert PC, Thompson JR, Weston CL, et al. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007; 8: 576-594. 2006/10/06. DOI: 10.1093/biostatistics/kxl030.
159. Pohar Perme M, Henderson R and Stare J. An approach to estimation in relative survival regression. *Biostatistics* 2009; 10: 136-146. 2008/07/05. DOI: 10.1093/biostatistics/kxn021.
160. Lambert PC, Dickman PW, Weston CL, et al. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2010; 59: 35-55. DOI: <https://doi.org/10.1111/j.1467-9876.2009.00677.x>.
161. Andersson TML, Dickman PW, Eloranta S, et al. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Medical Research Methodology* 2011; 11: 96. DOI: 10.1186/1471-2288-11-96.
162. Hinchliffe SR, Dickman PW and Lambert PC. Adjusting for the proportion of cancer deaths in the general population when using relative survival: a sensitivity analysis. *Cancer Epidemiol* 2012; 36: 148-152. 2011/10/18. DOI: 10.1016/j.canep.2011.09.007.
163. Andersson TM, Dickman PW, Eloranta S, et al. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Stat Med* 2013; 32: 5286-5300. 2013/09/17. DOI: 10.1002/sim.5943.
164. Seppä K, Hakulinen T and Läärä E. Regional variation in relative survival—quantifying the effects of the competing risks of death by using a cure fraction model with random effects. *Journal of*

the Royal Statistical Society: Series C (Applied Statistics) 2014; 63: 175-190. DOI: <https://doi.org/10.1111/rssc.12034>.

165. Jakobsen LH, Andersson TML, Biccler JL, et al. Estimating the loss of lifetime function using flexible parametric relative survival models. *BMC Medical Research Methodology* 2019; 19: 23. DOI: 10.1186/s12874-019-0661-8.
166. Rubio FJ, Rachet B, Giorgi R, et al. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics* 2019; 22: 51-67. DOI: 10.1093/biostatistics/kxz017.
167. Turner ME, Hazelrig JB and Blackstone EH. Bounded survival. *Mathematical Biosciences* 1982; 59: 33-46. DOI: [https://doi.org/10.1016/0025-5564\(82\)90107-9](https://doi.org/10.1016/0025-5564(82)90107-9).
168. Hwang JS and Wang JD. Monte Carlo estimation of extrapolation of quality-adjusted survival for follow-up studies. *Stat Med* 1999; 18: 1627-1640. 1999/07/17. DOI: 10.1002/(sici)1097-0258(19990715)18:13<1627::aid-sim159>3.0.co;2-d.
169. Nelson CL, Sun JL, Tsiatis AA, et al. Empirical estimation of life expectancy from large clinical trials: use of left-truncated, right-censored survival analysis methodology. *Stat Med* 2008; 27: 5525-5555. 2008/07/10. DOI: 10.1002/sim.3355.
170. Ishak KJ, Caro JJ, Drayson MT, et al. Adjusting for patient crossover in clinical trials using external data: a case study of lenalidomide for advanced multiple myeloma. *Value Health* 2011; 14: 672-678. 2011/08/16. DOI: 10.1016/j.jval.2011.02.1182.
171. Hawe E, Pearson I, Wolowacz S, et al. Use Of External Data To Guide Long-Term Survival Extrapolations Of Trial Data For Chronic Lymphocytic Leukemia. *Value Health* 2014; 17: A563. 2014/11/01. DOI: 10.1016/j.jval.2014.08.1867.
172. Hwang JS, Hu TH, Lee LJ, et al. Estimating lifetime medical costs from censored claims data. *Health Econ* 2017; 26: e332-e344. 2017/05/13. DOI: 10.1002/hec.3512.
173. Pennington M, Grieve R, der Meulen JV, et al. Value of External Data in the Extrapolation of Survival Data: A Study Using the NJR Data Set. *Value Health* 2018; 21: 822-829. 2018/07/15. DOI: 10.1016/j.jval.2017.12.023.
174. Vickers A. An Evaluation of Survival Curve Extrapolation Techniques Using Long-Term Observational Cancer Data. *Medical Decision Making* 2019; 39: 926-938. DOI: 10.1177/0272989x19875950.
175. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; 372: n71. DOI: 10.1136/bmj.n71.
176. Buxton M, Caine N, Chase D, et al. A review of the evidence on the effects and costs of implantable cardioverter defibrillator therapy in different patient groups, and modelling of cost-effectiveness and cost-utility for these groups in a UK context. *Health Technol Assess* 2006; 10: iii-iv, ix-xi, 1-164. 2006/08/15. DOI: 10.3310/hta10270.
177. Connolly SJ, Hallstrom AP, Cappato R, et al. Meta-analysis of the implantable cardioverter defibrillator secondary prevention trials. AVID, CASH and CIDS studies. Antiarrhythmics vs Implantable Defibrillator study. Cardiac Arrest Study Hamburg . Canadian Implantable Defibrillator Study. *Eur Heart J* 2000; 21: 2071-2078. 2000/12/05. DOI: 10.1053/euhj.2000.2476.
178. GUSTO investigators. An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction. *New England Journal of Medicine* 1993; 329: 673-682. DOI: 10.1056/nejm199309023291001.
179. Rosati RA, McNeer JF, Starmer CF, et al. A new information system for medical practice. *Arch Intern Med* 1975; 135: 1017-1024. 1975/08/11.
180. Mark DB, Hlatky MA, Califf RM, et al. Cost Effectiveness of Thrombolytic Therapy with Tissue Plasminogen Activator as Compared with Streptokinase for Acute Myocardial Infarction. *New England Journal of Medicine* 1995; 332: 1418-1424. DOI: 10.1056/nejm199505253322106.
181. Nelson CP, Lambert PC, Squire IB, et al. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 2007; 26: 5486-5498. DOI: <https://doi.org/10.1002/sim.3064>.

182. Lambert PC and Royston P. Further Development of Flexible Parametric Models for Survival Analysis. *The Stata Journal* 2009; 9: 265-290. DOI: 10.1177/1536867x0900900206.
183. Rai KR, Peterson BL, Appelbaum FR, et al. Long-Term Survival Analysis of the North American Intergroup Study C9011 Comparing Fludarabine (F) and Chlorambucil (C) in Previously Untreated Patients with Chronic Lymphocytic Leukemia (CLL). *Blood* 2009; 114: 536-536. DOI: 10.1182/blood.V114.22.536.536.
184. National Institute for Health and Care Excellence (NICE). NICE TA268: Ipilimumab for previously treated unresectable malignant melanoma - Company submission, <https://www.nice.org.uk/guidance/ta268/documents/melanoma-stage-iii-or-iv-ipilimumab-manufacturers-submission-bristol-myerssquibb2> (2011, accessed 2025/08/29).
185. Balch CM, Buzaid AC, Soong SJ, et al. Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma. *J Clin Oncol* 2001; 19: 3635-3648. 2001/08/16. DOI: 10.1200/jco.2001.19.16.3635.
186. Bonner JA, Harari PM, Giralt J, et al. Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival. *Lancet Oncol* 2010; 11: 21-28. 2009/11/10. DOI: 10.1016/s1470-2045(09)70311-0.
187. Gelber RD, Goldhirsch A and Cole BF. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. International Breast Cancer Study Group. *Control Clin Trials* 1993; 14: 485-499. 1993/12/01. DOI: 10.1016/0197-2456(93)90029-d.
188. National Joint Registry (NJR). About the NJR, <https://www.njrcentre.org.uk/about-the-njr/> (2020, accessed 2025/08/29).
189. NHS Digital. Hospital Episode Statistics (HES), <https://digital.nhs.uk/services/hospital-episode-statistics> (2020, accessed 2025/08/29).
190. Grupp SA, Laetsch TW, Buechner J, et al. Analysis of a Global Registration Trial of the Efficacy and Safety of CTL019 in Pediatric and Young Adults with Relapsed/Refractory Acute Lymphoblastic Leukemia (ALL). *Blood* 2016; 128: 221-221. DOI: 10.1182/blood.V128.22.221.221.
191. Ganti AK, Shostrom V, Alorabi M, et al. Early Stage Non-Small-Cell Lung Cancer in Octogenarian and Older Patients: A SEER Database Analysis. *Clin Lung Cancer* 2016; 17: 285-291. 2016/01/05. DOI: 10.1016/j.clcc.2015.11.014.
192. Vickers A and Hawe E. PRM159 - An Evaluation of Survival Curve Extrapolation Techniques Using Long-Term Observational Cancer Data. *Value in Health* 2018; 21: S383. DOI: <https://doi.org/10.1016/j.jval.2018.09.2279>.
193. Bagust A and Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. *Med Decis Making* 2014; 34: 343-351. 2013/08/01. DOI: 10.1177/0272989x13497998.
194. Bach PB, Cramer LD, Warren JL, et al. Racial differences in the treatment of early-stage lung cancer. *N Engl J Med* 1999; 341: 1198-1205. 1999/10/16. DOI: 10.1056/nejm199910143411606.
195. Herring W, Pearson I, Purser M, et al. Cost Effectiveness of Ofatumumab Plus Chlorambucil in First-Line Chronic Lymphocytic Leukaemia in Canada. *Pharmacoeconomics* 2016; 34: 77-90. 2015/11/01. DOI: 10.1007/s40273-015-0332-5.
196. Jackson CH, Sharples LD and Thompson SG. Survival models in health economic evaluations: balancing fit and parsimony to improve prediction. *Int J Biostat* 2010; 6: Article 34. 2010/01/01. DOI: 10.2202/1557-4679.1269.
197. Murray TA, Hobbs BP, Lystig TC, et al. Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics* 2014; 70: 185-191. 2013/12/07. DOI: 10.1111/biom.12115.
198. Lewis CJ, Sarkar S, Zhu J, et al. Borrowing from Historical Control Data in Cancer Drug Development: A Cautionary Tale and Practical Guidelines. *Stat Biopharm Res* 2019; 11: 67-78. 2019/08/23. DOI: 10.1080/19466315.2018.1497533.

199. Soikkeli F, Hashim M, Ouwens M, et al. Extrapolating Survival Data Using Historical Trial-Based a Priori Distributions. *Value Health* 2019; 22: 1012-1017. 2019/09/13. DOI: 10.1016/j.jval.2019.03.017.
200. Mba RD, Goungounga JA, Grafféo N, et al. Correcting inaccurate background mortality in excess hazard models through breakpoints. *BMC Medical Research Methodology* 2020; 20: 268. DOI: 10.1186/s12874-020-01139-z.
201. Felizzi F, Paracha N, Pöhlmann J, et al. Mixture Cure Models in Oncology: A Tutorial and Practical Guidance. *Pharmacoecon Open* 2021; 5: 143-155. 2021/02/28. DOI: 10.1007/s41669-021-00260-z.
202. van Oostrum I, Ouwens M, Remiro-Azócar A, et al. Comparison of Parametric Survival Extrapolation Approaches Incorporating General Population Mortality for Adequate Health Technology Assessment of New Oncology Drugs. *Value Health* 2021; 24: 1294-1301. 2021/08/29. DOI: 10.1016/j.jval.2021.03.008.
203. San Miguel JF, Schlag R, Khuageva NK, et al. Bortezomib plus melphalan and prednisone for initial treatment of multiple myeloma. *N Engl J Med* 2008; 359: 906-917. 2008/08/30. DOI: 10.1056/NEJMoa0801479.
204. Shustik C, Belch A, Robinson S, et al. A randomised comparison of melphalan with prednisone or dexamethasone as induction therapy and dexamethasone or observation as maintenance therapy in multiple myeloma: NCIC CTG MY.7. *British Journal of Haematology* 2007; 136: 203-211. <https://doi.org/10.1111/j.1365-2141.2006.06405.x>. DOI: <https://doi.org/10.1111/j.1365-2141.2006.06405.x>.
205. Botta L, Goungounga J, Capocaccia R, et al. A new cure model accounting for extra non-cancer mortality: Validation and application to real data. *Revue d'Épidémiologie et de Santé Publique* 2021; 69: S24. DOI: <https://doi.org/10.1016/j.respe.2021.04.039>.
206. Aballéa S, Chancellor JV, Raikou M, et al. Cost-effectiveness analysis of oxaliplatin compared with 5-fluorouracil/leucovorin in adjuvant treatment of stage III colon cancer in the US. *Cancer* 2007; 109: 1082-1089. 2007/02/01. DOI: 10.1002/cncr.22512.
207. Hisashige A, Yoshida S and Kodaira S. Cost-effectiveness of adjuvant chemotherapy with uracil-tegafur for curatively resected stage III rectal cancer. *Br J Cancer* 2008; 99: 1232-1238. 2008/09/16. DOI: 10.1038/sj.bjc.6604666.
208. Howard DH, Tangka FK, Seeff LC, et al. The impact of detection and treatment on lifetime medical costs for patients with precancerous polyps and colorectal cancer. *Health Econ* 2009; 18: 1381-1393. 2009/01/15. DOI: 10.1002/hec.1434.
209. Caro JJ, Ishak KJ and Migliaccio-Walle K. Estimating Survival for Cost-Effectiveness Analyses: A Case Study in Atherothrombosis. *Value in Health* 2004; 7: 627-635. <https://doi.org/10.1111/j.1524-4733.2004.75013.x>. DOI: <https://doi.org/10.1111/j.1524-4733.2004.75013.x>.
210. CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. *Lancet* 1996; 348: 1329-1339. 1996/11/16. DOI: 10.1016/s0140-6736(96)09457-3.
211. Che Z, Baio G and Paly VP. At First I Was Afraid, I Was Petrified... Issues and Possible Solutions to the Problems of Extrapolating Survival Curves from Limited Trial Data, <https://www.ispor.org/heor-resources/presentations-database/presentation/euro2021-3402/13243> (2021, accessed 2025/08/29).
212. Briggs A, Pollock C, Ouwens M, et al. Application of an Innovative Survival Extrapolation and Validation Method That Combines Clinical Trial Data, General Population Mortality and Expert Opinion, <https://www.ispor.org/heor-resources/presentations-database/presentation/euro2021-3376/13278> (2021, accessed 2025/08/29).
213. Pham HA, Smalbugge D, Kroi F, et al. The Use of Historical Clinical Trial Data to Inform Survival Extrapolation, <https://www.ispor.org/heor-resources/presentations-database/presentation-paper/euro2021-3406/13464/the-use-of-historical-clinical-trial-data-to-inform-survival-extrapolation> (2021, accessed 2025/08/29).

214. Che Z, Green N and Baio G. Blended Survival Curves: A New Approach to Extrapolation for Time-to-Event Outcomes from Clinical Trials in Health Technology Assessment. *Med Decis Making* 2023; 43: 299-310. 2022/11/01. DOI: 10.1177/0272989x221134545.
215. Che Z, Baio G and Green N. Blended Survival Curves: A New Approach to Extrapolation for Time-to-Event Clinical Trial Data in Health Technology Assessment, <https://www.ispor.org/heor-resources/presentations-database/presentation-paper/euro2021-3406/13465/blended-survival-curves-a-new-approach-to-extrapolation-for-time-to-event-clinical-trial-data-in-health-technology-assessment> (2021, accessed 2025/08/29).
216. Jackson Christopher H. survextrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology* 2023; 23: 282. DOI: 10.1186/s12874-023-02094-1.
217. Gallacher D. SurvInt: a simple tool to obtain precise parametric survival extrapolations. *BMC Medical Informatics and Decision Making* 2024; 24: 76. DOI: 10.1186/s12911-024-02475-6.
218. Langford B, Diamantopoulos A, Maher TM, et al. Using Data on Survival with Idiopathic Pulmonary Fibrosis to Estimate Survival with Other Types of Progressive Fibrosis Interstitial Lung Disease: A Bayesian Framework. *Advances in Therapy* 2022; 39: 1045-1054. DOI: 10.1007/s12325-021-02014-z.
219. Sanchez Alvarez J, Jaber M and Blanchet Zumofen M-H. Multiple Real-World Data Sources in a Bayesian Framework to Inform Long-Term Survival Estimates of Mosunetuzumab in Patients with Follicular Lymphoma. *Oncology and Therapy* 2023; 11: 495-511. DOI: 10.1007/s40487-023-00245-4.
220. Traore S, Sashegyi A, Winfree KB, et al. Bayesian survival extrapolation for cost-effectiveness analysis: a case study of RELAY for ramucirumab in combination with erlotinib in the treatment of non-small-cell lung cancer. *Journal of Medical Economics* 2023; 26: 1479-1488. DOI: 10.1080/13696998.2023.2272534.
221. Sweeting MJ, Rutherford MJ, Jackson D, et al. Survival Extrapolation Incorporating General Population Mortality Using Excess Hazard and Cure Models: A Tutorial. *Med Decis Making* 2023; 43: 737-748. 2023/07/14. DOI: 10.1177/0272989x231184247.
222. Lunn DJ, Thomas A, Best N, et al. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 2000; 10: 325-337.
223. Kearns B, Stevenson MD, Triantafyllopoulos K, et al. Dynamic and Flexible Survival Models for Extrapolation of Relative Survival: A Case Study and Simulation Study. 2022; 42: 945-955. DOI: 10.1177/0272989x221107649.
224. Bullement A. Incorporating external evidence into survival extrapolations for HTA: learnings from the literature and previous NICE appraisals. In: *SCHARR PGR Conference 2023* Sheffield, UK, 15 June 2023 2023.
225. Bullement A, Meng Y, Cooper M, et al. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available? *Journal of Medical Economics* 2019; 22: 205-214. DOI: 10.1080/13696998.2018.1547303.
226. Cooper M, Smith S, Williams T, et al. How accurate are the longer-term projections of overall survival for cancer immunotherapy for standard versus more flexible parametric extrapolation methods? *Journal of Medical Economics* 2022; 25: 260-273. DOI: 10.1080/13696998.2022.2030599.
227. Grumberg V, Roze S, Chevalier J, et al. A Review of Overall Survival Extrapolations of Immune-Checkpoint Inhibitors Used in Health Technology Assessments by the French Health Authorities. *International Journal of Technology Assessment in Health Care* 2022; 38: e28. 2022/03/25. DOI: 10.1017/S0266462322000125.
228. Reitsma F, Pham HA and Heeg B. The Use of External Data to Inform Survival Extrapolation in Oncology Technology Appraisals: A Comprehensive Review, <https://www.ispor.org/heor-resources/presentations-database/presentation/euro2021-3408/114205> (2021, accessed 2025/08/29).

229. National Institute for Health and Care Excellence (NICE). NICE TA742: Selpercatinib for treating advanced thyroid cancer with RET alterations (company submission), <https://www.nice.org.uk/guidance/ta742/> (2021, accessed 2025/08/29).
230. National Institute for Health and Care Excellence (NICE). NICE TA746: Nivolumab for adjuvant treatment of oesophageal or gastro-oesophageal junction cancer (company submission), <https://www.nice.org.uk/guidance/ta746/> (2021, accessed 2025/08/29).
231. National Institute for Health and Care Excellence (NICE). NICE TA689: Acalabrutinib for treating chronic lymphocytic leukaemia, <https://www.nice.org.uk/guidance/ta689/> (2021, accessed 2025/08/29).
232. Adam R, Delvart V, Pascal G, et al. Rescue surgery for unresectable colorectal liver metastases downstaged by chemotherapy: a model to predict long-term survival. *Annals of surgery* 2004; 240: 644-657; discussion 657-648. 2004/09/24. DOI: 10.1097/01.sla.0000141198.92114.f6.
233. Cucchetti A, Ferrero A, Cescon M, et al. Cure model survival analysis after hepatic resection for colorectal liver metastases. *Annals of surgical oncology* 2015; 22: 1908-1914. 2014/11/16. DOI: 10.1245/s10434-014-4234-0.
234. National Institute for Health and Care Excellence (NICE). NICE TA721: Abiraterone for treating newly diagnosed metastatic hormone-naïve prostate cancer (company submission), <https://www.nice.org.uk/guidance/ta721/> (2018, accessed 2025/08/29).
235. Lou F, Sima CS, Adusumilli PS, et al. Esophageal cancer recurrence patterns and implications for surveillance. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* 2013; 8: 1558-1562. 2014/01/07. DOI: 10.1097/01.JTO.0000437420.38972.fb.
236. National Institute for Health and Care Excellence (NICE). NICE TA695: Carfilzomib with dexamethasone and lenalidomide for previously treated multiple myeloma, <https://www.nice.org.uk/guidance/ta695/> (2021, accessed 2025/08/29).
237. National Institute for Health and Care Excellence (NICE). NICE TA724: Nivolumab with ipilimumab and chemotherapy for untreated metastatic non-small-cell lung cancer, <https://www.nice.org.uk/guidance/ta724/> (2021, accessed 2025/08/29).
238. Makuch RW. Adjusted survival curve estimation using covariates. *J Chronic Dis* 1982; 35: 437-443. 1982/01/01. DOI: 10.1016/0021-9681(82)90058-3.
239. Chang IM, Gelman R and Pagano M. Corrected group prognostic curves and summary statistics. *J Chronic Dis* 1982; 35: 669-674. 1982/01/01. DOI: 10.1016/0021-9681(82)90019-4.
240. National Institute for Health and Care Excellence (NICE). NICE TA684: Nivolumab for adjuvant treatment of completely resected melanoma with lymph node involvement or metastatic disease (company submission), <https://www.nice.org.uk/guidance/ta684/> (2020, accessed 2025/08/29).
241. National Institute for Health and Care Excellence (NICE). NICE TA704: Trastuzumab deruxtecan for treating HER2-positive unresectable or metastatic breast cancer after 2 or more anti-HER2 therapies, <https://www.nice.org.uk/guidance/ta704/> (2021, accessed 2025/08/29).
242. National Institute for Health and Care Excellence (NICE). NICE ID5121: Trastuzumab deruxtecan for treating HER2-positive unresectable or metastatic breast cancer after 1 or more anti-HER2 therapies (MA review TA704 and TA862), <https://www.nice.org.uk/guidance/topic-selection/gid-ta11674> (2025, accessed 2025/08/29).
243. National Institute for Health and Care Excellence (NICE). NICE TA720: Chloromethine gel for treating mycosis fungoides-type cutaneous T-cell lymphoma, <https://www.nice.org.uk/guidance/ta720/> (2021, accessed 2025/08/29).
244. Agar NS, Wedgeworth E, Crichton S, et al. Survival outcomes and prognostic factors in mycosis fungoides/Sézary syndrome: validation of the revised International Society for Cutaneous Lymphomas/European Organisation for Research and Treatment of Cancer staging proposal. *J Clin Oncol* 2010; 28: 4730-4739. 2010/09/22. DOI: 10.1200/jco.2009.27.7665.

245. National Institute for Health and Care Excellence (NICE). NICE TA741: Apalutamide for treating prostate cancer (company submission), <https://www.nice.org.uk/guidance/ta741/> (2021, accessed 2025/08/29).
246. National Institute for Health and Care Excellence (NICE). NICE TA673: Niraparib for maintenance treatment of advanced ovarian, fallopian tube and peritoneal cancer after response to firstline platinum-based chemotherapy (company submission), <https://www.nice.org.uk/guidance/ta673/> (2021, accessed 2025/08/29).
247. National Institute for Health and Care Excellence (NICE). NICE ID6403: Niraparib for maintenance treatment of advanced ovarian, fallopian tube and peritoneal cancer after response to first-line platinum-based chemotherapy (review of TA673), <https://www.nice.org.uk/guidance/topic-selection/gid-ta11675> (2025, accessed 2025/08/29).
248. National Institute for Health and Care Excellence (NICE). NICE TA677: Autologous anti-CD19-transduced CD3+ cells for treating relapsed or refractory mantle cell lymphoma (company submission), <https://www.nice.org.uk/guidance/ta677/> (2021, accessed 2025/08/29).
249. National Institute for Health and Care Excellence (NICE). NICE TA737: Pembrolizumab with platinum-based chemotherapy for untreated advanced oesophageal cancer (company submission), <https://www.nice.org.uk/guidance/ta737/> (2021, accessed 2025/08/29).
250. National Institute for Health and Care Excellence (NICE). NICE TA725: Abemaciclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy (company submission), <https://www.nice.org.uk/guidance/ta725/> (2021, accessed 2025/08/29).
251. National Institute for Health and Care Excellence (NICE). History of NICE, <https://www.nice.org.uk/about-us/history-of-nice> (2025, accessed 2025/08/29).
252. National Institute for Health and Care Excellence (NICE). NICE real-world evidence framework [ECD9], <https://www.nice.org.uk/corporate/ecd9/chapter/overview> (2022, accessed 2025/08/29).
253. Cooney P and White A. Direct Incorporation of Expert Opinion into Parametric Survival Models to Inform Survival Extrapolation. *Med Decis Making* 2023; 43: 325-336. 2023/01/18. DOI: 10.1177/0272989x221150212.
254. Bullement A, Latimer NR and Bell Gorrod H. Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study. *Value Health* 2019; 22: 276-283. 2019/03/06. DOI: 10.1016/j.jval.2018.10.007.
255. Brilleman S. *simsurv: Simulate Survival Data*. The R Foundation, 2017.
256. Davies A, Briggs A, Schneider J, et al. The Ends Justify the Mean: Outcome Measures for Estimating the Value of New Cancer Therapies. *Health Outcomes Research in Medicine* 2012; 3: e25-e36. DOI: <https://doi.org/10.1016/j.ehrm.2012.01.001>.
257. National Institute for Health and Care Excellence (NICE). NICE TA145 Cetuximab for the treatment of locally advanced squamous cell cancer of the head and neck, <https://www.nice.org.uk/guidance/ta145> (2008, accessed 2025/08/29).
258. National Institute for Health and Care Excellence (NICE). NICE TA531: Pembrolizumab for untreated PD-L1-positive metastatic non-small-cell lung cancer, <https://www.nice.org.uk/guidance/ta531> (2018, accessed 2025/08/29).
259. Stroup AM, Cho H, Scoppa SM, et al. The impact of state-specific life tables on relative survival. *Journal of the National Cancer Institute Monographs* 2014; 2014: 218-227. 2014/11/25. DOI: 10.1093/jncimonographs/lgu017.
260. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; 38: 2074-2102. 2019/01/18. DOI: 10.1002/sim.8086.
261. Hoaglin DC and Andrews DF. The Reporting of Computation-Based Results in Statistics. *The American Statistician* 1975; 29: 122-126. DOI: 10.2307/2683438.

262. Boulesteix AL, Lauer S and Eugster MJ. A plea for neutral comparison studies in computational sciences. *PLoS One* 2013; 8: e61562. 2013/05/03. DOI: 10.1371/journal.pone.0061562.
263. Pawel S, Kook L and Reeve K. Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal* 2024; 66: 2200091. DOI: <https://doi.org/10.1002/bimj.202200091>.
264. Morris T. Strengthening simulation studies in statistical research. In: *RSS conference* Manchester, UK, 2021.
265. Boulesteix AL, Baillie M, Edelmann D, et al. Editorial for the special collection "Towards neutral comparison studies in methodological research". *Biometrical journal Biometrische Zeitschrift* 2024; 66: e2400031. 2024/02/18. DOI: 10.1002/bimj.202400031.
266. Gray J, Sullivan T, Latimer NR, et al. Extrapolation of Survival Curves Using Standard Parametric Models and Flexible Parametric Spline Models: Comparisons in Large Registry Cohorts with Advanced Cancer. 2021; 41: 179-193. DOI: 10.1177/0272989x20978958.
267. Hess KR. muhaz: Hazard Function Estimation in Survival Analysis. Comprehensive R Archive Network (CRAN)2022.
268. Alshreef A, Latimer N, Tappenden P, et al. Assessing Methods for Adjusting Estimates of Treatment Effectiveness for Patient Nonadherence in the Context of Time-to-Event Outcomes and Health Technology Assessment: A Simulation Study. *Med Decis Making* 2025; 45: 60-73. 2024/11/13. DOI: 10.1177/0272989x241293414.
269. Michael JC and Paul L. STMIX: Stata module to fit two-component parametric mixture survival models. S457339 ed.: Boston College Department of Economics, 2011.
270. Paul L. STPM2: Stata module to estimate flexible parametric survival models. S457128 ed.: Boston College Department of Economics, 2010.
271. StataCorp. *Stata 18 Survival Analysis Reference Manual*. College Station, TX: Stata Press, 2023.
272. Lambert PC. Modeling of the Cure Fraction in Survival Studies. *The Stata Journal* 2007; 7: 351-375. DOI: 10.1177/1536867x0700700304.
273. Clements M, Liu X-R and Christoffersen B. rstpm2: Smooth Survival Models, Including Generalized Survival Models. 2015.
274. Wickham H, François R, Henry L, et al. dplyr: A Grammar of Data Manipulation. The R Foundation, 2014.
275. Wickham H, Chang W, Henry L, et al. Ggplot2: Create elegant data visualisations using the grammar of graphics. The R Foundation, 2007.
276. Csárdi G and FitzJohn R. progress: Terminal Progress Bars. The R Foundation, 2015.
277. Gasparini A and White IR. rsumsum: Analysis of Simulation Studies Including Monte Carlo Error. The R Foundation, 2018.
278. Birnie R. survtools: A collection of useful functions for survival analysis. 0.1 ed. 2019.
279. Ushey K, Allaire J, Wickham H, et al. rstudioapi: Safely Access the RStudio API. R package version 0.17.1 ed. 2024.
280. Wickham HH, Lionel. purrr: Functional Programming Tools. R package version 1.0.2 ed. 2023.
281. Soares M, Colson A, Bojke L, et al. Recommendations on the Use of Structured Expert Elicitation Protocols for Healthcare Decision Making: A Good Practices Report of an ISPOR Task Force. *Value Health* 2024; 27: 1469-1478. 2024/11/07. DOI: 10.1016/j.jval.2024.07.027.
282. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *Bmj* 1996; 312: 71-72. 1996/01/13. DOI: 10.1136/bmj.312.7023.71.

283. Bell Gorrod H, Latimer N and Abrams K. NICE DSU Technical Support Document 24: Adjusting survival time estimates in the presence of treatment switching: An update to TSD 16, <https://www.sheffield.ac.uk/nice-dsu/tsds/survival-analysis> (2024, accessed 2025/08/29).
284. National Institute for Health and Care Excellence (NICE). NICE TA1017: Pembrolizumab with chemotherapy before surgery (neoadjuvant) then alone after surgery (adjuvant) for treating resectable non-small-cell lung cancer, <https://www.nice.org.uk/guidance/ta1017/history> (2024, accessed 2025/08/29).
285. Incerti D, Thom H, Baio G, et al. R You Still Using Excel? The Advantages of Modern Software Tools for Health Technology Assessment. *Value in Health* 2019; 22: 575-579. DOI: <https://doi.org/10.1016/j.jval.2019.01.003>.
286. Metry A, Latimer N, Wailoo A, et al. *Real-world evidence (RWE) at NICE: Demonstrating methods for using the systemic anti-cancer therapy (SACT) dataset in technology appraisals*. 2025.
287. HTA Lab project working group. *Improving the use of real-world evidence from managed access: the Systemic Anti Cancer Therapy data set: HTA Innovation Laboratory report*. 2025.
288. Chen EY-T, Leontyeva Y, Lin C-N, et al. Comparing Survival Extrapolation within All-Cause and Relative Survival Frameworks by Standard Parametric Models and Flexible Parametric Spline Models Using the Swedish Cancer Registry. *Medical Decision Making* 2024; 44: 269-282. DOI: 10.1177/0272989X241227230.
289. Timmins IR, Torabi F, Jackson CH, et al. Simulation-based assessment of a Bayesian survival model with flexible baseline hazard and time-dependent effects. 2025.
290. Jiu L, Wang J, Versteeg JW, et al. Roadmap to Innovation of HTA Methods (IHTAM): insights from three case studies of quantitative methods. *Int J Technol Assess Health Care* 2024; 40: e49. 2024/11/05. DOI: 10.1017/s0266462324000564.
291. National Institute for Health and Care Excellence (NICE). NICE ID4067: Isatuximab with pomalidomide and dexamethasone for treating relapsed and refractory multiple myeloma [Review of TA658] <https://www.nice.org.uk/guidance/indevelopment/gid-ta10979> (2025, accessed 2025/08/29).
292. Kearns B, Stevens J, Ren S, et al. How Uncertain is the Survival Extrapolation? A Study of the Impact of Different Parametric Survival Models on Extrapolated Uncertainty About Hazard Functions, Lifetime Mean Survival and Cost Effectiveness. *Pharmacoeconomics* 2020; 38: 193-204. 2019/11/26. DOI: 10.1007/s40273-019-00853-x.
293. Hooijmans CR, Tillema A, Leenaars M, et al. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Lab Anim* 2010; 44: 170-175. 2010/06/17. DOI: 10.1258/la.2010.009117.
294. Guinness World Records. Oldest person ever (female), [https://www.guinnessworldrecords.com/world-records/oldest-person-\(female\)](https://www.guinnessworldrecords.com/world-records/oldest-person-(female)) (2021, accessed 2025/08/29).
295. Heinävaara S and Hakulinen T. Relative survival of patients with subsequent cancer. *J Cancer Epidemiol Prev* 2002; 7: 173-179. 2003/07/09.
296. Boag JW. Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society Series B (Methodological)* 1949; 11: 15-53.
297. Estève J, Benhamou E, Croasdale M, et al. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med* 1990; 9: 529-538. 1990/05/01. DOI: 10.1002/sim.4780090506.
298. Pastore G, Mosso ML, Carnevale F, et al. Survival trends of childhood cancer diagnosed during 1970-1994 in Piedmont, Italy: a report from the Childhood Cancer Registry. *Med Pediatr Oncol* 2001; 36: 481-488. 2001/03/22. DOI: 10.1002/mpo.1113.
299. MacLennan IC and Cusick J. Objective evaluation of the role of vincristine in induction and maintenance therapy for myelomatosis. Medical Research Council Working Party on Leukaemia in Adults. *Br J Cancer* 1985; 52: 153-158. DOI: 10.1038/bjc.1985.171.

300. MacLennan IC, Chapman C, Dunn J, et al. Combined chemotherapy with ABCM versus melphalan for treatment of myelomatosis. The Medical Research Council Working Party for Leukaemia in Adults. *Lancet* 1992; 339: 200-205. 1992/01/25. DOI: 10.1016/0140-6736(92)90004-m.
301. Olujohungbe AB, Dunn JA, Drayson M, et al. Prednisolone added to the ABCM as treatment for multiple myeloma increases serological responses but not overall survival or the number of stable clinical responses. *British Journal of Haematology* 1996; 93: 296-296.
302. Augustson BM, Begum G, Dunn JA, et al. Early mortality after diagnosis of multiple myeloma: analysis of patients entered onto the United kingdom Medical Research Council trials between 1980 and 2002--Medical Research Council Adult Leukaemia Working Party. *J Clin Oncol* 2005; 23: 9219-9226. 2005/11/09. DOI: 10.1200/jco.2005.03.2086.
303. Drayson M, Begum G, Basu S, et al. Effects of paraprotein heavy and light chain types and free light chain load on survival in myeloma: an analysis of patients receiving conventional-dose chemotherapy in Medical Research Council UK multiple myeloma trials. *Blood* 2006; 108: 2013-2019. 2006/05/27. DOI: 10.1182/blood-2006-03-008953.
304. Shkolnikov V, Barbieri M and Wilmoth J. The Human Mortality Database, <https://www.mortality.org/> (2021, accessed 2025/08/29).
305. Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* 1991; 47: 467-485. 1991/06/01.
306. Ibrahim JGC, Ming-Hui and Sinha D. *Bayesian survival analysis*. New York: New York : Springer, c2001, 2001.
307. Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 2011; 67: 1047-1056. 2011/03/03. DOI: 10.1111/j.1541-0420.2011.01564.x.
308. Touraine C, Grafféo N and Giorgi R. More accurate cancer-related excess mortality through correcting background mortality for extra variables. *Statistical Methods in Medical Research* 2020; 29: 122-136. DOI: 10.1177/0962280218823234.

Appendices

Appendix A	Additional survival modelling information	345
Appendix B	Web of Science animal filter	346
Appendix C	Data extraction for systematic review	349
Appendix D	Methods excluded from the systematic review	359
Appendix E	Data extraction for appraisals review	399
Appendix F	Miscellaneous code	420
Appendix G	Additional results from neutral comparison study	422

Appendix A. Additional survival modelling information

Appendix Table 1: Overview of commonly used survival models (from Kearns *et al.*, [2020])

Model (parameters)	Survival function S(t)	Cumulative hazard function H(t)	Hazard function h(t)	Possible shapes of h(t)
Exponential $\lambda > 0$	$e^{-\lambda t}$	λt	λ	Constant
Weibull $\lambda > 0$ $\gamma > 0$	$e^{-\lambda t^\gamma}$	λt^γ	$\lambda \gamma t^{\gamma-1}$	- Constant - Increasing monotonically - Decreasing monotonically
Lognormal $\mu \in (-\infty, \infty)$ $\sigma > 0$	$1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)$	$-\log\left(1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)\right)$	$\frac{1}{\sigma\sqrt{2\pi}} t^{-1} e^{-\frac{(\log(t)-\mu)^2}{2\sigma^2}}$ $S(t)$	Increasing then decreasing
Log-logistic $\alpha > 0$ $\beta > 0$	$\frac{1}{1 + (\alpha t)^\beta}$	$-\log\left(\frac{1}{1 + (\alpha t)^\beta}\right)$	$\frac{\alpha\beta(\alpha t)^{\beta-1}}{1 + (\alpha t)^\beta}$	- Decreasing monotonically - Increasing then decreasing
Gamma $\lambda > 0$ $\beta > 0$	$1 - \Gamma[\lambda^{-2}(e^{-\beta}t); \lambda^{-2}]$	$-\log(S(t))$	$\frac{\beta^\lambda t^{\lambda-1} e^{-\beta t}}{\Gamma(\lambda)S(t)}$	- Constant - Increasing monotonically - Decreasing monotonically
Generalised gamma $\lambda \in (-\infty, \infty)$ $\beta > 0$ $\sigma > 0$	$\begin{cases} 1 - \Gamma[\lambda^{-2}(e^{-\beta}t)^{\frac{\lambda}{\sigma}}; \lambda^{-2}] & \text{if } \lambda > 0 \\ \Gamma[\lambda^{-2}(e^{-\beta}t)^{\frac{\lambda}{\sigma}}; \lambda^{-2}] & \text{if } \lambda < 0 \end{cases}$	$-\log(S(t))$	$\frac{ \lambda }{\sigma t \Gamma(\lambda^{-2})} \alpha \lambda^{-2} e^{-\alpha}$ $S(t)$ where $\alpha = \lambda^{-2}(e^{-\beta}t)^{\frac{\lambda}{\sigma}}$	- Constant - Increasing monotonically - Decreasing monotonically - Bathtub
Gompertz $\lambda > 0$ $\theta \in (-\infty, \infty)$	$e^{-\frac{\lambda}{\theta}(e^{\frac{\lambda}{\theta}} - 1)}$	$\frac{\lambda}{\theta}(e^{\theta t} - 1)$	$\lambda e^{\theta t}$	- Constant - Increasing monotonically - Decreasing monotonically

Key: Φ , cumulative standard normal distribution; $\Gamma(t; \lambda) = \int_0^t \frac{x^{\lambda-1} e^{-x}}{\Gamma(\lambda)} dx$; e , exponential function. Allowing $\theta < 0$ for the Gompertz implies that the survival function will never equal 0.

Source: Table re-produced from Kearns *et al.*, (2020).²⁹² Table presented in this thesis under the terms of the Creative Commons Attribution 4.0 International License (available at: <http://creativecommons.org/licenses/by/4.0/>). Table re-formatted for ease of presentation, with no changes made to the content of the table. Kearns B, Stevens J, Ren S, Brennan A. How Uncertain is the Survival Extrapolation? A Study of the Impact of Different Parametric Survival Models on Extrapolated Uncertainty About Hazard Functions, Lifetime Mean Survival and Cost Effectiveness. *Pharmacoeconomics*. 2020 Feb;38(2):193-204. doi: 10.1007/s40273-019-00853-x.

Appendix B. Web of Science animal filter

A search filter was used within the first iteration of searching via the WOS database to remove animal studies. The filter is based on a previously-published PubMed filter by Hooijmans *et al.*, (2010).²⁹³ The published filter was modified to focus on the specific animal terms that may feature within either the title ('TI') or abstract ('AB') of animal studies, and aligned with the syntax required to search WOS. The filter used is shown below:

(AB = (animal OR animals OR pisces OR fish OR fishes OR catfish OR catfishes OR sheatfish OR silurus OR arius OR heteropneustes OR clarias OR gariepinus OR fathead minnow OR fathead minnows OR pimephales OR promelas OR cichlidae OR trout OR trouts OR char OR chars OR salvelinus OR salmo OR oncorhynchus OR guppy OR guppies OR millionfish OR poecilia OR goldfish OR goldfishes OR carassius OR auratus OR mullet OR mullets OR mugil OR curema OR shark OR sharks OR cod OR cods OR gadus OR morhua OR carp OR carps OR cyprinus OR carpio OR killifish OR eel OR eels OR anguilla OR zander OR sander OR lucioperca OR stizostedion OR turbot OR turbots OR psetta OR flatfish OR flatfishes OR plaice OR pleuronectes OR platessa OR tilapia OR tilapias OR oreochromis OR sarotherodon OR common sole OR dover sole OR solea OR zebrafish OR zebrafishes OR danio OR rerio OR seabass OR dicentrarchus OR labrax OR morone OR lamprey OR lampreys OR petromyzon OR pumpkinseed OR pumpkinseeds OR lepomis OR gibbosus OR herring OR clupea OR harengus OR amphibia OR amphibian OR amphibians OR anura OR salientia OR frog OR frogs OR rana OR toad OR toads OR bufo OR xenopus OR laevis OR bombina OR epidalea OR calamita OR salamander OR salamanders OR newt OR newts OR triturus OR reptilia OR reptile OR reptiles OR bearded dragon OR pogona OR vitticeps OR iguana OR iguanas OR lizard OR lizards OR anguis fragilis OR turtle OR turtles OR snakes OR snake OR aves OR bird OR birds OR quail OR quails OR coturnix OR bobwhite OR colinus OR virginianus OR poultry OR poultries OR fowl OR fowls OR chicken OR chickens OR gallus OR zebra finch OR taeniopygia OR guttata OR canary OR canaries OR serinus OR canaria OR parakeet OR parakeets OR grasskeet OR parrot OR parrots OR psittacine OR psittacines OR shelduck OR tadorna OR goose OR geese OR branta OR leucopsis OR woodlark OR lullula OR flycatcher OR ficedula OR hypoleuca OR dove OR doves OR geopelia OR cuneata OR duck OR ducks OR greylag OR graylag OR anser OR harrier OR circus pygargus OR red knot OR great knot OR calidris OR canutus OR godwit OR limosa OR lapponica OR meleagris OR gallopavo OR jackdaw OR corvus OR monedula OR ruff OR philomachus OR pugnax OR lapwing OR peewit OR plover OR vanellus OR swan OR cygnus OR columbianus OR bewickii OR gull OR chroicocephalus OR ridibundus OR albifrons OR great tit OR parus OR aythya OR fuligula OR streptopelia OR risoria OR spoonbill OR platalea OR leucorodia OR blackbird OR turdus OR merula OR blue tit OR cyanistes OR pigeon OR pigeons OR columba OR pintail OR anas OR starling OR sternus OR owl OR athene noctua OR pochard OR ferina OR cockatiel OR nymphicus OR hollandicus OR skylark OR alauda OR tern OR sterna OR teal OR crecca OR oystercatcher OR haematopus OR ostralegus OR shrew OR shrews OR sorex OR araneus OR crocidura OR russula OR european mole OR talpa OR chiroptera OR bat OR bats OR eptesicus OR serotinus OR myotis OR dasycneme OR daubentonii OR pipistrelle OR pipistrellus OR cat OR cats OR felis OR catus OR feline OR dog OR dogs OR canis OR canine OR canines OR otter OR otters OR lutra OR badger OR badgers OR meles OR fitchew OR fitch OR foumart OR fowlmart OR ferrets OR ferret OR polecat OR polecats OR mustela OR putorius OR weasel OR weasels OR fox OR foxes OR vulpes OR common seal OR phoca OR vitulina OR grey seal OR halichoerus OR horse OR horses OR equus OR equine OR equidae OR donkey OR donkeys OR mule OR mules OR pig OR pigs OR swine OR swines OR hog OR hogs OR boar OR boars OR porcine OR piglet OR piglets OR sus OR scrofa OR llama OR llamas OR lama OR glama OR deer OR deers OR cervus OR elaphus OR cow OR cows OR bos taurus OR bos indicus OR bovine OR bull OR bulls OR cattle OR bison OR bisons OR sheep OR sheeps OR ovis aries OR ovine OR lamb OR lambs OR mouflon OR moufflons OR goat OR goats OR capra OR caprine OR chamois OR rupicapra OR leporidae OR lagomorpha OR lagomorph OR rabbit OR rabbits OR oryctolagus OR cuniculus OR laprine OR hares OR lepus OR rodentia OR rodent OR rodents OR murinae OR mouse OR mice OR mus OR musculus OR murine OR woodmouse OR apodemus OR rat OR rats OR rattus OR norvegicus OR guinea pig OR guinea pigs OR cavia OR porcellus OR hamster OR hamsters OR mesocricetus OR cricetus OR gerbil OR gerbils OR jird OR jirds OR meriones OR unguiculatus OR jerboa OR jerboas OR jaculus OR chinchilla OR chinchillas OR beaver OR beavers OR castor fiber OR castor canadensis OR sciuridae OR squirrel OR squirrels OR sciurus OR chipmunk OR chipmunks OR marmot OR marmots OR marmota OR suslik OR susliks OR spermophilus OR cynomys OR cottonrat OR cottonrats OR sigmodon OR vole OR voles OR microtus OR myodes OR glareolus OR primate OR primates OR prosimian OR prosimians OR lemur OR lemurs OR lemuridae OR loris OR bush baby OR bush babies OR bushbaby OR bushbabies OR galago OR galagos OR anthropoidea OR anthropoids OR simian OR simians OR monkey OR monkeys OR marmoset OR marmosets OR callithrix OR cebuella OR tamarin OR tamarins OR saguinus OR leontopithecus OR squirrel monkey OR squirrel monkeys OR saimiri OR night monkey OR night monkeys OR owl monkey OR owl monkeys OR douroucoulis OR aotus OR spider monkey OR spider monkeys OR ateles OR baboon OR baboons OR papio OR rhesus monkey OR macaque OR macaca OR mulatta OR cynomolgus OR fascicularis OR green monkey OR green monkeys OR chlorocebus OR vervet OR vervets OR pygerythrus OR hominoidea OR ape OR apes OR hylobatidae OR gibbon OR gibbons OR siamang OR siamangs OR nomascus OR symphalangus OR hominidae OR orangutan OR orangutans OR pongo OR chimpanzee OR chimpanzees OR pan troglodytes OR bonobo OR bonobos OR pan paniscus OR gorilla OR gorillas OR troglodytes) OR TI = (animal OR animals OR pisces OR fish OR fishes OR catfish OR catfishes OR sheatfish OR silurus OR arius OR heteropneustes OR clarias OR gariepinus OR fathead minnow OR fathead minnows OR pimephales OR promelas OR cichlidae OR trout OR trouts OR char OR chars OR salvelinus OR salmo OR oncorhynchus OR guppy OR guppies OR millionfish OR poecilia OR goldfish OR goldfishes OR carassius OR auratus OR mullet OR mullets OR mugil OR curema OR shark OR sharks OR cod OR cods OR gadus OR morhua OR carp OR carps OR cyprinus OR carpio OR killifish OR eel OR eels OR anguilla OR zander OR sander OR lucioperca OR stizostedion OR turbot OR turbots OR psetta OR flatfish OR flatfishes OR plaice OR pleuronectes OR platessa OR tilapia OR tilapias OR oreochromis OR sarotherodon OR common sole OR dover sole OR solea OR zebrafish OR zebrafishes OR danio OR rerio OR seabass OR dicentrarchus OR labrax OR morone OR lamprey OR lampreys OR petromyzon OR pumpkinseed OR pumpkinseeds OR lepomis OR gibbosus OR herring OR clupea OR harengus OR amphibia OR amphibian OR amphibians OR anura OR salientia OR frog OR frogs OR rana OR toad OR toads OR bufo OR xenopus OR laevis OR bombina OR epidalea OR

calamita OR salamander OR salamanders OR newt OR newts OR triturus OR reptilia OR reptile OR reptiles OR bearded dragon OR pogona OR vitticeps OR iguana OR iguanas OR lizard OR lizards OR anguis fragilis OR turtle OR turtles OR snakes OR snake OR aves OR bird OR birds OR quail OR quails OR coturnix OR bobwhite OR colinus OR virginianus OR poultry OR poultries OR fowl OR fowls OR chicken OR chickens OR gallus OR zebra finch OR taeniopygia OR guttata OR canary OR canaries OR serinus OR canaria OR parakeet OR parakeets OR grasskeet OR parrot OR parrots OR psittacine OR psittacines OR shelduck OR tadorna OR goose OR geese OR branta OR leucopsis OR woodlark OR lullula OR flycatcher OR ficedula OR hypoleuca OR dove OR doves OR geopelia OR cuneata OR duck OR ducks OR greylag OR graylag OR anser OR harrier OR circus pygargus OR red knot OR great knot OR calidris OR canutus OR godwit OR limosa OR lapponica OR meleagris OR gallopavo OR jackdaw OR corvus OR monedula OR ruff OR philomachus OR pugnax OR lapwing OR peewit OR plover OR vanellus OR swan OR cygnus OR columbianus OR bewickii OR gull OR chroicocephalus OR ridibundus OR albifrons OR great tit OR parus OR aythya OR fuligula OR streptopelia OR risoria OR spoonbill OR platalea OR leucorodia OR blackbird OR turdus OR merula OR blue tit OR cyanistes OR pigeon OR pigeons OR columba OR pintail OR anas OR starling OR sturnus OR owl OR athene noctua OR pochard OR ferina OR cockatiel OR nymphicus OR hollandicus OR skylark OR alauda OR tern OR sterna OR teal OR crecca OR oystercatcher OR haematopus OR ostralegus OR shrew OR shrews OR sorex OR araneus OR crocidura OR russula OR european mole OR talpa OR chiroptera OR bat OR bats OR eptesicus OR serotinus OR myotis OR dasycneme OR daubentonii OR pipistrellus OR pipistrellus OR cat OR cats OR felis OR feline OR dog OR dogs OR canis OR canine OR canines OR otter OR otters OR lutra OR badger OR badgers OR meles OR fitchew OR fitch OR foumart OR foulmart OR ferrets OR ferret OR polecat OR polecats OR mustela OR putorius OR weasel OR weasels OR fox OR foxes OR vulpes OR common seal OR phoca OR vitulina OR grey seal OR halichoerus OR horse OR horses OR equus OR equine OR equidae OR donkey OR donkeys OR mule OR mules OR pig OR pigs OR swine OR swines OR hog OR hogs OR boar OR boars OR porcine OR piglet OR piglets OR sus OR scrofa OR llama OR llamas OR lama OR glama OR deer OR deers OR cervus OR elaphus OR cow OR cows OR bos taurus OR bos indicus OR bovine OR bull OR bulls OR cattle OR bison OR bisons OR sheep OR sheeps OR ovis aries OR ovine OR lamb OR lambs OR mouflon OR mouflons OR goat OR goats OR capra OR caprine OR chamois OR rupicapra OR leporidae OR lagomorpha OR lagomorph OR rabbit OR rabbits OR oryctolagus OR cuniculus OR laprine OR hares OR lepus OR rodentia OR rodent OR rodents OR murinae OR mouse OR mice OR mus OR musculus OR murine OR woodmouse OR apodemus OR rat OR rats OR rattus OR norvegicus OR guinea pig OR guinea pigs OR cavia OR porcellus OR hamster OR hamsters OR mesocricetus OR cricetus OR cricetus OR gerbil OR gerbils OR jird OR jirds OR meriones OR unguiculatus OR jerboa OR jerboas OR jaculus OR chinchilla OR chinchillas OR beaver OR beavers OR castor fiber OR castor canadensis OR sciuridae OR squirrel OR squirrels OR sciurus OR chipmunk OR chipmunks OR marmot OR marmots OR marmota OR suslik OR susliks OR spermophilus OR cynomys OR cottonrat OR cottonrats OR sigmodon OR vole OR voles OR microtus OR myodes OR glareolus OR primate OR primates OR prosimian OR prosimians OR lemur OR lemurs OR lemuridae OR loris OR bush baby OR bush babies OR bushbaby OR bushbabies OR galago OR galagos OR anthropoidea OR anthropoids OR simian OR simians OR monkey OR monkeys OR marmoset OR marmosets OR callithrix OR cebuella OR tamarin OR tamarins OR saguinus OR leontopithecus OR squirrel monkey OR squirrel monkeys OR saimiri OR night monkey OR night monkeys OR owl monkey OR owl monkeys OR douroucoulis OR aotus OR spider monkey OR spider monkeys OR ateles OR baboon OR baboons OR papio OR rhesus monkey OR macaque OR macaca OR mulatta OR cynomolgus OR fascicularis OR green monkey OR green monkeys OR chlorocebus OR vervet OR vervets OR pygerythrus OR hominoidea OR ape OR apes OR hylobatidae OR gibbon OR gibbons OR siamang OR siamangs OR nomascus OR symphalangus OR hominidae OR orangutan OR orangutans OR pongo OR chimpanzee OR chimpanzees OR pan troglodytes OR bonobo OR bonobos OR pan paniscus OR gorilla OR gorillas OR troglodytes))

Databases= WOS, BCI, BIOSIS, CCC, DRCI, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English

For the second iteration, the original filter was simplified to instead search based on topic ('TS'), which searches the title, abstract, author keywords, and Keywords Plus® (a set of key words unique to the WOS database, which includes words and phrases that have been harvested from the titles of cited articles). This edit was made in the interest of efficiency in case the search was to be repeated in the future. The revised animal filter is shown below:

(TS = (animal OR animals OR pisces OR fish OR fishes OR catfish OR catfishes OR sheatfish OR silurus OR arius OR heteropneustes OR clarias OR gariepinus OR fathead minnow OR fathead minnows OR pimephales OR promelas OR cichlidae OR trout OR trouts OR char OR chars OR salvelinus OR salmo OR oncorhynchus OR guppy OR guppies OR millionfish OR poecilia OR goldfish OR goldfishes OR carassius OR auratus OR mullet OR mullets OR mugil OR curema OR shark OR sharks OR cod OR cods OR gadus OR morhua OR carp OR carps OR cyprinus OR carpio OR killifish OR eel OR eels OR anguilla OR zander OR sander OR lucioperca OR stizostedion OR turbot OR turbot OR psetta OR flatfish OR flatfishes OR plaice OR pleuronectes OR platessa OR tilapia OR tilapias OR oreochromis OR sarotherodon OR common sole OR dover sole OR solea OR zebrafish OR zebrafishes OR danio OR rerio OR seabass OR dicentrarchus OR labrax OR morone OR lamprey OR lampreys OR petromyzon OR pumpkinseed OR pumpkinseeds OR lepomis OR gibbosus OR herring OR clupea OR harengus OR amphibia OR amphibian OR amphibians OR anura OR salientia OR frog OR frogs OR rana OR toad OR toads OR bufo OR xenopus OR laevis OR bombina OR epidalea OR calamita OR salamander OR salamanders OR newt OR newts OR triturus OR reptilia OR reptile OR reptiles OR bearded dragon OR pogona OR vitticeps OR iguana OR iguanas OR lizard OR lizards OR anguis fragilis OR turtle OR turtles OR snakes OR snake OR aves OR bird OR birds OR quail OR quails OR coturnix OR bobwhite OR colinus OR virginianus OR poultry OR poultries OR fowl OR fowls OR chicken OR chickens OR gallus OR zebra finch OR taeniopygia OR guttata OR canary OR canaries OR serinus OR canaria OR parakeet OR parakeets OR grasskeet OR parrot OR parrots OR psittacine OR psittacines OR shelduck OR tadorna OR goose OR geese OR branta OR leucopsis OR woodlark OR lullula OR flycatcher OR ficedula OR hypoleuca OR dove OR doves OR geopelia OR cuneata OR duck OR ducks OR greylag OR graylag OR anser OR harrier OR circus pygargus OR red knot OR great knot OR calidris OR canutus OR godwit OR limosa OR lapponica OR meleagris OR gallopavo OR jackdaw OR corvus OR monedula OR ruff OR philomachus OR pugnax OR lapwing OR peewit

OR plover OR vanellus OR swan OR cygnus OR columbianus OR bewickii OR gull OR chroicocephalus OR ridibundus OR albifrons OR great tit OR parus OR aythya OR fuligula OR streptopelia OR risoria OR spoonbill OR platalea OR leucorodia OR blackbird OR turdus OR merula OR blue tit OR cyanistes OR pigeon OR pigeons OR columba OR pintail OR anas OR starling OR sturnus OR owl OR athene noctua OR pochard OR ferina OR cockatiel OR nymphicus OR hollandicus OR skylark OR alauda OR tern OR sterna OR teal OR crecca OR oystercatcher OR haematopus OR ostralegus OR shrew OR shrews OR sorex OR araneus OR crocidura OR russula OR european mole OR talpa OR chiroptera OR bat OR bats OR eptesicus OR serotinus OR myotis OR dasycneme OR daubentonii OR pipistrelle OR pipistrellus OR cat OR cats OR felis OR catus OR feline OR dog OR dogs OR canis OR canine OR canines OR otter OR otters OR lutra OR badger OR badgers OR meles OR fitchew OR fitch OR foudmart or foulmart OR ferrets OR ferret OR polecat OR polecats OR mustela OR putorius OR weasel OR weasels OR fox OR foxes OR vulpes OR common seal OR phoca OR vitulina OR grey seal OR halichoerus OR horse OR horses OR equus OR equine OR equidae OR donkey OR donkeys OR mule OR mules OR pig OR pigs OR swine OR swines OR hog OR hogs OR boar OR boars OR porcine OR piglet OR piglets OR sus OR scrofa OR llama OR llamas OR lama OR glama OR deer OR deers OR cervus OR elaphus OR cow OR cows OR bos taurus OR bos indicus OR bovine OR bull OR bulls OR cattle OR bison OR bisons OR sheep OR sheeps OR ovis aries OR ovine OR lamb OR lambs OR mouflon OR mouflons OR goat OR goats OR capra OR caprine OR chamois OR rupicapra OR leporidae OR lagomorpha OR lagomorph OR rabbit OR rabbits OR oryctolagus OR cuniculus OR laprine OR hares OR lepus OR rodentia OR rodent OR rodents OR murinae OR mouse OR mice OR mus OR musculus OR murine OR woodmouse OR apodemus OR rat OR rats OR rattus OR norvegicus OR guinea pig OR guinea pigs OR cavia OR porcellus OR hamster OR hamsters OR mesocricetus OR cricetus OR gerbil OR gerbils OR jird OR jirds OR meriones OR unguiculatus OR jerboa OR jerboas OR jaculus OR chinchilla OR chinchillas OR beaver OR beavers OR castor fiber OR castor canadensis OR sciuridae OR squirrel OR squirrels OR sciurus OR chipmunk OR chipmunks OR marmot OR marmots OR marmota OR suslik OR susliks OR spermophilus OR cynomys OR cottonrat OR cottonrats OR sigmodon OR vole OR voles OR microtus OR myodes OR glareolus OR primate OR primates OR prosimian OR prosimians OR lemur OR lemurs OR lemuridae OR loris OR bush baby OR bush babies OR bushbaby OR bushbabies OR galago OR galagos OR anthropoidea OR anthropoids OR simian OR simians OR monkey OR monkeys OR marmoset OR marmosets OR callithrix OR cebuella OR tamarin OR tamarins OR saguinus OR leontopithecus OR squirrel monkey OR squirrel monkeys OR saimiri OR night monkey OR night monkeys OR owl monkey OR owl monkeys OR douroucoulis OR aotus OR spider monkey OR spider monkeys OR ateles OR baboon OR baboons OR papio OR rhesus monkey OR macaque OR macaca OR mulatta OR cynomolgus OR fascicularis OR green monkey OR green monkeys OR chlorocebus OR vervet OR vervets OR pygerythrus OR hominoidea OR ape OR apes OR hylobatidae OR gibbon OR gibbons OR siamang OR siamangs OR nomascus OR symphalangus OR hominidae OR orangutan OR orangutans OR pongo OR chimpanzee OR chimpanzees OR pan troglodytes OR bonobo OR bonobos OR pan paniscus OR gorilla OR gorillas OR troglodytes))

Databases= WOS, BCI, BIOSIS, CCC, DRCL, DIIDW, KJD, MEDLINE, RSCI, SCIELO, ZOOREC Timespan=All years Search language=English

Appendix C. Data extraction for systematic review

Appendix Table 2: SLR first iteration data extraction (Part 1 of 2)

Article info			Summary information				Requirements			
#	Author(s)	Year	Category	Name of method(s) used	Key reference(s)	Description of method	Extension to previous method	Data required	Key assumptions	Statistical software
1	Messori <i>et al.</i>	1999	Piecewise	Piecewise: Individual-level piecewise	Vaidya & Mittra (1997/98)	Piecewise approach - for surviving patients, adds life expectancy assuming all surviving patients are cured. Identifies the cut-off based on the timing of the last death event - observations after this time assume to be cured	No (or none described)	Patient-level data from main study + ability to estimate life expectancy for age- and sex-adjusted counterparts	People that are still alive are assumed cured	None listed
2	Viscomi <i>et al.</i>	2006	Piecewise	Piecewise: Individual-level piecewise	Messori <i>et al.</i>	Piecewise approach - overall life expectancy estimated as sum of time since diagnosis + mean survival from study period (parametric model) + assumed cure general population survival after 13.2 years (end of study period)	Yes, references study by Messori <i>et al.</i>	Patient-level study data (+ age at baseline) + general population survival statistics	Survival after cut-off assumed equivalent to general population	Models fitted in Stata, but no code presented
3	Larkin <i>et al.</i>	2015	Piecewise	Piecewise: Standard method	None	Three-part curve fit was used: from 0–1.5 years, Kaplan–Meier estimate from the trial; from 1.5–5 years, standard parametric curve fits; after 5 years, long-term data from the AJCC registry.	No, but it is a type of piecewise model	Trial data + observational data source (digitised KM estimate)	None described by authors, but main assumption is generalisability of source for long-term hazards from registry data	None listed
4	Barlev <i>et al.</i>	2017	Piecewise	Piecewise: Standard method	None	Three stages: KM from trial + conditional survival from an observational source + life tables assuming cure after 5 years	No, but it is a type of piecewise model	Trial data + observational data source + life tables	Assumes no benefit in terms of HR after end of KM period + assumes cure after 5 years	None listed
5	Fisher <i>et al.</i>	2017	Piecewise	Piecewise: Standard method	None	Two stages: KM from pooled trial data + parametric model fitted to registry data	No, but it is a type of piecewise model	Trial data + observational data source	Assumes generalisability of conditional survival between trial + observational study sources	Fitted using Stata and MS Excel. No code given
6	Thornton Snider <i>et al.</i>	2017	Piecewise	Piecewise: Option value approach	None	Piecewise approach - established a base curve using SEER/HMD data. Then, at various time points, adjusted the curve by specifying HRs (e.g., between 1 and 6 years) based on clinical trials.	No, but it is a type of piecewise model	Data for a base curve (could be trial, registry, etc.) + HRs from a trial	PH assumption applies, treatment effect will start and stop at given time points	None listed
7	Thornton Snider <i>et al.</i>	2018	Piecewise	Piecewise: Option value approach	None	Per previous row describing method by same authors	Per previous row describing method by same authors	Per previous row describing method by same authors	Per previous row describing method by same authors	None listed
8	Demiris & Sharples	2006	Bayesian	Bayesian: Additive hazard and multiplicative hazard models	Aalen (1989), Lin & Ying, (1994) Ibrahim <i>et al.</i> , 2001.	Looked at models based on parametric/semi-parametric versions of additive/multiplicative hazard models. Two sources of data: 'trial' and 'life tables'. Priors set for baseline hazard using life table data, and for effect of ICD based on the 'trial' data. Meta-analysis output used to obtain an estimate of the HR for ICD versus AAD. PH assumed	Yes, method is described as building on work previously done by Lin & Ying, (1994)	Two data sources: (1) patient-level data for the population of interest, and (2) reference data for the general population. Also published HRs to allow curves to be produced for other treatment arms.	Method assumes that long term survival distributions for patients have a similar shape to age–sex matched controls from an external source (such as census data), but have a fixed increase in risk, either additive or multiplicative. For relative effect, approach assumes PH by specifying a constant HR.	Authors note that all the methods described in the paper can be fitted using WinBUGS.
9	Benaglia <i>et al.</i>	2015	Bayesian	Bayesian: Two component polyhazard model	Berger & Sun (1993), Louzada-Neto (1999), Demiris <i>et al.</i> , (2015)	Paper predominantly concerned with the poly-Weibull or poly-Gompertz models (i.e., models that can be fitted within a PH framework). The method involves considering that there are two potential causes of death (e.g., disease-related, or non-disease-related). Method assumes the non-disease-related hazards are identical to an external data set. Then, assuming PH hold between the risks of death (i.e., risk of death in the study group is proportionally higher than in the other data set), can fit a poly-hazard model.	Yes, method may be considered an extension to the standard Weibull (or Gompertz) models.	Two individual-level survival datasets, one from the general population and one from patients with some disease of interest. Method assumes that the different causes of death are known in the non-trial data source (and can be quantified).	For an individual in the general population, the distributions of the times to death for two causes come from the same parametric family but with different parameters. Poly-Weibull assumes that hazards for different causes are additive and independent of each other, an assumption that is impossible to test and may require sensitivity analysis.	Fitted in WinBUGS

Article info			Summary information				Requirements			
#	Author(s)	Year	Category	Name of method(s) used	Key reference(s)	Description of method	Extension to previous method	Data required	Key assumptions	Statistical software
10	Guyot <i>et al.</i>	2014	Bayesian	Bayesian: Range of methods explored, though not listed in full and not explicitly named	None	Not clear from abstract	No (or none described)	RCT evidence, external registry data (in this case, data from a US cancer database [SEER]), general population data and expert opinion	Not clear from abstract	Not clear from abstract
		2017	Bayesian	Bayesian: Bayesian multi-parameter evidence synthesis	None	Survival models fitted within a Bayesian framework, with various priors/ restrictions set: >> OS at 40 years <= General population survival at 40 years >> Conditional survival after 6 years assumed identical between trial and SEER data >> Conditional survival after 33 years assumed equal to general population (to stop general population and SEER sources conflicting) >> HR up to year 5 based on a spline model >> HR from year 6 onwards was roughly =1, with some uncertainty built into the estimation	No (or none described)	RCT data, external evidence on general population survival, conditional survival from cancer registry databases, and expert opinion.	Several key assumptions made: >> Survival in the control arm of the trial will remain less than survival in the matched general population cohort over the entire time horizon >> 1-year conditional survival in the trial control arm will converge to that of a matched cancer cohort, five or six years after diagnosis. >> HR changes over time as a smooth function with one turning point (expected to decrease initially and then increase) >> HR assumed to 'kick in' at 1 year, when treatment begins, and return to unity by 6 years.	WinBUGS, custom code
11	Cope <i>et al.</i>	2019	Bayesian	Bayesian: Integration of expert opinion within survival analysis	Guyot <i>et al.</i> , (2017)	>> Elicit expert opinion for survival at 2, 3, 4, and 5 years via a SHELF-like tool >> Synthesize estimates and combine with observed data using parametric models (accounting for uncertainty) >> FP models used (due to the fact that there are no IPD for expert opinion)	No (or none described)	Data from a trial + expert opinion (point estimates and uncertainty)	People alive after 5 years assumed to be long-term survivors	JAGS and R
12	Heinavaara & Hakulinen	2002	Relative survival	Relative survival: Parametric mixture model for multiple cancers	Boag (1949), Esteve (1990)	Not applicable - not a method to estimate overall survival	Yes, extension of a relative survival mixture-cure model	Two data sources: life tables + data with dates of diagnosis (including later diagnoses), survival times, and baseline demographics	Conditional relative survival rates for the first and second cancers are assumed to be independent	Gauss 4.0
13	Maetani <i>et al.</i>	2004	Relative survival	Relative survival: Competing risks model, using a Boag model for disease-specific survival and age- and sex- matched controls for the overall survival curve	Boag (1949)	Estimate disease-related survival by fitting a Boag (mixture-cure) model. Estimate general population survival by using life tables. Take the product of these curves evaluated at each point in time to estimate OS	Yes, application of the Boag lognormal mixture cure model	Patient-level data for a cohort + demographic information to estimate general population survival	Proportion of patients 'cured' of the disease, who otherwise achieve the survival of the general population	None listed
14	Lambert <i>et al.</i>	2005	Relative survival	Relative survival: Additive relative survival model + fractional polynomial relative survival model	Dickman <i>et al.</i> , (2004)	Additive relative survival model: software for a GLM where covariates are reflected as additive rather than multiplicative. FP relative survival model: use FP model to reflect baseline excess hazard + potential time-varying covariates	Yes, builds on work of Dickman <i>et al.</i> in relative survival more generally	Patient-level data for a cohort + demographic information to estimate general population survival	Additive model assumes covariates act additively on the excess hazard.	Stata
15	Stare <i>et al.</i>	2005	Relative survival	Relative survival: Outcome measure for individualised relative survival	Dickman <i>et al.</i> , (2004), Hakulinen (1982)	Not applicable - not a method to estimate overall survival	No (or none described)	Patient-level data for a cohort + demographic information to estimate general population survival	Same assumptions per relative survival methods in general. Also requires assumptions related to timing of events within a given calendar year (given reporting of life tables)	None listed
16	Lambert <i>et al.</i>	2007	Relative survival	Relative survival: Non-mixture cure model	De Angelis <i>et al.</i> (1999)	Fit a relative survival model assuming that a proportion of the cohort are cured, but instead of this being a mixture, assume the disease-related survival curve eventually asymptotes at this value	Yes, this extends the standard non-mixture cure model to include relative survival	Patient-level data for a cohort + demographic information to estimate general population survival	Proportion of patients 'cured' of the disease, who otherwise achieve the survival of the general population	Stata
17	Pohar Perme <i>et al.</i>	2009	Relative survival	Relative survival: EM algorithm for fitting an additive hazards model	Hakulinen & Tenkanen, (1987), Dickman <i>et al.</i> , (2004)	Offers an alternative way of estimating the parameters of an additive hazards model. Not applicable to the review, as it is not a novel means of estimating overall survival	Yes, builds upon established relative survival methods (Ederer I, II, and Hakulinen)	Patient-level data for a cohort + demographic information to estimate general population survival	Not applicable	Can be fitted using 'standard software' as uses same principles as Cox model
18	Lambert <i>et al.</i>	2010	Relative survival	Relative survival: Mixture of 2 Weibull models for uncured group in MCM/NMCM	Lambert <i>et al.</i> , (2007)	Extends MCM/NMCM approach to consider a mix of two Weibulls for the uncured group. Otherwise, approach is same as per standard MCM/NMCM within a relative survival framework.	Yes, this extends the standard MCM/NMCM with relative survival	Patient-level data for a cohort + demographic information to estimate general population survival	Proportion of patients 'cured' of the disease, who otherwise achieve the survival of the general population. Uncured group exhibits a complex hazard function, requiring a flexible two-component Weibull	Stata

Article info			Summary information					Requirements		
#	Author(s)	Year	Category	Name of method(s) used	Key reference(s)	Description of method	Extension to previous method	Data required	Key assumptions	Statistical software
19	Andersson <i>et al.</i>	2011	Relative survival	Relative survival: Flexible parametric relative survival models with a cure fraction	Royston & Parmar, (2002) Nelson <i>et al.</i> , (2007) Lambert & Parmar, (2002)	Fits a flexible relative survival model but forces the log cumulative excess hazard to have a zero slope after the last knot (i.e., imposing a cure proportion). To do this, the algorithm calculates the spline variables 'backwards'. Authors have developed a package to fit the models.	Yes, extends flexible parametric models + mixture-cure models to combine, all within a relative survival framework	Patient-level data for a cohort + demographic information to estimate general population survival	Proportion of patients 'cured' of the disease, who otherwise achieve the survival of the general population. Uncured group exhibits a complex hazard function, requiring a spline-based modelling approach. Assumes zero risk of disease-related death after final knot	Stata
20	Hinchcliffe <i>et al.</i>	2012	Relative survival	Relative survival: Adjustment for deaths in general population within estimation of relative survival	Ederer (1961), Hakulinen (1982)	Simple sensitivity analysis where relative survival estimates were adjusted to account for an arbitrary proportion of cause-specific deaths in the external group. Not a novel extrapolation method	Yes, considered a sensitivity analysis of already established relative survival methods	Patient-level data for a cohort + demographic information to estimate general population survival + an assumption concerning the proportion of deaths in the external group were due to the cause of interest	Not applicable	None listed
21	Andersson <i>et al.</i>	2013	Relative survival	Relative survival: Flexible parametric relative survival models	Hakama &, Hakulinen, (1977), Nelson <i>et al.</i> , (2007), Royston & Parmar, (2002)	Method is an extension to the restricted cubic spline, described in detail by Royston and Parmar (2002). To extend to the relative survival framework, consider $S(t) = S^*(t) \times R(t)$, where $S(t) = OS$, $S^*(t)$ = expected survival, $R(t)$ = relative survival. The overall hazards $h(t)$ are therefore estimated as the sum of the expected hazards ($h^*(t)$) and the relative hazards ($\lambda_{\text{rel}}(t)$). Three alternative assumptions considered for extrapolation of the relative survival model: 1. assume log cumulative excess hazard beyond the last boundary knot follows the linear trend given by the estimated model parameters 2. assume statistical cure beyond the last boundary knot, by imposing constraints on spline parameters 3. assume constant excess hazard beyond the last boundary knot, by imposing constraints on spline parameters	Yes, method is an extension to restricted cubic splines, fitted within a relative survival framework. See key references for further detail.	Two data sources required: first, patient-level data from the target population; and second, all-cause mortality data to calculate the expected survival curve $S^*(t)$.	Requires the full all-cause survival curve of patients to be either known or estimated. Need to assume longer-term pattern of cause-specific hazards. In this study, three options were presented: (1) constant, (2) zero (i.e., 'cure'), or (3) linear trend, akin to a 'Weibull-like' tail.	Stata command stpm2 used to fit the flexible parametric models.
22	Seppa <i>et al.</i>	2014	Relative survival	Relative survival: Cure fraction model with random effects to explain regional differences	Standard relative survival references (Ederer, Hakulinen, Pohar Perme, etc.) + cure model by Lambert <i>et al.</i> + Chen <i>et al.</i>	As per approach to a mixture-cure model, with regional random effects modelled by a bivariate normal distribution, which is then introduced into the log-likelihood function	Yes, extends authors previous work on cause-specific survival to instead consider net survival	Patient-level data for a cohort + demographic information to estimate general population survival, including information of patient region in both data sources	Not applicable	Combination of C and R code
23	Jakobsen <i>et al.</i>	2019	Relative survival	Relative survival: Flexible parametric relative survival models	Nelson <i>et al.</i> , (2007), Andersson <i>et al.</i> (2011)	Three different methods considered: 1) Nelson <i>et al.</i> relative survival model, which is linear on the log cumulative excess hazard scale after the last knot, 2) Andersson <i>et al.</i> relative survival model, which is constant on the log cumulative excess hazard scale after the last knot and thereby incorporates statistical cure 3) a flexible mixture cure model, which incorporates statistical cure, but is not restricted to a constant log cumulative excess hazard after the last knot	Yes, considers application of Andersson <i>et al.</i> , (2013) and Nelson <i>et al.</i> , (2007) + new method that addresses issues with Andersson method	Patient-level data for a cohort + life tables	Assumptions specific to each method RE: the extrapolated tail. Else, assumptions per standard relative survival approaches	R, packages: rstpm2 and cuRe
24	Rubio <i>et al.</i>	2019	Relative survival	Relative survival: Correction for misspecification of background mortality rates in relative survival analysis	Cheuvart and Ryan (1991) + standard relative survival references (Ederer, Hakulinen, Pohar Perme, etc.)	Not applicable - not a method to estimate overall survival. Extensive detail provided in the article, but not relevant to this review	Yes, extends 'standard' relative survival modelling methods to adjust expected survival	Patient-level data for a cohort + life tables	Per standard relative survival analyses, plus the expectation of some missing covariates in life tables that would impact background mortality rates, and can be explained through specification of a correction to the background mortality rates	R
25	Turner <i>et al.</i>	1982	Other	Other methods: Bounded survival	None	Specify a parametric model wherein a parameter tau represents plausible maximum lifetime	Yes, extension of simple parametric models	Standard patient-level data + expectation of lifetime survival limit	There is a point in time at which all patients are expected to have died	None listed

Article info			Summary information				Requirements			
#	Author(s)	Year	Category	Name of method(s) used	Key reference(s)	Description of method	Extension to previous method	Data required	Key assumptions	Statistical software
26	Hwang <i>et al.</i>	1999	Other	Other methods: Monte Carlo extrapolation	Mark <i>et al.</i> , (1995)	Three stages: First, obtain a reference curve using a population with 'known' hazard function. Second, fit a simple linear regression to the logit transform of the ratio of survival curves for both the index and the reference populations up to the end of follow-up. Finally, use the regression + reference curve to estimate rest of survival curve. 1) create a reference population of patients using life tables data, via the 'Monte Carlo method' 2) fit a simple linear regression to the logit transform of the survival ratio (denoted $W(t)$ in the paper) between the cancer population and the reference population up to the end of the follow-up period. Regression obtains intercept and slope parameters. 3) use this regression to extrapolate beyond the end of follow-up, using the formula $S(t) = S\_ref(t) * (\exp(\text{intercept} + \text{slope} * t)) / (1 + (\exp(\text{intercept} + \text{slope} * t)))$, where $S\_ref(t)$ = the reference curve	No (or none described)	Two data sources: (1) patient-level data for the population of interest, and (2) reference data for the general population (e.g., national statistics) from which counterfactual survival times (i.e., survival without the disease) may be simulated.	Assumes proportionality of hazards, and that reference curve has a lower hazard of death versus trial group	None provided
27	Nelson <i>et al.</i>	2008	Other	Other methods: Age-based extrapolation model	None	Two-part model: First, a study model based on a Cox PH model fitted to study data. Second, an extrapolation model using external data to describe hazard as a function of age, to then inform remainder of time horizon being modelled	Yes, references studies by authors where method used before, but this is the first methods paper	Two data sources required: first, patient-level data from the target population; and second, an external data source to obtain an estimate of an age-based hazard function, controlling also for covariates (so ideally patient-level data)	Hazards stabilise after a given time point such that they can be represented as a function of patients current age + baseline covariates (not including age)	Custom SAS code
28	Ishak <i>et al.</i>	2011	Other	Other methods: Crossover adjustment using external data	None	Calibrates PPS by making use of an external data source. Obtain an adjustment factor and use this to edit PPS such that median OS matches that of an external source (covariate adjusted).	No (or none described)	Patient-level data from two sources: the main study, and an external source not affected by crossover	Generalisability of study populations (after covariate adjustment)	Suggested SAS, but not explicitly stated
29	Hawe <i>et al.</i>	2014	Other	Other methods: 3-stage approach using data from 2 RCTs	None	Stage 1: Fit parametric models for the control arm in both studies with an indicator for study Stage 2: Obtain treatment effect for active versus control by fitting models with covariates for both arm and study Stage 3: Combine estimate of survival for control from Stage 1 with treatment effect from Stage 2 to obtain survival for active arm	No (or none described)	Data 2 RCTs of the same control arms - ideally patient-level, but could digitise	Differences in survival can be described by covariate adjustment	None listed
30	Hwang <i>et al.</i>	2017	Other	Other methods: Rolling extrapolation	Hwang <i>et al.</i> (1999)	Very similar to Monte Carlo method, except extrapolation based on a spline model, instead of a linear regression	Yes, extension of the Monte Carlo method	Two data sources: (1) patient-level data for the population of interest, and (2) reference data for the general population (e.g., national statistics) from which counterfactual survival times (i.e., survival without the disease) may be simulated.	Assumes that reference curve has a lower hazard of death versus trial group	Links to R program developed by authors (iSQoL2): http://sites.stat.sinica.edu.tw/isqol/
31	Pennington <i>et al.</i>	2018	Other	Other methods: Joint modelling with adjustment to long-term hazards using summary data	Demiris & Sharples (2006), Guyot <i>et al.</i> , (2011).	Couple of options. If only summary stats, can estimate a constant hazard for long-term. If IPD, fit joint models, and consider summary stats for long-term HR estimation.	Yes, uses elements of work by Demiris & Sharples + Benaglia <i>et al.</i> , but not in a Bayesian framework. Similar to Hawe <i>et al.</i>	Varies, but as a minimum IPD for index population + external data either at summary level, with potential IPD for the control arm at least	Assumes similarity of shape in hazard functions across populations. Differences expected to be explained via PH, log odds of event, or acceleration of time (model choice depending)	Used Stata
32	Vickers	2019	Other	Other methods: Multiple, but broadly piecewise models with different early components	Guyot 2017 + Jackson review.	3-part model: First part informed by short-term data (target population). Second part by using adjusted external data until end of follow-up, then last part informed by adjusted general population data. For active treatment, further adjustment by tapering the HR (i.e., treatment waning effect)	Yes, one method adapts the Guyot (2017) approach, else methods use piecewise-style approaches	Data from 2 sources - a short-term (target population) study and long-term external study that is similar + general population survival estimates	Assumed similarity in populations, including constant AF describing differences between disease-specific sources, adjustment of HRs based on assumed time points	R and JAGS. Packages used: eha, flexsurv, survPresmooth, MESS, ggplot2, and loo.

Appendix Table 3: SLR first iteration data extraction (Part 2 of 2)

Article info			Application			Commentary		
#	Author(s)	Year	Context(s) applied	Validation approach(es)	Code provided	Advantages (from author)	Limitations (from author)	Any other comments
1	Messori <i>et al.</i>	1999	Sample of 17 pts with acute leukaemia (patient-level data) + 'vital statistics' for US population	None presented	Not provided, but calculations presented in the paper	Allows extrapolation to a lifetime	Cannot interpret results on an individual level	This is a really simplified method to extrapolate to lifetime survival, essentially imputing general population survival to the end of the observed survival time for survivors.
2	Viscomi <i>et al.</i>	2006	Dataset of people with childhood leukaemia. Patient-level study data (+ age at baseline) + general population survival statistics	None presented	No, not reported	None described	Method less relevant in diseases that are terminal, because of assumed general population mortality after follow-up	Method does not ultimately yield a curve, because the approach estimates means and sums them to obtain overall estimate of life expectancy
3	Larkin <i>et al.</i>	2015	Melanoma - trial data + registry data + life tables. Used in context of explaining approach taken in HTA	Validated extrapolations by speaking to clinical experts. Also compared projections to published long-term survival data for ipilimumab from Prieto <i>et al</i> (56 patients from a clinical trial of previously treated patients) and the recently published pooled analysis of 1,861 patients over all lines of therapy reported by Schadendorf <i>et al.</i> , which was not available at the time of analysis).	No, not reported	More realistic extrapolation based on evident plateau + aligned with clinical opinion versus conventional parametric approaches. Reduced mean absolute error in predictions.	None described	Very similar to method of Barlev <i>et al.</i> , but different sources of evidence and assumptions.
4	Barlev <i>et al.</i>	2017	Adults with leukaemia - trial data + observational study + life tables. Used to estimate survival in this group, not explicitly linked to HTA	Repeated analysis with a later data cut - commented that extrapolations were consistent	No, not reported	No other approaches discussed.	Limitations primarily related to use of historical data, including assumption of no additional effects of blinatumomab beyond two years, potential changes in treatment practices over time in the clinical vs historical populations and differences in transplant realisation rates.	Classic example of a piecewise approach - nothing 'novel' in terms of the method itself, but a clear worked example
5	Fisher <i>et al.</i>	2017	IPF - trial data + observational study. Used to estimate survival in this group, not explicitly linked to HTA	Mentioned not possible to validate as no other external data sources available	No, not reported	No other options mentioned, but implied that parametric models fitted to 1 year of data unlikely to do well	Potential biases with Inova being a registry. Definition of BSC may differ between trial + registry and may have changed over time. Registry KM prior to 52 weeks not same as trial	Another clear worked example, but nothing 'novel'. Interesting aspects are: (a) addressed problem with crossover, and (b) pooled data from multiple Phase 3 studies
6	Thornton Snider <i>et al.</i>	2017	Non-small cell lung cancer + renal cell carcinoma - SEER data + HRs from trials of nivolumab. Used in context of a value demonstration research project	None reported	No, not reported	None described (focus of paper was not placed on the survival estimation method itself)	Lack of long-term follow-up for recently diagnosed melanoma patients required authors to construct survival curves using several sources. Had to assume that 2L started 1 year after diagnosis, respectively. SEER population could not be made perfectly comparable to the clinical trial populations, potentially influencing life expectancy.	A relatively basic (albeit crude) way of adjusting curves using HRs to reflect subsequent therapies
7	Thornton Snider <i>et al.</i>	2018	Melanoma - SEER data + HRs from trials of nivolumab and ipilimumab. Used in context of a value demonstration research project	None reported	No, not reported	Per previous row describing method by same authors	Per previous row describing method by same authors	Per previous row describing method by same authors
8	Demiris & Sharples	2006	Authors present an example concerning implantable cardioverter defibrillator (ICD) therapy. Data come from a cohort study, UK life tables, and a meta-analysis of three RCTs.	No specific validation approaches presented, but authors do compare the proposed methods to a standard Weibull and compare fits.	No. Code is cited but hosted on webpage that no longer exists. Author contacted but no response	Method performs well, compared to a standard Weibull model. Bayesian methods are 'convenient' when prior and likelihood information is to be considered	The gamma process adopted for modelling the baseline hazard considered a basic approach but stated by authors to appear to work very well for the example presented. Authors acknowledge that other more sophisticated approaches exist in the literature which might be useful alternatives should more complicated prior structures for the baseline hazard be considered.	Complex approach compared to others, as code would be helpful to facilitate understanding
9	Benaglia <i>et al.</i>	2015	Applied to a cohort of patients that received implantable cardioverter defibrillators.	Simulation study. Poly-hazard model was compared to two 'incorrectly-specified' models. The incorrectly specified models did not account for the cause-specific effect of an intervention on the hazard of death but assume that the study patients have a proportionally increased hazard for overall survival compared to the general population. Outside the context of the simulation study, the authors fit two different poly-hazard models (poly-Weibull and poly-Gompertz) to the implantable cardioverter defibrillators data. The models are compared in terms of mean overall life-years, and their 'fit' to the Kaplan-Meier estimate and an empirical hazard plot (methods to produce this unclear).	Yes, provided as an appendix within the article ('Appendix A: WinBUGS model code')	Able to address potential bias where simply assuming proportional hazards for all-cause mortality when estimating expected lifetime survival.	Need to consider sensitivity analyses, and careful and consistent categorisation of causes of death in survival data, as results can be highly sensitive to assumed reference population causes of death.	Paper notes that the authors have focused on Weibull and Gompertz, which have the PH property which makes estimation and interpretation slightly simpler (and allows for use of published HRs if relevant).

Article info			Application			Commentary		
#	Author(s)	Year	Context(s) applied	Validation approach(es)	Code provided	Advantages (from author)	Limitations (from author)	Any other comments
10	Guyot <i>et al.</i>	2014	RCT conducted in head and neck cancer	Not clear from abstract	No, Not reported in abstract	The method enabled authors to estimate models consistent with all evidence	No limitations highlighted explicitly within the abstract, though authors note that other flexible models than splines could be investigated.	Only available as an abstract. Full paper published in 2017, which is assumed to supersede this research
		2017	RCT conducted in head and neck cancer	Compared to standard models with no external evidence included	Yes, included in an appendix	Allows for fitting of a single, consistent model that acknowledges all data sources	Issues of generalisability, potential age of studies, method would be difficult to use where data are limited	Really interesting approach, but very computationally challenging. Note that this approach tries to use everything, theoretically leaving nothing to validate with
11	Cope <i>et al.</i>	2019	Trial in leukaemia (young adults, curative potential)	Validated using a later data cut	No, not reported	Leads to an increase in the precision of survival estimates versus trial-data based estimates alone	Limited ability to validate approach given small sample + censoring	Good to be aware of, and the concept is intuitive. However, the use of FP models is non-standard and complex
12	Heinavaara & Hakulinen	2002	Registry for colorectal and lung cancers	None reported	No, not reported	Not applicable	Not applicable	Not applicable owing to specific context
13	Maetani <i>et al.</i>	2004	Gastric cancer cohort	Models were fitted to 5-year follow-up data, and compared to a later data cut ('lifetime follow up')	No, not reported	None explicitly stated	Difficult to estimate/ verify the proportion of people cured. The model generally over-estimated mean survival by a small quantity	This is essentially a mixture-cure model fitted in a relative survival framework
14	Lambert <i>et al.</i>	2005	Large breast cancer cohort (>100,000) to illustrate different ways of estimating relative survival	None reported	No. Code is cited, but hosted on webpage that no longer exists	Additive hazards: Ease of interpretation, can be used to capture zero or potential negative covariates (e.g., covariates meaning that after cure patients have better survival than general population). FP models: a wide range of plausible curves can be explored, use fewer parameters than splines/ piecewise models.	None explicitly described	Different ways of specifying a relative survival model. FP approach very data hungry
15	Stare <i>et al.</i>	2005	Cohort of patients who have previously had a myocardial infarction	None reported	No, not reported	Not applicable	Not applicable	Not an applicable method as nothing to do with extrapolation
16	Lambert <i>et al.</i>	2007	Illustrative examples presented for patients with ovarian cancer	Compared to mixture-cure models	None reported	The NMCM has a PH model as a special case for the whole population (the MCM does not have this property)	None described	A relatively minor extension, but included for completeness
17	Pohar Perme <i>et al.</i>	2009	Long-term survival data for patients that have experienced an MI	Not applicable	None reported	Not applicable	Not applicable	Mostly not applicable as these are estimators that contribute to OS, not estimating OS directly, so not considered further
18	Lambert <i>et al.</i>	2010	Illustrative examples presented for patients with colon cancer	None reported	Yes, link to Stata package included (Lambert, 2007, Stata journal)	Increased flexibility in the shape of the excess hazard or relative survival function versus conventional MCMs/NMCMs	None described	An extension, but very minor
19	Andersson <i>et al.</i>	2011	Colon cancer cohort	Compared with a Weibull non-mixture cure model + the non-parametric Ederer II method	Yes, link to Stata package included (Lambert, 2009, Stata journal)	Improves upon a Weibull based cure model when the Weibull model gives biased estimates, otherwise looks similar	Suffers from same issues faced by all cure models. Selection of the cure point is a major limitation for some contexts.	Assumed this is superseded by the Jakobsen paper
20	Hinchcliffe <i>et al.</i>	2012	Registry for breast, colon, and prostate cancers	Not applicable	None reported	Not applicable	Not applicable	An extension, but not worth including as very minor within context of HTA

Article info			Application			Commentary		
#	Author(s)	Year	Context(s) applied	Validation approach(es)	Code provided	Advantages (from author)	Limitations (from author)	Any other comments
21	Andersson <i>et al.</i>	2013	Applied to data from the Swedish Cancer Registry. Data from four different cancers.	Method evaluated and presented using Swedish cancer registry data (but no validation of these extrapolations separately)	No, but references updates made to the stpm2 command in Stata.	Using a relative survival approach gives better estimates of the full all-cause survival function. Other approaches when applied to grouped data have assumed either cure or a long-term fixed hazard of death. However, in this approach, the extrapolation can also be carried out by extrapolating the linear trend given by the estimated model parameters. Estimation can be carried out on individual level data and not only on grouped data.	Potential problems include sensitivity to the number of knots uses, and the amount of follow-up needed for the extrapolation to perform well. Authors used naive assumption that the population mortality rates stay the same from 1980 onwards. Extrapolations could be poor if background mortality rates change substantially over time. Approach has not been evaluated on small datasets or data from clinical trials or cost-effectiveness studies (but could be useful in these kind of studies)	Paper focused predominantly on estimating the loss in expectation of life due to cancer, as opposed to the all-cause mortality curve itself. However, authors note that this method can be used for estimating the survival curve itself (but the metric of interest to the study may be considered useful especially within the context of defining the burden of disease and comparing cancers).
22	Seppa <i>et al.</i>	2014	Colon cancer cohort	Not applicable	Code provided, but no dummy data set - available via link provided in the paper	Not applicable	Not applicable	Specific edit to a MCM to consider regional differences - not directly relevant for review
23	Jakobsen <i>et al.</i>	2019	Danish lymphoma registry	Simulation study	No, not reported	None described	None described	Essentially a mixture-cure model fitted in a relative survival framework
24	Rubi o <i>et al.</i>	2019	Lung cancer cohort	Not applicable (but did a simulation study)	Yes - available on request from authors, and some hosted via GitHub/ personal website	Not applicable	Not applicable	An extension, but not worth including as very minor within context of HTA
25	Turner <i>et al.</i>	1982	Life tables	None reported	No, not reported	Addresses a 'serious deficiency' of standard approaches wherein death is not guaranteed to occur by a fixed time.	None described	Cannot verify method with example provided.
26	Hwang <i>et al.</i>	1999	Simulation study + illustrative examples. Later studies have applied the method in practice	Simulation study	No, not reported	Can make use of registry databases or life tables, so is quite flexible	Assumes constant excess hazard, which is important to test, and if this assumption does not hold, method may not extrapolate well	Could be argued as a piecewise method, but is very specific and has been referenced a lot
27	Nelson <i>et al.</i>	2008	Patients that have had an MI	Some later 5-year data reported	Yes, included in an appendix	Ability to adjust for covariates. Doesn't assume a functional form because Cox PH rather than Weibull, Gompertz etc. Makes predominantly 'weak' assumptions. Can be implemented in standard software (e.g., SAS). Use of age metric makes efficient use of available data	Complex methodology. Not applicable to all situations (e.g., small sample sizes)	Interesting approach, but very data hungry, and complex to implement (even if code is available)
28	Ishak <i>et al.</i>	2011	Multiple myeloma - two RCTs + evidence from 4 other MRC trials	None presented (though not explicitly possible given crossover research question). Did comment on similar findings to improvement in TTP versus OS post crossover	No, not reported	Uses more 'standard' statistical methods, versus complex crossover methods (e.g., RPSFTM)	Relies on external data being 'free from bias'. Cannot remove impact of crossover in the external study. Other generalisability concerns with respect to treatment received, age of study etc.	Very specific application within the context of crossover + adjustment to OS in the context of PFS and PPS.
29	Hawe <i>et al.</i>	2014	Two RCTs in CLL	None reported	No, not reported	Can be used to make use of external data when absolute outcomes differ	Assumes PH/AFT (depending on model fitted) between the study arms Choice of model could influence extrapolations markedly Difficult to validate given lack of separate long-term data	Potentially helpful within the context of having access to the two studies with the same control arm, but unlikely to occur in practice.
30	Hwang <i>et al.</i>	2017	Simulation study + illustrative example	Simulation study + compared to a standard Weibull model	Yes, in the form of the iSQoL2 package	Does not require excess hazards to be constant over time	Logit W(t) values may be unstable if small n + limited follow up, which could cause issues in the rolling extrapolation algorithm	Modification to the Monte Carlo method
31	Pennington <i>et al.</i>	2018	Data set from the National Joint Registry, about knee replacements	Compared estimates relative to 'best possible model' using all available data	Yes, included in an appendix	Could help improve estimates of cost-effectiveness (assuming estimates without external data are pessimistic)	Risk of selection bias when using observational data. Some specific assumptions made on relationship between curves, which may not be possible to verify.	Outcome not OS - actually prosthesis survival
32	Vickers	2019	Semi-simulated data from patients with NSCLC. External data from a published paper + general population data	Compared extrapolations to external data, but no formal validation presented	Some code included in an appendix, but not the full code for all reported analyses	External data needed to produce accurate extrapolations in presence of a 'large' treatment effect	Relies on similarity of external data. Need to assume time points to taper hazard ratio.	A series of quite complex adjustments to 3 data sources to ultimately give a piecewise-style model. Best method identified by the author was the Guyot (2017) method

Appendix Table 4: SLR second iteration data extraction (Part 1 of 2)

Article info			Summary information				Requirements			
#	Author(s)	Year	Category	Name of method(s) used	Key reference(s)	Description of method	Extension to previous method	Data required	Key assumptions	Statistical software
1	Jackson <i>et al.</i>	2010	Bayesian	Bayesian: Semi-parametric Bayesian model with constant hazard extrapolation	Kalbfleisch, (1978), Demiris & Sharples, (2006), Nelson <i>et al.</i> , (2008)	The method itself is similar to a Cox model, in that a semi-parametric model for survival is estimated with various covariates. Hazards are assumed to be piecewise constant, changing only upon the occurrence of a death event. After the final death event, the hazard of death is assumed to remain constant for the remainder of the modelled time horizon. Can then use the posterior survival probabilities to inform a model, by extracting relevant probabilities to apply in a state-transition model (e.g., risk of death for patients in a particular age/sex/stage group).	Builds upon a Cox model and a Bayesian semi-parametric model without extrapolation	Patient-level data from index population + prior beliefs around model parameters and/or outcomes in general (e.g., 5-year survival)	Assumes long-term hazard of death is constant	WinBUGS
2	Murray <i>et al.</i>	2014	Bayesian	Bayesian: Semi-parametric Bayesian commensurate survival model	Ibrahim <i>et al.</i> , (2001), Plante, (2009), Hobbs <i>et al.</i> , (2012), Hobbs <i>et al.</i> , (2013)	Jointly fits a semi-parametric Bayesian model based on two data sources (current and historical), to borrow strength from historical data. Model fitting process is computationally intensive, so please refer to the main paper for a detailed description.	Builds on work by Hobbs <i>et al.</i> whose method relied on exchangeability of data sources	Two data sets - one from a historical and one from a current study, both including patient-level characteristics	Assumes the current and historical data sources are assumed to be non-exchangeable. Also assumes covariates are captured in a Cox-like manner, and baseline hazards are captured via a piecewise exponential model	JAGS code (via R)
3	Lewis <i>et al.</i>	2019	Bayesian	Bayesian: Bayesian hierarchical parametric survival model	Lunn <i>et al.</i> , (2000)	Fits a Bayesian hierarchical model to borrow strength from historical control data. The strength of borrowing is contingent on the specification of a parameter, tau, under the commensurate prior framework.	Builds on a standard Bayesian hierarchical modelling method used in other settings	At least two data sets - one from a historical and one from a current study, both including patient-level characteristics (if wanting to adjust for these). Method could consider more sources if pooling historical control data.	Assumes patient survival time follow the same distribution type across sources (in this example, Weibull). Also need to determine how influential historical control data are through specification of priors	JAGS code
4	Soikkeli <i>et al.</i>	2019	Bayesian	Bayesian: Specification of a prior on the shape parameter	Jackson review, Demiris & Sharples, (2006), Pennington <i>et al.</i> , (2018)	Fits a parametric model with an informative prior for the shape parameter informed via historical data	Bayesian alternative of the Hawe <i>et al.</i> method	Two data sets - one from a historical study for long-term shape parameter, and another to produce extrapolations for. Could perform method on pseudo-IPD or 'true' IPD	Assumes PF/ constant AF if including treatment as a covariate. Otherwise, assumes 'similar' shape parameter/functional form of model across studies	WinBUGS
5	Nelson <i>et al.</i>	2007	Relative survival	Relative survival: Spline model	Royston & Parmar, (2002); standard relative survival references (e.g., Esteve); Lambert <i>et al.</i> , (2005).	Fits an additive hazards relative survival model, using a spline for disease-specific survival	Extends spline model to relative survival	Individual-level data from trial population + background mortality estimates	Assumes hazards are additive	Stata
6	Mba <i>et al.</i>	2020	Relative survival	Relative survival: Correction for inaccurate background mortality through specifying breakpoints	Esteve <i>et al.</i> , (1990) and Touraine <i>et al.</i> , (2020)	Additive hazards model extended to capture differences in background mortality between trial and general populations which also changes over time. Differences captured using a multiplicative variable which changes value over time according to the specification of breakpoints.	Extends the Touraine model, which itself was an extension of the Esteve model	Individual-level data from trial population + background mortality estimates	Assumes hazards are additive + difference exists in background mortality in study versus general population, which also changes over time	R
7	Felizzi <i>et al.</i>	2021	Other	Other: Informed MCM	Andersson <i>et al.</i> , (2013), Chen (2015), Othus <i>et al.</i> , (2017)	Fits a standard MCM but imposes the assumption that the cure fraction is fixed at a static value, obtained via some external evidence source	Yes, builds upon a standard MCM	Individual-level data from trial population + background mortality estimates + relevant information for cure fraction and epidemiological information in target population	Can obtain an estimate of the cure fraction from an external source, and consider this 'fixed'	R
8	van Oostrum <i>et al.</i>	2021	Other	Other: background mortality methods	TSD 21	Four different methods: (1) internal additive hazards (relative survival model), (2) external additive hazards (add on background mortality, assuming negligible during follow-up), (3) converging hazards model (cap hazards by background mortality), and (4) proportional hazards model (estimate survival by applying a hazard ratio to a background mortality curve)	No	Individual-level data from trial population + background mortality estimates	Internal additive hazards model: Hazards related to disease and other causes are additive (not multiplicative); external additive hazards model: In addition to the assumptions made for the internal additive hazards model, assumes general population hazards negligible during trial follow-up; converging hazards model: Without adjustment, model would yield estimates of long-term hazards that fall below general population. This method assumes it is reasonable to 'ignore' implausible long-term extrapolation in favour of general population hazards. Indirectly, method also assumes 'cure' when switching to general population hazards; proportional hazards model: relationship between general population survival and the survival of the group of interest can be adequately represented via a hazard ratio	R
9	Che <i>et al.</i>	2021	Bayesian	Bayesian: Blended survival model	None	Model fitted over three periods: (i) initial period where survival follows that of the index group, (ii) mid-range period ('mixing area') where survival is considered a weighted average of the index and external groups, and (iii) a long-term period where survival is based on the external group only. The model is fitted making assumptions about the 'mixing area'	Considered to take learnings from mixture-cure and poly-hazard models	Some form of external data such that a relevant time point (e.g., 10 years) and survival proportion (e.g., 20%) can be determined	Relies on several important parameter assumptions, including how 'quickly' the population follows the survival experience of the external cohort	R (based on presentation, but not explicitly stated)

Appendix Table 5: SLR second iteration data extraction (Part 2 of 2)

Article info			Application			Commentary		
#	Author(s)	Year	Context(s) applied	Validation approach(es)	Code provided	Advantages (from author)	Limitations (from author)	Any other comments
1	Jackson <i>et al.</i>	2010	Illustrated with an example in oral cancer screening	Compared approach to other methods, all fitted in a Bayesian framework, but no formal validation presented.	Paper explains that is available as an appendix on the publisher's web site, but no hosted code is available.	Better than a 'standard' Cox model approach because expressing uncertainty is 'more straightforward' under a Bayesian model	Assumption of a constant hazard is unrealistic for extrapolation for older populations. Authors comment that additional data/ expert judgement would be needed in practice to inform long-term models. When fitting the Bayesian semi-parametric model, some smoothing of the hazard function is required to obtain adequate predictive ability and avoid sensitivity to the choice of prior.	Very limited emphasis on extrapolation, so excluded from further consideration
2	Murray <i>et al.</i>	2014	Method illustrated using post-market surveillance data concerning adverse events in people on dialysis that had a stent implanted during cardiac surgery	Conducted simulations to investigate the method (but not explicitly referred to as a simulation study, and provided only in a supplementary appendix)	Yes, included as online supplementary material	Can be used in the context of non-exchangeable data sources	Could be extended to consider more flexible functions for baseline hazards (e.g., splines).	Very complex method but no commentary on extrapolation, so not directly relevant to the review
3	Lewis <i>et al.</i>	2019	Tested in the context of three colorectal cancer trials (RCTs)	Simulation study	Yes, provided in an appendix	(Not explicitly stated by authors but inferred through presentation of results). Extent to which historical data influence model parameter estimates can be varied and tested as part of a sensitivity analysis	Inappropriate borrowing of historical control information could lead to erroneous conclusions	Extension of a Bayesian hierarchical model, but not focused on extrapolation, so further extensions would be needed to inform extrapolation. Included for now but may have limited applicability when attempting to use in an HTA setting.
4	Sokkeli <i>et al.</i>	2019	Newly diagnosed multiple myeloma (RCT population)	Extrapolations validated against a later data cut from the same trial	Yes, included as supplementary material	Improved upon extrapolations without specification of a prior in the example presented.	Not suitable for situations where a constant treatment effect cannot be specified (if capturing multiple treatment groups). Limited flexibility of the standard parametric models considered	Formalises the Hawe <i>et al.</i> method within a Bayesian framework
5	Nelson <i>et al.</i>	2007	Real-world population with coronary heart disease	Compared with 'standard relative survival analyses'	No, but 'Stata program available from the authors on request.'	More flexible approach to model excess hazard, avoids needs to split timescale (as would be needed in a piecewise approach).	Number of knots can be influential on results	Extension of relative survival to include spline models - approach already covered from first iteration, but this is the first study for this method specifically
6	Mba <i>et al.</i>	2020	French population-based colorectal cancer data	Simulation study	No, but 'programs available on request'	More flexible versus Touraine model as captures changes over time	Risk of over-parameterisation if several breakpoints specified, and assumes changes are piecewise (i.e., constant within time intervals)	Captures possible differences in background mortality between study and general populations accounting for time-varying effects, but likely far too complex for HTA and does not directly impact survival extrapolation. Excluded from further consideration.
7	Felizzi <i>et al.</i>	2021	Illustrated with an example in advanced melanoma	None presented	Yes, provided a link to a GitHub repository	Can address limited data available to robustly estimate an MCM via an uninformed approach	Requires IPD, cure fraction may be uncertain and not reflected in the model, and potential differences in trial versus background mortality populations	Minor modification to a standard MCM, just fixing the cure fraction according to some external evidence source
8	van Oostrum <i>et al.</i>	2021	Newly diagnosed multiple myeloma (RCT population) + breast cancer (another RCT)	Extrapolations validated against a later data cut from the same trial	No	General population mortality data are routinely available in the UK. Transparent and intuitive application of general population data to inform extrapolations.	None explicitly highlighted for the approach in general. For the PH model, the ability to justify PH holds is a limitation. The converging hazards require assumption of cure which may not hold, and if it does hold cure-based models may be preferred. External additive hazards requires general population hazards to be negligible during follow-up.	Range of background mortality methods, those authors comment that internal additive hazards seems to be the best
9	Che <i>et al.</i>	2021	Chronic lymphocytic leukaemia (CLL) - RCT	None presented	No	- Overcomes the overly-optimistic constant treatment effect - Sufficient flexibility with various plausible scenarios - Less concerned about model selection	- Elicited belief: probably a weak form of evidence, need long-term data - Does not account for other kind of external knowledge	Similar to a piecewise model, except there is a gradual movement away from data for the index population to the external population (as opposed to an immediate change), and the model is fitted in one consistent framework

Appendix D. Methods excluded from the systematic review

First iteration

Turner et al., (1982)

A study by Turner et al., (1982) described a ‘bounded survival’ method to address a “serious deficiency” of standard parametric models: “No allowance is made for the case where life is finitely limited, where death is guaranteed by a fixed time, where in the absence of overt disease or accident, life will still be terminated after a certain length of time” (Turner et al., 1982, p.34).¹⁶⁷ In other words, standard parametric models can (without adjustment) project a non-zero proportion of patients to still be alive beyond a realistic LE (e.g., beyond 122 years and 164 days of age – the age at which the oldest person ever to have lived, Jeanne Louise Calment, died according to Guinness World Records).²⁹⁴

To explain the bounded survival model, the authors of this study introduce a ‘master model’. First, an equation is defined for the derivative of $Q(t)$, the probability of dying by time t (i.e., the survival function $S(t) = 1 - Q(t)$). This is presented as a differential equation:

$$\frac{dQ}{dt} = \beta(t) \left[\frac{p}{|p|} \left(Q - \frac{|p| - p}{2|p|} \right) \right]^{1+n} \times \left(\frac{n}{|n|} \left\{ \left[\frac{p}{|p|} \left(Q - \frac{|p| - p}{2|p|} \right) \right]^{-n} - 1 \right\} \right)^{1+p}$$

Note: Notation in the equation above has been updated to reflect standard notation of differential equations.

In the master model, n and p are model shape parameters, and $\beta(t)$ is called the ‘modulus of mortality’. The authors explain that in previous work, $\beta(t)$ is assumed to be a constant; however, in the bounded survival model, $\beta(t)$ can vary over time, subject to some constraints:

- $\beta(t)$ must be integrable.
 - This means that $\beta(t) = \frac{d}{dt}B(t) > 0$, for $0 < t \leq \tau$, where τ is specified as the upper bound on survival (i.e., the maximum length of life).
- $B(0) = 0$ and $B(\tau) \rightarrow \infty$.

By considering these constraints, it is then possible to integrate the previous differential equation to obtain the mortality function $Q(t)$:

$$Q(t) = \frac{|p| - p}{2|p|} + \frac{p}{|p|} \left\{ 1 + \frac{n}{|n|} \left[\frac{|n| - n}{2|n|} + |np|B(t) \right]^{-\frac{1}{p}} \right\}^{-\frac{1}{n}}$$

This equation for $Q(t)$ then simplifies depending on the signs of n and p , as shown in Appendix Table 6. Of note, at least one of n and p must be positive.

Appendix Table 6: Special cases for bounded survival method

Case	Equation
Both n and p are positive	$Q(t) = 1 - \left\{ 1 + [n p B(t)]^{-\frac{1}{p}} \right\}^{-\frac{1}{n}}$
n is positive, p is negative	$Q(t) = \left\{ 1 - [1 + n pB(t)]^{-\frac{1}{p}} \right\}^{-\frac{1}{n}}$
n is negative, p is positive	$Q(t) = \left\{ 1 + [np B(t)]^{-\frac{1}{p}} \right\}^{-\frac{1}{n}}$

It is next necessary to determine an appropriate function for the modulus of mortality $\beta(t)$, and its integral $B(t)$. The authors suggest specification of a function termed the '*hourglass function*'. The full derivation of the hourglass function is presented by Turner *et al.*, (1982)¹⁶⁷; however, for brevity only the resultant formulae are presented for $\beta(t)$ and its integral $B(t)$:

$$\beta(t) = \beta \left(1 - \frac{t}{\tau} \right)^{-\delta\tau-1}, \quad B(t) = \frac{\beta}{\delta} \left[\left(1 - \frac{t}{\tau} \right)^{-\delta\tau} - 1 \right]$$

In the hourglass function, β and δ are constants, and τ is the (specified) upper bound on survival. Using the hourglass function, a bounded survival equation for $Q(t)$ can be produced by replacing $\beta(t)$ in the general equation for $Q(t)$:

$$Q(t) = \frac{|p| - p}{2|p|} + \frac{p}{|p|} \left\{ 1 + \frac{n}{|n|} \left[\frac{|n| - n}{2|n|} + \frac{\beta|np|}{\delta} \left[\left(1 - \frac{t}{\tau} \right)^{-\delta\tau} - 1 \right] \right]^{-\frac{1}{p}} \right\}^{-\frac{1}{n}}$$

This bounded master function for $Q(t)$ can then be used to reflect a variety of different survival models.

Key assumption(s)

The main assumption made by the authors in this approach is that there exists $\beta(t)$, the modulus of mortality, which is assumed to be represented by an hourglass function. Fundamentally, $\beta(t)$ is associated with specific constraints which in theory allow for a bounded lifespan to be reflected within the survival equation.

Strengths and limitations

The principal strength of the proposed approach is that it addresses a clear limitation of standard parametric models and represents an intuitive approach to do so – that is, a 'bound' is imposed to ensure estimates of survival are realistic, and that this bound is set by the analyst. However, the method is computationally complex, and it could be argued that simpler

means of ensuring extrapolations do not provide unrealistic extrapolations exist which would render this approach unnecessarily complex.

Rationale for exclusion

In the article, an example is provided wherein it states: “*The master model with p negative was fitted to the data by maximum likelihood... Estimates were found to be: $\beta = 1.95 \times 10^6$, $\hat{\delta} = 10.8$, $\hat{t} = 115$, $\hat{n} = 0.524$, and $\hat{p} = -3.04$.*” (Turner *et al.*, 1982, p.43).¹⁶⁷ However, when investigating the method further (i.e., not reported by the original study authors), substituting these values in the bounded master function for $Q(t)$, for all values of $t \in [0, \infty)$, $Q(t) = 0$. Consequently, it was not possible to produce a visual representation of the example presented by Turner *et al.*, and therefore it was excluded on the basis of it not being possible to verify the approach taken.

Hwang & Wang (1999)

Hwang and Wang present a method called the ‘Monte Carlo estimator’.¹⁶⁸ The method is proposed as a means of extrapolating quality-adjusted survival for a given population of patients under consideration (henceforth termed the index population), but in my illustration the method is described assuming a utility value of 1 for simplicity. The method is considered in three parts.

Part i)

The first part of the method requires the estimation of a reference curve. The authors explain that at least matching patients based on age and sex should be undertaken if using life tables. Then, if life tables are to be used, a reference curve can be generated using the following approach (assuming for now matching is only performed on age and sex):

1. For a given patient i , identify from the life table of the general population the proportion of individuals (matched for age and sex) that are alive at time $x + k$, but die before time $x + k + 1$, defined as p_{x+k}^{x+k+1} .
2. The probability of survival at time t for an equivalent reference patient given they have survived up until time x can therefore be estimated as:

$$S(t|x) = \prod_{k=0}^t 1 - p_{x+k}^{x+k+1}$$

3. After drawing a random number from a Uniform distribution between 0 and 1, defined as R , a given survival time t_x can be generated such that $S(t_x|x) = R$ (in other words, survival times are randomly drawn using the equation shown above in order to

construct a life table). The resulting survival time is therefore assumed to be the corresponding survival time for the reference population.

4. The survival curve of the reference population can then be produced by using the simulated survival times to produce a KM estimate of OS.

Part ii)

After establishing a reference curve, the next part of the method requires the estimation of the ratio of OS between the index and reference populations, evaluated at time t . This ratio is termed $W(t)$, and is expected to either slowly decrease or remain approximately constant after t is larger than a 'critical' or 'unstable' time point, denoted T_S . $W(t)$ may also be considered as a measure of relative survival.

Since the hazard of death is expected to be poorer in the index population versus the reference population, it is expected that $W(t)$ takes a value between 0 and 1 for all values of t . To consider a regression model to explore how $W(t)$ changes as t increases, a logit transformation is applied such that $\text{logit}(W(t))$ takes values on the scale $(-\infty, +\infty)$. The logit function is presented below for completeness:

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

The authors explain that the tail ends of the logit transform of $W(t)$ are expected to be unstable, but the central part of the transform will be approximately linear. Using this linear aspect of the logit transformed $W(t)$, a linear regression can then be fitted:

$$\text{logit}(W(t)) = \log\left(\frac{W(t)}{1-W(t)}\right) = \alpha + \beta t + N_t \quad \text{for } T_S \leq t \leq T_E$$

In the equation above, α and β are intercept and gradient terms within a simple linear regression, and N_t represents a noise term which is independently and normally distributed with mean $\mu = 0$ and variance σ^2 . T_S represents a time point after which $W(t)$ is expected to be approximately constant, whereas T_E represents a time point less than or equal to the last recorded follow-up time, T_F . The determination of both T_S and T_E is not automated, but the authors describe several ways they may be determined, including reference to a study by Gelber *et al.*, (1993).¹⁸⁷ Alternatively, the authors explain that different values of T_S and T_E may be explored and edited based on the fit of the linear regression.

Part iii)

After determining the least squares estimates of the regression parameters ($\hat{\alpha}$ and $\hat{\beta}$), an estimate of OS curve for the index population for $t \geq T_F$ can be obtained using the following equation:

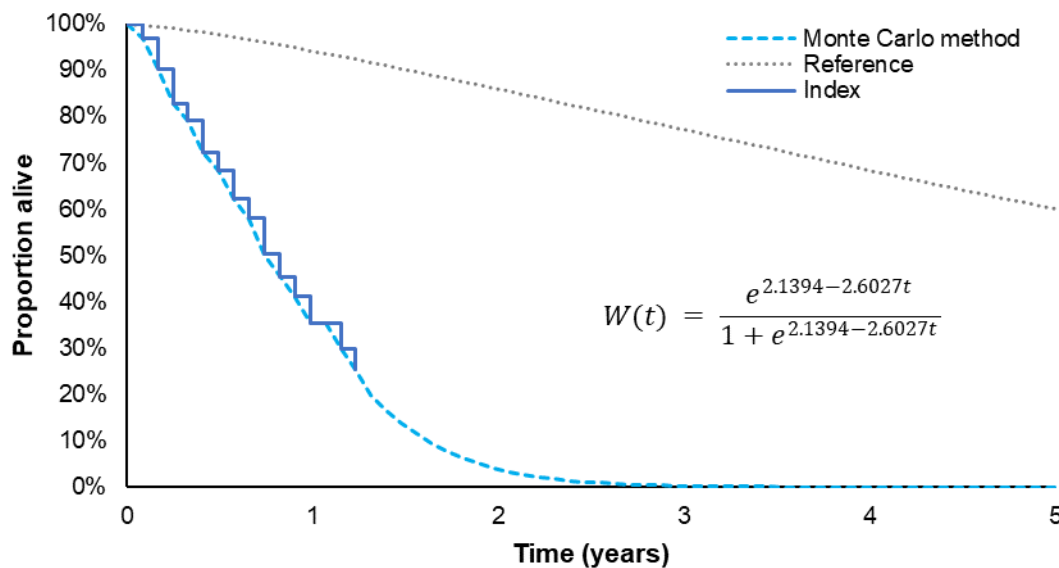
$$S_{Index}(t) = S_{Reference}(t) \left(\frac{\exp(\hat{\alpha} + \hat{\beta}t)}{1 - \exp(\hat{\alpha} + \hat{\beta}t)} \right)$$

In the equation above, $S_{Index}(t)$ denotes the survival curve for the index population, and $S_{Reference}(t)$ denotes the survival curve for the reference population. Using this relationship, the survival curve for the index population can be extrapolated after T_F using a piecewise approach.

A worked example/ illustration of the method (produced using the approach outlined by the authors using dummy data and assumptions) is provided in Appendix Figure 1, where three curves are presented:

- 'Index' is the estimated, incomplete survival curve for the population of interest.
- 'Reference' is a complete survival curve for the reference population.
- 'Monte Carlo method' is the resultant extrapolation obtained.

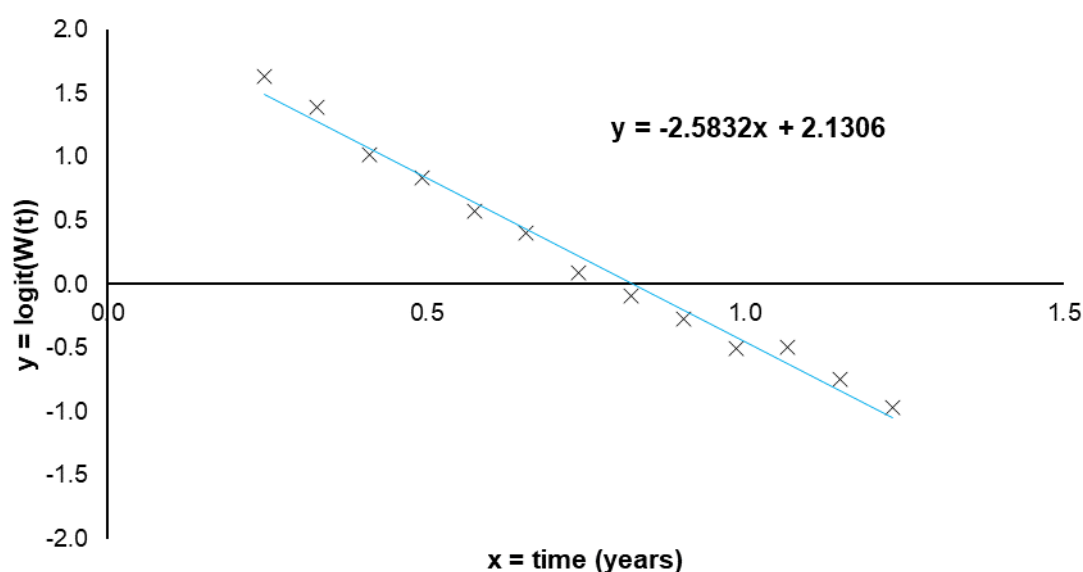
Appendix Figure 1: Illustration of Hwang *et al.* Monte Carlo method



Key: t, time (measured in days, but x-axis presented above in years); $W(t)$, ratio of survival between the index and reference populations at time t.

In this example, the ratio of OS between the index and reference populations was estimated based on the time interval [90, 450] days (i.e., T_S was set at 90 days, and T_E was set at 450 days [same value as T_F]). This yielded the linear regression presented in Appendix Figure 2 (alongside the regression equation), which was then used to inform the remainder of the survival curve presented in Appendix Figure 1.

Appendix Figure 2: Output of regression using Monte Carlo method



Key: $W(t)$, ratio of overall survival between the index and reference populations, evaluated at time t .

Key assumption(s)

A key assumption of the Monte Carlo method is that the excess hazard for the index population versus the reference population is modelled to be constant, or (as described by the authors) $W(t)$ is “assumed to behave in a stable manner” after T_S (Hwang *et al.*, 1999, p.1630).¹⁶⁸ This follows directly from the specification of a linear regression model to describe $W(t)$ (i.e., relative survival). A further assumption made is that $W(t)$ takes values between 0 and 1 for all timepoints, meaning that the reference population is expected to always have a lower hazard of death at all timepoints versus the index population.

Strengths and limitations

Compared with a standard piecewise approach, this method does not assume ‘cure’ beyond the duration of follow-up, but still makes use of general population data to extrapolate including an adjustment. Therefore, for populations where an assumption of cure is inappropriate, this approach provides a means of using general population data in a more suitable way. However, the method is not without limitations.

Firstly, owing to the assumption of a constant HR after T_F , the relationship between hazards in the index and reference populations is assumed to be fixed over time (i.e., the method assumes the PH assumption holds). In reality, this assumption may not hold in the longer term, and it is likely not possible to test the assumption (unless additional data were available in which case, they could be used directly instead of undertaking the approach proposed).

In addition, the method requires estimation of a reference curve, and proposes an approach using general population data including drawing random samples. For populations with small sample sizes, it may be necessary to adjust this part of the method to include some element of bootstrapping to ensure the reference curve is not adversely impacted by a particularly high or low random number draw. However, the resultant reference curve may imply extrapolations are more certain than they truly are (i.e., extrapolations are 'anchored' on the reference curve, and will therefore follow the overall shape projected by it).

Finally, it is unclear how the method would most appropriately be subject to uncertainty analysis. For example, the method requires assumptions related to the choice of a time interval to estimate $W(t)$, but it is unclear how this should (or could) be varied within sensitivity analysis. Furthermore, it is unclear how to best reflect parameter uncertainty within the estimation of the reference curve and/or the coefficients of the regression equation for $W(t)$.

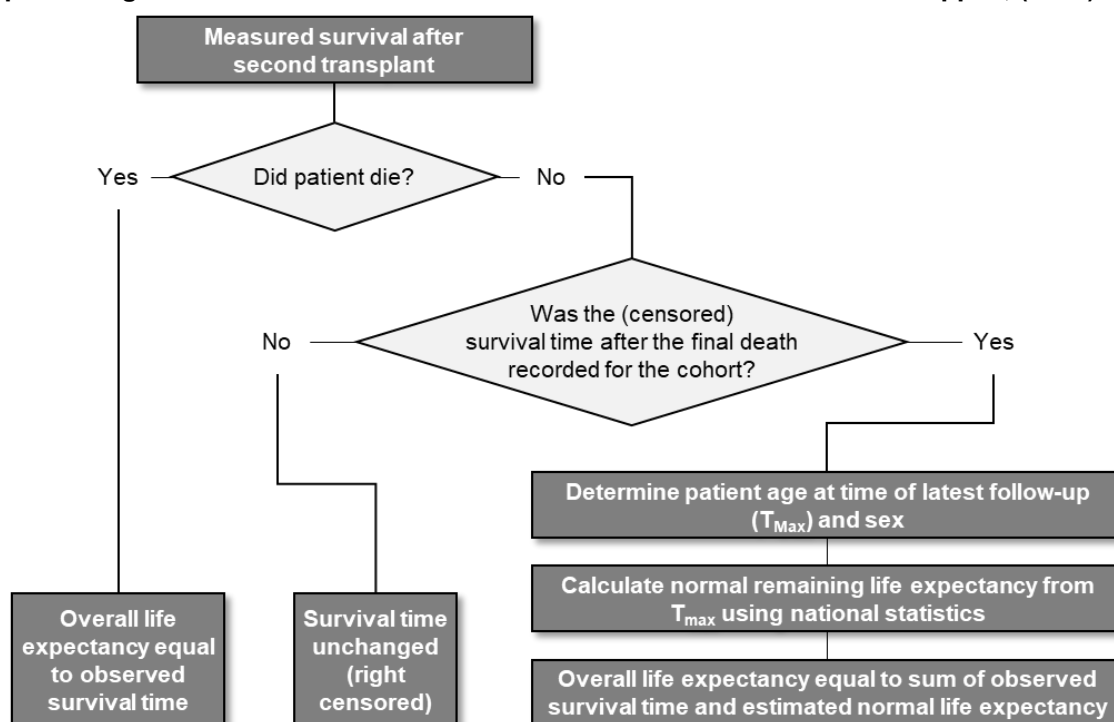
Rationale for exclusion

This method is excluded on the basis of the author later publishing the 'rolling extrapolation' method which supersedes this approach.

Messori & Trippoli, (1999)

In this study, the authors aimed to develop a new method to express lifetime survival for patients within the context of a cost-effectiveness analysis.¹⁴⁴ To illustrate their method, a worked example is presented, using a cohort of 17 patients with relapsed acute leukaemia after allogeneic bone marrow transplantation and then a subsequent (second) transplant. The authors combined complete survival data with imputed, incomplete survival data based on the process presented in Appendix Figure 3, to estimate the time in days since the time of a second transplant.

Appendix Figure 3: Process to determine survival in method of Messori & Trippoli, (1999)



After re-constructing the survival data for all patients, these data can then be analysed using a range of techniques, though given that the data may be considered near-complete (with the exception of patients that are right-censored), mean survival may be possible to evaluate via the KM estimate directly.

Key assumption(s)

Using this approach, the authors assumed that long-term survivors would have a remaining LE equivalent to that of the age- and sex- adjusted general population (estimated from the last date at which the patient was known to still be alive) – in other words, patients alive after the end of follow-up are assumed to be effectively ‘cured’. In the worked example presented by the authors, only 3 patients were still alive at the end of follow-up, all of whom had recorded survival times greater than the longest survival time of patients that had died. This means that no patients were ultimately right censored in the analysis, as in the total sample of $n=17$ patients, $n=14$ patients had recorded death events, and $n=3$ patients were extrapolated to a lifetime horizon assuming cure. However, the authors explain in their discussion that if some patients had censored observations that were recorded prior to the latest death, these should not be extrapolated to assume cure, and instead should be right censored at the time of observation.

Strengths and limitations

The key strengths of this method are that it is relatively simple to understand and could be readily explained. In addition, although the assumption of cure is strong, it is well-defined and the method itself is therefore easily re-producible. The fundamental limitation of this analysis lies within the assumption of cure, and when patients are assumed to have met the threshold of being considered cured. It is not possible to ascertain if the assumption of cure (in the sense that disease-related mortality is zero) is reasonable, and even if it is reasonable when it should apply from. Linking this to the pattern of data collection within the study, while pragmatic, is not an ideal means of determining whether or not patients are cured, as this is likely driven by the study design, rather than the underlying disease.

Rationale for exclusion

This method relies on the assumption that patients are effectively 'cured' if they survive longer than a given time point (which happens to be the latest timepoint of a recorded death in the sample). This assumption is unlikely to hold in other situations and may not truly hold within the context considered by the authors of this study. For example, if in another setting the study considered follow-up for a maximum of 6 months, and the final death occurred at 4 months, it is unlikely to be appropriate to assume all patients censored after 4 months are effectively cured. Elements of this method (e.g., making use of general population statistics, and the possibility of patients surviving in the longer term being 'cured') feature in other piecewise (and non-piecewise) approaches that are considered further.

Heinävaara & Hakulinen (2002)

Heinävaara & Hakulinen proposed an extension of an MCM fitted within a relative survival framework, as a means of estimating net survival to account for patients with multiple cancers (in the example considered, patients who were originally diagnosed with colorectal cancer, who went on to develop lung cancer).^{154, 295*} The authors acknowledge the original use of an MCM was developed by Boag (1949)²⁹⁶, which is discussed in further detail when describing a study identified by Maetani *et al.*, (2004).¹⁵⁵ To illustrate the approach of Heinävaara & Hakulinen, consider the following:

- Patients are first diagnosed with (localised) colorectal cancer (C_0) at age x_0 . Consider this group G .

* In the formal database searching, a different publication by the authors was identified. However, following an email request for this study from the author (owing to the journal no longer being possible to access online), an alternative study was provided in lieu of the original paper (published in the same journal, in the same year, by the same authors, on the same topic).¹⁵⁴

Heinävaara S and Hakulinen T. Parametric mixture model for analysing relative survival of patients with multiple cancers. *J Cancer Epidemiol Prev* 2002; 7: 147-153. 2003/04/01. This alternative paper is discussed henceforth given that it is describing the same overarching method and was available at full text, but both citations are included for completeness.

- These patients may later develop lung cancer (L_2) when aged x_2 .
- At a given follow-up time t , patients have a positive excess hazard because of C_0 , which can be defined as $h_{C_0}(t)$.
 - This excess hazard can be calculated by taking the difference between the total mortality hazard of the group itself, $h_T(t)$, and the expected mortality hazard in the general population taking into account current age, $h_e(x_0 + t)$; which can be expressed as: $h_{C_0}(t) = h_T(t) - h_e(x_0 + t)$.
- However, patients who have L_2 have an excess hazard $h_{L_2}(t)$ because of their cancer, which can be expressed as: $h_{L_2}(t) = h_T(t) - h_e(x_0 + t) - h_{C_0}(t)$.
- A separate group of patients, J , may be first diagnosed with lung cancer (L_1) at age x_1
 - At a given follow-up time t , patients in this group have a positive excess hazard because of L_1 , which can be defined as $h_{L_1}(t)$.

The above considerations of excess hazards hold if it can be assumed that the conditional survival from the first cancer is independent of the conditional survival from the second cancer. The authors acknowledge that this assumption may be reasonable, but that it should be considered in the context of empirical data being analysed.¹⁵⁴ Using similar notation to that highlighted above, the authors then define a parametric MCM of relative survival:

$$S_v(t_v, z_v) = [P_v + (1 - P_v)\exp(-\lambda_v t_v)^{\gamma_v}]^{\exp(\beta_v z_v)}$$

Here, the subscript v refers to cancers C_0 , L_1 , and L_2 . However, in the context of prognostic factors included (z_v), L_1 and L_2 are assumed to be equivalent. P_v is defined as the cure rate for cancer v , whereas $\lambda_v > 0$ and $\gamma_v > 0$ are shape and scale parameters, respectively. Finally, β_v is a vector of parameters related to the prognostic factors. This formula can then be used as the basis for comparing survival from lung cancer as a first or second cancer, for which further details are provided in the original study by Heinävaara & Hakulinen.¹⁵⁴

Rationale for exclusion

The authors cite the use of the ‘maximum likelihood method’ (published by (Estève *et al.*, [1990]²⁹⁷) to fit the relative survival model. However, Estève *et al.*, describe this as “*a maximum likelihood method for computing net survival when the causes of death of cohort members are not known or when it seems advisable not to use them*” (Estève *et al.*, 1990, p.536).²⁹⁷ Therefore, given that net survival methods were not considered of relevant to this review, this method is not considered further.

In addition, the method proposed here refers to the hypothetical situation wherein estimation of subsequent diagnoses is of importance to model, which is seldom the case in contemporary

HTAs, and so it is unlikely that this method would be useful in the majority of cancer assessments even if it were suitable for estimating OS.

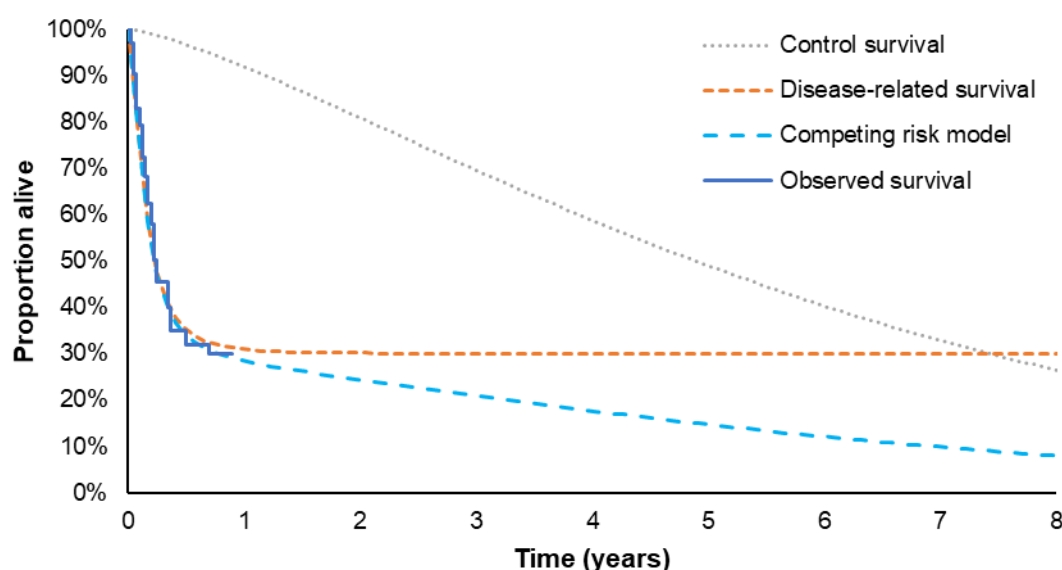
Maetani et al., (2004)

Maetani *et al.*, (2004) explored the use of parametric survival analyses to estimate the mean survival of patients with gastric cancer.¹⁵⁵ The authors present two different methods to estimate mean survival:

1. By taking the product of two survival curves: (i) a disease-related survival curve using a 3-parameter 'Boag' model (Boag, [1949]²⁹⁶), and (ii) a survival curve for a matched control group. The authors refer to this as the 'competing risks' model.
 - The Boag model²⁹⁶ was the first documented proposal to use an MCM framework to estimate a cure proportion associated with a given treatment at the same time as estimating the survival rate for patients not captured within the cure proportion.
 - As used in the original study by Boag, the application by Maetani *et al.* assumes a lognormal distribution can be used to reflect the survival of the 'uncured' proportion of patients.
 - The survival curve for the control group was constructed using estimates from Japanese life tables for the general population.
2. As an alternative approach to estimate mean survival, the authors evaluated the area under the disease-related survival curve up until a given time point T_{max} , and that this quantity was assumed to represent the mean survival for the cohort of interest.
 - Here, T_{max} was suggested to take the value of the mean survival experienced by the control group (which could be calculated by evaluating the area under the survival curve for the control group).
 - This approach could be described as a 'restricted mean survival time' analysis.

By definition, the first method is an application of relative survival, where the observed survival of the gastric cancer group, $S_O(t)$, is estimated using the Boag model, and the survival for the general population, $S_P(t)$, is estimated via Japanese life tables. However, the method is not described as being an application of relative survival by the authors. The second method is considered beyond the scope of this review, as it approximates OS by truncating an estimate of disease-specific survival. An illustration of the first approach taken by Maetani *et al.* (the competing risks model) is presented in Appendix Figure 4.

Appendix Figure 4: Illustration of Maetani *et al.* competing risks method



Key assumption(s)

The approach assumes that there exists a non-zero proportion of patients that will not die of their disease, and that their survival can be inferred from general population mortality statistics. Accordingly, patients that are assumed to be ‘cured’ do not experience any additional risk of death due to complications or comorbidities linked to their disease, nor is there any risk of relapse or recurrence.

Strengths and limitations

As with any MCM approach, the approach presented by Maetani *et al.* presents an intuitive means of distinguishing between long-term survivors and those that will die of their disease within a relatively short time period. However, a key limitation is that the estimation of the ‘Boag’ model likely assumes that the majority of patients still alive at the end of follow-up are effectively ‘cured’, though it is unclear if the assumption of cure truly holds in the long term.

Rationale for exclusion

This approach represents a relatively simple application of a cure-based model in a relative survival framework. The key elements of this method (i.e., the use of a cure model with a relative survival framework) are considered further, but the application of this specific method (i.e., a ‘Boag’ model to estimate disease-related survival) is not, as it is a specific type of relative survival MCM which is captured as part of a broader class of MCMs fitted in a relative survival framework.

Stare et al., (2005)

The majority of studies in relative survival focus on the estimation of the relative survival curve, $S_R(t)$, which can be interpreted as the proportion of patients within a given cohort that survive up until a given point in time out of the proportion of patients that would be expected to survive up until this time in the general population. For example, if $S_P(1) = 95\%$ and $S_O(1) = 50\%$, then the relative survival $S_R(t)$ can be estimated at 52.6% (i.e., 50% divided by 95%).

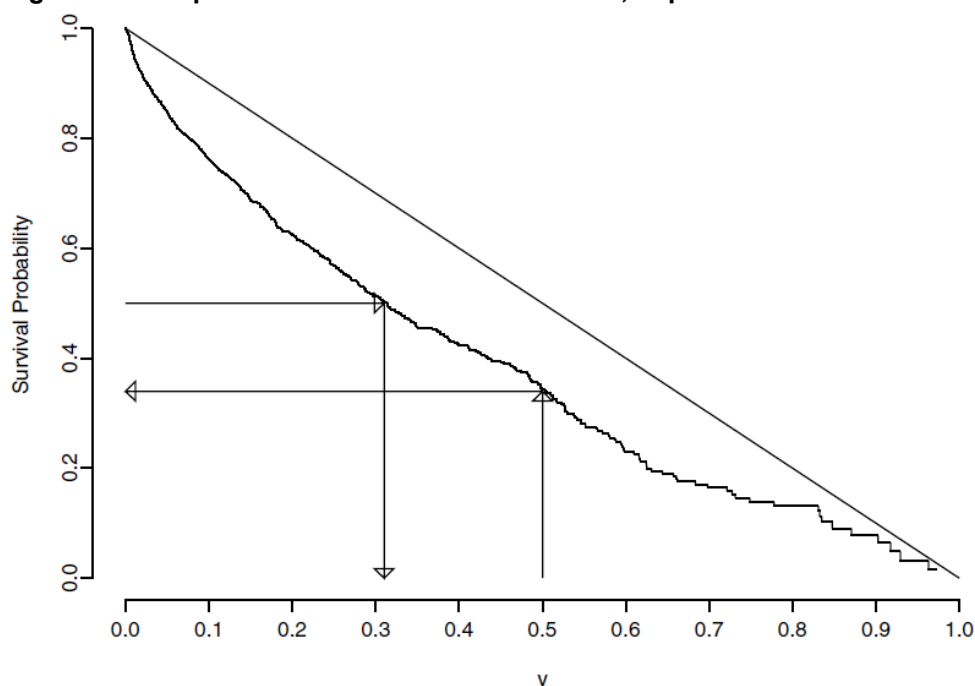
Stare *et al.* propose an alternative approach of considering relative survival by focusing on the individual rather than the cohort.¹⁵⁷ The new outcome measure is based around the concept of applying a transformation to the conditional cumulative distribution function for residual lifetime by age and sex. Using this measure, the authors explain that each individual would be assigned an outcome measure “*relative to the appropriate background population*” (Stare *et al.*, 2005, p.115).¹⁵⁷

The process to obtain the new outcome measure is described below:

- First, define a cumulative distribution function (based on age, sex, and calendar year) as F_{ASY} .
- The observed survival time of a given patient can be defined as t .
- Next, the transformation $y = F_{ASY}(t)$ can be considered, which combines the cumulative distribution function for the general population and relates (or converts) this to the observed survival time t .

This relationship can then be used to make inferences concerning the relative survival of a given patient versus the general population. Appendix Figure 5 shows an example plot provided by Stare *et al.* to show how this measure of individual relative survival could be used to make inferences about the population of interest.

Appendix Figure 5: Example of the individual-level measure, re-produced from Stare *et al.*, (2005)



Note(s): This figure illustrates how the outcome measure proposed by the authors, y , relates to the observed survival probability, presented in Stare *et al.*, (2005).¹⁵⁷ This is taken from Figure 3 from the original study publication. The straight line from (0,1) to (1,0) illustrates the hypothetical situation wherein the observed survival matches expected survival exactly, and the Kaplan-Meier estimate shows the relationship in the example cohort. In this example, the Kaplan-Meier estimate falls beneath the diagonal line, illustrating that patients in this population do not live for as long as their general population counterparts.

Source: Excerpted and reprinted with permission from the Journal of the Royal Statistical Society: Series C (Applied Statistics) (Copyright © 2004, John Wiley and Sons). Stare J, Henderson R, Pohar M. An individual measure of relative survival. *Appl. Statist.* 2005 Jan 54(1):115-126.

Two interpretations are shown in Appendix Figure 5. By reading up from the x-axis (labelled 'y') at 0.5, we can read off a value of 34% on the y-axis (labelled 'Survival Probability'). This means that 34% of patients lived longer than their expected median survival times in the general population. By reading right from the y-axis at 0.5, we can read off a value of 0.31. This means that the observed median corresponds to the 31st percentile of survival in the background population.

Rationale for exclusion

The outcome measure proposed by Stare *et al.* gives an alternative means of exploring relative survival, with a focus placed at the individual level. Therefore, this particular method does not meet the criteria to be defined as a novel method to extrapolate estimates of OS and is therefore not considered further.

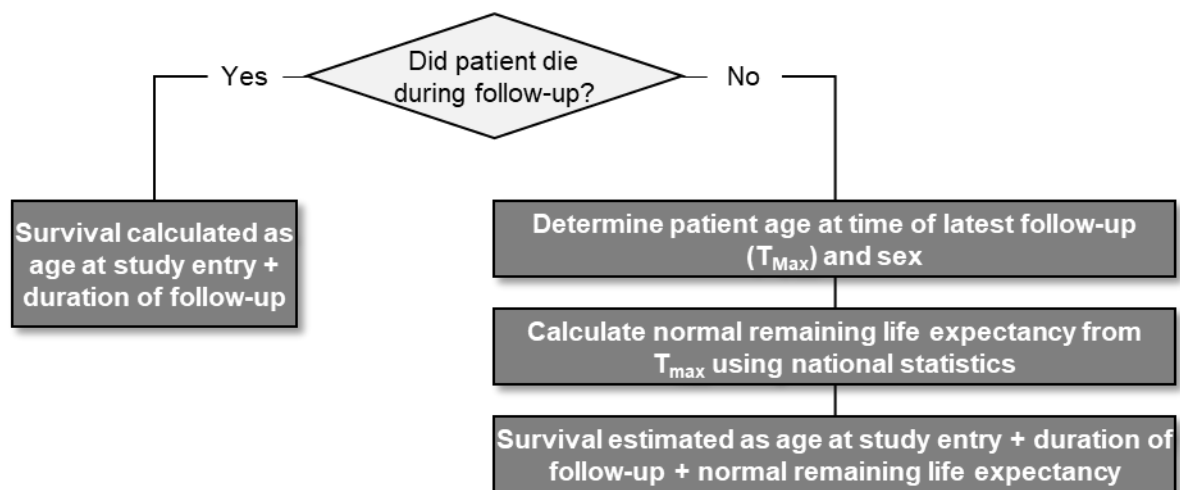
Viscomi *et al.*, (2006)

The researchers in this study aimed to calculate LE from diagnosis for patients with childhood leukaemia.¹⁴⁵ To do this, the authors calculated LE from birth as the sum of three components:

- **A:** Mean age at diagnosis – this covers the time period from birth until entry to the study (which in this example corresponds to the time of diagnosis).
- **B:** Mean LE during follow-up – this covers the time period while patients were enrolled within the study.
- **C:** Mean LE after follow-up – this covers the remaining LE for patients that were still alive at the end of follow-up.

The method was demonstrated using data from the Childhood Cancer Registry of Piedmont (CCRP) database (n=305 patients).²⁹⁸ Component A was reported as a baseline characteristic of each patient enrolled within the CCRP database (and so did not require any specific calculation), component B was estimated using three models (exponential, Gompertz, and Weibull) fitted to CCRP data, whereas component C was estimated using general population data, assuming no excess mortality risk attributable to disease. The process taken by the authors is summarised in Appendix Figure 6, which is similar to the approach taken by Messori & Trippoli, (1999).¹⁴⁴

Appendix Figure 6: Process to determine survival in method of Viscomi *et al.*, (2006)



The authors of this study acknowledge that their approach is similar to the approach presented by Messori & Trippoli, (1999)¹⁴⁴, and highlight that the approach taken can address the need to estimate lifetime survival, as opposed to survival within a restricted follow-up period. However, the method proposed by Viscomi *et al.*, (2006)¹⁴⁵ differs in two critical ways; first, that overall LE is estimated from birth by taking into consideration survival until diagnosis; and second, that mean LE was estimated instead of an OS curve.

Key assumption(s)

In this method, the authors assumed that any patient that is censored at the end of follow-up is effectively cured and would have a remaining LE equivalent to that of the age- and sex-adjusted general population (estimated from the date at which the patient was censored in the study). This assumption means that the time after which patients are assumed to be ‘cured’ is a direct consequence of the duration of follow-up for the study. In the example considered by Viscomi *et al.*, this meant that any patient that was still alive at 13.2 years after diagnosis was assumed to be cured, and their individual life expectancies were estimated based on their age at the end of follow up.

Strengths and limitations

As per the study by Messori & Trippoli, (1999)¹⁴⁴, the key strengths of this method are that it is relatively simple to understand and could be readily explained, with a strong (but clear) assumption of cure made after the end of follow-up. While the transparency of the method is a key strength, the primary limitation is the assumption of cure, which in the context of this study is especially important as any patient that is still alive at the end of follow-up is assumed to be cured. The authors explain that *“this approach can only be used for diseases and populations for which it is reasonable to assume that patient survival at the end of follow-up is similar to that in the general population”* (Viscomi *et al.*, 2006, p.170).¹⁴⁵ However, the authors do not explain how this assumption could be tested or justified.

Rationale for exclusion

The authors of this study were interested in the ability to estimate mean survival expectancy, rather than a survival curve. Therefore, the proposed method does not provide information concerning the proportion of patients expected to be alive at specific time points (as the method involves summing mean survival estimates over three time periods: before study, during study, and after study). As such, the explicit method described by Viscomi *et al.* is not considered further owing to its inability to provide information concerning the full distribution of expected survival for a cohort. However, elements of the method could be adapted to be considered as part of a piecewise analysis, similar to the approach of Messori & Trippoli (1999),¹⁴⁴ and other piecewise methods are considered further later in the thesis.

Pohar Perme *et al.*, (2009)

Pohar Perme *et al.* set out to improve the parameter estimation of an additive hazards model, and by also assuming a non-parametric baseline excess hazard function.¹⁵⁹ The additive hazards model was defined previously, but for ease of reading is shown again below:

$$\lambda_O(t) = \lambda_P(t) + \lambda_E(t)$$

In the above equation, $\lambda_o(t)$ refers to the observed (total) hazard of death (at time t), $\lambda_p(t)$ is the hazard in the general population, and $\lambda_E(t)$ is the excess hazard experienced by the cohort under consideration. When considering an additive regression model, with covariates Z , the same equation is often presented using the following:

$$\lambda_o(t, Z) = \lambda_p(t) + \lambda_0(t)e^{\beta Z}$$

Here, the excess hazard is assumed to be derived based on the baseline excess hazard, $\lambda_0(t)$, and a set of model coefficients, β , which relates to the covariates, Z . The authors explain that the baseline excess hazard function is usually assumed to be piecewise constant over specified intervals over the time period from baseline to the end of follow-up (i.e., a stepped function). However, the authors note that this is “*unrealistic and estimates can only be interpreted as averages over the specified intervals.*” (Pohar Perme et al., 2009, p.137).¹⁵⁹

To address the limitation, the authors advocate the use of an expectation-maximisation (EM) algorithm to model $\lambda_0(t)$, making no assumption about its form, and treating the cause of death as if it were missing data. This ultimately leads to an alternative way in which the additive hazards model can be fitted, which the authors purport to be a more flexible, and (under certain circumstances) better performing method of fitting an additive hazards model.

Rationale for exclusion

As the proposed method simplify offers an alternative means of estimating the parameters for an already-established OS modelling approach (i.e., an additive hazards model), this study is not considered further within the context of this review.

Andersson et al., (2011)

Andersson *et al.* consider a similar extension to the NMCM akin to the approach taken by Lambert *et al.*, by extending flexible parametric models to be fitted within a relative survival framework including a cure fraction or proportion.^{160, 161} In other words, the authors propose a spline-based model for modelling excess hazard, but beyond the final knot the log cumulative hazard function plateaus with zero slope indicating a cure proportion, and therefore no additional disease-related mortality. The cure proportion is estimated directly based on the placement of the last knot, and so the placement of this final knot is of critical importance when considering the plausibility of model estimates.

Rationale for exclusion

While this method represents a novel approach to estimating OS including reference to external evidence, a later study by Jakobsen *et al.*, (2019) (also co-authored by Andersson) compared this model to two other alternative flexible cure models, and highlighted that an assumption of cure immediately following the placement of the last knot was “*strong*”

(Jakobsen *et al.*, 2019, p.3).¹⁶⁵ While Jakobsen *et al.* did not identify any specific method which consistently outperformed the others, the method of Andersson *et al.* appeared to generally yield more biased estimates when compared to the other approaches. Therefore, the other two methods explored by Jakobsen *et al.* were assumed to supersede the approach proposed by Andersson *et al.*, and therefore this method is not considered further.

Ishak *et al.*, (2011)

A study by Ishak *et al.*, (2011) described an approach to making use of external data to address the issue of treatment crossover for patients in the MM-009/010 trials of lenalidomide + dexamethasone versus dexamethasone alone for MM ([NCT00056160](#) and [NCT00424047](#), respectively).¹⁷⁰ In addition to the data from the MM-009/010 trials, the authors highlight the availability of data from trials conducted by the Medical Research Council (MRC).²⁹⁹⁻³⁰³

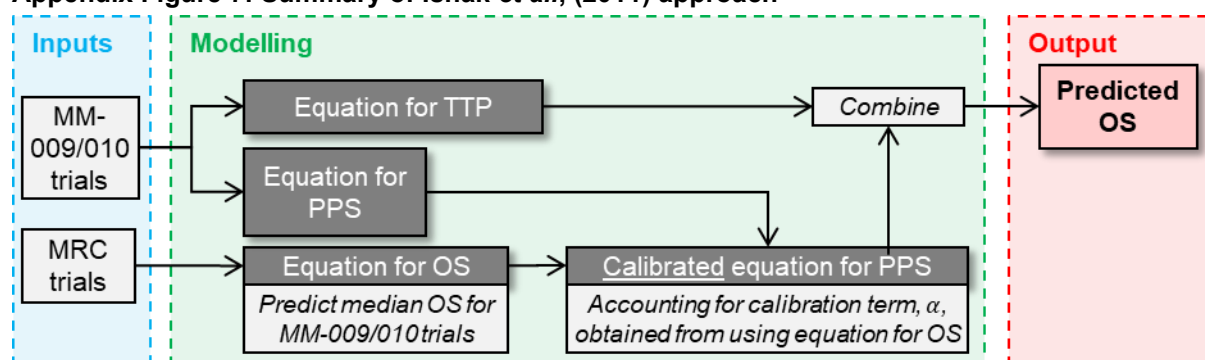
To adjust OS from the MM-009/010 trials for crossover, the authors considered OS to comprise of two parts: time to progression (TTP), and post-progression survival (PPS). Of these, PPS is expected to be predominantly affected by crossover owing to the timing of crossover in the MM-009/010 trials (i.e., at, or shortly after, progression for the majority of patients that did crossover). Therefore, the authors sought to re-calibrate the estimates of PPS in the MM-009/010 trials by making use of an equation derived for OS from the MRC trials.

To do this, the authors first derived an equation to estimate OS from the MRC trials, accounting for baseline characteristics; as well as equations for TTP and PPS from the MM-009/010 trials. Using the OS equation, median OS was predicted for the MM-009/010 studies by inputting the baseline characteristics of the study cohort. After obtaining a median OS estimate, the PPS equation was then adjusted using a calibration term, α , derived via the following formula:

$$\alpha = \frac{-\ln \left(\frac{\lambda^c(X)}{\lambda(X)} \right)}{\gamma}$$

Here, $\lambda(X)$ refers to the scale parameter of a Weibull model used to describe TTP pre-calibration, whereas $\lambda^c(X)$ denotes the corresponding scale parameter for the calibrated model. γ is a shape parameter, unaffected by baseline characteristics. The process to obtain α is described in detail by Ishak *et al.*, as it is not possible to estimate algebraically since $median\ TTP + median\ PPS \neq median\ OS$. The process followed is shown in Appendix Figure 7.

Appendix Figure 7: Summary of Ishak et al., (2011) approach



Key: MRC, Medical Research Council; OS, overall survival; PPS, post-progression survival; TTP, time to progression.

Key assumption(s)

The main assumption made in this approach is that there exists a calibration term, α , which can be used to adjust PPS from one source (in this case, the MM-009/010 trials) based on OS estimates from another source (here, the MRC trials). The calibration term, α , takes a constant value which does not vary according to different covariates, or over time.

Strengths and limitations

The main advantage afforded by this approach is that estimates of OS can be obtained that deal with the issue of crossover (as an unadjusted analysis of OS would lead to biased estimates of OS). The approach presented is simple in principle, as it 'borrows' information from another study to address a gap in evidence for the population of interest.

A limitation of the approach is that OS is not directly estimated; rather, OS is considered to comprise of sub-components, sub-divided by the time of disease progression (i.e., TTP and PPS). The estimation of OS as a combination of interim clinical events is beyond the scope of the review, as such approaches necessitate the determination of a clinically relevant event. In addition, it is also unclear how the method would perform in the presence of a non-negligible number of pre-progression deaths.

Rationale for exclusion

This method is excluded on the basis of it addressing a very specific issue (crossover), with specific data requirements for survival times linked to a measure of disease progression (which may not always be available for different therapeutic areas). It is therefore not a direct method of incorporating external evidence for OS within a method of estimating OS, and so is not considered further.

Hinchcliffe et al., (2012)

Hinchcliffe *et al.* consider the hypothetical situation wherein the disease of interest is responsible for “a high proportion of deaths due to a specific cancer in the external group” (Hinchcliffe *et al.*, 2012, p.148).¹⁶² In other words, the authors describe the possibility of a non-negligible proportion of deaths in the reference population (e.g., general population) being related to the disease of interest. The authors presented a range of analyses using data for patients with breast, colon, and prostate cancer; assuming that 2%, 5%, or 10% of deaths in the external population (i.e., expected survival) were due to the cancer under consideration. This proportion is termed α .

The authors propose an approach wherein three annual probabilities of death are used:

- q : the probability of dying from any cause in the external population.
- q_c : the probability of dying from the disease of interest in the external population.
- q_o : the probability of dying from other causes in the external population.

These probabilities may be expressed using the relationship:

$$q = q_c + q_o$$

In other words, the total probability of death is the sum of cause-specific and non-cause specific probabilities). Typically, q_c is assumed to be very small such that it is usually assumed to be zero. However, in this analysis the authors propose that if cause-specific deaths in the external population can be expressed as a proportion ($\alpha = q_c/q$), then the probability of dying from other causes in the external population (q_o) can be calculated as:

$$q_o = q(1 - \alpha)$$

The authors then use this simple adjustment to provide adjusted estimates of relative survival account for hypothetical α values of 2%, 5%, and 10%. The authors found that the largest discrepancies in results were noted for older populations in which a large proportion of deaths in the external population were due to the cause under consideration.

Key assumption(s)

The main assumption made within this analysis is that there exists a non-negligible proportion of deaths in the background mortality sources (for which relative comparisons are being made), such that adjustment is necessary to avoid double counting.

Strengths and limitations

This approach addresses a possible limitation of any relative survival exercises where causes of death are unknown, by incorporating an adjustment. However, a key limitation of the method is that it is extremely challenging to accurately estimate how many deaths in the general

population are related to a given disease. Such estimation would rely on reporting on death certificates to be entirely accurate, and it may be the case that multiple comorbidities were each linked to a given death.

Rationale for exclusion

The approach presented in this study is not in itself a novel means of incorporating external evidence into estimates of OS, as it represents a (relatively) minor adjustment to account for the possibility of a substantial proportion of deaths being linked to the cause of interest in the general population. The majority of HTAs of cancer treatments consider a relatively small population of patients (accounting for specific treatment lines and disease stage), and so adjusting for a large proportion of deaths observed in the general population is unlikely to be consequential for decision-making. Therefore, this method is excluded from further consideration.

Seppä et al., (2014)

Seppä *et al.* highlight the potential for there to be differences in relative survival between subgroups included as part of a whole cohort under consideration.¹⁶⁴ The authors consider regional variation in the relative survival of colon cancer patients in Finland, noting that findings of differences between regions would have important implications from a public health perspective. A method is then presented to incorporate random effects to explain regional differences within the estimation of a cure fraction model fitted within a relative survival framework. By undertaking such an approach, the authors explain that “*the potential heterogeneity in expected survival is taken into account and expected survival is considered as a random quantity*” (Seppä *et al.*, 2014, p.175).¹⁶⁴

Rationale for exclusion

While findings are likely important from a public health perspective, it is extremely unlikely that regional variation in relative survival would be possible to consider within the context of estimating OS for HTA (owing to the sample size of studies typically available to inform decision-making), and so this method is not considered further. However, other types of relative survival analysis that may be considered within the context of HTA are considered further.

Thornton Snider et al., (2017, 2018)

Two studies were identified by Thornton Snider *et al.* (published in 2017¹⁴⁹ and 2018¹⁵⁰) on the topic of the ‘option value’ of innovative treatments for patients with melanoma, non-small cell lung cancer, and renal cell carcinoma. Here, option value refers to the potential for patients to benefit from future advances in care if treatment with an innovative therapy enables them

to survive long enough. In these studies, the authors aimed to quantify such benefits, which required estimating differences in survival following the introduction of innovative treatments into the market.

Firstly, the authors considered a base curve which represented outcomes for patients treated with established care, prior to the availability of the innovative treatment(s) considered to have option value. The base curve was then adjusted to reflect various expected changes in the management of patients based on later availability of innovative treatments. In both studies, a base curve was estimated using data from the Surveillance, Epidemiology and End Results (SEER) cancer registry, up until the end of follow-up (approximately 20 years), after which general population mortality rates were appended to extrapolate the remainder of the survival curve (taken from the Human Mortality Database [HMD]).^{86, 304}

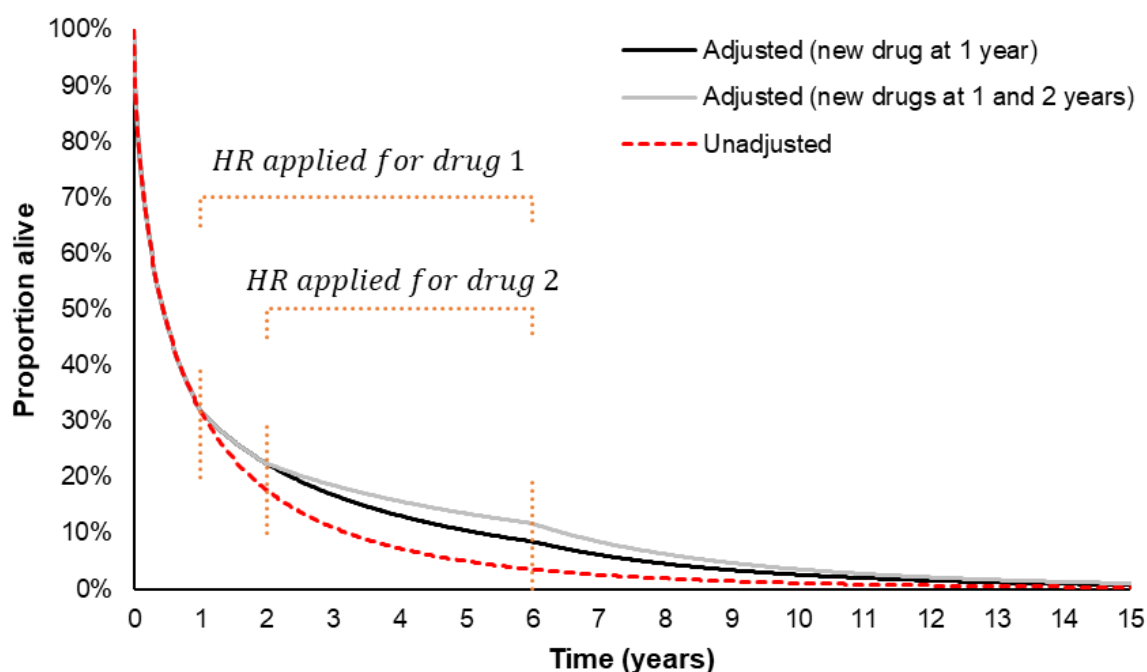
After obtaining the base curve, adjustments were made by specifying HRs to apply between specific time points reflecting the introduction of novel treatments, and the duration of time over which they are expected to impact the hazard of death. In the 2017 study¹⁴⁹, the innovative therapy was assumed to be administered 1 year after diagnosis, and all benefits from treatment were assumed to no longer apply 4 years after initiation of the therapy. In the 2018 study¹⁵⁰, two time points were considered reflecting the potential introduction of several novel treatment options:

- The first therapy was assumed to be administered at 1 year after diagnosis, and so the base survival curve was unadjusted prior to this time point.
- After 1 year, the base curve was adjusted using an HR.
 - This adjustment was applied up until 5 years, after which any survival benefit was assumed to dissipate.
- However, if patients were still alive at 2 years after diagnosis, another option was considered to switch to a different innovative therapy, and therefore a different HR was applied.
 - Like the other innovative treatment option, the beneficial effects were assumed to dissipate at 6 years after diagnosis.

An illustration of this modelling approach is presented in Appendix Figure 8. In this plot, the dashed red line shows the base curve, which is assumed to be a Weibull model with shape $k=0.6$ and scale $\lambda=0.8$. The adjusted curve accounting for a new treatment is shown with the solid black line, where a treatment given at 1 year is associated with an HR of 0.6 which applies for a duration of 5 years, after which projections revert to the unadjusted curve. An addition adjusted curve accounting for the first new treatment allowing patients to then receive a second new treatment is shown via a solid grey line, where the first treatment is given at 1

year with an HR of 0.6, and the second treatment is given at 2 years with an HR of 0.4 (versus the base curve). As with the first adjusted curve, all beneficial effects were assumed to cease by 5 years after receiving the first treatment.

Appendix Figure 8: Illustration of Thornton Snider *et al.* piecewise modelling approach



Key: HR, hazard ratio; t, time.

The approach of Thornton Snider *et al.* makes use of elements of the other piecewise approaches highlighted previously, but instead of using conditional survival probabilities derived from a different data source (e.g., an observational dataset), HRs are used to simply adjust aspects of a base curve.

Key assumption(s)

This approach makes a number of assumptions in order to obtain a survival curve for patients that receive up to two 'new' treatments. A base curve is established such that the impact of the new treatment(s) on survival can be expressed via the specification of an HR, and so hazards are assumed to be proportional. All patients are assumed to start a new treatment at a given (fixed) time point, which is linked to the availability of the treatment.

Strengths and limitations

This method provides a simple representation of how the introduction of new treatment(s) may affect survival, accounting for the availability of these treatments in the future. However, the method itself has a number of important limitations. There is no way of testing the PH assumption for which this method relies upon holding in order to adjust the base curve. The

impact of the second new treatment is assumed to be expressed via an HR versus the base curve, disregarding any residual benefits of the first new treatment. In addition, the method carries the risk of reflecting a sharp change in hazards between the piecewise components, as seen in the illustrative example presented in Appendix Figure 8 at $t=6$ years.

Rationale for exclusion

Application of this method requires several assumptions to be made, which are linked both to the expected difference in treatment effects for given therapies but are also the expected uptake of new treatments. The method has limited applicability outside of the context considered by the authors, and the latter assumptions related to market uptake are beyond the scope of extrapolation exercises typically required for contemporary HTA.

Rubio et al., (2019)

Rubio *et al.* explain that in typical relative survival analyses, expected mortality rates from general population life tables are usually at least based on age, sex, and region; however, other characteristics (such as deprivation and ethnicity) may not be available for some regions.¹⁶⁶ To address the limitation in relative survival analyses, the authors propose two different corrections (either a single parameter or a random effect [frailty]) to account for potential misspecification of background mortality rates.

Rationale for exclusion

The approach presented within this study is not a novel means of incorporating external evidence into estimates of OS and so is excluded from further consideration. Rather, it is a means of adjusting background mortality rates to account for differences in the population under consideration and available life table information to inform the estimation of background mortality.

Second iteration

Jackson et al., (2010)

Jackson *et al.* demonstrate how Bayesian semiparametric models can be used to estimate survival, where the baseline hazard function can be made arbitrarily complex while still enabling survival to be extrapolated.¹⁹⁶ The method allows for specification of prior distributions for the baseline hazard, but first the model fitting process is described in general.

The Bayesian semi-parametric model is implemented using a 'counting process' formulation, attributed to a study by Clayton, (1991).³⁰⁵ For a given individual i , let $N_i(t)$ be the number of deaths recorded for this individual up until time t . Then, define the counting process $dN_i(t)$, and consider the value it takes over a given time interval $[t, t + dt]$ – if the individual i dies

within this period, the increment of the counting process takes a value of 1, else it takes a value of 0. Then, the hazard function of the process, $\lambda_i(t)$, can be defined as:

$$\lambda_i(t) = \lim_{dt \rightarrow 0} E(dN_i(t)|\mathcal{F}_t)$$

Here, $\mathcal{F}_t$ is defined as the history of the process up to time t . Therefore, the hazard function is defined as the instantaneous expected value of the counting process, acknowledging the history of the counting process up until time t .

Next, it is necessary to determine the likelihood of the data, and to do this we must first define each of the recorded survival times – let t_j be defined as a unique death or censoring time for each individual $j \in \{1, 2, \dots, n\}$, and let $t_0 = 0$. Then, the likelihood can be computed by multiplying probabilities of survival or death over intervals up until a given individual's observed survival time (whether that is a death or censor):

$$\prod_i \prod_{j: t_j \leq \min(T_i, T_i^*)} \lambda_i(t_j)^{dN_i(t_j)} \exp(-\lambda_i(t_j)(t_j - t_{j-1}))$$

Here, T_i is defined as the 'true' survival time, and T_i^* is the censoring time, and so t_j takes the minimum of the two values. The authors highlight that this function is equivalent to a Poisson model for $dN_i(t_j)$ with mean $\lambda_i(t_j)(t_j - t_{j-1})$, and also note that the hazard, $\lambda_i(t)$, is assumed to be piecewise-constant (i.e., that is only changes at observed death times).

Then, for efficiency of computation, the $dN_i(t_j)$ values are aggregated for each individual i , such that patients with the same covariates at baseline (in the example presented by the authors, this is age, sex, and stage at diagnosis) are assigned the same identical hazard function $\lambda_i(t)$. The data are then analysed further in the format of dN_{jk} , which is defined as the number of deaths at each t_j (unique death or censoring times), in each of the covariate categories k . In the example presented by the authors, there are $5 \times 2 \times 4$ covariate categories, based on there being five categories for age (40-49, 50-59, 60-69, 70-79, or 80+ years), two for sex (male or female), and four for stage of disease (I-IV, inclusive). dN_{jk} can then be modelled based on the following:

$$dN_{jk} \sim \text{Poisson}(\mu_{jk}), \quad \mu_{jk} = n_{jk} \lambda_{0j} \exp(\beta_k)(t_j - t_{j-1})$$

Here, n_{jk} is defined as the number of individuals in covariate category k at risk at time t_{j-1} , λ_{0j} is the baseline hazard, and $\exp(\beta_k)$ is the HR for covariate category k (which is equal to the product of the relevant HRs for each age, sex, and disease stage HR, without any interactions included). This method represents what the authors describe as a model with 'hazards per death', but alternatives models are considered with 'stratified hazards' which change every 10 deaths within a given strata (see Table 1 in the study publication for details).

Using this model, priors can then be specified for different elements of the model, including the baseline hazard (which is the most relevant aspect for the purpose of this review). The authors explain that the prior for the baseline hazard, λ_{0j} , is varied across a range of alternative assumptions. Jackson *et al.* consider four different priors for λ_{0j} :

1. $\log(\lambda_{0j}) \sim N(0, 10000)$.
2. $\log(\lambda_{0j}) \sim N(-8, 10)$, which corresponds to a 5-year survival probability having a prior mean of 0.5 and a 95% credible interval (CrI) of (0.01, 0.99).
3. $\lambda_{0j} \sim \text{Gamma}(\mu c, c)$, with a prior mean $\mu = 0.00038$ (corresponding to a 5-year survival probability of 0.5), and low precision $c = 0.001$.
4. $\lambda_{0j} \sim \text{Gamma}(\mu c, c)$, with a prior mean $\mu = 0.00038$ and low precision $c = 100$, giving a 95% CrI of (0.000354, 1) for the 5-year survival probability.

The authors also explained that successive hazards are perhaps unlikely to be independent, and that in reality it may be that hazards closer together in time are more highly correlated versus those further apart in time. Accordingly, the authors estimated an additional four models in which λ_{0j} was constrained to be 'piecewise-constant' in time, changing based on several different events (e.g., every month in which a death was observed).

The Bayesian semi-parametric model yields posterior distributions of survival probabilities within the period of the observed data, accounting for the different covariate groupings, but does not provide extrapolated survival probabilities beyond the end of the follow-up period available. The authors therefore impose the assumption that extrapolations can be based on the estimated hazard at the final observed death time remaining constant in the future, allowing for incorporation within a cost-effectiveness model adopting a lifetime horizon.

Specific to the example highlighted by the authors, the outputs of the Bayesian semi-parametric analysis are ultimately used to inform a Markov model. In this model, *"the effect of increasing age as the same patient grows older is assumed to be the same as the effect of age at diagnosis between different patients"* (Jackson *et al.*, 2010, p.5).¹⁹⁶ This means that in this example, an all-encompassing survival curve is not obtained from the Bayesian semi-parametric analysis, but a series of probabilities are obtained which are later implemented within a separate economic model.

Key assumption(s)

This approach makes two key assumptions. First, the impact of any covariates on the estimated hazard of death are assumed to apply across the full extrapolation period, which may or may not be appropriate under specific circumstances. For example, if an impact of age is estimated based on the observed period, this impact is applied indefinitely for patients that

later 'age into' a particular covariate group. Second, the method assumes a constant hazard of death after the end of follow-up, based on the final calculated hazard. The authors themselves highlight that *"a constant hazard is unrealistic for extrapolating mortality risk to older ages, therefore additional data or expert judgements are required in practice to inform survival in long-term economic models"* (Jackson *et al.*, 2010, p.11).¹⁹⁶

Strengths and limitations

The authors argue that the Bayesian semi-parametric model improves upon an equivalent approach taking a 'standard' Cox model approach because expressing uncertainty is considered more straightforward under a Bayesian model. However, the method suffers from a number of limitations. First and foremost, the assumption of constant hazards in the long-term may be inappropriate under certain circumstances. In addition, when fitting the model, some smoothing of the hazard function may be required to obtain adequate predictive ability and avoid sensitivity to the choice of prior (which may seem counter intuitive as the specification of a seemingly non-informative prior could be influential on the end results).

Rationale for exclusion

While an aspect of the proposed method considers the use of external evidence in estimating survival (i.e., through specifying a prior on the proportion of patients alive at a given time point), the approach itself does not yield a final model for OS extrapolated over a lifetime horizon. Instead, the method gives a means of estimating probabilities for inclusion within a state-transition modelling framework, which falls outside the remit of this review. Furthermore, the emphasis of the use of external evidence in this example is placed on the period of follow-up for which data are already available, which is not the primary goal of introducing external evidence for HTA purposes. Therefore, this method is excluded from further consideration.

Murray et al., (2014)

Murray *et al.* developed a *"semiparametric Bayesian Cox-type survival model using a piecewise exponential baseline hazard with a correlated prior process, time dependent covariate adjustment, and commensurate priors to jointly model current and historical data"* (Murray *et al.*, 2014, p.186).¹⁹⁷ The method builds on a range of previous approaches previously developed to combine non-exchangeable current and historical survival data (noting that in this context, non-exchangeability refers to the expectation that data from the current and historical data are drawn from different distributions).

The derivation of the approach is presented in detail within the original paper. However, given that the derivation of the model is complex and lengthy (covering nearly 3 pages in the article)

it is not re-presented here, but instead a lay description of the individual components of the overall approach is presented below:

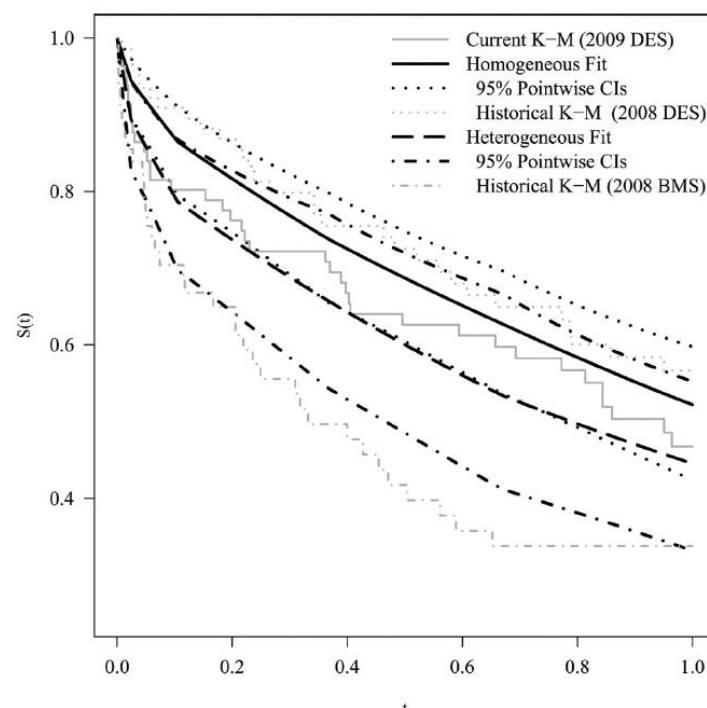
- **Semi-parametric:** The model makes some parametric assumptions (reflecting the baseline hazards, and the impact of covariates), but does not adopt a fully parametric framework (e.g., as would be the case in a Weibull model).
- **Bayesian Cox-type:** The model produces regression coefficient estimates that are similar to a Cox PH model fitted in a frequentist framework, but this model is fitted in a Bayesian framework instead.
- **Piecewise exponential:** Time is partitioned into a number of intervals, and within each interval the baseline hazard is assumed to remain constant.
- **Correlated prior process:** This process is used to 'smooth' baseline hazards through the introduction of a smoothing parameter, which is expected to 'resist overfitting' and 'increase flexibility'. The correlated prior process is cited from a book by Ibrahim *et al.*, (2001).³⁰⁶
- **Time dependent covariate adjustment:** The method could be fitted with or without the introduction of covariates which change over time and may have an impact on the estimated hazard of a given event occurring.
- **Commensurate priors:** By including commensurate priors, the method attempts to centre the parameters of the model in the current population likelihood on the parameters for their counterparts in the historical population (i.e., that the parameters in the current population are similar to those in the historical population).

To illustrate the method used, the authors consider a motivating example for patients on peritoneal dialysis. Two datasets were available to the authors:

- Data from the United States Renal Data System (USRDS) for patients that had a bare metal stent (N=62) or a drug-eluting stent (N=78) in 2008.
- Data from USRDS for patients that had a bare metal stent (N=57) or a drug-eluting stent (N=81) in 2009.

The authors were interested in comparing outcomes in patients in 2009 for bare metal versus drug-eluting stents, while 'borrowing strength' from patients that had a bare metal stent implanted in 2008. Two different model fits were produced, based on making use of data from each arm of the historical dataset – when using historical data from the drug-eluting stent cohort from 2008, the model fit was described as 'homogeneous' (same pair of interventions); whereas when using historical data from the drug-eluting stent cohort from 2008, the model fit was described as 'heterogenous' (different pairs of interventions). The posterior survival curves for both model fits are presented in Appendix Figure 9.

Appendix Figure 9: Posterior survival curve fits re-produced from Murray *et al.*, (2014)



Key: BMS, bare metal stent; CI, confidence interval; DES, drug-eluting stent; K-M, Kaplan-Meier; t , time; $S(t)$, survival function evaluated at time t .

Note(s): This figure illustrates the posterior survival curves for two different model fits ('homogenous' and 'heterogeneous'). On the figure, three different K-M estimates are presented: (1) 'Current K-M (2009 DES)' which refers to the current population and the group of direct interest for inference (i.e., the 2009 cohort that were implanted with a DES); (2) 'Historical K-M (2008 DES)' which refers to the historical population and those with a more similar set of characteristics to the current data, hence used to inform the 'Homogenous Fit'; and (3) 'Historical K-M (2008 BMS)' which refers to the historical population and those with a less similar set of characteristics to the current data, hence used to inform the 'Heterogeneous Fit'.

Source: Re-produced from Murray *et al.*, (2014): Murray TA, Hobbs BP, Lystig TC, Carlin BP. Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics*. 2014 Mar;70(1):185-91. Figure presented here with permission from John Wiley and Sons. Copyright for this figure is retained by John Wiley and Sons.

Key assumption(s)

The principal assumption made in the analysis is that the current and historical data sources are assumed to be non-exchangeable. There are also several other assumptions made within the analysis as a direct consequence of the semi-parametric functional form of the model (e.g., that covariates are captured in a Cox-like manner, and baseline hazards are assumed to be appropriately captured via the specification of a piecewise exponential model).

Strengths and limitations

The main advantage afforded by this approach versus those that precede it is that the historical data can be considered non-exchangeable with the current data, enabling a possibly broader range of sources to be considered when attempting to 'borrow strength' from historical data. However, the method itself is computationally complex, requiring in-depth knowledge of Bayesian statistical programming, and fundamentally it is not designed for research problems

concerning primarily with survival extrapolation. Furthermore, in relation to the assumption of non-exchangeability of data sources, it is unclear how the method may perform if data sources could be considered exchangeable.

Rationale for exclusion

This particular method of borrowing strength from historical data does not itself inform extrapolations of OS – rather, it focuses on “*making valid inference on safety outcomes*” by combining data from multiple sources (Murray *et al.*, 2014, p.191).¹⁹⁷ More specifically, in order to be useful for survival extrapolation, a novel approach to extrapolate survival would need to be implemented on top of the semi-parametric model used to combine the data sources in this method (which is not described by the authors). Therefore, while the method could (in principle) be extended to consider extrapolations of survival, the approach presented in this study is not relevant to this review and is thus not considered further.

Lewis *et al.*, (2019)

In this study, Lewis *et al.* develop a Bayesian hierarchical parametric survival model to estimate the survival outcomes for patients treated with a new intervention (‘Drug A’) versus SoC.¹⁹⁸ The over-arching motivation behind the example presented by the authors was to borrow information from historical control data to inform inferences concerning the relative effect of a given treatment (in this example, Drug A). However, the approach could (in principle) be extended to consider how borrowing from historical control data may influence the estimation of survival for the SoC arm.

The authors present an example in the context of three colorectal cancer studies:

- **CRC1:** Historical control cohort of N=350 patients that received SoC.
- **CRC2:** Historical control cohort of N=62 patients that received SoC.
- **CRC3:** RCT of N=125 patients that received either Drug A (N=63) or SoC (N=62).

The control arm of CRC3 is denoted the ‘concurrent control’, whereas the controls from CRC1 and CRC2 are denoted ‘historical control’.

Using Bayesian survival modelling methods, the authors consider two different possible comparisons to draw inferences concerning the relative effect of Drug A versus SoC borrowing information from the historical control data:

- A naïve 2-arm comparison of all SoC arms (i.e., data from CRC1, CRC2, and CRC3 for SoC) versus Drug A.
- A 3-arm comparison of historical control SoC (i.e., data from CRC1 and CRC2), concurrent control SoC (i.e., SoC in CRC3), and Drug A in CRC3.

The authors explain the approach in principle without considering extensions to involve additional covariates, and introduce the following notation:

- β = a coefficient vector which contains one element for each comparison group (i.e., two values for the 2-arm comparison, and three for the 3-arm comparison).
- z_i = covariate vector containing indicator variables (i.e., a value of 1 if applicable for a given subject i , and a value of 0 if not applicable).
- n = the total number of subjects in a given group.
- t_i = time-to-event for subject i . In this illustrative example, the authors assume that survival times follow a Weibull distribution with parameters:
 - r = shape parameter.
 - μ_i = scale parameter, which is assumed to take the value of $e^{z_i' \beta}$.
 - This means that $t_i \sim \text{Weibull}(r, \mu_i)$, $i = 1, \dots, n$.

This model is referred to the authors as the ‘base model’. In addition to the base model, the authors considered several extensions to consider additional covariates. This included extending the scale parameter of the Weibull model to include an additional continuous variable denoting the duration of oxaliplatin treatment, as well as the introduction of several other covariates. However, for simplicity, only the base model is considered further for the purpose of explaining how the approach works in general.

In the base model fitting process, all coefficients were assumed to have independent vague normal priors, though the shape parameter (r) was assumed to have an $\text{Exp}(1)$ prior (described by the authors to yield “*values between 0 and 6, a range more than wide enough to contain the values that are plausible for [their] data*” (Lewis *et al.*, 2019, p.69).¹⁹⁸

In addition to the base model fitting process, the 3-arm model was also re-fitted using a commensurate prior (attributed to a study by Hobbs *et al.*, 2011).³⁰⁷ When fitted using a commensurate prior, the coefficient vector for the concurrent control (β_{CC}) is assumed to be informed by the coefficient vector for the historical control (β_{HC}), or more formally:

$$\beta_{CC} | \beta_{HC} \sim N(\beta_{HC}, \tau)$$

In this notation, τ is a “*crucial precision parameter that controls the degree of borrowing*” (Lewis *et al.*, 2019, p.69).¹⁹⁸ The authors then explain that in their example, τ is either fixed at a large value (e.g., 1000) to encourage borrowing, or instead is assigned a moderately informative hyperprior[†] centred at the same value (e.g., $\text{Gamma}(1, 0.001)$).

[†]A ‘hyperprior’ is defined as a parameter of the prior distribution. In this case, the authors assign a distribution for τ which in effect ensures it is drawn from a distribution which is centred on a ‘large’ value.

The authors present results based on tabulated values of β (the coefficient vector for each relevant group) without presenting Kaplan-Meier estimates or modelled survival curves, and so interpretation is somewhat challenging. The values of β refer to the hazard rate in the modelled Weibull curves, and so differences in the values of β for group pairings can be exponentiated to obtain an estimated HR (e.g., $\exp(\beta_{CC} - \beta_{HC})$ would represent the estimated HR between the concurrent and historical control arms in the 3-arm model). However, the authors explain that “other choices” of the target of inference “could be contemplated” (Lewis *et al.*, 2019, p.70).¹⁹⁸

Key assumption(s)

A critical assumption made in this approach is that patient survival times follow the same distribution across each study. In the example presented by Lewis *et al.*, all survival times are assumed to be Weibull distributed (with a common shape parameter, r). In addition, when using the commensurate prior approach, the analyst is required to determine the strength of influence the historical control data has on the estimation of survival for the concurrent control (and, by extension, the new intervention if also considered as part of the analysis).

Strengths and limitations

While the determination of how influential historical control data may be on the overall target of inference, this is simultaneously a key strength of the approach – that the extent to which historical data influence model parameter estimates can be varied and tested as part of a sensitivity analysis. However, an important limitation of the approach is highlighted by the authors, noting that inappropriate borrowing of historical control information could lead to erroneous conclusions. Accordingly, the authors acknowledge that any borrowing of historical control data should be considered based on identifying and adjusting for important covariates, as well as specifying suitable priors.

Rationale for exclusion

The method presented in this study could be used to consider extrapolations of OS, but this is not the focus on the approach presented. Therefore, while the precise approach used by the authors of this study is not considered further as part of this review, Bayesian modelling is taken forward as a broader approach to integrate historical data within the estimation of a survival model from which extrapolations may be obtained.

Mba *et al.*, (2020)

Mba *et al.* propose a regression model for excess mortality to correct for potentially inaccurate background mortality rates.²⁰⁰ To do this, the authors extend the relative survival models proposed originally by Estève *et al.*, (1990)²⁹⁷ and Touraine *et al.*, (2020)³⁰⁸ to introduce “age-

*dependent multiplicative parameters through breakpoints, which gives some flexibility” (Mba et al., 2020, p.1).*²⁰⁰

To explain the proposed approach, the authors start by presenting the Estève *et al.* model, then how this was extended by Touraine *et al.*, before finally presenting their extension of the Touraine *et al.* model. Each of these models are discussed in turn below.

Estève model

In this model, it is assumed that the overall hazard of death λ_o at time t after diagnosis based on a vector of covariates z can be represented as:

$$\lambda_o(t|z) = \lambda_E(t|z) + \lambda_p^*(a + t|z_D)$$

Here, z_D is a vector of demographic variables included within the broader vector of covariates z (i.e., z_D is a subset of z). The overall hazard of death (λ_o) comprises the sum of two hazards: λ_E which represents the excess hazard, and λ_p^* which represents background mortality. In the Estève model, background mortality is assumed to be experienced by the study population as per the general population, regardless of any characteristics not available in the life tables.

Touraine model

In this approach, the background mortality experienced by the study population is allowed to differ from that of the general population. To allow for this to be factored into the model, Touraine *et al.* used a multiplicative parameter to correct a potentially inaccurate background mortality experienced by subjects belonging to a specific group. The formula for the Touraine model is presented below:

$$\lambda_o(t|z) = \lambda_E(t|z) + \sum_m \alpha_m I(x = m) \lambda_p^*(a + t|z_D),$$

for $m = 1, \dots, M$

This model introduces additional notation versus the Estève model. First, the model considers that there exists a categorical variable x with M levels (which is a covariate included in z [the vector of covariates available for the study population]), and that the available life table data are not reported for each x . Then, $I(x = m)$ represents the indicator function, which takes a value of 1 if the subject belongs to m (a category of the variable x). Finally, α_m is a vector of the aforementioned multipliers to correct for inaccurate background mortality. In essence, the Touraine model extends the Estève model by acknowledging that specific patients might experience a different background mortality to the general population based on a given covariate at baseline.

Mba model

Mba *et al.* note that the Touraine model assumes that the impact of the covariate on background mortality might not be constant over time. Therefore, Mba *et al.* extend the Touraine model further to consider age-dependent multiplicative parameters, made possible by specifying breakpoints (i.e., time points between which are time segments). Additional notation is therefore required, including a vector ε of a total of B breakpoints, containing values ($\varepsilon_1 < \varepsilon_2 < \dots < \varepsilon_B$). Each time segment between breakpoints is termed b , and so there are a total of $B + 1$ segments. The formula for the Mba model is presented below:

$$\lambda_O(t|z) = \lambda_E(t|z) + \sum_m \sum_b \alpha_{mb} I_b(a + t|x = m) \lambda_P^*(a + t|z_D),$$

for $m = 1, \dots, M$ and $b = 1, \dots, B + 1$

Key assumption(s)

The model makes two key assumptions. Firstly, the model assumes there is a categorical variable which can be used to describe differences in background mortality between the study and general populations (an assumption made by the Touraine model). However, the second assumption applies specifically to the Mba model – that this difference changes over time, and therefore an approach capturing these changes over time is needed, and can be appropriately reflected using breakpoints (analogous to a piecewise approach).

Strengths and limitations

A key strength of the approach proposed is that it extends the Touraine model to improve upon its flexibility accounting for changes in the impact of a given covariate on background mortality over time. By specifying this type of model, the authors avoid the need to identify life tables stratified by influential covariates. Two specific limitations are highlighted by the authors: first, that if several breakpoints are specified, there is a risk that the model could be over-parameterised; and second, that the model adopts a piecewise approach, and so it may be inappropriate to consider constant effects within a given interval (though this is nevertheless an improvement on no change at all).

Rationale for exclusion

The focus of this study was placed on the estimation of excess hazard ratios (EHRs), representing the difference in the risk of death related to the cause of interest between groups of patients (e.g., according to age group). The authors do not consider extrapolation of OS as a part of their analysis, and it is unclear how this approach could practically improve extrapolations for HTA beyond potentially increasing confidence in differences in survival within given subgroups of patients. More specifically, the authors state that their method

“provides an interesting alternative to answer epidemiological questions involving a specific variable affecting both the excess mortality and the background mortality in the absence of life table stratified by this variable and when no external evidence exists and/or is available to construct such a stratified life table” (Mba *et al.*, 2021, p.11)²⁰⁰ – in other words, a useful approach to be considered in epidemiological research under a number of specific circumstances. This method is therefore excluded on the basis of it not providing a means of incorporating external evidence to aid in the extrapolation of survival outcomes. In addition, the Touraine model is also not considered further given that it does not allow for an improvement in extrapolation of OS including external evidence specifically.

Additional searching methods

Briggs *et al.*, (2021)

Briggs *et al.* consider a Bayesian approach including the use of clinical trial data, general population mortality, and expert opinion.²¹² The proposed method considered three stages to produce long-term survival projections:

- **Stage 1 ('Data book')**: In the first stage, the authors reviewed published data on mortality for the population under consideration, including looking at clinical trial and observational study data. Then, they summarised all relevant information within a data book. This data book was then provided to experts as part of the next stage.
- **Stage 2 ('Expert elicitation')**: Next, the authors identified six experts in the disease area under consideration. The experts received training on the elicitation process, and possible risks of bias. Then, they were asked to complete an elicitation survey to obtain survival probability distributions at key time points (10 and 20 years in the example presented).
- **Stage 3 ('Bayesian modelling')**: Finally, using the short-term data from the clinical trial(s), the elicited survival estimates (specified as priors within a Bayesian framework), and information concerning general population mortality from published life tables, a Bayesian model was fitted to produce long-term survival projections.

However, no further technical details of the method were presented, and so this method was not considered further in the thesis. It should be noted, however, that this method bears resemblance to the MPES approaches of Guyot *et al.* and Jackson.^{153, 216} However, neither of these MPES approaches specifically included expert elicitation as part of the method itself.

Pham *et al.*, (2021)

Pham *et al.* consider two different approaches to incorporate historical trial data to inform survival extrapolation.²¹³ The first method involved specifying an informative prior for the shape

parameter of the survival curve for the immature data based on the historical trial data. This method has been previously identified based on a published study by Soikkeli *et al.*, (2019)¹⁹⁹, and is therefore not discussed further. The second method involved using the historical data as a ‘third arm’, though as no further details were presented, this method was also not considered further. The approach of using the data as a ‘third arm’ does however bear resemblance to a method previously identified by Hawe *et al.*, (2014).¹⁷¹

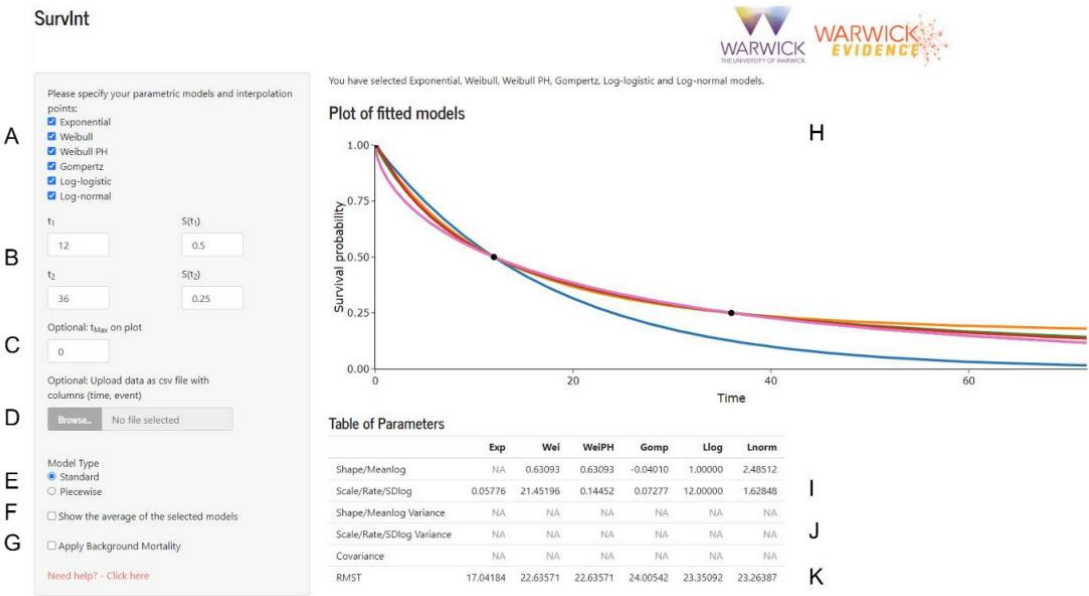
2024 update

Gallacher (2024)

Gallacher presents an R Shiny tool (called ‘SurvInt’).²¹⁷ SurvInt comprises a common parametric model based on the user selecting two time points (t_1 and t_2), two corresponding survival proportions ($S(t_1)$ and $S(t_2)$), and a given parametric model (e.g., Weibull). The tool will then back-calculate the model parameters needed to provide the best fit.

An illustration of the SurvInt R Shiny tool is presented in Appendix Figure 10. In this example, the user has inputted two survival time points at 12 and 36 months, with estimated survival probabilities of 50% and 25%, respectively. Based on these data points, the tool has generated various model fits that attempt to align with these estimates. Notably, the exponential model is not compatible with these data points, since there is a clearly a non-constant hazard of death implied by specifying these two data points (i.e., in very simplistic terms, the hazard of death up to 12 months must be greater than the hazard of death between 12 and 36 months).

Appendix Figure 10: Example of SurvInt



Key: PH, proportional hazard; RMST, restricted mean survival time; SD, standard deviation.
Note: The letters A, B, ..., K refers to annotations provided in the User Guide for SurvInt, details of which can be found in the electronic supplementary material with the original article by Gallacher (2024).
Source: Gallacher (2024). © The Author(s) 2024. Re-produced here in accordance with the Creative Commons Attribution 4.0 License (CC BY 4.0) (<https://www.creativecommons.org/licenses/by/4.0/>). Gallacher D. SurvInt: a

Key assumption(s)

This method assumes that there is a suitable parametric model which accurately captures the shape of the survival model, and that this can be achieved by identifying two time points to fit a model between.

Strengths and limitations

The main benefit of this approach over other methods is its relative simplicity, and in turn, its transparency. The end results will yield estimates of survival which align closely with the user-specified inputs. However, therein lies an important limitation of the approach – by specifying a model with estimates of survival which should be attained at specific time points, the model will inevitably ‘do as it is told’ (i.e., will output a model which reflects the exact data points used to inform it). The model fit is therefore a function of which data points are used as inputs.

Rationale for exclusion

Ultimately, this method does not involve fitting a model to patient-level data from a clinical trial, and then supplementing these data with external evidence. Instead, the method produces a model which aligns with user-specified inputs for times and proportions alive. This method is therefore not considered further in the thesis, as it sets aside the patient-level data from the clinical trial of relevance, in favour of user-specified timepoints which may or may not be within the period of trial follow-up. It is unlikely that such a method would be suitable for HTA purposes, but would be useful in other contexts (e.g., early models where data are not yet available).

Langford et al., (2022)

Langford *et al.* consider the use of data from one condition, idiopathic pulmonary fibrosis (IPF), to estimate survival for people with other conditions, grouped as progressive fibrosis interstitial lung disease (PF-ILD).²¹⁸

The authors used propensity score matching (PSM) methods to adjust data for patients in studies of IPF to resemble those of the target population (i.e., people with PF-ILD). The authors matched patients on variables that were collected for both populations; namely: age, sex, ethnicity, time since diagnosis, haemoglobin-corrected percentage of predicted diffusion capacity for carbon monoxide (DL_{CO}), percentage predicted forced vital capacity (FVC), and smoking status.

Following the PSM process, the authors then followed the approach of Soikkeli *et al.*, (2019)¹⁹⁹, which involved the specification of one or more informative priors on the shape parameter(s) of a given parametric survival model. This approach is described further in Chapter 3 (second iteration).

Key assumption(s)

In addition to the assumptions described for the original approach of Soikkeli *et al.*, (2019)¹⁹⁹, the approach of Langford *et al.* makes a further assumption that the PSM process adequately addresses any imbalances between the patient populations such that the shape parameter(s) estimated for the external data set (i.e., the population with IPF) can be generalised to the target population (i.e., the population with PF-ILD). This approach inherently assumes that there are no disease-specific variables that were omitted which would influence the generalisability of the shape parameter(s) across populations.

Strengths and limitations

The main strength of this approach is that it could be a helpful approach to consider in the context where data are available in a *similar* condition, but where data may not be available in the *same* condition. This could be especially relevant in the context of rare diseases. However, it may be noted that the key limitation is the assumption that the survival experiences can be generalised, which would likely remain an outstanding area of uncertainty in the resultant extrapolations.

Rationale for exclusion

The incorporation of external evidence in this method is identical to the approach of Soikkeli *et al.*, (2019)¹⁹⁹, and therefore this method is not considered further as it is already covered.

Sanchez Alvarez et al., (2023)

Sanchez Alvarez *et al.* present a Bayesian analysis aimed at estimating long-term survival for people with follicular lymphoma (FL) using multiple real-world evidence sources.²¹⁹ The authors used data from seven different real-world evidence sources plus data from an extension study of a clinical trial to obtain meta-analysed estimates of survival at two time points: 5 years and 8 years. These meta-analysed estimates of survival at 5 and 8 years were then used as informative priors when fitting of parametric survival models.

The authors explain how the point estimates of survival at 5 and 8 years were translated to prior information for the Bayesian model fitting process: “*The survival function for each statistical distribution (e.g., exponential, Weibull, etc.) were used at each time point by setting the survival function with the parameter for time and survival fixed at specific values. The equation is then solved for the scale or rate parameter (depending on the distribution) as a*

function of the specific survival value (e.g., 0.5), time point (e.g., 8 years), and remaining parameters if any (e.g., in the case of the exponential distribution there would be none as it only has 1 parameter)." (Sanchez Alvarez *et al.*, 2023, p.500).²¹⁹

This approach resembles the approach of Soikkeli *et al.*, (2019)¹⁹⁹, but differs in that the informative prior is placed on the scale or rate parameter of a given parametric model, rather than the shape parameter(s). The authors provide an explanation of how this could be achieved in the case of the exponential model in the supplementary materials, but no explanation is given for how the scale or rate parameter could be 'solved' in the presence of two (or more) unknowns (i.e., a scale/rate plus at least one shape parameter).

Key assumption(s)

The key assumptions underpinning this method are similar to the assumptions made by Soikkeli *et al.*, (2019)¹⁹⁹, except that the prior information is specified on the scale or rate parameter, rather than a shape parameter.

Strengths and limitations

An added benefit of this method over other methods that rely on informative prior specification is that the source of the prior information comes from multiple evidence sources, as opposed to a single source of evidence. However, this in turn introduces a limitation, in that the evidence sources need to be considered sufficiently similar.

Rationale for exclusion

The incorporation of external evidence in this method is conceptually similar to the approach of Soikkeli *et al.*, (2019)¹⁹⁹, though it is not clear from the authors reporting how all parameters could reliably be estimated in the case of distributions which require more than one parameter (i.e., anything other than an exponential model). Therefore, since (i) the method is methodologically similar to the method of Soikkeli *et al.*, and (ii) it is unclear how the method could be routinely applied given the lack of established methodology to solve for unknown parameters in multi-parameter distributions (which is not an issue for the method of Soikkeli *et al.*); it is not considered further.

Traore *et al.*, (2023)

Traore *et al.* present a Bayesian analysis aimed at addressing the context of a study where limited OS data were available for the key clinical trial, but historical data for both OS and progression-free survival (PFS) are available that may assist predictions of OS.²²⁰ The authors consider a case study in NSCLC where patient-level data from the RELAY trial ([NCT02411448](https://clinicaltrials.gov/ct2/show/study/NCT02411448)) were available, which provided mature estimates of PFS, but immature estimates of OS. The authors then carried out a literature review of Phase III trials in similar

populations, and extracted estimates of the HR for both OS and PFS endpoints. The authors then used a combination of these data sources to explore univariate and joint models to derive estimates of OS through a combination of the aforementioned data sources:

- The **univariate** model was based on a predicted HR and an acceleration factor for OS. A linear predictive model was developed using the findings from the literature review with the OS HR (on log scale) acting as a dependent variable and PFS HR (on log scale) as a predictor.
- The **joint** model was based on the correlation between PFS and OS in trials of similar populations. This was based on a 'Bayesian normal induced copula estimation model', and relied on the assumption that PFS could be considered similar to time to progression (TTP).

Further details for both methods are provided by Traore *et al.*²²⁰

Key assumption(s)

Both methods (i.e., univariate and joint models) rely on the assumption that a clear link between PFS and OS can be established, and that this link is generalisable across different clinical trials.

Strengths and limitations

The key strength of this method is that it is able to make use of a large number of evidence sources to inform a surrogacy relationship between the clinically-relevant endpoint of PFS and OS. The key limitation is the previously noted assumption concerning the surrogacy relationship between PFS and OS.

Rationale for exclusion

Both methods proposed by Traore *et al.* require data for the outcome of PFS, in addition to OS, and rely upon estimating treatment effects. The scope of this thesis covers methods that use only data for OS, and does not explicitly consider methods that are aimed at estimating treatment effects, and so both approaches proposed by Traore *et al.* are excluded from further consideration.

Appendix E. Data extraction for appraisals review

Appendix Table 7: STA review data extraction (Part 1 of 10), summary-level STA information

#	ID	Model type	Post CDF?	Title	Published	Recommendation
1	TA756	DES	No	Fedratinib for treating disease-related splenomegaly or symptoms in myelofibrosis	16-Dec-21	Recommended (CDF)
2	TA754	DT + PartSA	No	Mogamulizumab for previously treated mycosis fungoides and Sézary syndrome	15-Dec-21	Recommended
3	TA746	STM	No	Nivolumab for adjuvant treatment of resected oesophageal or gastro-oesophageal junction cancer	17-Nov-21	Recommended
4	TA742	PartSA	No	Selpercatinib for treating advanced thyroid cancer with RET alterations	03-Nov-21	Recommended (CDF)
5	TA740	PartSA	No	Apalutamide with androgen deprivation therapy for treating high-risk hormone-relapsed non-metastatic prostate cancer	28-Oct-21	Recommended
6	TA741	PartSA	No	Apalutamide with androgen deprivation therapy for treating hormone-sensitive metastatic prostate cancer	28-Oct-21	Optimised
7	TA739	PartSA	Yes	Atezolizumab for untreated PD-L1-positive advanced urothelial cancer when cisplatin is unsuitable	27-Oct-21	Recommended
8	TA736	PartSA	Yes	Nivolumab for treating recurrent or metastatic squamous cell carcinoma of the head and neck after platinum-based chemotherapy	20-Oct-21	Recommended
9	TA737	PartSA	No	Pembrolizumab with platinum- and fluoropyrimidine-based chemotherapy for untreated advanced oesophageal and gastro-oesophageal junction cancer	20-Oct-21	Recommended
10	TA728	PartSA	No	Midostaurin for treating advanced systemic mastocytosis	22-Sep-21	Recommended
11	TA725	PartSA	Yes	Abemaciclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy	15-Sep-21	Recommended
12	TA724	PartSA	No	Nivolumab with ipilimumab and chemotherapy for untreated metastatic non-small-cell lung cancer	08-Sep-21	Not recommended
13	TA722	PartSA	No	Pemigatinib for treating relapsed or refractory advanced cholangiocarcinoma with FGFR2 fusion or rearrangement	25-Aug-21	Recommended
14	TA720	STM	No	Chlormethine gel for treating mycosis fungoides-type cutaneous T-cell lymphoma	18-Aug-21	Optimised
15a	TA721	MSM (original)				
15b	TA721	PartSA (post-ACM2)	No	Abiraterone for treating newly diagnosed high-risk hormone-sensitive metastatic prostate cancer	18-Aug-21	Not recommended
15c	TA721	PartSA (post-appeal)				
16	TA716	PartSA	No	Nivolumab with ipilimumab for previously treated metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency	28-Jul-21	Recommended
17	TA712	PartSA	No	Enzalutamide for treating hormone-sensitive metastatic prostate cancer	07-Jul-21	Recommended
18	TA713	PartSA	Yes	Nivolumab for advanced non-squamous non-small-cell lung cancer after chemotherapy	07-Jul-21	Optimised
19a	TA709	PartSA	No	Pembrolizumab for untreated metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency	23-Jun-21	Optimised
19b	TA709	STM				
20	TA707	PartSA	No	Nivolumab for previously treated unresectable advanced or recurrent oesophageal cancer	15-Jun-21	Recommended
21	TA705	PartSA	No	Atezolizumab monotherapy for untreated advanced non-small-cell lung cancer	02-Jun-21	Recommended
22	TA704	PartSA	No	Trastuzumab deruxtecan for treating HER2-positive unresectable or metastatic breast cancer after 2 or more anti-HER2 therapies	26-May-21	Recommended (CDF)
23	TA692	PartSA	Yes	Pembrolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum-containing chemotherapy	28-Apr-21	Not recommended
24	TA693	PartSA	No	Olaparib plus bevacizumab for maintenance treatment of advanced ovarian, fallopian tube or primary peritoneal cancer	28-Apr-21	Optimised (CDF)
25	TA695	PartSA	No	Carfilzomib with dexamethasone and lenalidomide for previously treated multiple myeloma	28-Apr-21	Optimised
26a	TA689	STM (1L)	No	Acalabrutinib for treating chronic lymphocytic leukaemia	21-Apr-21	Optimised
26b	TA689	PartSA (2L+)				
27	TA691	PartSA	Yes	Avelumab for untreated metastatic Merkel cell carcinoma	21-Apr-21	Recommended
28a	TA687	STM (original)	Yes	Ribociclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy	31-Mar-21	Recommended
28b	TA687	PartSA (post-TE)				
29a	TA684	STM	Yes	Nivolumab for adjuvant treatment of completely resected melanoma with lymph node involvement or metastatic disease	17-Mar-21	Recommended
29b	TA684	PartSA				
30	TA683	PartSA	Yes	Pembrolizumab with pemetrexed and platinum chemotherapy for untreated, metastatic, non-squamous non-small-cell lung cancer	10-Mar-21	Recommended
31	TA680	PartSA	No	Lenalidomide maintenance treatment after an autologous stem cell transplant for newly diagnosed multiple myeloma	03-Mar-21	Optimised
32	TA677	PartSA	No	Autologous anti-CD19-transduced CD3+ cells for treating relapsed or refractory mantle cell lymphoma	24-Feb-21	Recommended (CDF)
33	TA673	PartSA	No	Niraparib for maintenance treatment of advanced ovarian, fallopian tube and peritoneal cancer after response to first-line platinum-based chemotherapy	17-Feb-21	Recommended (CDF)
34	TA669	PartSA	No	Trifluridine–tipiracil for treating metastatic gastric cancer or gastro-oesophageal junction adenocarcinoma after 2 or more therapies	27-Jan-21	Not Recommended
35	TA670	PartSA	No	Brigatinib for ALK-positive advanced non-small-cell lung cancer that has not been previously treated with an ALK inhibitor	27-Jan-21	Recommended
36	TA668	PartSA	No	Encorafenib plus cetuximab for previously treated BRAF V600E mutation-positive metastatic colorectal cancer	06-Jan-21	Recommended

Note: Appraisals highlighted in yellow above are the n=13 appraisals of ‘further interest’ described in Chapter 4.

Appendix Table 8: STA review data extraction (Part 2 of 10), scope of appraisal and parties involved

#	Population	Intervention	Comparator(s)	Company, EAG, and Committee		
1	Adults with primary myelofibrosis, post polycythaemia vera myelofibrosis or post essential thrombocythaemia myelofibrosis	Fedratinib	• No previous treatment with ruxolitinib - o Ruxolitinib (for people with intermediate-2 risk or high risk disease) - o Established clinical practice (including but not limited to hydroxycarbamide, other chemotherapies, androgens, splenectomy, radiation therapy, erythropoietin and red blood cell transfusion) • Previous treatment with ruxolitinib or if ruxolitinib is not appropriate - o Established clinical practice (including but not limited to hydroxycarbamide, other chemotherapies, androgens, splenectomy, radiation therapy, erythropoietin and red blood cell transfusion)	Celgene	Sheffield	C
2	Adults with mycosis fungoides or Sézary syndrome cutaneous T-cell lymphoma following at least one prior systemic therapy.	Mogamulizumab	Established clinical management without mogamulizumab	Kyowa Kirin	Kleijnen	D
3	Adults with resected oesophageal or gastro-oesophageal junction cancer	Nivolumab	Routine surveillance	Bristol-Myers Squibb	Sheffield	A
4	• People with advanced RET fusion-positive thyroid cancer who require systemic therapy and whose disease has progressed following prior treatment • People with advanced RET mutation-positive medullary thyroid cancer (MTC) who require systemic therapy	Selpercatinib	• For advanced RET fusion-positive thyroid cancer which has progressed following prior treatment: - o lenvatinib or sorafenib for differentiated thyroid cancer which did not respond to radioactive iodine (adults only) - o best supportive care or palliative care • For advanced RET mutation-positive MTC: - o cabozantinib (adults only) - o best supportive care or palliative care	Eli Lilly	Kleijnen	D
5 6	Adults with prostate cancer	Apalutamide with androgen deprivation therapy	Androgen deprivation therapy • For people with metastatic hormone-sensitive prostate cancer: - o Androgen deprivation therapy alone (including orchidectomy, luteinising hormone-releasing hormone agonist therapy or monotherapy with bicalutamide) - o Docetaxel with androgen deprivation therapy - o Abiraterone with prednisone or prednisolone and androgen deprivation therapy (subject to ongoing NICE appraisal) - o Enzalutamide with androgen deprivation therapy (subject to ongoing NICE appraisal)	Janssen	Southampton	B
7	Adults with untreated PD-L1 positive locally advanced or metastatic urothelial cancer, for whom cisplatin based chemotherapy is unsuitable and whose tumours have a PD-L1 expression ≥ 5	Atezolizumab	• Gemcitabine plus carboplatin • Best supportive care	Roche	Southampton	D
8	Adults with recurrent or metastatic squamous-cell carcinoma of the head and neck who have previously received platinum-based chemotherapy	Nivolumab	• Docetaxel • Paclitaxel • Methotrexate	Bristol-Myers Squibb	Kleijnen	D
9	Adults with untreated, unresectable locally advanced or metastatic oesophageal cancer or gastroesophageal junction adenocarcinoma	Pembrolizumab with platinum-based chemotherapy	• Platinum-based chemotherapy without pembrolizumab, such as: - o Doublet treatment with fluorouracil or capecitabine plus cisplatin or oxaliplatin - o Triplet treatment with fluorouracil or capecitabine plus cisplatin or oxaliplatin epirubicin	Merck Sharp & Dohme	Exeter	A
10	Adults with aggressive systemic mastocytosis, systemic mastocytosis with associated haematological neoplasm or mast cell leukaemia.	Midostaurin	Current clinical management including but not limited to interferon alpha, cladribine, imatinib, nilotinib, or dasatinib (these treatments do not currently have a marketing authorisation in the UK for this indication)	Novartis	Liverpool	C
11	People with advanced hormone-receptor positive HER2-negative breast cancer that has progressed after prior endocrine therapy	Abemaciclib in combination with fulvestrant	Everolimus and exemestane	Eli Lilly	BMJ	A
12	Adults with untreated metastatic NSCLC without sensitizing EGFR mutations or ALK fusions	Nivolumab with ipilimumab and standard chemotherapy	• For adults with non-squamous histology: - o Pemetrexed in combination with a platinum drug (carboplatin or cisplatin) (for people with adenocarcinoma or large-cell carcinoma only) - with (following cisplatin-containing regimens only) or without pemetrexed maintenance treatment - o Chemotherapy (docetaxel, gemcitabine, paclitaxel or vinorelbine) in combination with a platinum drug (carboplatin or cisplatin) - with or without pemetrexed maintenance treatment - o Atezolizumab with bevacizumab, carboplatin and paclitaxel (for people whose tumours express PDL1 with less than 50% tumour proportion score) - o Pembrolizumab (for people whose tumours express PD-L1 with at least a 50% tumour proportion score) • For adults with squamous histology: - o Chemotherapy (docetaxel, gemcitabine, paclitaxel or vinorelbine) in combination with a platinum drug (carboplatin or cisplatin) - o Pembrolizumab (for people whose tumours express PD-L1 with at least a 50% tumour proportion score)	Bristol-Myers Squibb	York	D
13	People with advanced cholangiocarcinoma with FGFR2 fusion or rearrangement that is relapsed or refractory after at least 1 prior systemic therapy	Pemigatinib	• Chemotherapy • Best supportive care (including stent insertion)	Incyte	Kleijnen	C
14	Adults with mycosis fungoides-type cutaneous T-cell lymphoma	Chlormethine gel	• Skin directed therapies such as photo therapy (PUVA, UVB) and total skin electron beam therapy • In patients for whom the above skin directed therapies are contraindicated: - o Established clinical management without chlormethine gel (including systemic therapies such as interferons and retinoids)	Recordati Rare diseases/ Helsinn Healthcare SA	Aberdeen	A
15a				Janssen	Aberdeen	B

15b	Adults with newly diagnosed high risk metastatic hormone-naïve prostate cancer	Abiraterone in combination with prednisone or prednisolone and androgen deprivation therapy	- Androgen deprivation therapy alone (including orchidectomy, luteinising hormone-releasing hormone agonist therapy or monotherapy with bicalutamide) - Docetaxel in combination with androgen deprivation therapy			
16	Adults with previously treated recurrent or metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency	Nivolumab with ipilimumab	- Single-agent irinotecan (after FOLFOX) - FOLFIRI (after either FOLFOX or CAPOX) - FOLFOX (after either FOLFIRI or CAPOX) - Raltitrexed (if 5-fluorouracil and folinic acid are not suitable) - Trifluridine–tipiracil - Best supportive care	Bristol-Myers Squibb	BMJ	B
17	People with metastatic hormone-sensitive prostate cancer	Enzalutamide in combination with androgen deprivation therapy	- Androgen deprivation therapy alone (including orchidectomy, luteinising hormone-releasing hormone agonist therapy) or monotherapy with bicalutamide - Docetaxel with androgen deprivation therapy - For people with newly diagnosed high-risk disease: - Abiraterone with prednisone or prednisolone and androgen deprivation therapy (subject to ongoing NICE appraisal)	Astellas	Aberdeen	B
18	People with previously treated locally advanced or metastatic non-squamous non-small cell lung cancer	Nivolumab	- Non-squamous EGFR-TK mutation positive tumours: - After one prior therapy: - Platinum therapy (in combination with gemcitabine, vinorelbine, pemetrexed or a taxane) - Single agent gemcitabine and vinorelbine (for people for whom platinum therapy is not appropriate) - Afatinib (for people who have not previously had an EGFR-TKI) or erlotinib (if no previous EGFR-TKI therapy received due to delayed confirmation of mutation status) - After two prior therapies (an EGFR-TKI and one other therapy): - Docetaxel monotherapy - Erlotinib (if not received previously) - Nintedanib in combination with docetaxel - Best supportive care - Non-squamous EGFR-TK mutation negative or unknown tumours: - After one prior therapy: - Docetaxel monotherapy - Erlotinib (only for unknown EGFR-TK mutation status) - Nintedanib in combination with docetaxel - Crizotinib (only for patients with ALK positive mutation status) - Ceritinib (only for patients with ALK positive mutation status) - Best supportive care - After two prior therapies: - Docetaxel monotherapy - Erlotinib (if not received previously) - Best supportive care	Bristol-Myers Squibb	Liverpool	C
19a	Adults with metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency	Pembrolizumab	- For all patients - Folinic acid plus fluorouracil plus oxaliplatin (FOLFOX) - Folinic acid plus fluorouracil plus irinotecan (FOLFIRI) - Capecitabine plus oxaliplatin (CAPOX) - Capecitabine - Tegafur with uracil (in combination with folinic acid) - Raltitrexed (only when folinic acid and fluorouracil are not tolerated or unsuitable) - For patients with RAS wild-type metastatic colorectal cancer - Panitumumab in combination with FOLFOX or folinic acid plus fluorouracil plus irinotecan (FOLFIRI) - For patients with EGFR expressing, RAS wild-type metastatic colorectal cancer - Cetuximab in combination with FOLFOX or FOLFIRI	Merck Sharp & Dohme	BMJ	B
20	Adults with previously treated advanced or recurrent unresectable oesophageal cancer that is refractory or intolerant to standard therapy	Nivolumab	- Chemotherapy including taxanes (docetaxel/paclitaxel) or irinotecan - Best supportive care (including but not limited to antiemetics, blood transfusions, oesophageal stents)	Bristol-Myers Squibb	Exeter	A
21	Adults with non-squamous or squamous untreated metastatic NSCLC with PD-L1 positive tumour expression and without EGFR- or ALK-positive mutations	Atezolizumab	- For people whose tumours express PD-L1 with at least a 50% tumour proportion score: - Pembrolizumab - For people with non-squamous NSCLC whose tumours express PD-L1 with a tumour proportion score below 50%: - Atezolizumab plus bevacizumab, carboplatin and paclitaxel - Chemotherapy (docetaxel, gemcitabine, paclitaxel or vinorelbine) in combination with a platinum drug (carboplatin or cisplatin) - with or without pemetrexed maintenance	Roche	Aberdeen	D
22	People with HER2-positive, unresectable or metastatic breast cancer who have received 2 or more prior anti-HER2 therapies	Trastuzumab deruxtecan	- Capecitabine - Vinorelbine - Eribulin (for people who have had 2 or more chemotherapy regimens)	Daiichi-Sankyo	Liverpool	A
23	Adults with locally advanced and unresectable or metastatic urothelial cancer that has progressed on or after platinum-containing chemotherapy	Pembrolizumab	- Retreatment with 1st line platinum-based chemotherapy (only for people whose disease has had an adequate response) - Docetaxel - Paclitaxel - Best supportive care	Merck Sharp & Dohme	Warwick	D
24	People with newly diagnosed advanced ovarian, fallopian tube, or primary peritoneal cancer	Platinum-based chemotherapy with bevacizumab (15 mg/kg every 3 weeks) followed with olaparib and bevacizumab maintenance therapy only in responders	- Platinum based chemotherapy followed with routine surveillance - In addition, for people who would receive bevacizumab through the Cancer Drugs Fund: platinum-based chemotherapy with bevacizumab (7.5 mg/kg every 3 weeks) followed with bevacizumab maintenance therapy	AstraZeneca	BMJ	A
25	Adults with multiple myeloma who have had at least 1 previous therapy	Carfilzomib plus lenalidomide and dexamethasone	- For people who have had 1 previous therapy: - Carfilzomib plus dexamethasone - Lenalidomide plus dexamethasone	Amgen	BMJ	C

			- o Bortezomib • For people who have had 2 previous therapies: - o Lenalidomide plus dexamethasone - o Panobinostat plus bortezomib and dexamethasone • For people who have had 3 or more previous therapies: - o Lenalidomide plus dexamethasone - o Panobinostat plus bortezomib and dexamethasone - o Pomalidomide plus dexamethasone			
26a 26b	People with chronic lymphocytic leukaemia (includes untreated and untreated)	Acalabrutinib alone or with obinutuzumab	• For untreated CLL, including (but not limited to): - o Ibrutinib (17p deletion or TP53 mutation) - o Idelalisib with rituximab (17p deletion or TP53 mutation) - o Chlorambucil with or without rituximab - o Obinutuzumab with chlorambucil - o Bendamustine with or without rituximab - o Rituximab with fludarabine and cyclophosphamide - o Venetoclax with obinutuzumab (subject to NICE appraisal) • For treated CLL, including (but not limited to): - o Bendamustine with or without rituximab - o Venetoclax with rituximab - o Ibrutinib - o Rituximab with fludarabine and cyclophosphamide - o Idelalisib with rituximab	AstraZeneca	Sheffield	C
27	People with untreated metastatic Merkel cell carcinoma	Avelumab	• Chemotherapy (such as cisplatin or carboplatin with or without etoposide) • Best supportive care	Merck Serono/Pfizer	BMJ	A
28a 28b	• People with untreated advanced hormone-receptor positive HER2-negative breast cancer • People with advanced hormone-receptor positive HER2-negative breast cancer that has progressed after prior endocrine therapy	Ribociclib in combination with fulvestrant	• For people with untreated advanced hormone-receptor positive HER2-negative breast cancer: - o Palbociclib in combination with an aromatase inhibitor - o Ribociclib in combination with an aromatase inhibitor - o Tamoxifen (in accordance with NICE guidance CG81) • For people with advanced hormone-receptor positive HER2-negative breast cancer that has progressed after one line of prior endocrine therapy: - o Exemestane - o Everolimus and exemestane - o Tamoxifen - o Fulvestrant - o Chemotherapy (in accordance with NICE guidance CG81)	Novartis	BMJ	A
29a 29b	People with completely resected stage III or IV melanoma	Nivolumab	Routine surveillance	Bristol-Myers Squibb	BMJ	A
30	Adults with untreated metastatic non-squamous NSCLC	Pembrolizumab in combination with pemetrexed and platinum chemotherapy	• Pemetrexed in combination with a platinum drug (carboplatin or cisplatin) (for people with adenocarcinoma or large-cell carcinoma only) - o with (following cisplatin-containing regimens only) or without pemetrexed maintenance treatment • Chemotherapy (docetaxel, gemcitabine, paclitaxel or vinorelbine) in combination with a platinum drug (carboplatin or cisplatin) - o with or without pemetrexed maintenance treatment • Pembrolizumab monotherapy (for people whose tumours express PD-L1 with at least a 50% tumour proportion score) (recommended for use within Cancer Drugs Fund; guidance update in progress)	Merck Sharp & Dohme	Exeter	D
31	People with newly diagnosed multiple myeloma who have had autologous stem cell transplantation	Lenalidomide	Established clinical management without lenalidomide maintenance therapy (including monitoring and follow up)	Celgene	Exeter	B
32	People with relapsed or refractory mantle cell lymphoma who have received at least two previous lines of therapy	KTE-X19	Established clinical management including but not limited to: • Chemotherapy with or without rituximab • Allogeneic haemopoietic stem cell transplant	Kite	York	A
33	People with advanced ovarian, fallopian tube, or primary peritoneal cancer that has responded (complete or partial) to first-line platinum-based chemotherapy	Niraparib	Routine surveillance	GlaxoSmithKline	BMJ	A
34	Adults with metastatic gastric or gastro-oesophageal junction cancer, who have had 2 or more previous therapies	Trifluridine–tipiracil hydrochloride	• Chemotherapy (such as docetaxel or paclitaxel monotherapy or combination chemotherapy) • Best supportive care (including but not limited to antiemetics, blood transfusions, oesophageal stents, palliative radiotherapy and palliative surgery)	Servier	Sheffield	C
35	Adults with ALK-positive advanced NSCLC that has not been previously treated with an ALK inhibitor	Brigatinib	• Alectinib • Ceritinib • Crizotinib	Takeda	Liverpool	D
36	People with previously treated BRAF V600E mutation positive metastatic colorectal cancer	• Encorafenib with cetuximab • Encorafenib with cetuximab and binimetinib	• Folinic acid plus fluorouracil plus irinotecan (FOLFIRI) • Irinotecan • Trifluridine-tipiracil (only after treatment with fluoropyrimidine-, oxaliplatin- or irinotecan-based chemotherapies or where these are not tolerated or unsuitable) • Best supportive care	Pierre Fabre (encorafenib) and Merck (cetuximab)	Warwick	B

Appendix Table 9: STA review data extraction (Part 3 of 10), direct or indirect estimation of OS

#	How is OS modelled in the CS?	Description (based on reporting in the CS or other appraisal documentation)
1	Both indirectly and directly	CS, p172 "The base case overall survival for fedratinib is based on the Gompertz extrapolation of the JAKARTA-2 Int-2/high risk survival data. " However, company later prepared an addendum in which OS for the intervention was calculated separately for responders versus non-responders after discontinuation. CS addendum: "Addition of functionality for fedratinib OS to be estimated from time of treatment discontinuation (TTD), to resolve concerns regarding face validity of the stopping rule and the independence of OS and TTD."
2	Indirectly	CS, p88-95 describe the model structure. Structure is complex, but diagrams illustrate how OS is estimated to inform model transitions, but broken down based on a decision tree
3	Indirectly	CS, p61-62 "Pre-recurrence mortality (i.e. the transition directly from disease free to death) for the first three years in the disease free health state was modelled using a logistic regression of prerecurrence death events in the CheckMate 577 [IPD]; specifically any events that were death without recurrence" and "Due to immaturity of the OS data, the [IPD] were not able to provide death events for modelling an OS curve to inform this transition. Therefore, the modelled death events post recurrence are informed by the literature"
4	Directly	CS, p123 "The proportion of patients in each health state at each monthly model cycle was then determined for each therapy directly from cumulative survival probabilities from PFS and OS curves"
5	Directly	CS, p49 presents tabulated summary of scenario analyses, including use of alternative models as the survival extrapolation of OS
6		CS, p50 presents tabulated summary of scenario analyses, including use of informed curve fits for the outcome of OS
7	Directly	CS, p22 "OS, PFS and TTD results from IMvigor130 were extrapolated to the time-horizon of the model as lifetime results are not available for subjects in the IMvigor130 study"
8	Directly	CS, p28 "In addition to standard parametric distributions (exponential, Weibull, loglogistic, lognormal, Gompertz, and generalised gamma) and spline models, piecewise analyses (lognormal and exponential) were also explored for OS in line with the committee's preferred assumptions in TA490"
9	Directly	CS, p2 "The follow-up period of KEYNOTE-590 was shorter than the time horizon of the economic model, therefore, extrapolation of the OS and PFS curves was required. Parametric models were fitted to the KEYNOTE-590 KM data."
10	Directly	CS p112, "Key efficacy data for midostaurin come from the pivotal D2201 trial where individual patient-level clinical data (IPD) from the latest data cut-off: 1st December 2014 for PFS/DoR and 24th August 2017 for OS (Section B.2.6) were used to inform the base case economic analysis."
11	Directly	ERG report, p46 "The company's model structure remains unchanged from TA579. The cohort-based partitioned survival model (presented in Figure 8) includes three health states: progression-free survival (PFS), progressed disease (PD), and death" - model estimates OS directly as part of this structure, but limited explanations provided in CS owing to this appraisal being a CDF resubmission
12	Directly	CS, p100: "The partitioned survival approach allows for modelling of OS and PFS based on study-observed events, which is expected to reflect disease progression and the long-term expected survival profile of patients treated with nivolumab in combination with ipilimumab"
13	Directly	CS p86: "OS for patients treated with pemigatinib was informed by parametric survival models fitted to the FIGHT-202 OS KM data for Cohort A patients (participants with FGFR2 fusions/rearrangements; Table 34)"
14	Indirectly	CS p113: "Transitions to the Death health state (disease stage specific) were derived from Agar et al. (2010)"
15a 15b 15c	Indirectly	CS p27: "Transition probabilities were estimated from LATITUDE trial data using multi-state modelling (MSM)" Multi-state modelling approach, so OS is not estimated directly
16	Directly	CS p97: "Clinical data to inform NIVO+IPI PFS and OS can be derived from CheckMate 142. However, follow-up was substantially less than the maximum time horizon of the model. Therefore, parametric extrapolation of survival data from the study was required to inform long-term outcomes..."
17	Directly	CS p118: "Finally, all health states are subject to mortality, informed by OS, with death being the absorbing final health state of the model" and "In the model, the area under the TTD curve represents the mHSPC on treatment health state, everything between rPFS and OS represents mHRPC and everything above OS represents death"
18	Directly	CS p12: "The observed Kaplan-Meier followed by the exponential model is an appropriate method for extrapolating OS."
19a 19b	Directly	CS p97: "In the partitioned survival model, the proportion of patients within each health state at each point in time is based directly on each treatment's progression free survival (PFS) and overall survival (OS) estimates"
20	Directly	CS p92 "Health state occupancy is defined by treatment-specific PFS and OS extrapolations, derived from available data..."
21	Directly	CS p86 "The CUA model was developed in Microsoft Excel and is an Area-Under-the-Curve (AUC) or PartSA model" and "Further, the direct correspondence between the key endpoints of the IMpower110 trial (i.e. PFS and OS) and the survival functions required to determine state occupancy allows for full use of the IMpower110 data"
22	Directly	CS p110 " The OS curve is used to inform the proportion of individuals in the 'Dead' health state over time"
23	Directly	CS p14 "The model effectiveness parameters were estimated from the KEYNOTE-045 patient-level data for OS, PFS and ToT [9-11]. Parametric models were fitted to the KM data and the survival curve fitting was carried out in line with the NICE Decision Support Unit (DSU) guidelines..."
24	Directly	CS p114 "The death health state captures patient deaths from both cancer and non-cancer related causes; the proportion of patients occupying the death state is estimated as one minus the cumulative survival of OS"
25	Directly	CS p82 "A Weibull curve was selected for OS of CRd and Rd for the duration of ASPIRE"
26a 26b	Indirectly	CS p117: "This technique allows the model to track the survival of each cohort entering the PD state which is vital, as upon progression to the R/R setting, patients experience deterioration in HRQoL arising from disease progression and its associated impact on anxiety and symptom burden."
27	Directly	CS p17: " For OS and PFS, simple and spline-based parametric models were fitted, which cover the range of extrapolation approaches discussed by the committee in TA517"
28a 28b	Indirectly	Committee slides: "Company changed model structure to allow use of trial OS data, as requested by ERG" - original model did not use OS data, but later company switched to a PartSA model (at technical engagement) which then used OS directly
29a 29b	Indirectly	CS p33: "The previous submission presented two modelling approaches: a partitioned survival model and a state-transition model. [IPD] meta-regression with survival analyses were conducted using the latest data from CheckMate 238 (RFS and OS) in order to extrapolate these outcomes over the modelled lifetime horizon. " In the PartSA (which was preferred by the committee, as acknowledged in the FAD), OS was estimated directly. In the state-transition model, OS was estimated indirectly via pre- and post-recurrence survival estimates.
30	Directly	CS p23: "Guidance from the NICE DSU document was followed to identify the base case parametric survival models for OS and PFS extrapolation (see sections A.8.2 and A.8.3 respectively)"
31	Directly	CS p64: "The OS curve is used to estimate the proportion of patients alive over time; the area under the extrapolated OS curve provides an estimate of mean life expectancy"
32	Directly	CS p98: "In a partitioned survival model, OS and PFS are modelled independently and the proportions of patients in each health state over time are derived directly from the OS and PFS projections"
33	Directly	CS p113: "A de novo partitioned survival model was developed to assess the cost-effectiveness of niraparib versus RS."
34	Directly	CS p72: "The probability of patients transitioning between the alive health states is inferred via extrapolated PFS and OS curves that are fitted to the clinical trial data; however, transition probabilities are not explicitly calculated for all possible transitions"
35	Directly	CS p92: "To inform the inputs for brigatinib and crizotinib in the economic model, the data from the ALTA-1L trial were extrapolated for the following outcomes: OS, PFS BIRC and intracranial PFS"
36	Directly	CS p98: "OS and PFS data from the clinical trial can be used directly in the model"

Appendix Table 10: STA review data extraction (Part 4 of 10), main data sources

#	Pivotal trial name(s)	Noteworthy trial features (intervention and/or comparator)?	Survival derived from RCT?	Survival derived from another source?	Explanation
1	JAKARTA-2	Relatively small sample (n=97 in the ITT population) combined with single-arm design	No	Yes	Single-arm trial for estimating outcomes for intervention arm (JAKARTA-2). For comparator, four different options were included based on studies identified via a literature review: • Schain et al., (2019) - retrospective observational study *Base case* • Kuykendall et al., (2017) - retrospective observational study • COMFORT-II (NICE 2016) - Phase III RCT • Palandri et al., (2019) - retrospective observational study
2	MAVORIC	One-direction switching was permitted (comparator to intervention), affecting 72.6% of the comparator group. Company employed crossover adjustment methods to account for this feature of the trial.	Yes	Yes	MARVORIC was an RCT, but compared to another active therapy which is not standard of care. However, the company made a case for the comparator arm being a reasonable proxy for the comparator ('Established clinical management'). Outcomes for patients after stem cell transplant were taken from NICE TA577 - these data were derived from published KM curves based on real-world data from the London supra-regional centre using the minimal intensity Stanford Protocol.
3	CheckMate-577	Limited follow-up from trial with respect to long-term outcomes needed for model owing to adjuvant setting	Yes	Yes	CHECKMATE-577 used to derive rate of disease recurrence and deaths prior to recurrence. However, after recurrence, deaths were informed by analysis of data from a published study by Lou et al., (2013). This was a retrospective cohort study of a single centre.
4	LIBRETTO-001	Single-arm study, covering a large number of patients but with different characteristics. Appraisal considers two populations, one of which has only n=19 patients in this study. n=212 used for the other population	Yes	Yes	Used propensity score weighted data from LIBRETTO-001 for the intervention. For comparator, used either unweighted data from published study (EXAM and SELECT studies, depending on subgroup). OS from SELECT was based on RPSFT-adjusted data. HR from EXAM used to obtain OS for another comparator. In other words, an unanchored MAIC was performed for both populations
5	SPARTAN	OS possibly affected by crossover. Company discusses a 'one novel therapy' commissioning rule, meaning there was an issue of generalisability with the study. Company used external data from another study (COU-AA-302) to adjust trial data to account for the impact of this on outcomes.	Yes	No	Standard parametric models used for OS
6	TITAN	OS possibly affected by crossover + data considered immature. Company discusses a 'one novel therapy' commissioning rule, meaning there was an issue of generalisability with the study	Yes	Yes	Company cites Pennington et al., (2018) for using 'informed fits' wherein historical trial data were used to inform extrapolations of OS. No other details given as only summary submission presented. HR generated for another comparator
7	IMvigor130	Longer follow-up versus original appraisal (around 4 years)	Yes	No	Parametric models fitted to data from IMvigor130. Used KM + extrapolated tail. Treatment waning included by assuming full effect until 5 years, gradually waning completely by 7 years
8	CheckMate-141	Longer follow-up versus original appraisal (around 4 years)	Yes	No	Parametric models fitted to data from CheckMate-141. Range of models, including standard parametric models, spline-based models, and piecewise KM + extrapolated tail models
9	KeyNote-590	No crossover permitted, treatment with intervention stopped at 2 years. Generalisability issue flagged by ERG with respect to the trial population (region)	Yes	No	Parametric models fitted to data from KeyNote-590. Two approaches: fully parametric, or KM + extrapolated tail
10	D2201	Single-arm trial with small sample size (n=166). Results considered for the full population, and a subgroup (SM-AHN + MCL)	No	Yes	Used D2201 data for the intervention. For the comparator, applied a hazard ratio (HR) based on registry data in the base-case analysis (Reiter et al., 2017)
11	MONARCH 2	Data from MONARCH 2 were originally deemed immature in the previous assessment, and originally made use of an external data source (CONFIRM trial) for long-term extrapolation. Data were considered more mature after the CDF data collection period	Yes	Yes	Fractional polynomial (FP) network meta-analysis (NMA) undertaken for a range of studies. A 'base curve' for a non-scoped comparator (but a treatment in the network) was produced using data from MONARCH 2 and CONFIRM studies. Methodology to combine data from MONARCH 2 and CONFIRM studies unclear, but it is implied that these data were synthesized as part of the broad FP-NMA approach
12	CheckMate 9LA	OS data considered immature (minimum follow up of approximately 12.7 months). Comparator dependent on subgroup, but company considered 'all comers' as the population of primary interest	Yes	Yes	Data from another study (CheckMate 227 - Part 1) used to inform long-term survival. Broadly, a piecewise approach was taken using KM + extrapolated tail, with the tail informed by external data
13	FIGHT-202	Single-arm trial with three cohorts: A (n=107, with a specific mutation), B (n=20, with a different but still specific mutation), and C (n=18, without mutation)	Yes	Yes	Due to single-arm trial design of FIGHT-202, an ITC was needed. Company undertook a MAIC to obtain relative treatment effects, based on another study (ABC-06)
14	Study 201	Not designed to evaluate OS	No	Yes	From the CS: "Within the cost-effectiveness model, patients could transition to the Death health state from any other health state within any disease stage. Transition probabilities were treatment independent and were based on disease stage only (i.e. independent of skin burden)." Transitions were estimated from a published study by Agar et al., (2010) (a published, single-centre, retrospective cohort study of 1,502 patients). No treatment effect was captured in the company's model related to OS (equal LYG in company's base-case analysis).
15a	LATITUDE	Large sample size (n=1,199)	Yes	Yes	In the model, survival after progression is informed by external trial data (COU-AA-302 - a population with metastatic castration-resistant prostate cancer, mCRPC), whereas prior to progression patients are considered to have metastatic hormone-sensitive prostate cancer (mHSPC). There was also some 'calibration' of the survival curves to adjust for differences in populations (between LATITUDE and COU-AA-302). Comparator efficacy informed via an ITC
15b			Yes	No	In the PartSA model, the company fitted standard parametric models (e.g., Weibull). Comparator efficacy informed via an ITC
15c			No	No	As above
16	CheckMate 142	Within-trial comparator not relevant to scope	Yes	Yes	CheckMate 142 used for intervention. Matching-adjusted indirect comparison (MAIC) performed to estimate outcomes for the comparator arm
17	ARCHES and ENZAMET	Both studies were deemed similar, so in the base-case analysis the company naively pooled data from both sources to inform OS	Yes	Yes	Data pooled from two studies were used as the source from which to fit models. Both are RCTs. For the non-trial comparator, an ITC was used to derive a hazard ratio
18	CheckMate 057	Pre-CDF OS data were immature, but post-CDF provided data with up to 5 years of follow-up	Yes	No	Models fitted to each arm of the CheckMate 057 study
19a	KeyNote-177	Patients could crossover, and 80.5% of patients in the comparator group that had progressed did indeed switch to receive the intervention. In addition, OS was immature - at the end of the KM for the intervention, 60.6% were still alive (49.2% for the comparator)	Yes	No	No details presented in main submission, so remainder of extraction not completed
19b			Yes	Yes	Most transitions based on the KeyNote-177 study. However, post-surgery PFS and OS data were obtained from the literature (due to small numbers of patients and limited follow-up). Two studies were considered: Adam et al. and Cucchetti et al., with the latter used in the base-case analysis. Both studies were observational studies. Post-surgery was modelled as a 'mini' PartSA. Outcomes for the comparator arm were based on a fractional polynomial ITC (for KeyNote-177 based outcomes)
20	ATTRACTION-3	Predominantly Asian trial population (96%), so some concerns raised with respect to generalisability. Further to this, OS curves for intervention and comparator cross at ~3 months	Yes	No	Estimates based on ATTRACTION-3 data only

21	IMpower110	Within-trial comparator not relevant to scope	Yes	Yes	OS estimates for intervention based on pivotal trial. For comparator, model used time-varying hazard ratios derived from an ITC
22	DESTINY-Breast01	OS data deemed too immature to inform the cost-effectiveness model by itself	Yes	Yes	OS for the intervention estimated based on a hazard ratio applied to an extrapolated curve based on a different study (TH3RESA) for a "similar" treatment (based on mechanistic properties). TH3RESA had more mature data considered suitable for extrapolation. For comparators, Kaplan-Meier estimates were taken from external clinical trial data (Cortes 2011 and Fumoleau 2004) without adjustment, and patient-level data were digitised to allow for the fitting of parametric survival models
23	KeyNote-045	Control arm went on to receive subsequent PD-1/PD-L1 treatments, and company explained this would dilute the estimated survival benefit of the intervention, so crossover adjustment was undertaken to remove the impact of subsequent treatment on survival	Yes	No	Fitted piecewise KM + extrapolated tail models to both arms (with crossover-adjusted data used for the comparator)
24	PAOLA-1	OS data from PAOLA-1 considered to be immature	Yes	No	Standard parametric models fitted, but when the OS model crossed the PFS (or PFS2) model OS was assumed to be equal to PFS (or PFS2). Company used a mixture-cure model for PFS, hence when OS curve crosses PFS, eventually this means that OS is assumed to be the same as the general population
25	ASPIRE	ASPIRE has a median follow-up of 5.5 years, which was noted to be double the length of follow-up versus what was available when NICE originally assessed this intervention as part of a broader (earlier) appraisal (TA457)	Yes	Yes	Hybrid model fitted including data from ASPIRE until 72 months, then data from a registry study (MyelomaToul) for remainder of time horizon. Registry data reflective of patients treated with the control intervention only, and included a relatively large sample size (n=1,890 patients that received the relevant treatment in the relevant line). Efficacy of intervention was assumed to be the same as the efficacy of a specific comparator based on clinical input
26a	ELEVATE-TN	OS data deemed "too immature to provide robust parametric modelling estimates"	Yes	Yes	Death prior to progression was based on data from the ELEVATE-TN study. Death after progression was based on data from two other clinical trials: RESONATE and MURANO
26b	ASCEND	Follow-up in ASCEND not as long as in RESONATE (RCT of ibrutinib - the comparator, but not against acalabrutinib [the intervention])	Yes	Yes	Base-case approach was a cost-minimisation analysis, where curves were fitted to the RESONATE trial (ibrutinib arm) and assumed to be the same for acalabrutinib
27	JAVELIN Merkel 200 (Part B)	Single-arm trial with small sample size (n=116)	No	Yes	Pivotal trial was single-arm, and so company's model made use of data collected from observational studies to populate the comparator arm.
28a	MONALEESA-3	Comprised two subpopulations, of which subpopulation B is the group deemed relevant to this appraisal	Yes	Yes	Most transitions informed by the MONALEESA-3 study, including post-progression survival which was assumed equal across arms. However, PFS for the comparator was informed by data from the BOLERO-2 trial (which indirectly affects projections of OS)
28b			Yes	Yes	OS curve fitted to data from MONALEESA-3, but for the comparator a hazard ratio was applied based on a Bucher indirect comparison to the BOLERO-2 study
29a			Yes	Yes	Deaths were modelled separately pre-recurrence and post-recurrence. Pre-recurrence deaths were captured based on data from CheckMate-238 and an ITC. Post-recurrence deaths were based on weighted survival curves based on different treatments that may be offered in this setting, with weights based on subsequent treatment data from CheckMate-238. From the CS: "long-term survival estimates were estimated from the AJCC 8th edition registry data applied at 10 years... The 10-year time point was selected as this is approximately when long-term melanoma survival outcomes begin to plateau, suggesting that the risk of death due to melanoma is reduced at this time"
29b	CheckMate-238	Due to the setting in which the trial was undertaken, OS data were considered immature even though there was a minimum follow-up of 48 months - median OS not reached	Yes	Yes	A patient-level meta-regression was undertaken to compare nivolumab to routine surveillance (with data from CheckMate-238 and CA184-029, with ipilimumab as a link between the studies). This yielded parametric model parameters which included a multi-dimensional treatment effect (i.e., did not assume proportional hazards or accelerated failure times). To then implement these curves within the model, the company used the corrected group prognosis (CGP) method to appropriately weight the curves to ensure they were aligned with the population of interest (most notably, patients with Stage IIIA - Stage IV and no evidence of disease [NED]). From the CS: "long-term survival estimates were estimated from the AJCC 8th edition registry data applied at 10 years... The 10-year time point was selected as this is approximately when long-term melanoma survival outcomes begin to plateau, suggesting that the risk of death due to melanoma is reduced at this time"
30	KeyNote-189	Median duration of follow-up pre-CDF was 18.8 months, which increased by an extra 8.3 months based on data collected after the original submission was made	Yes	No	Company fitted standard parametric models (e.g., Weibull) as well as piecewise models (KM + extrapolated tail). Scenarios were also explored with 'treatment waning effects' applied where hazards were assumed equal across arms from a given time point
31	Myeloma XI and CALGB 100104	Several issues affecting both studies, including use of a dose of the intervention that was not aligned with the marketing authorisation and subsequent therapies not reflecting practice (as well as treatments only available via the CDF)	Yes	No	Independent parametric models fitted to data from Myeloma XI - log-logistic for intervention, Weibull for comparator
32	ZUMA-1	Single arm trial, and due to nature of intervention (single infusion), company submitted with a modified ITT population in the base case	No	Yes	Data for OS for the comparator arm was taken from published literature. Long-term survival estimates for both arms were adjusted based on a standardised mortality ratio (SMR)
33	PRIMA	Immature data (<10% of patients died on intervention arm)	Yes	Yes	KM for OS from trial was deemed too immature for extrapolation. OS was estimated for the intervention using the following relationship: mean intervention OS = mean comparator OS + mean difference in PFS * 2. OS for the comparator was estimated from the PRIMA trial data
34	TAGS	Relatively mature, given prognosis poor for this cohort	Yes	No	Phase III placebo-controlled RCT. Comparator assumed to serve as proxy for BSC
35	ALTA-1L	Data from ALTA-1L trial were noted to be both immature and confounded by crossover	Yes	Yes	For the intervention, standard parametric models were fitted to the relevant arm of the ALTA-1L study. For the comparators, indirect treatment comparisons were conducted to produce equivalent OS estimates (based on estimated hazard ratios (HRs), with the ALEX trial of most relevance to allow for a comparison to alectinib. Given that patients were permitted to crossover in the ALTA-1L study, analyses were conducted to remove the potential bias associated with treatment switching as a part of the indirect comparison reflected within the company's economic model.
36	BEACON CRC	Trial comparator not relevant to scope of appraisal, so company relies on ITCs	Yes	Yes	BEACON CRC provides data for the intervention, but ITC needed to compare to FOLFIRI. Considers data from two studies by Peeters <i>et al.</i> , (2010 and 2015), with the assumption that FOLFIRI + panitumumab is equivalent to irinotecan + cetuximab (i.e., node was grouped in the ITC)

Appendix Table 11: STA review data extraction (Part 5 of 10), background mortality and informal use

#	Was background mortality data used to inform estimation of OS? If so, how was this accounted for?		Was there any informal use of external evidence? If so, description provided based on reporting in the CS or other appraisal documentation. If not, explanation provided for why this was not the case	
1	No or not stated		Yes	CS p173: "Clinical opinion was that fedratinib OS would not be expected to be worse than BAT OS at any time point."
2	Yes or implied	From the CS: "Additionally, in the model the weekly probability of death has been limited to be no greater than the latest available age- and gender-equivalent general population values from Office for National Statistics (ONS) data for England."	Yes	CS p119-120: "The Gompertz model was preferred ... This behaviour was validated by clinical experts who considered that after 12 months of receiving an aSCT, patients would enter a long-term remission where it would be unlikely to observe new relapse events."
3	Yes or implied	Background mortality explicitly modelled as an additional cause of death in each model cycle. The company acknowledges that this likely introduces some form of double counting, but that this is expected to have a minimal impact on predicted survival (and by extension, cost effectiveness).	Yes	CS p66: "After three years the mortality in the disease free health state is assumed as general population, and assumption which was validated by clinicians."
4	Yes or implied	From the CS: "Mean age and the percentage female were used alongside UK life tables to calculate the natural mortality of the general population"	Yes	CS p142: "Based on feedback from clinical experts and for consistency with TA535, the piecewise exponential model was used to extrapolate OS in the base case analysis, as shown in Figure 40."
5	No or not stated		Yes	CS p39: "The suitability of independent versus dependent curve fitting approaches and each of the parametric distributions were assessed using the following criteria: ... Clinical plausibility for both short- and long-term estimates of survival, based on clinical expert validation"
6	No or not stated		Yes	CS p39: "The suitability of independent versus dependent curve fitting approaches and each of the parametric distributions were assessed using the following criteria: ... Clinical plausibility for both short- and long-term estimates of survival, based on clinical expert validation"
7	No or not stated		Yes	CS p22: "Clinical expert opinion was also utilised to validate the extrapolation approach taken"
8	No or not stated		Yes	CS p26: "The striking similarity in OS observed between the SACT data cohort and the CheckMate 141 trial, despite the SACT data cohort including 7% patients with ECOG performance status 2-3, supports the generalisability of the OS data from the CheckMate 141 trial to patients who might receive nivolumab in UK clinical practice."
9	No or not stated		Yes	CS p95-96: "The log-logistic curve's tail was considered by clinical experts to be more credible based on the expectation that a percentage of patients would derive a long-term survival benefit from treatment with pembrolizumab in combination with chemotherapy"
10	Yes or implied	From the CS: "The mean age and gender distribution was used in the model alongside UK life tables..." "... to incorporate general population mortality when extrapolating survival curves (to avoid the extrapolated hazard to be less than the expected hazard of death in the general population)"	Yes	CS p117: "Assessment of the long-term extrapolation for OS was informed by both (i) clinical expert opinion and (ii) external validation using long-term evidence from the A2213 trial."
11	No or not stated		No or unclear	No evidence explicitly presented, though this was a condensed submission as it was a CDF review.
12	Yes or implied	From the ERG report: "Survival is adjusted according to the ONS general population mortality life tables... to ensure it is never better than the general population"	Yes	CS p169: "Clinical experts reviewed the survival curves for the best-fitting models for OS ... The considerations of the experts were used as part of the validation process, described in Section B.3.3, to identify the most appropriate survival models to use in the economic model."
13	Yes or implied	From the CS: "General population mortality was included to ensure that the modelled mortality risk did not fall below the general population mortality risk at any given age. To do so, the hazards of PFS and OS events were always equal to or exceeded the general population mortality hazard."	Yes	CS p92: "Therefore, the log-normal distribution was chosen as it has the better statistical fit and is closely aligned with clinical expert opinion"
14	Yes or implied	From the CS: "A built-in constraint was applied to ensure that the modelled (i.e. disease-specific) mortality did not drop below that of the general population mortality at any time point."	No or unclear	Information related to clinical expert input was provided in an appendix which is not publicly available. No other discussion presented suggesting informal use of external evidence.
15a	Yes or implied	From the pre-meeting briefing slides: "Trial data + age-adjusted general population mortality from Office for National Statistics"	Yes	CS p28: "Survival curves were taken from the base case analysis in the previous appraisal of AAP and BSC in mCRPC [TA387], which were extrapolated from the COU-AA302 study. All life-extending subsequent therapies in mCRPC were assumed to have the same relative effectiveness as AAP, as validated by clinical experts in prostate cancer."
15b	Yes or implied	From the company's submission addendum: "All extrapolations were also adjusted for general population mortality; if the predicted hazard based on the parametric survival curves fell below that of the general population, the general population mortality hazard was applied."		
15c	Yes or implied	As above		
16	Yes or implied	From the CS: "Therefore, the model includes age and gender-adjusted mortality based on information from UK life tables..." "...These values are included in every cycle in addition to the disease-related mortality values and are applied multiplicatively. While some form of double counting occurs, this effect applies equally to all comparators and is likely to have a minimal impact on predicted survival (and hence cost-effectiveness)."	Yes	CS p98: "Additionally, clinical expert opinion was sought to ensure that the survival extrapolation approach can be considered appropriate."
17	Yes or implied	From the CS: "The OS curves were adjusted for background mortality, by increasing the mortality rate in any cycle to the age-matched population mortality rate, if the population mortality was higher. Consequently, the exponential, lognormal, log-logistic and generalised gamma curves are virtually identical, as all have crossed"	Yes	CS p126: "Longterm survival estimates from STAMPEDE, CHAARTED and GETUG suggested a 7-year survival in the standard of care arm of 27%-34%. Based on these data, log-logistic provided the most clinically plausible fit..."
18	No or not stated		Yes	CS p34: "To allow comparison of the predicted survival from the extrapolations with the long term trial data, Table 12 presents the survival estimates generated by the survival curves considered for the updated company base case alongside the observed survival from both CheckMate 057 and CheckMate 003."
19a	No or not stated			
19b	Yes or implied	From the ERG report: "General population mortality was applied to TTP estimates as a minimum, such that the probability of a patient being alive and progression-free could never be below background mortality."	Yes	CS p117: "Input from clinical experts suggested the exponential curve was the most clinically plausible estimation."
20	Yes or implied	From the CS: "... Therefore, the model includes age and gender-adjusted mortality based on information from UK life tables... These values are included in every cycle in addition to the disease-related mortality values and are applied multiplicatively. While some form of double counting occurs, this effect applies equally to all comparators and is likely to have a minimal impact on predicted survival (and hence cost-effectiveness)."	Yes	CS p104: "The only other source available to validate the estimates for the taxane arm are from Auzolle et al. This study reported an OS median of 7.5 months and 3.9 for PFS although ITT numbers were low (n=29 in the taxane arm). These estimates are lower for OS than the observed and predicted values for the taxane arm but are broadly in line with the estimates for the taxane arm that were modelled."
21	No or not stated		Yes	CS p93: "Curves were visually inspected and validated against relevant long-term data sources available to help identify the most plausible survival model. Clinical expert opinion was also sought to validate the extrapolation approach taken and determine which of the extrapolations better represent UK clinical practice."

22	Yes or implied	From the CS: "Life tables are used in the model to ensure the weekly probability of mortality never falls below that of the general population"	Yes	CS p116: "OS for eribulin and capecitabine is estimated by fitting parametric survival curves to the digitized KM data from the relevant studies; given that available OS data for vinorelbine were not considered plausible or reflective of survival outcomes in UK patients in this setting by clinical experts at the August advisory board and PFS estimates for vinorelbine were similar to/lower than for capecitabine..."
23	No or not stated		Yes	CS p23: "Additionally, extrapolation with this choice of curve provides results that are consistent with the Expert clinical opinion concerning 5-year OS in the SoC arm"
24	Yes or implied	From the ERG report: "After year 5 in the model, the proportion of long-term survivors in the PFS curve incurred the background mortality rate for the UK general population matched by age and gender." and "At 5 and 6 years, respectively, all short-term survivors have progressed in the company's base case."	Yes	Comparison of ratio of PFS:OS in previous appraisals presented as external data. CS p135: "Specifically, the mean PFS:OS ratio of 1:0.93 predicted by the model (i.e. one month of PFS translates to 0.93 month of OS) is conservative when compared to ratios accepted in previous NICE appraisals in the more advanced relapsed ovarian cancer setting (TA528, TA611, TA620) (Table 47)"
25	Yes or implied	From the CS: "Survival in the model is also capped by age-sex-matched general population mortality"	Yes	CS p84: "Since most models had a similar fit to the ASPIRE data, assessment of the plausibility of the extrapolations based on clinical expert opinion and comparisons to long-term data were used to select the base case survival extrapolation approach."
26a	Yes or implied	From the CS: "To ensure survival did not exceed that expected for the general population, TP2 was restricted by general population mortality which was applied as a competing risk" and "As with TP2, general population mortality was applied as a competing risk to prevent survival exceeding that of the general population"	Yes	CS p127: "For all curves presented in the following sections, the following key criteria were applied: - Clinical plausibility of long-term extrapolation compared to real world data..."
26b	Yes or implied	From the CS: "To avoid unrealistic survival projections in the model, OS was constrained by age and gender adjusted general population mortality. At each model cycle, the highest mortality risk (from the OS curve or general population mortality) was used to estimate survival for the next cycle. The mortality risks of the general population were sourced from UK National Statistics"	Yes	CS p177: "Based on the results of the MAIC and the findings from the clinical experts (See Section B2.), it was determined that a cost-minimisation analysis would be conducted for the comparison against irutrinib in R/R CLL."
27	Yes or implied	From the CS: "The following model corrections were made: ... corrected an error in the calculation of background mortality"	Yes	CS p17-18: "More specifically, the majority of the simple parametric models (e.g. lognormal and log-logistic) resulted in extrapolations wherein the curve crossed the latest Kaplan-Meier curve from JM200: Part A (see Appendix 3 for further information)"
28a	Yes or implied	From the CS: "It should be noted that these projections do not include general population mortality, which is captured separately in the model". The submission does not explain precisely how general population mortality was accounted for, but it is implied elsewhere in the CS that general population mortality was added to disease-specific mortality: "... such that general population mortality is considered in addition to mortality from breast cancer represented by the hazard rate for PFS" and "The model includes two death states: one for death due to breast cancer, and another for death due to other causes"	Yes	CS p69: "These updated data were not available at the time of the CDF entry, but are consistent with the data reported in the data cut (3 November 2017) used to inform the earlier assessment, along with data generated and captured within the SACT dataset (17 July 2019 and 16 January 2020)."
28b	No or not stated		Yes	TE response p7: "The base case OS was modelled using a Weibull (R) distribution, based on statistical fit, visual fit to the KM data, and with clinical experts' expectations."
29a	Yes or implied	From the CS: "In addition to registry data, all OS projections were adjusted for background mortality so that the estimated hazard of mortality was never less than that of the age-adjusted general population"	Yes	CS p35: "The generalised gamma was chosen as the base case curve given that it had the best fitting curve statistically and visually, and gave plausible outcomes compared with literature sources (see Section A.7)."
29b	Yes or implied		Yes	CS p38: "[Retaining original approach for transitions] was shown to be a conservative assumption given that SACT data already show a large proportion of anti-PD-1 use within the year of data collection."
30	Yes or implied	From the CS: "However, consistently with other NSCLC submissions, the modelled OS has been capped by the survival rate for the general population"	Yes	CS p23: "The choice of base case parametric models was validated in terms of clinical plausibility accepted by the committee during the appraisal at the CDF entry point ("A 5-year survival rate of 5% to 11% for standard care is reasonable for decision-making")."
31	Yes or implied	From the CS: "[Rules were applied in the model to] ensure that the hazards of OS and PFS do not fall below that of the age- and gender-matched general population risk of mortality: when this occurs, the matched general population mortality applies. For greater transparency, the model reports the time point at which the PFS and OS curves would have crossed, had they been permitted to. In the base case, this is estimated as occurring after 30 years and this was therefore not considered a significant problem within the analysis"	Yes	CS p77-78: "CALGB 100101 validation... For observation, the Weibull appears the best and most plausible distribution as it closely fits both Myeloma XI and the CALGB 100101 data over the entire follow-up of this study... For lenalidomide, the most plausible distributions are the log-logistic, generalised gamma and log-normal (Figure 17) although none closely fits CALGB 100101..."
32	Yes or implied	From the CS: "UK life table data to inform age- and gender-matched background mortality"	Yes	CS p127: "Given that the log-normal model was selected for KTE-X19 PFS, this was deemed most appropriate to capture the same underlying trends across the two endpoints. This was explained at NHS Consultant review, with little constructive feedback given the visual similarity across extrapolations"
33	Yes or implied	From the CS: "OS for all treatments was then implemented such that the mortality risk of the modelled population was never below the mortality risk observed in the age and gender matched general population"	Yes	CS p216: "In order to validate the OS model outputs, the long-term extrapolations for RS (ITT) were compared with external real world data, via the Edinburgh study as previously described and also with published 5-year survival rates from Cancer Research UK (Stage III and IV) data in Table 60."
34	Yes or implied	From the CS: "All extrapolations of OS are capped within the model according to the age- and sex adjusted background mortality rates, taken from the Office for National Statistics Life Tables (released September 2018). Should the predicted hazard of death for either treatment arm be less than the general population, the model will apply the hazard of death for the general population instead." Background mortality adjustment undertaken using a 'capping' approach, rather than being integrated within the survival model fitting process.	No or unclear	Information related to a clinical advisory board attended by 12 UK practicing oncologists provided as a reference, but not included in public documentation. No other discussion presented suggesting informal use of external evidence.
35	No or not stated		Yes	CS p92: "The fit of each parametric model to the survival data was assessed through AIC/BIC statistics, comparison with the Kaplan-Meier curves and experts' judgements on long-term clinical plausibility"
36	No or not stated	Company's response to clarification question B6 confirms that background mortality estimates were included via a 'capping' approach, but in the base-case analysis these did not influence results	Yes	CS p112: "Following confirmatory consultation with clinical experts, it was determined that it was appropriate to assume the same efficacy for both FOLFIRI and irinotecan, and that data from the control arm could be used to inform the parametric models for FOLFIRI."

Appendix Table 12: STA review data extraction (Part 6 of 10), summary of external evidence used

#	Any for intervention?	Quick summary	Any for comparator?	Quick summary	Background mortality	ITC	Trial data	Non-trial data	Opinion/ assumption
1	Yes	Technically uses comparator curve after a given timepoint, which is from another study	No		x	✓	x	✓	✓
2	Yes	Uses background mortality for adjustment, and uses external data for post-transplant survival	Yes	Same as intervention, but adjusted for crossover	✓	x	x	✓	x
3	Yes	External data used to populate post-recurrence survival. Background mortality included at each cycle for all death transitions (not adjusted, added on)	Yes	Same as intervention	✓	x	x	✓	x
4	Yes	Life tables used to account for background mortality + external trial data used for comparator efficacy (via MAIC)	Yes	Same as intervention + ITC	✓	✓	x	x	x
5	No		No		x	x	x	x	x
6	Yes	Historical data used for 'informed fits'. Unclear if trial and/or non-trial data, but based on ERG report it suggests there is a mixture of trial and non-trial evidence: "Seven published studies selected in a systematic literature review... had longer follow-up relative to that in TITAN". Here, ERG uses the term "studies", whereas earlier it uses "trials" when only discussing clinical trial data	Yes	Same as intervention + ITC	x	✓	✓	✓	x
7	No		No		x	x	x	x	x
8	No		No		x	x	x	x	x
9	No		No		x	x	x	x	x
10	Yes	Life tables used to account for background mortality	Yes	Same as intervention + ITC	✓	✓	x	x	x
11	Yes	External trial data used within FP-NMA	Yes	Same as intervention	x	✓	x	x	x
12	Yes	External trial data used to inform death after a cut point, plus use of life tables for background mortality	Yes	Same as intervention + ITC	✓	✓	✓	x	x
13	Yes	Life tables used to account for background mortality	Yes	Same as intervention + ITC	✓	✓	x	x	x
14	Yes	OS derived entirely from a combination of a cohort study and background mortality	Yes	Same as intervention	✓	x	x	✓	x
15a	Yes	External trial data used to inform death after progression, plus use of life tables for background mortality	Yes	Same as intervention + ITC	✓	✓	✓	x	x
15b	Yes	Life tables used to account for background mortality	Yes	Same as intervention + ITC	✓	✓	x	x	x
15c	Yes	As above	Yes	As above	✓	✓	x	x	x
16	Yes	Life tables used to account for background mortality	Yes	Same as intervention + ITC	✓	✓	x	x	x
17	Yes	Life tables used to account for background mortality + use of pooled data as basis for OS modelling (but both RCTs)	Yes	Same as intervention + ITC	✓	✓	x	x	x
18	No		No		x	x	x	x	x
19a	N/A	N/A	N/A	N/A	x	x	x	x	x
19b	Yes	External data used for modelling post-surgery survival + life tables used to account for background mortality	Yes	Same as intervention + ITC	✓	✓	x	✓	x
20	Yes	Life tables used to account for background mortality	Yes	Same as intervention	✓	x	x	x	x
21	No		Yes	ITC-derived time-varying hazard ratio applied to intervention curve	x	✓	x	x	x
22	Yes	Other trial data (TH3RESA trial + HR) + life tables	Yes	Other trial data (naïve comparison based on digitising) + life tables + assumed same OS for some comparators on basis of clinical opinion/ assumption	✓	x	✓	x	✓
23	No		No		x	x	x	x	x
24	Yes	Life tables used to account for background mortality + assumption of cure	Yes	Same as intervention	✓	x	x	x	✓
25	Yes	Life tables used to account for background mortality + registry data for long-term survival + assumption of equivalent efficacy made	Yes	Same as intervention	✓	x	x	✓	✓
26a	Yes	External trial data for post-progression deaths, and life tables used to account for background mortality	Yes	Same as intervention	✓	x	✓	x	x
26b	Yes	Life tables used to account for background mortality	Yes	Same as intervention	✓	x	x	x	x
27	Yes	Life tables used to account for background mortality	Yes	Use of observational study data + life tables for background mortality	✓	x	x	✓	x
28a	Yes	Life tables used to account for background mortality	Yes	External trial data + life tables for background mortality + post-progression survival assumed equal across arms	✓	x	✓	x	✓
28b	No		Yes	External trial data + possibly life tables for background mortality (but not explicitly stated)	x	✓	✓	x	x
29a	Yes	External trial data used for post-recurrence deaths, life tables for background mortality, registry data for long-term survival after 10 years	Yes	Same as intervention + ITC	✓	✓	✓	✓	x
29b	Yes	External trial data for meta-regression and CGP weighting, life tables for background mortality, registry data for long-term survival after 10 years	Yes	Same as intervention + ITC	✓	✓	✓	✓	x
30	Yes	Life tables used to account for background mortality	Yes	Same as intervention	✓	x	x	x	x
31	Yes	Life tables + data from the CALGB 100104 study were used to determine which parametric model(s) provided the most plausible extrapolations	Yes	Same as intervention	✓	x	x	x	x
32	Yes	Life tables for background mortality + adjustment for long-term survival based on an SMR	Yes	External trial data + life tables for background mortality + adjustment for long-term survival based on an SMR	✓	x	✓	✓	x
33	Yes	Background mortality adjustment + surrogacy relationship assumed for basis of estimating HR against comparator	Yes	Background mortality adjustment	✓	x	x	x	✓
34	Yes	Background mortality adjustment	Yes	Background mortality adjustment	✓	x	x	x	x
35	No		Yes	HR derived from ITC	x	✓	x	x	x
36	No		Yes	ITC-derived hazard ratio applied to intervention curve (for within-trial comparator and ITC-only comparator)	x	✓	x	x	x

Appendix Table 13: STA review data extraction (Part 7 of 10), model structure and company's analyses

#	Description of model structure	Intervention base case OS	Comparator base case OS	Company's justification	Approach to sensitivity analysis
1	DES. Patients can experience the following events over the course of the model time horizon: • Response assessment • Loss of response (if response achieved) • Treatment discontinuation (to palliative care, or to 'best available therapy' if on the intervention arm) • Progression to acute myeloid leukaemia • Death	Gompertz model, but assumes comparator OS once curves crossed	Weibull model	For intervention, clinicians suggested exponential, Weibull, or Gompertz could be plausible. Clinical opinion was prioritised over statistical goodness-of-fit. For comparator, clinical opinion determined Weibull or exponential were best fit, and of these Weibull provided the better statistical goodness-of-fit.	Not clear as full details provided in an appendix which is unavailable. It is implied however, that scenarios included alternative parametric models and disabling the adjustment for when the base-case curves cross
2	Model considers a decision tree with three branches, each of which has its own PartSA model	Lognormal model, independent fit. Lognormal model also chosen for the stem cell transplant survival curve (applies to both arms)	Exponential model, independent fit. Lognormal model also chosen for the stem cell transplant survival curve (applies to both arms)	"After consideration of the goodness-of-fit-statistics, visual inspection of the curves and diagnostic plots, and validation with external data, independently fitted models were chosen for the base case: exponential model for vorinostat/ECM and lognormal for mogamulizumab." For the stem cell transplant model: "Following TA577, the log-normal model was used in the base case."	The compared explore a selection of alternative models for OS, but all the scenarios explored applied the same model for both arms (different to the base-case approach taken). No sensitivity analysis was presented exploring uncertainty in survival beyond stem cell transplant
3	3-state (pre-recurrence, post-recurrence, death) state transition model. Chosen to enable use of external data to inform deaths after recurrence.	N/A	N/A	N/A	N/A
4	3-state PartSA	RET-mutant MTC: Unstratified Weibull model (i.e., fitted with a common shape parameter) RET fusion-positive TC: KM + exponential tail, cut-off at 6 months	Same as intervention, but with HR used for one of the comparators in the RET-mutant MTC subgroup. For the non-HR comparison, models were fitted per the base-case analysis to re-created patient-level data	Visual fit, clinical plausibility and "external validation", but no details presented	Range of alternatives explored as scenarios
5	5-state PartSA (PF on treatment, PD 1L, PD 2L, PD 3L, dead)	Weibull model	Same as intervention	Inspection of log-cumulative hazard plot, visual fit, statistical goodness of fit, assessment of changing conditional survival probabilities over time, consideration of treatment effect over time, and clinical plausibility of extrapolations. Limited details presented as only summary available	Generalised gamma model considered as a scenario
6	5-state PartSA (PF on treatment, PD 1L, PD 2L, PD 3L, dead)	'Informed' Weibull model	Same as intervention, but with HR for additional comparator	Inspection of log-cumulative hazard plot, visual fit, statistical goodness of fit, assessment of changing conditional survival probabilities over time, consideration of treatment effect over time, and clinical plausibility of extrapolations. Limited details presented as only summary available	Alternative models explored, including non-informed Weibull model
7	3-state PartSA	KM + exponential tail, no cut-off stated but based on plot appears to be ~2 years. In the TE response, it is explained the cut-point was chosen when 20% of patients were still at risk	KM + exponential tail, no cut-off stated but based on plot appears to be ~2 years. In the TE response, it is explained the cut-point was chosen when 20% of patients were still at risk	Based on "NICE guidance", with TSD 14 referenced	Scenario using a log-logistic tail presented
8	3-state PartSA	KM + lognormal tail, cut-point at 96 weeks	KM + lognormal tail, cut-point at 96 weeks	Piecewise model chosen based on committee preference prior to CDF data collection. Lognormal tail selected based on visual fit, and 96-week cut-off chosen to 'maximise' use of the 'observed trial data'	Scenarios presented to consider alternative curve fits, with and without initial KM period, plus scenarios accounting for a potential treatment waning effect (both a 'full' and 'partial' treatment waning effect)
9	3-state PartSA	KM + log-logistic tail, cut-point at 40 weeks	KM + log-logistic tail, cut-point at 40 weeks	Statistical fit, visual inspection, clinical opinion, and for the comparator arm was considered to be aligned with estimates from external data	Alternative parametric models, including different cut-points to switch from the KM to the extrapolated tail, and exploration of treatment waning effects
10	4-state PartSA (PF split by 'sustained response' and 'lack or loss of response', progressed, dead)	1-knot hazard spline	Hazard ratio applied to intervention model	Best statistical fit, within a PH framework (given use of HR for comparator efficacy)	Variety of alternative curve fits explored, and alternative HRs
11	3-state PartSA	First-order FP-NMA	First-order FP-NMA	Evidence to reject PH assumption, hence uses FP-NMA. First-order FP-NMA chosen based on providing the 'best fit', with deviance information criteria (DIC) cited	Scenario provided using the first order RE FP-NMA model with $p1 = -1.0$ as an alternative for the long-term extrapolation of OS (instead of the first-order FE FP-NMA model with $p1 = -1.0$)
12	3-state PartSA	KM from CheckMate 9LA up to 13 months + 2-knot normal spline model based on data from CheckMate 227	Within-trial comparator: KM from CheckMate 9LA up to 13 months + log-logistic model based on data from CheckMate 227. Outside-trial comparator uses relative risks obtained from a fractional polynomial NMA	Piecewise approach justified based on immature data from CheckMate 9LA. Base-case models selected on basis of statistical fit and long-term plausibility (citing landmark estimates from a range of external sources)	Alternative models based on CheckMate 9LA data only considered as scenarios, as well as changing the cut-point to 18 months, and exploration of a potential treatment waning effect at 3 years after end of treatment
13	5-state PartSA (PF on treatment, PF off treatment, PD on treatment, PD off treatment, dead). Base-case analysis rendered it a 4-state PartSA as no occupancy of PD on treatment	Log-logistic model	Based on ITC	Log-logistic model chosen for intervention as it provided second-best statistical fit but did not lead to crossing of OS and PFS curves (which occurred if using best-fitting Weibull model)	Weibull model considered in scenario analysis. In addition, scenarios were explored for the MAIC, including adjustment for a potential prognostic impact of the mutation of interest, using a non-MAIC HR, using curves fitted to unadjusted external ABC-06 data, and by weighting FIGHT-202 data to match ABC-06

14	State-transition model, with risks of death based on disease stage (not affected by treatment)	N/A	N/A	N/A	N/A
15a	State transition model, using multi-state modelling methodology. No single over-arching OS curve	N/A	N/A	N/A	N/A
15b	3-state PartSA	Between Weibull (under-estimate) and log-logistic (over-estimate) models	Based on ITC	For intervention, clinicians suggested Weibull may under-estimate, and log-logistic may over-estimate, so the truth may fall between these curves. Statistical goodness-of-fit scores were noted to be similar across a number of models, including Weibull and log-logistic	Alternative parametric models presented
15c	As above	Log-logistic model	As above	Based on STAMPEDE data, discussed as part of the appeal hearing	Not clear, but FAD implies sensitivity analyses were limited
16	3-state PartSA	KM until 6.44 months followed by log-logistic model	Exponential model based on estimated mean OS from MAIC	For intervention, complex hazard profile necessitated a more flexible modelling approach. Base-case approach was considered based on statistical fit and plausibility of extrapolations. For comparator, simplified approach taken based on MAIC-estimated mean survival	Full range of alternative extrapolations explored for intervention. For comparator, a threshold analysis was presented showing the mean survival needed for the ICER to hit a given value
17	6-state PartSA (mHSPC stable on treatment, mHSPC stable off treatment, progressed disease 1L, progressed disease 2L, progressed disease 3L, dead)	Weibull model	Weibull model	Trade-off between statistical fit and plausibility of long-term survival estimates	Scenario using log-logistic presented (labelled the "second best fit")
18	3-state PartSA	Lognormal model, independent fit	Lognormal model, independent fit	Statistical fit, visual inspection, and comparison with long-term data from another study (CheckMate 003)	Subgroup analysis presented with alternative curve fits, and one alternative extrapolation of OS explored as a scenario (likely due to limitations of CDF resubmission template)
19a	N/A	N/A	N/A	N/A	N/A
19b	5-state STM (progression free, post-surgery + progression free, progressed disease, post-surgery + progressed disease, dead). Later amended to 3-state STM (removed surgery states)	N/A	N/A	N/A	N/A
20	4-state PartSA (progression-free split by on and off treatment, progressed, dead)	KM + log-logistic tail, cut-point at 2.99 months	KM + exponential tail, cut-point at 2.99 months	No specific rationale provided other than commenting that the selected models were 'considered to be most appropriate'. Extensive discussion was provided concerning the specification of a 'semi-parametric' modelling approach (i.e., KM + extrapolated tail)	Alternative parametric models presented, including fully-parametric as well as KM + parametric tail functions. However, options for each treatment arm considered independently
21	3-state PartSA	Weibull model	Based on ITC	Choice aligned with clinical expert opinion consensus and 'had' statistical and visual fit to the data	Alternative parametric models presented, including some scenarios where adjustment for capping survival benefit from approximately 8 years
22	4-state PartSA (progression-free split by on and off treatment, progressed, dead)	Generalised gamma model + hazard ratio	Generalised gamma model (eribulin) and Gompertz model (capecitabine)	For intervention, selected was model was considered to better reflect the shape of OS curves observed in studies of similar drugs. For comparator, chosen based on statistical goodness of fit	Full range of alternative models considered
23	3-state PartSA	KM until 24 weeks, followed by log-logistic model	KM until 24 weeks, followed by log-logistic model	Statistical goodness of fit, visual fit, aligned with clinical expert opinion, choice of model consistent with ERG's preferred approach pre-CDF data collection period	Considered alternative cut-point for the KM + extrapolated tail approach (40 weeks instead of 28), and explored treatment waning scenarios
24	4-state PartSA (pre-progression, first progression, second progression, and dead)	Weibull model until 5 years after which OS assumed per general population (when PFS and OS curves cross)	Weibull model until 6 years after which OS assumed per general population (when PFS and OS curves cross)	Company justifies base-case approach on the basis of goodness of fit scores, visual fit, and clinical plausibility. Assumption of cure is justified based on published literature and clinical opinion. Company also highlighted that the modelled PFS to OS ratio (1:0.93) was deemed "conservative" versus the ratios implied by other past appraisals	Two alternative OS models chosen (lognormal and generalised gamma)
25	3-state PartSA	Hybrid model using a Weibull distribution fitted to ASPIRE data until 72 months, after which extrapolations were based on a log-logistic model fitted to an external French registry (MyelomaToul) of patients receiving the comparator treatment, adjusted accounting for mortality differences in the observed periods of ASPIRE and MyelomaToul, with a hazard ratio applied to obtain estimates for the intervention group	Same as intervention, but without applying the hazard ratio	Company considered standard Weibull model extrapolations from the ASPIRE trial data may underestimate long-term survival, and so sought external evidence to inform long-term survival extrapolations	Alternative extrapolation assumptions are presented in scenario analyses including the use of jointly fitted Weibull and lognormal curves without external data, alternative cut-off points for starting the use of external data, alternative cut-off points for piecewise exponential model (60 months and 48 months), and use of a constant HR for ASPIRE vs. MyelomaToul
26a	3-state semi-Markov (but not a PartSA) model. Model made use of tunnel states to track timing of progression	N/A	N/A	N/A	N/A
26b	3-state PartSA, but implemented as a cost-minimisation model	Exponential model	Same as intervention - assumed no difference in survival	Exponential chosen on basis of statistical goodness of fit	None presented
27	3-state PartSA	"1-knot odds" spline-based model	Weibull model	Comparator curve unchanged from prior appraisal (as this is a CDF review). For the intervention, this model was chosen on the basis of statistical goodness-of-fit, clinical plausibility of projections, and for consistency with the model selected in the assessment of treatment experienced mMCC (TA517)	Range of alternative models explored in sensitivity analysis
28a	State transition model. No single over-arching OS curve	N/A	N/A	N/A	N/A

28b	3-state PartSA	Weibull model fitted with a covariate for treatment arm from the MONALEESA-3 trial, but selecting only the intervention group	Hazard ratio applied to intervention model	Chosen based on the Weibull model having the best fit statistical fit, good visual fit to the KM data, yielded projections of OS that were consistent with clinical experts' expectations and met PH assumptions	Range of alternative models explored in sensitivity analysis
29a	3-state semi-Markov (but not a PartSA) model	N/A	N/A	N/A	Scenarios presented varying the source for long-term survival (registry versus SACT data), the cut-off used to switch to the external data (10 years versus 5 years), and the extrapolation chosen (generalised gamma versus log-logistic or lognormal). Also, a sensitivity analysis was presented with treatments used after recurrence weighted based on SACT data collected over the course of the CDF period (which impacts outcomes in the STM but not the PartSA model)
29b	3-state PartSA	Generalised gamma model	Generalised gamma model	From the CS: "... best fitting curve statistically and visually, and gave plausible outcomes compared with literature sources... This was also consistent with the chosen curve used to extrapolate OS from CA184-029 in the previous submission"	Scenarios presented varying the source for long-term survival (registry versus SACT data), the cut-off used to switch to the external data (10 years versus 5 years), and the extrapolation chosen (generalised gamma versus log-logistic or lognormal)
30	3-state PartSA	Log-logistic model	Log-logistic model	From the CS: "Goodness of fit statistics and visual inspection suggests that the loglogistic is the best fitting extrapolation for the updated clinical data. Additionally, this is consistent with the ERG preferred method of extrapolation in the original submission."	Scenarios presented using an alternative parametric model and using a piecewise extrapolation. Parametric model inputs were also included within a one-way sensitivity analysis
31	3-state PartSA	Log-logistic model	Weibull model	Chosen based on plausibility of extrapolations, based on comparison to data from the CALGB 100104 study	Use of generalised gamma model for OS for the intervention arm is reported as a scenario analysis, but is not presented in the tabulated scenario results
32	3-state PartSA	Lognormal mixture-cure model, with SMR of 1.09 applied to background mortality reflecting excess mortality experience	Meta-analysis of standard lognormal models fitted to OS data from four different published studies, with SMR of 1.09 applied to background mortality reflecting excess mortality experience	Standard models rejected for intervention based on poor fit and unrealistic projections. Lognormal selected on basis of statistical goodness-of-fit as projections were similar across models. Meta-analysis undertaken based on studies identified via a systematic literature review	Range of alternative models explored in sensitivity analysis, with focus placed on the most optimistic and most pessimistic models
33	3-state PartSA	HR applied to comparator curve such that the overall modelled difference in PFS was half of the modelled difference in OS (assessed based on extrapolated means)	Log-logistic model	Approach for intervention justified on basis of limited data from PRIMA trial and precedence set in past appraisals (with a note that outcomes should be just as good, if not better in a maintenance population). For the comparator, the log-logistic curve was chosen as it was the only curve fitted that did not 'substantially under- or over-estimate' OS when compared to real-world data presented by the company	Considered lognormal instead of log-logistic curve for comparator OS. Varied the assumed ratio of PFS:OS from as little as 1:1 to as much as 1:3 (base-case 1:2). Also explored the possibility of basing estimate on restricted mean PFS rather than modelled mean PFS.
34	3-state PartSA	Jointly-fitted lognormal model, with capping for background mortality based on hazards	Jointly-fitted lognormal model, with capping for background mortality based on hazards	Based on visual fit, statistical goodness of fit, and clinical plausibility of longer-term projections	Full range of alternative models considered, including independent models
35	4-state PartSA (splits CNS and non-CNS progression, progression-free, dead)	Exponential model	Based on ITC	Exponential chosen on basis of statistical goodness of fit, consistency with a prior appraisal, and based on clinical expert opinion	Other standard models explored as sensitivity analyses. Also, scenarios were considered which assume same mortality rate across arms after 7-, 10- and 20-years.
36	3-state PartSA	Log-logistic model	Hazard ratio applied to intervention model (for within-trial comparator and ITC comparator)	Based on visual fit, statistical goodness of fit, and clinical plausibility of longer-term projections. Company also noted that TSD 14 guidance suggests that the same type of parametric model should be used across all arms, and so for statistical goodness-of-fit, elected to choose the model that provided the best general fit across three different sources	Use of Weibull model instead of log-logistic, as well as varying HRs based on confidence interval limits. The company provided additional scenarios within its submitted model that were not described in the CS (but were picked up by the ERG)

Appendix Table 14: STA review data extraction (Part 8 of 10), EAG/ERG critique

#	Was OS identified as being a 'key issue' by the EAG/ERG? If so, what did the key issue encompass?	If OS was not considered a 'key issue', was anything similar considered a key issue?		EAG/ERG's preferred base case
1	Yes ERG questioned implementation of OS in the company's original model because it is independent of the other events, most notably discontinuation and response. This means, for example, that responders die at the same rate as non-responders in the model. Furthermore, the ERG criticised the fact that the OS comparison effectively is a naïve comparison, since no adjustments were made to either arm in light of the data coming from two different studies. The ERG also commented on the lack of mature data available from the JAKARTA-2 study.			Generalised gamma model for the intervention, and set survival for the comparator to be equal to the intervention
2	Yes ERG preferred using an exponential curve for both arms (i.e., not the lognormal chosen by the company for the intervention). This was because of a lack of justification the ERG felt was provided by the company to select different parametric models for each treatment arm, though this was also linked to the ERG's preferred crossover adjustment method (where the ERG explained the same parametric model should be chosen for both arms).			Exponential models for both arms
3	No	Yes	ERG mostly focused on the estimation of DFS, which is linked closely to OS	No change made with respect to OS modelling
4	Yes ERG considered estimation of OS to be generally associated with 'many uncertainties', driven by limitations in the evidence base (mostly, reliance on the MAIC outputs). Company provided some additional data at response to clarification questions, and ERG considered base-case model for OS for intervention in the RET-mutant MTC to be optimistic, and too focused on the PH assumption.			RET-mutant MTC: Stratified Weibull model (i.e., fitted separately to each arm) RET fusion-positive TC: No change
5	Yes ERG clinical advice suggested company's preferred modelling led to an underestimate of OS at 10 years, and possibly also at 15 years			ERG preferred a jointly-fitted generalised gamma model based on alternative clinical expert opinion
6	No ERG did not consider OS to be a key issue for this population, with its changes instead focused on treatment durations, costs, and utilities			ERG did not make any specific changes to the company's modelling of OS
7	Yes ERG considered it more appropriate to fit a model to the full data, not just the tail			Exponential model for both arms, without the KM used
8	Yes ERG preferred inclusion of a treatment waning effect, so factored this into its preferred base-case analysis based on smoothed hazard plots which appeared to converge towards the end of follow up. Also, ERG commented that the choice of model had not been subjected to 'external validation', described as either expert opinion or external data), and so the extrapolations were deemed uncertain			Waning effect added in, such that from year 5 hazards for the intervention arm were assumed to be identical to those for the comparator arm
9	Yes ERG considered that a range of alternative models could be considered plausible, yet had marked impacts on cost-effectiveness results. ERG included treatment waning within its preferred base-case analysis			Waning effect added in, such that between years 5 and 7 hazards gradually approached those of the comparator arm
10	Yes ERG raised concern with the HR applied to obtain comparator OS, stating it lacked a robust source. The ERG also questioned the plausibility of the lifetime treatment effect assumed within the model (as the HR is assumed to apply indefinitely)			None provided
11	No	Yes	ERG was mostly concerned with the population informing the FP-NMA, which impacted all outcomes (including OS) but was not a criticism specific to OS alone	ERG did not make any specific changes to the company's modelling of OS
12	Yes ERG criticised survival estimates for the within-trial comparator, the outside-trial comparators (critique of the fractional polynomial NMA), estimates for subgroups, and duration of treatment effect for the intervention			ERG presented three preferred base-case analyses based on three different subpopulations. • Sq-PD-L1 <50%: Assumed treatment effect until 5 years • NSq-PD-L1 <50%: Changed the FPNMA to include CheckMate 227 + assumed treatment effect until 5 years • Mixed histology PD-L1 ≥50%: Changed the FPNMA to include CheckMate 227 + assumed treatment effect until 5 years
13	Yes ERG criticised the company's choice of model for OS based on the clinical opinion it obtained which suggested that 5% of patients would be expected to be alive at 5 years. The ERG preferred the model which produced estimates closer to this value - the generalised gamma			Generalised gamma model for the intervention. No change made to comparator estimate
14	No -	No	ERG focused on other, non-survival issues that influenced cost-effectiveness estimates, given lack of survival benefit being modelled. The ERG did, however, note that estimates from this study are inherently uncertain given that an exponential curve is assumed (i.e., a curve estimated based on the median only).	ERG made one small change to modelling survival, by assuming the risk of death in Stage IA would be the same as the general population, rather than deriving a transition probability based on the median estimate reported by Agar <i>et al.</i>
15a	Yes ERG criticised the company's overarching modelling approach, which by extension meant that the ERG was not convinced by the company's projections of OS benefit.			ERG did not make any specific changes to the company's modelling of OS
15b	Yes ERG's response to the company's addendum states: "Since the population is a group with high-risk metastatic disease, the ERG's clinical advisor believed that the log-logistic curves offered overly optimistic projections of rPFS and OS, and that the Weibull extrapolations were more plausible. Further, the ERG's clinical expert was of the opinion that even the most pessimistic projections offered by the Gompertz distributions were not completely implausible for a high-risk cohort with metastatic disease"			Not stated
15c	No ERG did not comment on this, as the committee aligned with company's preferred approach (outcome of the appeal hearing)			Not stated
16	Yes ERG preferred an earlier cut-point (2.99 months instead of 6.44 months). ERG also noted that long-term projections are impossible to predict and that "all curves are clinically plausible". ERG also criticised the approach taken to account for background mortality			Cut-point changed to 2.99 months
17	Yes ERG considered OS data from ARCHES and ENZAMET to be immature (median OS not reached for intervention), which led to a high degree of uncertainty around the lifetime extrapolations which inform the model			ERG did not make any specific changes to the company's modelling of OS for the within-trial comparator, but for the intervention instead preferred to use the output of the ITC instead of the model fitted to the trial data
18	Yes ERG considered company's approach to be reasonable, but flagged it was misaligned with the committee's preference pre-CDF, and so it was therefore deemed a key issue			Per company's base-case
19a	N/A ERG focused its critique on the 3-state STM	N/A	N/A	N/A

19b	Yes	ERG changed its preferred estimates for TTP and PFS			Alternative options for TTP and PFS (but OS not modelled directly)
20	Yes	ERG did not agree with choice of base-case OS models, and did not agree with the approach taken to account for background mortality (yet ultimately did not adjust this within its preferred analysis)			KM + generalised gamma tail for both arms, with a later cut-point of 5.75 months
21	No	-	Yes	ERG was unconvinced by the company's base-case choice of OS model, but ultimately considered it an appropriate base-case option. ERG's main areas of concern related to the assumption of a duration of treatment effect and the ITC, both of which are linked closely to OS	Per company's base-case + scenarios around duration of treatment effect
22	Yes	ERG considered modelling of OS for both the intervention and comparators to not be robust. In fact, OS modelling was the only key issue raised by the ERG			None provided
23	Yes	ERG unclear if choice of log-logistic tail is the most suitable model, but ultimately accepted it in their base-case analysis. However, ERG preferred a 3-year cap on treatment effect (company preferred 5 years)			Per company's base-case, but with treatment waning applied from 3 years instead of 5 years
24	Yes	ERG expresses concern that the company's combination of approaches to estimate OS and PFS (+ PFS2) leads to estimates of OS which "considerably overestimate" the relative effect of the intervention			ERG's preferred base case removes the assumption of 'cure' from 5 (or 6) years, relying on the Weibull model extrapolation but without the adjustment according to PFS (or PFS2)
25	Yes	ERG raised issues with the clinical plausibility of the company's preferred extrapolations of OS, suggesting that the overall approach taken led to over-inflated estimates for the intervention. The ERG also noted that had the company simply selected the model fitted to the ASPIRE data that yielded results in keeping with the MyelomaToul registry data, then the exponential model might have been chosen which produced estimates that were deemed similar to the base-case hybrid approach			ERG preferred use of exponential models fitted to the ASPIRE data only to inform its base-case analysis
26a	Yes	ERG raised concerns with various aspects of the model which influenced the estimation of OS, owing to the semi-Markov modelling approach taken. Notably, the ERG explained that the company's base-case approach modelled OS projections for the acalabrutinib arm that were "very similar to that of the general population without CLL"			ERG changed various aspects of the company's base-case analysis which had an impact on how OS was modelled, but owing to the model structure none of these were changes specifically made to adjust OS
26b	Yes	ERG requested the company change the CMA to a CUA by applying relative effects obtained from a MAIC to capture differences in OS. However, in the ERG's preferred analysis, there is no modelled difference in OS.			ERG did not make any specific changes to the company's modelling of OS
27	Yes	ERG preferred the more conservative of the three 1-knot spline based models owing to the uncertainty inherent within estimating long-term treatment effects on the basis of immature survival data. For the comparator, the ERG considered it more suitable to account for potential imbalances through a propensity score analysis it requested at clarification stage, as opposed to a naive comparison which was previously the base-case analysis.			"1-knot hazard" spline-based model for intervention. For comparator, ERG preferred a propensity score weighting analysis provided by the company at clarification stage
28a	Yes	ERG notes that due to the availability of more mature OS data (given this is a CDF review), the company could have opted to change to using a PartSA model.			ERG did not make any specific changes to the company's modelling of OS, but explained it would have preferred a PartSA model
28b	Yes	ERG explained it had some reservations regarding use of the Weibull model, owing mostly to the plausibility of its longer-term projections			ERG preferred use of a Gompertz model as it was considered a more conservative choice
29a	Yes	ERG highlighted three key issues relating to survival: (1) the impact of censoring within the context of the ITC (in brief, treatment with ipilimumab was permitted beyond 1 year in CA184-029 but was stopped at 1 year in CheckMate 238, and so scenarios were considered where patients were either censored or not censored if treatment continued beyond 1 year in CA184-029), (2) the choice of model structure (where the ERG preferred the PartSA), and (3) that the hazard of death for the comparator arm may be over-estimated given that data for the control arm "[do] not reflect the current survival outcomes for routine surveillance patients due to advances in the subsequent treatment pathway"			ERG preferred PartSA model, so no base-case stated for the STM
29b					A range of scenarios presented which the ERG considered the 'true' ICER to fall within. The scenarios explored (with respect to OS modelling) were: (1) assuming no difference in hazards after 3 years, (2) assuming no difference in hazards after 2 years, (3) assuming no difference in hazards after 3 years + 1-year censoring of ipilimumab OS patients, and (4) assuming no difference in hazards after 2 years + 1-year censoring of ipilimumab OS patients
30	Yes	ERG highlighted choice of model for OS as a key issue, highlighting that an alternative model which appeared clinically plausible was not considered by the company			ERG preferred generalised gamma model for base-case analysis
31	Yes	ERG preferred use of a jointly-fitted log-logistic model for OS. Outside of the choice of model, the ERG criticised the choice of model structure as it did not allow for a detailed exploration of alternative settings relating to the use of subsequent therapies			Joint-fitted log-logistic model
32	Yes	ERG raised concerns with the extrapolation of OS, adjustment of long-term survival based on an SMR of 1.09, and the choice of data source to inform OS for the comparator group			2-knot normal spline model for the within-trial period, followed by general population estimates for the remainder of the modelled time horizon, adjusted based on an alternative hazard ratio (HR) approach, with a value of either 2.36 or 4.37
33	Yes	ERG did not consider there to be enough robust evidence to justify the use of PFS as a surrogate for OS, and this approach to modelling OS was identified as the key model driver. The ERG presented three alternative scenarios to explore the possible impact on results, showing that the ICER ranged from £38,252 to £950,200 depending on the approach taken			A range of scenarios presented, which the ERG implied the 'true' ICER was likely to fall within. These were: (1) assuming a PFS:OS ratio of 1:0.66, (2) assuming a PFS:OS ratio of 1:1, and (3) assuming a hazard ratio (HR) of 0.70 for OS (based on the reported HR from the PRIMA study) - this is equivalent to assuming an OS:PFS ratio of 1:1.13
34	Yes	ERG preferred specification of independent models, rather than a jointly-fitted model			Independent models, but possibly either lognormal, log-logistic, or generalised gamma (range of scenarios presented)
35	Yes	ERG noted that the main driver of uncertainty in the model was the validity of OS estimates. The ERG was mostly concerned with the ITC approach, rather than the choice of model for the intervention arm			None provided
36	Yes	ERG highlighted issues with the ITCs conducted, as well as the choice of OS model in the base-case analysis			ERG preferred use of a piecewise model for OS, using the KM until 2.8 months followed by an exponential model for both arms in the BEACON CRC study (approach for ITC comparator unchanged)

Appendix Table 15: STA review data extraction (Part 9 of 10), technical engagement

#	Changes made by company	New external evidence at TE?	Changes made by ERG +/- critique
1	Survival modelling was changed such that for the intervention, survival after discontinuation was modelled separately for responders versus non-responders (i.e., no longer one overall model for survival per the original approach taken)	No	ERG highlighted that the company's edits introduce a differential approach to estimating survival for each group
2	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
3	N/A	No	N/A
4	RET-mutant MTC: Stratified Gamma model - based on 'additional clinical expert feedback' sought by the company RET fusion-positive TC: No change	Yes	Aligned with company view, in light of no evidence to contradict view (but ERG noted there were no other data to inform choice of model, so it had no reason to disagree with the company's clinical experts)
5	Company accepted ERG edit	No	No changes made to preferred survival modelling approach
6	Final TITAN analysis provided in response to technical engagement + NMA was updated, so analyses were updated but no change in approach	No	No changes made to preferred survival modelling approach
7	Company changed cut-point to be earlier (24 weeks, instead of ~2 years) for consistency with the pre-CDF committee preference. 24 weeks was when 48% of patients were still at risk	No	No changes made to preferred survival modelling approach
8	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
9	Company changed population on basis of regulatory decision, so results changed, but the overall approach was per the original population	Yes	No changes made to preferred survival modelling approach
10	Company provided latest data from D2201 (main trial, an addition year of follow-up) to compare to the registry data cited by Reiter et al., (2017), and produced updated HRs, including propensity score matching. Company also explored different durations of treatment effect (3, 5, and 10 years). Company did not change base-case analysis, but provided additional analyses as scenarios	Yes	No changes made to preferred survival modelling approach
11	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
12	Company revised its FPNMA to include CheckMate 227 data. Company also confirmed selected models for each subgroup, which differed to the approach taken for the 'all-comers' population: • Sq-PD-L1 <50%: intervention: log-logistic; comparator: log-logistic • NSq-PD-L1 <50%: intervention: log-logistic; comparator: spline odds 2 knot • Mixed histology PD-L1 ≥50%: intervention: generalised gamma; comparator: spline odds 1 knot	No	No changes made to preferred survival modelling approach
13	Additional follow-up data from FIGHT-202 provided in response to technical engagement. Company proceeded to re-fit the survival models, but maintained its preference for a log-logistic model in its base-case analysis	Yes	ERG maintained preference for generalised gamma model, but fitted to the updated FIGHT-202 data
14	The company accepted the ERG's edit for Stage IA	No	N/A
15a	No technical engagement step	N/A	N/A
15b	No technical engagement step	N/A	N/A
15c	No technical engagement step	N/A	N/A
16	Company updated model with updated data from pivotal CheckMate 142 trial, and changed its base-case to use a log-normal model after 6.44 months instead of a log-logistic per the previous data cut	Yes	ERG accepted 6.44 month cut point based on later data provided, but preferred log-logistic tail (i.e., the tail selected by the company for the previous data cut) instead of the lognormal tail
17	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
18	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
19a	N/A	N/A	N/A
19b	Company accepted ERG changes to modelling TTP and PFS	No	No changes made to preferred survival modelling approach
20	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
21	No changes made to preferred survival modelling approach. Company provided various sensitivity analyses, but none were concerned with the approach to estimating OS specifically	No	No changes made to preferred survival modelling approach
22	Company edited its MAIC analysis for the comparator outcomes, but otherwise did not change how OS was modelled for the intervention. The company did, however, update the input data from the study to make use of additional follow-up data	Yes	No changes made to preferred survival modelling approach
23	No changes made to preferred survival modelling approach. However, company provided additional commentary and sensitivity analyses to justify its preferred approach	No	No changes made to preferred survival modelling approach
24	No changes made to preferred survival modelling approach. Company provided data from another study (SOLO1) to support 'cure' at 5 (or 6) years	Yes	ERG revised its preferred assumption to permit mixture-cure modelling for PFS, but cap this with OS (i.e., the opposite to the company's base-case in which OS is capped by PFS). Revised OS model comprised a piecewise approach with KM applied until 30 months, then exponential model (difference in curves by arm reflected via HR)
25	No changes made to preferred survival modelling approach. However, company provided commentary and sensitivity analyses as further justification	No	No changes made to preferred survival modelling approach
26a	No technical engagement step	N/A	N/A
26b	No technical engagement step	N/A	N/A
27	No changes made to preferred survival modelling approach. However, company provided analysis of Part A of the JAVELIN Merkel 200 study (treatment-experienced mMCC) to demonstrate differences in long-term projections, in defence of its preferred approach.	Yes	No changes made to preferred survival modelling approach
28a	The company switched to a PartSA model at technical engagement	N/A	N/A
28b	Described previously given this model was provided at technical engagement	N/A	N/A
29a	The company provided a scenario where post-recurrence survival was informed based on data from CheckMate 238, instead of the literature as an additional scenario. However, no new external data or information was provided	No	No changes made to preferred modelling approach (PartSA)
29b	The company provided a scenario adjusting for post-recurrence treatments given to patients in the CA184-029 study based on what "would have been observed if the same subsequent therapies received by patients in CheckMate 238 were available", and assuming no difference in hazards beyond the median duration of recurrence-free survival (though the value itself is redacted, but was later revealed to 4.36 years)	No	ERG expanded its range of scenarios to include a scenario where subsequent therapy for routine surveillance was assumed to be nivolumab for all patients, which expanded the range of ICERs versus previous ERG range
30	No changes made to preferred survival modelling approach	Yes	Company referred to external data for the plausibility of landmark survival probabilities (e.g., 5-year OS)
31	Company accepted specification of a joint model for OS, but opted to naively pool data from the Myeloma XI and CALGB 100104 studies to inform its revised base-case analysis. Company chose a jointly-fitted Weibull model for OS for both arms, based on the pooled data set	Yes	ERG made no changes, but criticised the decision to naively pool data from both studies
32	Company accepted the ERG's edits, except for the adjustment to survival in the long term (company preferred its original 1.09 SMR adjustment)	No	N/A
33	No changes made to preferred survival modelling approach	No	N/A
34	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach
35	No changes made to preferred survival modelling approach	No	No changes made to preferred survival modelling approach, but scenarios provided based on alternative ITCs
36	Company updated model with additional follow-up data, but no changes made to preferred extrapolation approach for OS	Yes	Company provided analyses using a later data cut from the BEACON CRC trial

Appendix Table 16: STA review data extraction (Part 10 of 10), committee discussion

#	Is survival modelling discussed in the ACD? If yes, what is covered?		Is survival discussed by the company in its ACD response? If yes, what is covered?		New external evidence in ACD response?	Is survival discussed by the ERG in its critique of the ACD response? If yes, what is covered?		Is survival modelling discussed in the FAD? If yes, what is covered?	
1	Yes	The committee considered the company's model to be too complex, including how survival was estimated. However, it wasn't the approach to survival specifically that the committee seemed to have the biggest concern with, it was a broader concern about the modelling in general being too complex.	Yes	Company provides a new scenario called a 'simplified pooled analysis', in which survival is estimated from time 0 for both the intervention and comparator	No	Yes	Nothing specific about extrapolation that has not been previously stated in an earlier document, but general issues around survival are discussed again. Notably, the ERG does not comment on the company's new scenario	Yes	Same content as ACD
2	Yes	The committee preferred the ERG's exponential approach for survival due to a lack of evidence in support of selecting different models per the company's base-case analysis	Yes	Company provided a new scenario comparing the MAJORIC trial data to Hospital Episode Statistics data for the comparator. This involved a complete new set of survival analyses, including weighting based on a MAIC. The company selected an exponential model for the intervention and the comparator. The overall approach taken to fitting the models was otherwise the same as the original submission.	Yes	Yes	The ERG was concerned with the choice of an exponential model for the comparator, given that other models (including the generalised gamma and lognormal) provided a better statistical fit. The ERG selected the lognormal model for the comparator arm in its preferred analysis, but left the exponential model for the intervention arm unchanged.	Yes	The committee noted that the comparison to the Hospital Episode Statistics data provided in response to the ACD addressed some concerns of the original within-trial analysis, but it was also subject to lots of other uncertainties. It did, however, side with the company in choosing the same model fitted to both arms.
3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee noted that OS data were unavailable, and so decision making relies on use of disease-free survival (which the committee considered to be reasonable)
4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee discussed how uncertain OS estimates were, which ultimately informed the decision for the intervention to be made available via the CDF
5	Yes	The committee noted that adjustments related to the 'one novel therapy' rule introduced uncertainty, and may not have been necessary. Committee considered generalised gamma model preferred by ERG to be reasonable, but was still uncertain particularly in the long term	Yes	Adjustment for crossover / 'one novel therapy' rule included, but no clear stated preferences	No	Yes	Adjustments discussed briefly, but no clear preference stated	Yes	The committee considered the adjustments made for crossover/ 'one novel therapy' to be suitable for decision making, yet still potentially unnecessary and still uncertain. Other commentary as per ACD, with generalised gamma model deemed suitable for decision making but still uncertain
6	Yes	The committee noted that adjustments related to the 'one novel therapy' rule introduced uncertainty, as well as more broadly adjustment for crossover. Committee considered choice of a Weibull model to not be flexible enough, and that a more flexible model was warranted, but did not raise any issues with the 'informed fits' approach taken	Yes	Company decided not to provide further flexible models as it felt its clinical expert input did not warrant such models	No	Yes	ERG noted that company chose not to provide more flexible models for OS	Yes	The committee considered the adjustments made for crossover/ 'one novel therapy' to be suitable for decision making, yet still potentially unnecessary and still uncertain. Committee would have preferred to see more flexible models, but considered the 'informed' Weibull model to be suitable for decision making
7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee considered estimates of survival preferred by both the company and the ERG to be similar, and that both were clinically plausible. Committee also commented on reasons for differences in survival seen for the trial versus SACT populations, accepting use of the trial data to populate the economic model
8	Yes	The committee considered extrapolations of OS to be uncertain, particularly for the comparator arm. Treatment waning at 5 years was deemed plausible	Yes	Company presents a revised base-case analysis wherein a treatment stopping rule was introduced at 5 years, but also that treatment waning was disabled. Company also provides additional analyses to consider mean survival and survival extension within the context of the end-of-life criteria	No	Yes	ERG maintained its preference for a 5-year waning effect, adding that the stopping rule at 5 years should not be used as an argument for the waning effect not being included.	Yes	The committee considered the company's base case and a range of other scenarios to provide plausible estimates of OS. Committee preferred no stopping rule, but a 5-year waning effect
9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee acknowledged uncertainty in the estimates of OS, but ultimately determined that both the company's and ERG's preferred analyses led to acceptable ICERs
10	Yes	The committee considered the OS HR to be uncertain but it preferred the propensity-score matched HR provided by the company at technical engagement, and that a duration of treatment effect of 3 years was most suitable to consider in its preferred base-case analysis	Yes	Company commented that the combination of the committee's preferred assumptions (propensity score matched HR and treatment effect until 3 years) led to an 'overly pessimistic overall assessment of the evidence'. Company preferred treatment benefit until 5 years	No	Yes	ERG commented on uncertainty with duration of treatment effect, noting limitations of a PartSA model to incorporate such an effect	Yes	The committee maintained its preferences per the first committee meeting, with regards to the HR used and the 3-year duration of treatment effect
11	Yes	The committee discussed the issue of the correct population, and hence its expected impact on modelling overall survival. However, no other commentary presented with respect to the methodology used	No	N/A	No	No	N/A	Yes	The committee expressed the same views as per the ACD. Additional commentary was provided commenting on the differences in outcomes by population, but no commentary on the methods used
12	Yes	The committee discussed the uncertainty inherent within the modelling undertaken, in	Yes	Company maintained its preference for using CheckMate 9LA data within the comparator	Yes	Yes	ERG aligned with committee's preferences in terms of duration of treatment effect and use of	Yes	The committee discussed the uncertainty inherent within the ITCs conducted, noting at least some of these were not considered robust for

		particular with respect to either considering the 'all comers' population, or separate analyses by subgroup. It concluded that the subgroups should be considered separately. For one comparator in particular, it was considered more appropriate to only use data from CheckMate 227 (and not combine these data with data from CheckMate 9LA)		survival estimation. Presented outcomes from a <i>de novo</i> SLR and NMA of first-line therapies to comment further on outcome differences by study in this setting			CheckMate 9LA data for the comparator group. General commentary provided for other aspects of company's response, but no other changes made to its preferred analysis		decision making. Ultimately however, commentary was focused mostly on the broader topics of population and comparators, and that in all groups the ICERs were not deemed cost effective
13	Yes	The committee acknowledge the range of scenarios presented, highlighting that clinical opinion of 5% alive at 5 years may be based on historical estimates, and that an estimate of 10% may better reflect the likely survival of patients treated with the intervention. The committee stated it would have preferred for independent models to be fitted to the comparator arm data, instead of using the MAIC-HRs, and wanted to see external data to 'validate' comparator survival estimates	Yes	Company compared survival estimates for the control arm to data from another study (ClarIDHy) which was identified as another potential source for validation. Company provided additional curve fits to the ABC-06 data to inform comparator outcomes (instead of using HRs obtained via the MAIC)	Yes	Yes	ERG aligned with committee's preference for independent models (in favour of using the MAIC HRs), and so updated its preferred base-case analysis as a range of options using different variations of weighting/non-weighting	Yes	The committee preferred the independent model fits, and preferred use of a log-logistic model for both arms
14	No	N/A	No	N/A	No	No	N/A	No	N/A
15a	Yes	The ACD states: "The company's economic model does not ... give plausible estimates for modelled survival". It could therefore not determine a "plausible cost-effectiveness estimate", and requested the company to make a number of changes to its model, including "reflecting plausible post-progression and overall survival across modelled treatment arms"	Yes	Company provided additional sensitivity analyses varying survival beyond progression, including setting a hazard ratio for post-progression survival to be at unity.	No	Yes	ERG still did not accept the company's approach	N/A	N/A
15b	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee sided with the ERG's preference towards the Weibull model as it was considered more clinically plausible than the log-logistic model. This led to a negative FAD, which the company then appealed
15c	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee accepted the use of the log-logistic model for OS in the full population, but commented that it would have preferred the ability to consider further sensitivity analyses
16	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee ultimately preferred the log-logistic extrapolation to inform the model, and commented that it was appropriate to incorporate background mortality into estimates (but did not comment on the methodology used)
17	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee preferred the ERG's approach of using the hazard ratio from the NMA to obtain OS for the intervention, but otherwise accepted the base-case approach to estimating OS (including the naive pooling of data to inform the within-trial comparator). The committee did, however, request a scenario where hazards were assumed equal from 8.5 years for all treatments being compared
18	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee considered both the base-case and scenario analyses presented by the company to be plausible
19a	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	The committee preferred the semi-Markov model over the PartSA based on its concern that the OS data from the KeyNote-177 study were likely to be biased
19b	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee highlighted that the company's and ERG's preferred extrapolations for progression-free survival and time-to-progression were acceptable, and that the assumption of equal post-progression survival for all treatments was likely conservative, but also unlikely to reflect clinical practice. Nevertheless, the committee noted the intervention was cost-effective under its preferred assumptions
20	Yes	The committee noted uncertainty in OS extrapolation, but commented that a more recent data cut from the ATTRACTION-3 study was available but was not provided to the committee ahead of the committee meeting	Yes	Company provided updated analyses with models fitted to a later data cut from ATTRACTION-3. Base choices of models were KM + log-logistic for intervention and KM + Weibull for comparator (with 5.75 months cut-off)	No	Yes	ERG considered change to OS for intervention to be acceptable, but did not agree with choice for comparator. However, model was not updated so ERG could not explore this further, and so its preferred base-case analysis was unchanged	Yes	The committee highlighted the uncertainty within the model estimates (from both the company and ERG), but ultimately considered a range of scenarios to inform its decision making which, paired with a revised commercial arrangement, led to a positive recommendation
21	N/A	N/A	N/A	N/A	N/A	N/A	N/A	No	N/A
22	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee broadly discussed the limitations highlighted by the ERG, including limited OS data from the pivotal study, and a reliance on naive comparisons
23	Yes	Crossover adjustment was considered uncertain, treatment effect was preferred to stop from 3 years (aligned with ERG), and three different extrapolations were considered plausible and suitable for decision making (log-logistic, lognormal, and generalised gamma)	Yes	Company re-iterates its position with respect to the use of the crossover-adjusted data and 5-year treatment effect duration	No	Yes	Similar to the company, the ERG comments on the crossover adjustment and duration of treatment effect, but its preferred base-case analysis remains unchanged	Yes	The committee expressed the same views as per the ACD. An additional scenario provided by the company in response to the ACD was considered inappropriate

24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee considered the OS data to be uncertain but promising, and that it was unclear if a cure assumption was appropriate. Ultimately, the committee felt the mixture-cure model was not justified and could over-estimate benefits, and considered the ERG's analyses to demonstrate the uncertainty in results
25	Yes	The committee preferred the ERG's approach to extrapolating OS (i.e., an exponential model based on ASPIRE data only). Of note, the ACD states: "The committee preferred the exponential model for estimating overall survival for both treatment arms because it used data entirely from the ASPIRE trial."	Yes	Company provided additional commentary to justify its preferred approach, citing previous NICE guidance in a similar setting to highlight inconsistency in landmark survival estimates for the comparator group. Company provided a sensitivity analysis where estimates of survival were 'uplifted' to align with estimates from a past appraisal.	Yes	Yes	ERG noted that the past appraisal emphasised the uncertainty in OS estimates, and considered it inappropriate to favour an uncertain past estimate when mature data from ASPIRE are available. ERG's preferred approach remains unchanged	Yes	The committee expressed the same views as per the ACD, acknowledging that no new evidence was provided by the company in response to the ACD to change its view as to which extrapolation was preferred
26a	Yes	The committee "concluded that there was considerable uncertainty in the overall survival estimates for acalabrutinib because of the extrapolation using data from trials for other treatments and the immature data from ELEVATE-TN." In addition, the ACD states: "The data from ELEVATE-TN showed a trend towards improved overall survival for acalabrutinib compared to chlorambucil plus obinutuzumab. But the data were immature, with a median follow-up at 28 months, and the difference between the groups was not statistically significant"	Yes	Company argued that the committee's preference for post-progression survival was not supported by clinical rationale and was clinically implausible	No	Yes	The ERG re-emphasised the limited data/evidence in support of the OS advantages reflected in the company's preferred analysis	Yes	The committee did not discuss the overall approach at length, but general commentary is provided concerning the uncertainty associated with OS estimation
26b	No	N/A	No	N/A	No	No	N/A	No	N/A
27	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee agreed with the ERG that any of the three 1-knot spline-based models would be reasonable to inform decision making, and that the choice of any one model over the others would be unlikely to make a difference to decision making.
28a	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
28b	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee noted uncertainty in OS extrapolation, and explained that the most appropriate extrapolation is likely to lie between the Weibull (company preferred) and Gompertz (ERG preferred) extrapolations
29a	Yes	The committee preferred the PartSA, so the STM is not discussed further	N/A	N/A	N/A	N/A	N/A	N/A	N/A
29b	Yes	The committee considered the scenario presented by the company at technical engagement was likely too optimistic, and preferred the ERG's "more conservative approach", reflecting a range of ICERs with assumed no difference in hazards from a given time point (2 or 3 years)	Yes	Company did not amend its preferred base-case analysis, but provided a range of additional scenario analyses including setting hazards to be equal at yearly intervals from 3 to 10 years, as well as including or excluding censoring within the ITC. Company suggested earliest time point for assuming no difference in hazards should be 4 years, based on additional analyses provided (including smoothed hazard plots and recurrence-free survival data)	No	Yes	ERG noted maximum time point for assuming no difference in hazards should be 5 years (but reasons were redacted), and that the "absolute minimum" should be 3 years. This means that the ERG's range of analyses meant that hazards would be assumed equal from a certain time point somewhere between 3 and 5 years. ERG preferred censoring in the ITC. ERG's most plausible ICER applied equal hazards from 3 years, censoring in the ITC, and subsequent therapy for the routine surveillance arm based on use of nivolumab arm of the CheckMate 238 study	Yes	The committee considered OS modelling to be highly uncertain, and assumed equal hazards of death from between 3 to 5 years to be most appropriate for decision making. For other assumptions, the committee sided with the ERG's preferences which were deemed conservative
30	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee concluded that the ERG's OS estimates for the ITT population were pessimistic but the company's estimates were too optimistic, and preferred the ERG's approach to capturing a treatment waning effect (gradual from 2 to 5 years, instead of instant at 5 years)
31	Yes	The committee noted the limitations of the model structure with respect to exploring the impact of subsequent therapies. Rationale for pooling Myeloma XI and CALGB 100104 data was considered unclear, and would have instead preferred an approach to estimating survival which was based mainly on the Myeloma XI study, but supported with the CALGB 100104 study. Treatment effect waning was highlighted as a missing feature of the base-case analyses presented by both the company and ERG	Yes	Company provided additional analyses to re-weight patients from CALGB 100104 to 'match' those from Myeloma XI. Models were then fitted to the Myeloma XI data until 5 years, after which the re-weighted CALGB 100104 data were used to inform the remainder of the extrapolations. Preferred extrapolation model changed to a jointly-fitted generalised gamma	Yes	Yes	ERG maintained its preference for its Myeloma XI only analysis, but accepted that the company's revised approach was more appropriate versus the approach presented in the previous committee meeting	Yes	The committee acknowledged the high degree of uncertainty in results, especially noting that the final analyses provided by the company relied on crossover adjustment and re-weighting of patients from CALGB 100104 to 'match' the Myeloma XI study. The committee ultimately accepted a range of scenarios as suitable to inform decision making.
32	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee discussed the alternative approaches presented for adjusting long-term survival, noting that there was substantial uncertainty. Of the SMR option from the company and the two HRs presented by the ERG, the committee preferred the lower HR from the ERG (2.36) to inform its preferred ICER.

33	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee noted that the estimates of OS from the model were highly uncertain, owing to the immature data from the PRIMA trial. It ultimately concluded that further data were needed to reduce this uncertainty
34	Yes	The committee preferred the ERG's approach of using independent models fitted separately to each treatment arm	Yes	Company reaffirms its preferred base case	No	Yes	ERG reaffirms its preferred base case	Yes	The committee preferred the ERG's approach of using independent models fitted separately to each treatment arm
35	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes	The committee noted that there was uncertainty about the magnitude of survival benefit, and rejected the base-case approach taken to the ITC (an unanchored matching-adjusted indirect comparison). However, the committee ultimately considered that in the intervention was likely to be cost-effective across the range of scenarios presented
36	Yes	The committee preferred the ERG's approach of using a piecewise extrapolation approach, but considered that further work was required to consider use of the updated data cut provided in technical engagement as well as exploration of a range of alternative curve extrapolations	Yes	Company provided analyses in line with those requested in the ACD (i.e., using a piecewise approach with the updated data cut)	No	N/A	ERG noted that alternative extrapolations provided very similar visual fits to the available data from the BEACON CRC study, and so it is difficult to distinguish between them	Yes	The committee preferred a piecewise approach using a generalised gamma tail after 2.8 months for the intervention. For the comparator, the committee noted that the most likely 'true' OS estimate was likely to fall somewhere between the approach using an HR and an approach using extrapolations based on the control arm of the BEACON CRC study

Key: ACD, appraisal consultation document; CRC, colorectal cancer; ERG, Evidence Review Group; FAD, Final Appraisal Determination; HR, hazard ratio; ICER, incremental cost-effectiveness ratio; ITC, indirect treatment comparison; KM, Kaplan-Meier; MAIC, matching-adjusted indirect comparison; NMA, network meta-analysis; OS, overall survival; PartSA, partitioned-survival analysis; SLR, systematic literature review; SMR, standardised mortality ratio; STM state-transition model.

Appendix F. Miscellaneous code

Code to sense check interval argument in simsurv

When using `simsurv`, an issue was identified when sampling a large number of data points. Based on correspondence with the package developer, the cause of the issue was identified as the 'interval' argument in the 'simsurv' function. To check if this had any impact on the simulated data, the following code was generated, which check that the interval argument does not materially influence the simulated data.

```
# Load simsurv package
library(simsurv)

# Set seed for reproducibility and input simulation parameters
set.seed(123)
n <- 1000
covs <- data.frame(id = 1:n, x = rnorm(n))
lambda_in <- 1
gamma_in <- 0.5

# Reset seed for reproducibility and simulate survival times without interval
set.seed(123)
sim1 <- simsurv(dist = "weibull", lambda = lambda_in, gamma = gamma_in, x = covs, beta = c(x = 0.5),
maxt = 10)

# Reset seed for reproducibility and simulate survival times with wide interval
set.seed(123)
sim2 <- simsurv(dist = "weibull", lambda = lambda_in, gamma = gamma_in, x = covs, beta = c(x = 0.5),
maxt = 10, interval = c(1E-20, 1E20))

# Compare the produced data sets
table(sim1 == sim2) == 3*n
```

Code to use UK life tables in Sweeting tutorial

To include UK life tables in the relative survival code by Sweeting *et al.*, the input data need to be converted in a specific format (a 'rate table'). Converting life table data from a .txt file to a rate table is not straightforward, and so the code below demonstrates how this was achieved. The US portion of the code mirrors the code of Sweeting *et al.*, whereas the UK portion of the code was developed from first principles to align with the US specification.

```
if(location == "us"){
  #us code - survexp.us comes from the 'survival' package

  survexp.us.df <- as.data.frame.table(survexp.us, responseName = "exprate") %>%
    mutate(exprate = 365.25 * exprate)
  survexp.us.df$age <- as.numeric(as.character(survexp.us.df$age))
  survexp.us.df$year <- as.numeric(as.character(survexp.us.df$year))
  ratetable.in <- survexp.us
  ratetable.df.in <- survexp.us.df
} else {
  #uk code
  #use UK life tables from HMD - downloaded from the websites below (note: these need a login so cannot
  #automate fully)
  #https://www.mortality.org/File/GetDocument/hmd.v6/GBR_NP/STATS/fltper_1x1.txt
  #https://www.mortality.org/File/GetDocument/hmd.v6/GBR_NP/STATS/mltper_1x1.txt
  fltper1x1 <- read.table("data/fltper_1x1.txt", header = TRUE, skip = 2)
  mltper1x1 <- read.table("data/mltper_1x1.txt", header = TRUE, skip = 2)

  survexp.uk.df_f <- data.frame(age = fltper1x1$Age,
                                sex = "female",
                                year = fltper1x1$Year,
                                exprate = fltper1x1$qx/365.25)
```

```

survexp.uk.df_m <- data.frame(age = mltper1x1$Age,
                             sex = "male",
                             year = mltper1x1$Year,
                             exprate = mltper1x1$qx/365.25)
survexp.uk.df <- rbind(survexp.uk.df_f,survexp.uk.df_m)

survexp.uk.df[survexp.uk.df == "110+"] <- 110
survexp.uk.df$age <- as.numeric(survexp.uk.df$age)

survexp.uk <- structure(survexp.uk.df$exprate/1, dim = c(111L, 2L, 99L), dimnames = list(
  age = as.character(seq(0,110,1)),
  sex = c("male", "female"),
  year = as.character(seq(1922,2020,1))),
  type = c(2, 1, 4),
  cutpoints = list(seq(0,365.25*110,365.25),
                    NULL,
                    structure(as.numeric(as.Date(paste0(seq(1922,2020,1),"-01-01"))),
                              class = "Date")),
  summary = function (R){
    x = c(format(round(min(R[, 1])/365.25, 1)),
          format(round(max(R[, 1])/365.25, 1)),
          sum(R[, 2] == 1), sum(R[, 2] == 2))
    x2 = as.Date(c(min(R[, 3]),
                  max(R[, 3])),
                origin = "1970/01/01")
    paste(" age ranges from", x[1], "to", x[2], "years\n", " male:",
          x[3], " female:", x[4], "\n", " date of entry from",
          x2[1], "to", x2[2], "\n")
  },
  class = "ratetable")

survexp.uk.df <- as.data.frame.table(survexp.uk, responseName = "exprate") %>%
  mutate(exprate = 365.25 * exprate)
survexp.uk.df$age <- as.numeric(as.character(survexp.uk.df$age))
survexp.uk.df$year <- as.numeric(as.character(survexp.uk.df$year))
ratetable.in <- survexp.uk
ratetable.df.in <- survexp.uk.df
}

```

Appendix G. Additional results from neutral comparison study

Logic checking for main results (mean survival)

Appendix Table 17 shows the results of the logic error checking for the blended survival method, described further in Section 6.3.1 of the thesis.

Appendix Table 17: Logic error checking for blended survival method

#	Sample size	Follow-up	Hazard profile*	Survival magnitude	External similarity	Logic error
1	100	2 years	Monotonically decreasing	Low	Identical prognosis	738
2					Poorer prognosis	409
3				Medium	Identical prognosis	872
4					Poorer prognosis	556
5			Non-monotonic	Low	Identical prognosis	591
6					Poorer prognosis	35
7				Medium	Identical prognosis	901
8					Poorer prognosis	543
9		5 years	Monotonically decreasing	Low	Identical prognosis	371
10					Poorer prognosis	58
11				Medium	Identical prognosis	525
12					Poorer prognosis	123
13			Non-monotonic	Low	Identical prognosis	522
14					Poorer prognosis	57
15				Medium	Identical prognosis	447
16					Poorer prognosis	117
17	500	2 years	Monotonically decreasing	Low	Identical prognosis	615
18					Poorer prognosis	205
19				Medium	Identical prognosis	815
20					Poorer prognosis	493
21			Non-monotonic	Low	Identical prognosis	35
22					Poorer prognosis	0
23				Medium	Identical prognosis	313
24					Poorer prognosis	27
25		5 years	Monotonically decreasing	Low	Identical prognosis	170
26					Poorer prognosis	2
27				Medium	Identical prognosis	251
28					Poorer prognosis	23
29			Non-monotonic	Low	Identical prognosis	30
30					Poorer prognosis	0
31				Medium	Identical prognosis	80
32					Poorer prognosis	2

Note: *Parameters for the hazard profiles are presented in Table 6-1.

Additional results for primary estimand (mean survival)

Tabulated results including MCSE for the primary estimand are presented in Appendix Table 18 (trial as truth) and Appendix Table 19 (external as truth). Additional NLPs are presented in Appendix Figure 11 and Appendix Figure 12, for absolute and percentage bias, respectively, where the *external* population is assumed to represent the truth (instead of the *trial* population).

Appendix Table 18: Bias in mean survival, all scenarios, trial as truth

#	Mean bias (MCSE)							
	Standard parametric	Blended survival	Expert opinion	Informative prior on shape	M-spline MPES	Relative survival	Standard piecewise	Summary data piecewise
1	0.2167 (0.0230)	-0.2186 (0.0081)	0.2150 (0.0160)	0.2940 (0.0194)	0.0138 (0.0090)	0.2689 (0.0239)	-0.0468 (0.0094)	0.0256 (0.0099)
2	0.1919 (0.0228)	-0.5832 (0.0070)	-0.0637 (0.0138)	0.3207 (0.0200)	-0.2897 (0.0082)	0.2430 (0.0237)	-0.3003 (0.0083)	-0.2050 (0.0091)
3	1.0348 (0.0649)	0.0557 (0.0129)	1.0598 (0.0352)	1.3920 (0.0524)	0.0902 (0.0204)	1.3157 (0.0680)	0.0147 (0.0199)	-0.0826 (0.0184)
4	1.0947 (0.0618)	-1.0729 (0.0125)	-0.0887 (0.0285)	1.5517 (0.0519)	-0.7580 (0.0175)	1.3812 (0.0649)	-0.8170 (0.0168)	-0.8155 (0.0163)
5	-0.4624 (0.0086)	-0.1957 (0.0068)	-0.3968 (0.0075)	-0.3373 (0.0089)	0.0672 (0.0074)	-0.4517 (0.0089)	-0.0574 (0.0078)	0.0247 (0.0084)
6	-0.4663 (0.0088)	-0.4392 (0.0062)	-0.4987 (0.0067)	-0.3604 (0.0088)	-0.1719 (0.0065)	-0.4556 (0.0090)	-0.2353 (0.0069)	-0.1649 (0.0076)
7	-2.2950 (0.0226)	0.1419 (0.0126)	-1.4358 (0.0208)	-1.4610 (0.0269)	0.2290 (0.0194)	-2.2349 (0.0239)	0.0008 (0.0199)	0.0816 (0.0184)
8	-2.2889 (0.0225)	-0.9689 (0.0114)	-2.0946 (0.0161)	-1.5152 (0.0264)	-0.5562 (0.0161)	-2.2290 (0.0237)	-0.7130 (0.0165)	-0.6284 (0.0159)
9	0.1534 (0.0178)	-0.1136 (0.0082)	0.1679 (0.0142)	0.2219 (0.0163)	-0.0314 (0.0111)	0.2292 (0.0190)	-0.0160 (0.0117)	-0.0363 (0.0117)
10	0.1398 (0.0179)	-0.4112 (0.0077)	0.0213 (0.0130)	0.2448 (0.0169)	-0.1618 (0.0106)	0.2142 (0.0191)	-0.1242 (0.0115)	-0.1393 (0.0112)
11	0.8483 (0.0484)	0.3256 (0.0134)	0.8851 (0.0295)	1.1131 (0.0434)	0.0246 (0.0229)	1.2765 (0.0530)	0.0564 (0.0261)	-0.0452 (0.0236)
12	1.0066 (0.0465)	-0.6870 (0.0119)	0.1067 (0.0248)	1.3106 (0.0415)	-0.5168 (0.0196)	1.4502 (0.0510)	-0.3819 (0.0221)	-0.4633 (0.0202)
13	-0.1430 (0.0097)	-0.0006 (0.0076)	-0.0944 (0.0090)	-0.1016 (0.0095)	0.0315 (0.0088)	-0.1174 (0.0101)	-0.0706 (0.0091)	0.0129 (0.0092)
14	-0.1525 (0.0096)	-0.2010 (0.0070)	-0.1541 (0.0083)	-0.1200 (0.0093)	-0.0725 (0.0082)	-0.1276 (0.0100)	-0.1310 (0.0085)	-0.0643 (0.0086)
15	-0.6499 (0.0305)	0.5759 (0.0125)	-0.1995 (0.0208)	-0.3761 (0.0296)	0.1542 (0.0208)	-0.4452 (0.0333)	-0.1320 (0.0233)	0.0970 (0.0218)
16	-0.5704 (0.0340)	-0.4447 (0.0122)	-0.8007 (0.0190)	-0.3556 (0.0317)	-0.3567 (0.0197)	-0.3586 (0.0370)	-0.4924 (0.0214)	-0.3181 (0.0203)
17	0.1551 (0.0096)	-0.1769 (0.0043)	0.1739 (0.0090)	0.1758 (0.0090)	0.0037 (0.0037)	0.2047 (0.0100)	-0.0477 (0.0041)	0.0277 (0.0043)
18	0.1457 (0.0094)	-0.5423 (0.0036)	0.0921 (0.0084)	0.1778 (0.0089)	-0.3329 (0.0034)	0.1952 (0.0098)	-0.3067 (0.0036)	-0.2036 (0.0040)
19	1.0612 (0.0282)	0.1398 (0.0074)	1.0884 (0.0235)	1.1236 (0.0250)	0.0692 (0.0080)	1.3567 (0.0297)	0.0305 (0.0087)	-0.0707 (0.0082)
20	1.0944 (0.0296)	-1.0035 (0.0065)	0.6854 (0.0232)	1.1855 (0.0261)	-0.8976 (0.0068)	1.3894 (0.0311)	-0.8248 (0.0072)	-0.8228 (0.0070)
21	-0.4746 (0.0036)	-0.0518 (0.0039)	-0.4564 (0.0036)	-0.4378 (0.0037)	0.0523 (0.0030)	-0.4649 (0.0037)	-0.0505 (0.0034)	0.0365 (0.0036)
22	-0.4819 (0.0036)	-0.3148 (0.0033)	-0.4854 (0.0034)	-0.4497 (0.0036)	-0.2014 (0.0027)	-0.4724 (0.0037)	-0.2315 (0.0030)	-0.1480 (0.0032)
23	-2.3655 (0.0089)	0.4299 (0.0063)	-2.1462 (0.0096)	-2.0903 (0.0098)	0.2858 (0.0075)	-2.3120 (0.0094)	0.0042 (0.0087)	0.0753 (0.0081)
24	-2.3605 (0.0087)	-0.6962 (0.0056)	-2.2850 (0.0084)	-2.1045 (0.0094)	-0.5819 (0.0064)	-2.3070 (0.0092)	-0.7100 (0.0073)	-0.6293 (0.0070)
25	0.1327 (0.0080)	-0.0707 (0.0038)	0.1475 (0.0077)	0.1497 (0.0078)	-0.0352 (0.0050)	0.2064 (0.0086)	-0.0078 (0.0053)	-0.0270 (0.0052)
26	0.1237 (0.0076)	-0.3837 (0.0035)	0.1039 (0.0071)	0.1472 (0.0074)	-0.1690 (0.0047)	0.1963 (0.0081)	-0.1262 (0.0050)	-0.1356 (0.0050)
27	0.9085 (0.0220)	0.3777 (0.0063)	0.9301 (0.0193)	0.9643 (0.0208)	0.0072 (0.0105)	1.3495 (0.0242)	0.0878 (0.0119)	-0.0209 (0.0108)
28	0.8720 (0.0206)	-0.6726 (0.0054)	0.6199 (0.0172)	0.9548 (0.0194)	-0.5759 (0.0085)	1.3097 (0.0226)	-0.4139 (0.0097)	-0.5096 (0.0089)
29	-0.1721 (0.0042)	0.0281 (0.0035)	-0.1576 (0.0042)	-0.1615 (0.0042)	-0.0013 (0.0040)	-0.1490 (0.0044)	-0.0876 (0.0041)	0.0001 (0.0042)
30	-0.1653 (0.0043)	-0.1710 (0.0033)	-0.1616 (0.0042)	-0.1575 (0.0042)	-0.0851 (0.0037)	-0.1418 (0.0044)	-0.1341 (0.0039)	-0.0633 (0.0040)
31	-0.6642 (0.0142)	0.6325 (0.0060)	-0.5187 (0.0132)	-0.5862 (0.0140)	0.1419 (0.0098)	-0.4603 (0.0156)	-0.1011 (0.0112)	0.1215 (0.0104)
32	-0.6696 (0.0143)	-0.4166 (0.0055)	-0.7040 (0.0124)	-0.6057 (0.0140)	-0.4167 (0.0084)	-0.4664 (0.0156)	-0.5214 (0.0098)	-0.3541 (0.0091)

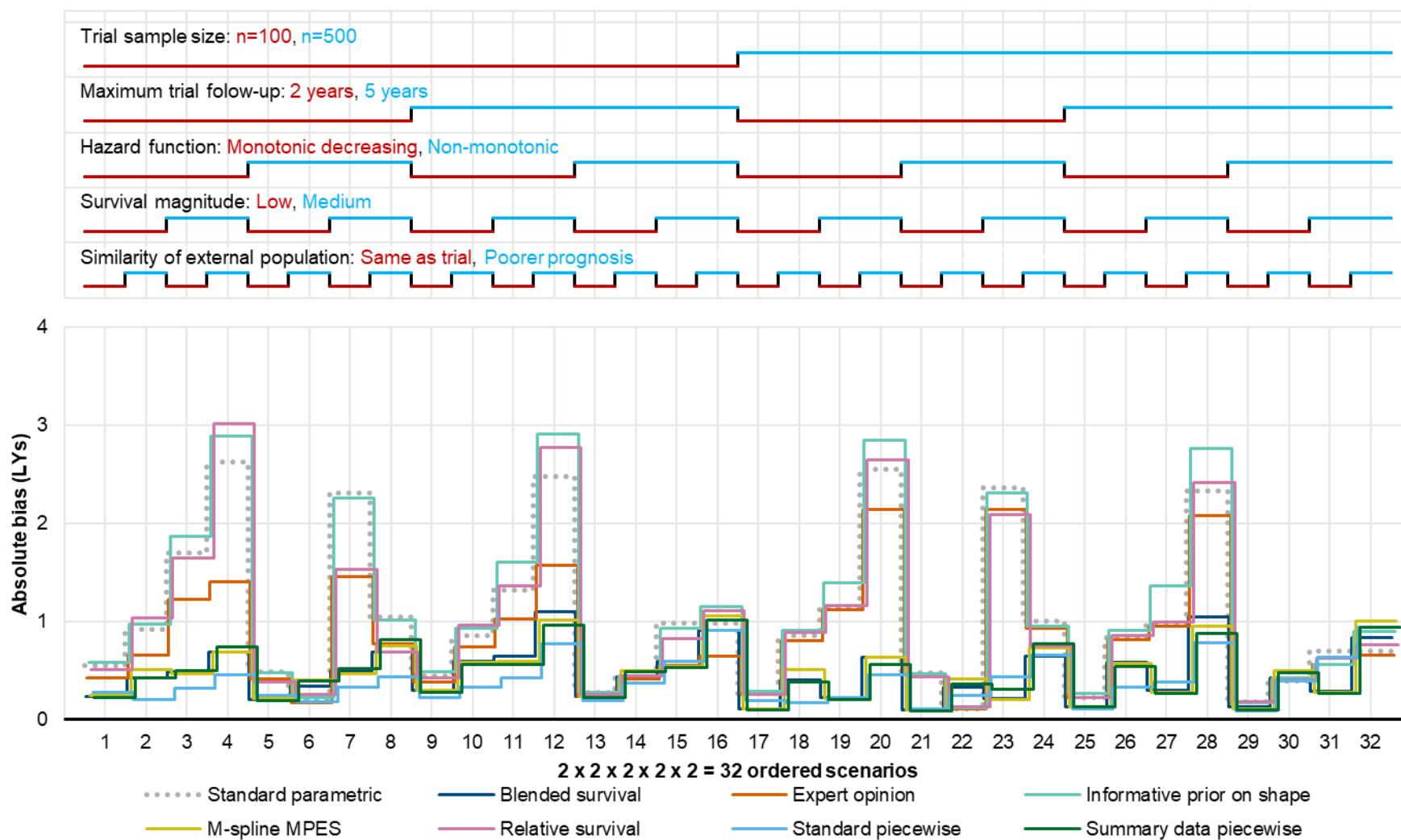
Key: MCSE, Monte Carlo simulation error; MPES, multi-parameter evidence synthesis.

Appendix Table 19: Bias in mean survival, all scenarios, external as truth

#	Mean bias (MCSE)							
	Standard parametric	Blended survival	Expert opinion	Informative prior on shape	M-spline MPES	Relative survival	Standard piecewise	Summary data piecewise
1	0.2167 (0.0230)	-0.2186 (0.0081)	0.2150 (0.0160)	0.2940 (0.0194)	0.0138 (0.0090)	0.2689 (0.0239)	-0.0468 (0.0094)	0.0256 (0.0099)
2	0.9065 (0.0228)	0.1314 (0.0070)	0.6509 (0.0138)	1.0353 (0.0200)	0.4249 (0.0082)	0.9576 (0.0237)	0.4143 (0.0083)	0.5095 (0.0091)
3	1.0348 (0.0649)	0.0557 (0.0129)	1.0598 (0.0352)	1.3920 (0.0524)	0.0902 (0.0204)	1.3157 (0.0680)	0.0147 (0.0199)	-0.0826 (0.0184)
4	2.5547 (0.0618)	0.3871 (0.0125)	1.3713 (0.0285)	3.0117 (0.0519)	0.7020 (0.0175)	2.8412 (0.0649)	0.6429 (0.0168)	0.6444 (0.0163)
5	-0.4624 (0.0086)	-0.1957 (0.0068)	-0.3968 (0.0075)	-0.3373 (0.0089)	0.0672 (0.0074)	-0.4517 (0.0089)	-0.0574 (0.0078)	0.0247 (0.0084)
6	0.0973 (0.0088)	0.1244 (0.0062)	0.0649 (0.0067)	0.2032 (0.0088)	0.3917 (0.0065)	0.1080 (0.0090)	0.3282 (0.0069)	0.3987 (0.0076)
7	-2.2950 (0.0226)	0.1419 (0.0126)	-1.4358 (0.0208)	-1.4610 (0.0269)	0.2290 (0.0194)	-2.2349 (0.0239)	0.0008 (0.0199)	0.0816 (0.0184)
8	-0.9314 (0.0225)	0.3886 (0.0114)	-0.7371 (0.0161)	-0.1577 (0.0264)	0.8013 (0.0161)	-0.8715 (0.0237)	0.6445 (0.0165)	0.7291 (0.0159)
9	0.1534 (0.0178)	-0.1136 (0.0082)	0.1679 (0.0142)	0.2219 (0.0163)	-0.0314 (0.0111)	0.2292 (0.0190)	-0.0160 (0.0117)	-0.0363 (0.0117)
10	0.8544 (0.0179)	0.3034 (0.0077)	0.7359 (0.0130)	0.9594 (0.0169)	0.5528 (0.0106)	0.9288 (0.0191)	0.5904 (0.0115)	0.5753 (0.0112)
11	0.8483 (0.0484)	0.3256 (0.0134)	0.8851 (0.0295)	1.1131 (0.0434)	0.0246 (0.0229)	1.2765 (0.0530)	0.0564 (0.0261)	-0.0452 (0.0236)
12	2.4666 (0.0465)	0.7730 (0.0119)	1.5667 (0.0248)	2.7706 (0.0415)	0.9432 (0.0196)	2.9102 (0.0510)	1.0781 (0.0221)	0.9966 (0.0202)
13	-0.1430 (0.0097)	-0.0006 (0.0076)	-0.0944 (0.0090)	-0.1016 (0.0095)	0.0315 (0.0088)	-0.1174 (0.0101)	-0.0706 (0.0091)	0.0129 (0.0092)
14	0.4110 (0.0096)	0.3626 (0.0070)	0.4095 (0.0083)	0.4436 (0.0093)	0.4911 (0.0082)	0.4360 (0.0100)	0.4326 (0.0085)	0.4993 (0.0086)
15	-0.6499 (0.0305)	0.5759 (0.0125)	-0.1995 (0.0208)	-0.3761 (0.0296)	0.1542 (0.0208)	-0.4452 (0.0333)	-0.1320 (0.0233)	0.0970 (0.0218)
16	0.7871 (0.0340)	0.9128 (0.0122)	0.5568 (0.0190)	1.0019 (0.0317)	1.0008 (0.0197)	0.9989 (0.0370)	0.8651 (0.0214)	1.0394 (0.0203)
17	0.1551 (0.0096)	-0.1769 (0.0043)	0.1739 (0.0090)	0.1758 (0.0090)	0.0037 (0.0037)	0.2047 (0.0100)	-0.0477 (0.0041)	0.0277 (0.0043)
18	0.8603 (0.0094)	0.1723 (0.0036)	0.8067 (0.0084)	0.8924 (0.0089)	0.3817 (0.0034)	0.9098 (0.0098)	0.4079 (0.0036)	0.5110 (0.0040)
19	1.0612 (0.0282)	0.1398 (0.0074)	1.0884 (0.0235)	1.1236 (0.0250)	0.0692 (0.0080)	1.3567 (0.0297)	0.0305 (0.0087)	-0.0707 (0.0082)
20	2.5544 (0.0296)	0.4565 (0.0065)	2.1454 (0.0232)	2.6454 (0.0261)	0.5624 (0.0068)	2.8494 (0.0311)	0.6352 (0.0072)	0.6372 (0.0070)
21	-0.4746 (0.0036)	-0.0518 (0.0039)	-0.4564 (0.0036)	-0.4378 (0.0037)	0.0523 (0.0030)	-0.4649 (0.0037)	-0.0505 (0.0034)	0.0365 (0.0036)
22	0.0816 (0.0036)	0.2488 (0.0033)	0.0782 (0.0034)	0.1139 (0.0036)	0.3622 (0.0027)	0.0911 (0.0037)	0.3321 (0.0030)	0.4155 (0.0032)
23	-2.3655 (0.0089)	0.4299 (0.0063)	-2.1462 (0.0096)	-2.0903 (0.0098)	0.2858 (0.0075)	-2.3120 (0.0094)	0.0042 (0.0087)	0.0753 (0.0081)
24	-1.0030 (0.0087)	0.6613 (0.0056)	-0.9275 (0.0084)	-0.7470 (0.0094)	0.7756 (0.0064)	-0.9495 (0.0092)	0.6475 (0.0073)	0.7282 (0.0070)
25	0.1327 (0.0080)	-0.0707 (0.0038)	0.1475 (0.0077)	0.1497 (0.0078)	-0.0352 (0.0050)	0.2064 (0.0086)	-0.0078 (0.0053)	-0.0270 (0.0052)
26	0.8383 (0.0076)	0.3309 (0.0035)	0.8185 (0.0071)	0.8618 (0.0074)	0.5456 (0.0047)	0.9109 (0.0081)	0.5884 (0.0050)	0.5790 (0.0050)
27	0.9085 (0.0220)	0.3777 (0.0063)	0.9301 (0.0193)	0.9643 (0.0208)	0.0072 (0.0105)	1.3495 (0.0242)	0.0878 (0.0119)	-0.0209 (0.0108)
28	2.3320 (0.0206)	0.7874 (0.0054)	2.0799 (0.0172)	2.4148 (0.0194)	0.8841 (0.0085)	2.7697 (0.0226)	1.0461 (0.0097)	0.9503 (0.0089)
29	-0.1721 (0.0042)	0.0281 (0.0035)	-0.1576 (0.0042)	-0.1615 (0.0042)	-0.0013 (0.0040)	-0.1490 (0.0044)	-0.0876 (0.0041)	0.0001 (0.0042)
30	0.3982 (0.0043)	0.3926 (0.0033)	0.4020 (0.0042)	0.4061 (0.0042)	0.4785 (0.0037)	0.4217 (0.0044)	0.4295 (0.0039)	0.5003 (0.0040)
31	-0.6642 (0.0142)	0.6325 (0.0060)	-0.5187 (0.0132)	-0.5862 (0.0140)	0.1419 (0.0098)	-0.4603 (0.0156)	-0.1011 (0.0112)	0.1215 (0.0104)
32	0.6879 (0.0143)	0.9409 (0.0055)	0.6535 (0.0124)	0.7518 (0.0140)	0.9408 (0.0084)	0.8910 (0.0156)	0.8361 (0.0098)	1.0034 (0.0091)

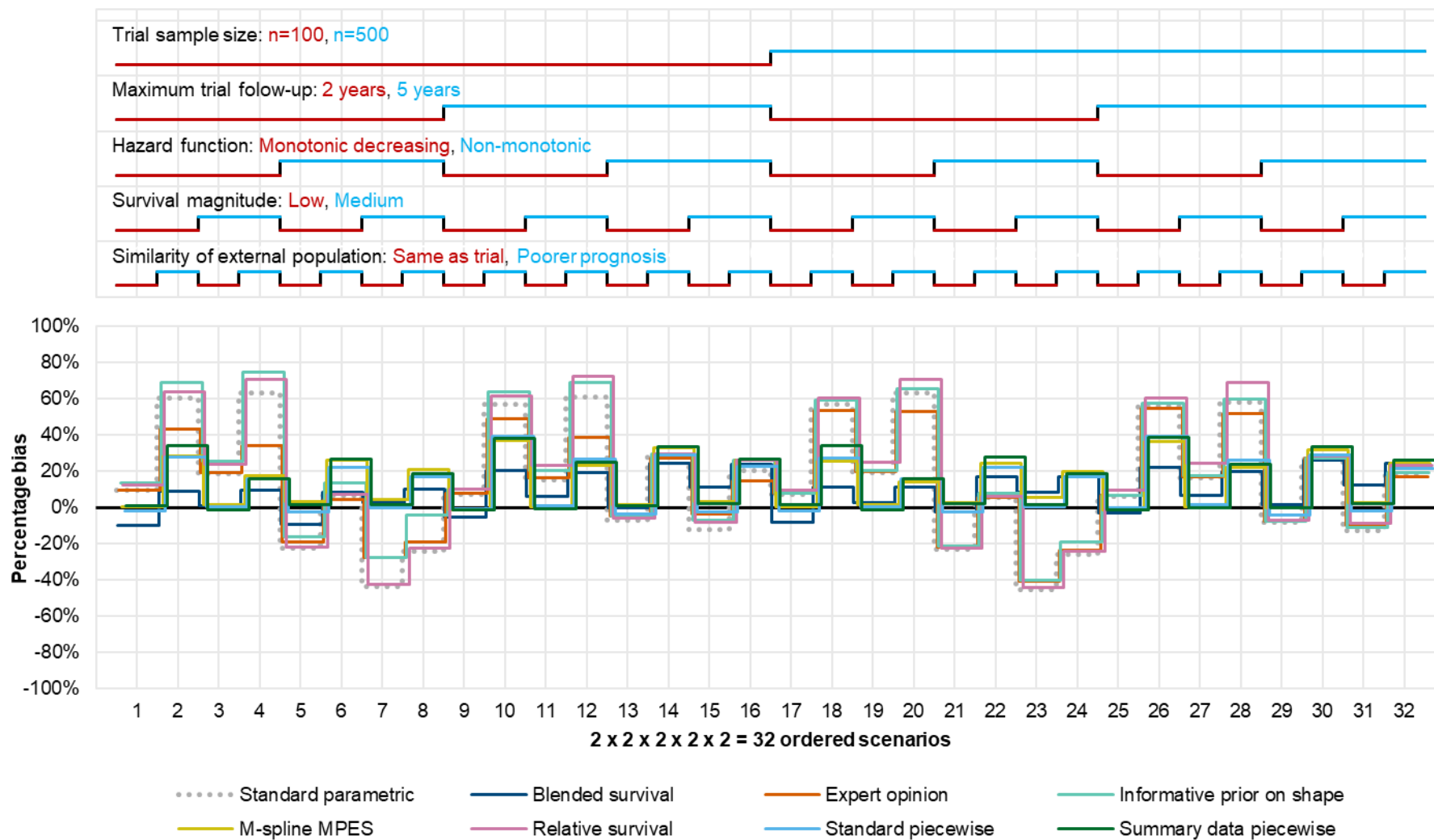
Key: MCSE, Monte Carlo simulation error; MPES, multi-parameter evidence synthesis.

Appendix Figure 11: NLP: Absolute bias in mean survival, all scenarios, external as truth



Key: LYs, life-years; MPES, multi-parameter evidence synthesis.

Appendix Figure 12: NLP: Percentage bias in mean survival, all scenarios, external as truth



Key: MPES, multi-parameter evidence synthesis.

--- End ---