A methodological framework to account for the impact of non-adherence on the cost-effectiveness of chronic medications

Abualbishr Alshreef
BPharm (Hons), MPH, MSc.

Supervised by

Professor Simon Dixon, Professor Paul Tappenden, Dr Nicholas Latimer and Dr William McKane

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Medicine, Dentistry and Health
School of Health and Related Research (ScHARR)

August 2021

Acknowledgements

I would like to thank:

- My supervisors; Professor Simon Dixon, Professor Paul Tappenden, Dr Nicholas Latimer and Dr William McKane for their valued advice, support and comments on the individual chapters and the full draft thesis.
- My advisors; Professor Dyfrig Hughes, Professor Ian White, Professor Allan Wailoo, Dr Michael Crowther, Dr James Fotheringham, Dr Ruth Wong and Mr Duncan Gray for their advice on my doctoral research plan, the systematic review, the simulation study, the kidney transplantation case study, the methodological framework and/or comments on the individual chapters of my thesis.
- Dr Monica Hernandez Alava, my line manager for the past three years, for her support and encouragement.
- My wife Sara for her incredible support, encouragement and understanding. This journey would not have been possible without her.

Funding

This PhD thesis is independent research supported by the National Institute for Health Research **(NIHR Doctoral Research Fellowship, Mr Abualbishr Alshreef, DRF-2017-10-025)**. The views expressed in this thesis are those of the author and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Dissemination

I have published part of the research presented in this thesis in the following peer-reviewed journal article:

Alshreef A, Latimer N, Tappenden P, Wong R, Hughes D, Fotheringham J, Dixon S. Statistical Methods for Adjusting Estimates of Treatment Effectiveness for Patient Non-adherence in the Context of Time-to-Event Outcomes and Health Technology Assessment: A Systematic Review of Methodological Papers. *Medical Decision Making.* 2019; 39(8)910-925. DOI: 10.1177/0272989X19881654

I have presented part of this research at the following international conference:

Alshreef A, Latimer N, Tappenden P, McKane W, Wong R, Hughes D, Wailoo A, White, I, Fotheringham J, Dixon S. Adjusting the Causal Effects of Treatments in the Presence of Patient Non-adherence: A Systematic Review of Methodological Papers. International Society for Medication Adherence (ESPACOMP) Conference. Dublin, Ireland. November 2018

I have submitted part of this research for presentation at the following international conference:

Alshreef A, Latimer N, Tappenden P, Dixon S. Performance of Methods for Adjusting Estimates of Treatment Effectiveness for Patient Non-adherence in the Context of Time-To-Event Outcomes and Health Technology Assessment: A Simulation Study. *Society for Medical Decision Making (SMDM) 43rd Annual North American Meeting*. Virtual Meeting. October 2021

Dedication

In memory of my father

Abstract

This thesis investigates the problem of estimating the effectiveness and cost-effectiveness of prescribed medications in the presence of patient non-adherence, which is a major issue when making reimbursement decisions. Poor reimbursement decisions can have significant negative consequences for both patient outcomes and costs.

This research aimed to develop a methodological framework to account for the impact of non-adherence on the cost-effectiveness of chronic medications in the context of health technology assessments (HTA) using time-to-event outcomes. Such a framework did not exist when this research started. The framework put forward in this thesis was informed by four linked stages of this research: (1) a systematic review of methodological papers that identified 12 non-adherence adjustment methods and assessed their suitability for use in HTA; (2) a simulation study that assessed the performance of four adjustment methods in 90 scenarios; (3) a case study that applied two generalised methods (g-methods) to a trial of maintenance immunosuppressants in kidney transplantation in order to produce cost-effectiveness estimates based on external 'real world' non-adherence data; and (4) the development of an analytical framework for incorporating non-adherence into cost-effectiveness models.

The review suggests that g-methods and pharmacometrics-based methods using pharmacokinetics and pharmacodynamics (PKPD) analysis appear to be more appropriate to estimate effectiveness in the presence of real-world non-adherence. The simulation study demonstrates that g-methods and per-protocol (PP) analysis are the best-performing methods for adjusting estimates of treatment effectiveness for non-adherence compared to intention-to-treat analysis, although the PP estimand is not theoretically appropriate. The case study findings show reduced net health benefits and increased cost per patient when real-world non-adherence is taken into account. The framework provides guidance on the steps that could be followed to account for real-world non-adherence when undertaking cost-effectiveness analyses. Future research is recommended to use the framework in different case studies for evaluation.

Table of Contents

List of tables

List of figures

List of abbreviations

ABC	Ascertaining Barriers to Compliance
ACE	Average Causal Effect
ADR	Adverse Drug Reaction
AFTM	Accelerated Failure Time Model
AIDS	Acquired Immunodeficiency Syndrome
ATE	Average Treatment Effect
AT	As Treated
AZT	Azidothymidine
BAS	Basiliximab
BD	Twice a Day
BMI	Body Mass Index
BPAR	Biopsy-Proven Acute Rejection
CACE	Complier Average Causal Effect
CESP	Compliers Effect on Survival Probability
CD4	Cluster of Differentiation 4
CNI	Calcineurin Inhibitors
CPG	Comprehensive Pearl Growing
CPH	Cox Proportional Hazards
CSR	Clinical Study Report
CsA	Cyclosporine
CRT	Cluster Randomised Trial
CRF	Case Report Form
C-PROPHET	Compliers PROPortional Hazards Effect of Treatment
CV	Coefficient of Variation
DAG	Directed-Acyclic Graph
DWFG	Death with a Functioning Graft
DES	Discrete Event Simulation
DGM	Data-Generating Mechanism
DSA	Donor-Specific Antibodies

ESPACOMP	International Society for Medication Adherence
EM	Expectation Maximisation
EW	Efficient one-step Weights
eGFR	Estimated Glomerular Filtration Rate
EMP	Electronic Medical Packaging
EmpSE	Empirical Standard Error
FG	Functioning Graft
FL	Full Likelihood
FMPR	Fixed Medication Possession Ratio
FPM	Flexible Parametric Model
GL	Graft Loss
GP	General Practitioner
HEAP	Health Economics Analysis Plan
HEOR	Health Economics and Outcomes Research
HTA	Health Technology Assessment
HR	Hazard Ratio
HIV	Human Immunodeficiency Virus
ITT	Intention to Treat
ICER	Incremental Cost Effectiveness Ratio
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
IPCW	Inverse Probability of Censoring Weighting
IPTW	Inverse Probability of Treatment Weighting
IPW	Inverse Probability Weighting
IPV	Intra-Patient Variability
IPD	Individual Patient-level Data
IV	Instrumental Variable
KM	Kaplan-Meier
LOCF	Last Observation Carried Forward
MAR	Missing at Random
MCAR	Missing Completely at Random
MCSE	Monte Carlo Standard Error

MEMS	Medication Event Monitoring System
MNA	Medication Non-Adherence
MSM	Marginal Structural Model
MLE	Maximum Likelihood Estimator
MH	Mantel Haenszel
MSE	Mean Square Error
MCC	Markov Compliance Class
MCMC	Markov Chain Monte Carlo
MMA	Multiple Medication Adherence
MMF	Mycophenolate
MPR	Medication Possession Ratio
ModSE	Model-based Standard Error
MNAR	Missing Not at Random
NEC	Nuisance Estimation from Control
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NHS	National Health Service
NHB	Net Health Benefit
NMA	Network Meta-Analysis
OD	Once Daily
OR	Odds Ratio
OTC	Over the Counter
PCP	Pneumocystis Carinii Pneumonia
PKPD	Pharmacokinetics and Pharmacodynamics
PNEMLE	Plug-in Non-Parametric Empirical Maximum Likelihood Estimator
PH	Proportional Hazards
PLE	Partial Likelihood Estimator
PE	PharmacoEconomics
PP	Per Protocol
PSA	Probabilistic Sensitivity Analysis
PSM	Parametric Survival Model

QALY	Quality-Adjusted Life year
RMST	Restricted Mean Survival Time
RCM	Rubin Causal Model
RPSFTM	Rank-Preserving Structural Failure Time Model
RCT	Randomised Controlled Trial
RWD	Real-World Data
SAFT	Structural Accelerated Failure Time Model
SAP	Statistical Analysis Plan
SE	Standard Error
ST	Corticosteroids
SEM	Structural Equation Modelling
SNM	Structural Nested Model
SNFTM	Structural Nested Failure Time Model
SNMM	Structural Nested Mean Model
SNDM	Structural Nested Distribution Model
SNLM	Structural Nested Logistic Model
SMP	Shared-Memory Parallelism
ShARC	Sheffield Advanced Research Computer
TAC	Tacrolimus
TDM	Therapeutic Drug Monitoring
TDS	Three Times a Day
TAG	Technology Assessment Group
T2D	Type 2 Diabetes
UKTR	UK Transplant Registry
VMPR	Variable Medication Possession Ratio
Wtd PP	Weighted Per Protocol
WTP	Willingness to Pay
3SM	3-Stage Method

Declaration

I, Abualbishr Alshreef, 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.

Chapter 1: Thesis introduction and background

1.1 Introduction

Economic evaluations and Health Technology Assessments (HTA) are increasingly used to inform decision making around the adoption of new treatments worldwide. An economic evaluation typically assesses the cost-effectiveness (value for money) of a new treatment compared to standard treatment using evidence on both clinical effectiveness and costs of adopting each treatment option. In practice, clinical effectiveness evidence usually comes from randomised controlled trials (RCTs) where the intention-to-treat (ITT) analysis is typically used as a conventional analytical approach. The ITT analysis mixes the benefit of receiving the treatment among adherent patients with the lack of (or suboptimal) benefit among patients who experience non-adherence, which may generate biased estimates of treatment effectiveness compared to that seen in standard clinical practice. In addition, non-adherence to treatment would be expected to alter the costs borne by the health service. In the HTA context, decision makers are generally concerned with costs and effectiveness in standard clinical practice (or the 'real world') as an important issue. Both clinical effectiveness and costs have a direct impact on cost-effectiveness; and therefore, adjustment of effectiveness and costs for non-adherence is needed if adherence in standard clinical practice differs from that observed in RCTs.

The subject of this doctoral research study is the development of a methodological framework to account for the impact of non-adherence in the context of HTAs in chronic conditions. The focus is on methods for adjusting the causal effect of treatment on time-to-event outcomes and cost-effectiveness. This involves a systematic review to identify non-adherence adjustment methods and the application of these in a simulation study and a case study on maintenance immunosuppressive therapy after kidney transplantation. Although the focus is on kidney transplantation, I envisage that the methodological framework will be applicable in any chronic disease area with a time-to-event outcome. Chronic conditions are chosen because non-adherence to medications taken over a long period of time is more likely to have a substantial impact on costs and outcomes. Non-adherence to medication was raised as an issue in a recent NICE appraisal of immunosuppression following kidney transplant and this was thought to provide an important and topical case study. Given that this was based on time-to-event outcomes, my work has focussed on this, however, it is recognised that the same issue will be relevant to all forms of patient outcomes (e.g. continuous and categorical outcomes). However, non-adherence is also an important issue in acute conditions, and the methodological framework may also be applicable in these conditions where a time-to-event outcome is used. Many aspects of the methodological framework put forward in this thesis will be applicable

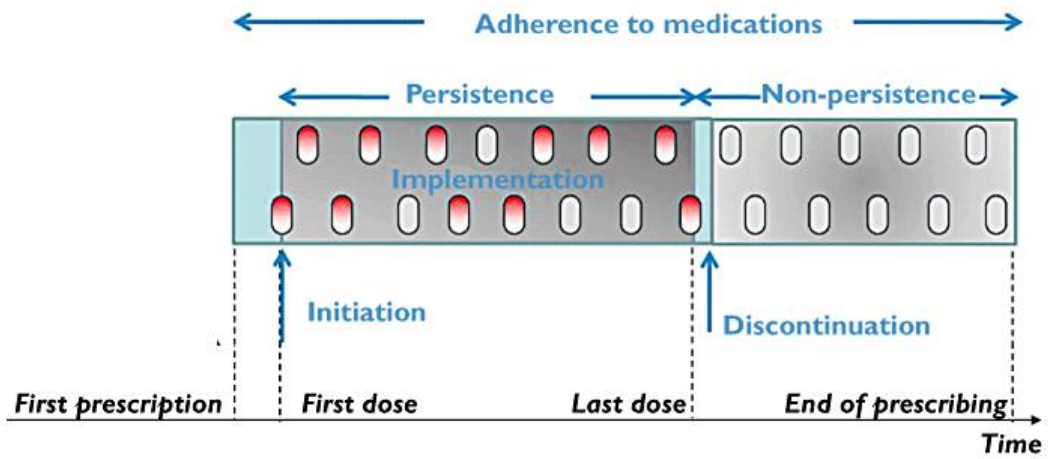
for handling non-adherence in instances in which a time-to-event outcome is not used and this is discussed in further detail in Chapter 7.

This introductory chapter defines non-adherence and key concepts used throughout the thesis; specifies the problem and its importance; identifies the gaps in the health economics literature; and outlines the research questions, and the aim and objectives of the study.

1.2 Definition of non-adherence to medications

Adherence to medication is defined as *'the process by which patients take their medications as prescribed'*.[1] An influential taxonomy of adherence to medications set out by Vrijens et al.[1] (see Figure 1) identifies three components: (a) initiation of the treatment (when the first dose is taken by the patient); (b) implementation of the dosing regimen (to what extend the actual dosage of a patient corresponds to the prescribed dosing regimen); and (c) discontinuation (end of therapy).[1] Based on this taxonomy, non-adherence can occur in one or a combination of three situations: late or non-initiation, sub-optimal implementation, and/or early discontinuation (non-persistence).[1, 2] A number of other terms are commonly used in the literature to describe non-adherence including 'non-compliance' and 'non-concordance'.[3] Whilst some authors suggest differences between the precise meanings of these different terms, for the purpose of this thesis, I will treat them as being synonymous with non-adherence, and the Vrijens et al. 'ABC' taxonomy will be used as standard. The acronym ABC stands for 'Ascertaining Barriers to Compliance' project, which is an international research collaboration in the field of adherence to medications that led to the development of the taxonomy.

Figure 1: Adherence to medications taxonomy



Adapted from Vrijens *et al.*[1]

There are some arguments in the literature considering other medication-taking behaviours such as treatment switching as a type of non-adherence. The counter-argument to this is that switching prescribed medication is a different type of change in therapy as it must be initiated by a medical practitioner. Another related term is 'early discontinuation' which happens before the end of prescribing by the patient's own behaviour, which I would consider as non-adherence (see Figure 1).

Evidence suggests that patients' preferences to adhere to their prescribed medications are consistently affected by side effects, age and experience.[4] Adherence may also be affected by the impact of a specific disease, the complexity of the dosing regimen,[4] and potentially other medications taken by the patient (i.e. polypharmacy).[5]

It is important to measure adherence using the three components specified by the ABC taxonomy (initiation, implementation, persistence) and to be explicit about what is being measured.[1] For instance, initiation and persistence can be measured using time-to-event variables (i.e. time to initiation and time to discontinuation). Conversely, implementation as a continuous process requires a different approach to quantify. Implementation can be measured as a summary statistic using proportions of prescribed doses taken over a particular period. However, summary statistics have their own limitations and can sometimes be misleading, especially if the medication adherence trend over a longer period of time is considered important. As an alternative, longitudinal comparisons using a medication event-monitoring device such as the medication event monitoring system (MEMS), as an example, can provide a better estimate of non-adherence at the implementation stage.

In practice, there have been considerable inconsistencies in measuring non-adherence, with many researchers considering it as a unidimensional problem.[6] Many tools are available for measuring medication adherence including both subjective and objective measures. Existing medication adherence measures can be broadly classified into five groups:[7]

(i) Direct measures such as drug concentration levels in the blood

(ii) Measures involving secondary data analysis such as the medication possession ratio (MPR) calculated as the days of medication supply for all prescription refills divided by the days of the interval period

(iii) Electronic measures such as MEMS

(iv) Pill counting

(v) Measures involving clinical assessments and self-reporting such as questionnaires.

Other high-level groupings are: (a) objective versus subjective measures; (b) cross-sectional versus longitudinal measures; and (c) direct versus indirect measures.[7] There are different approaches for calculating each of these adherence measures. For instance, MPR can be calculated using the fixed MPR approach (FMPR) using a fixed days interval as a denominator, or the variable MPR (VMPR) using the time between the initiation of therapy and the end of supply for the last acquisition as a denominator.[8] Each tool has its own strengths and weaknesses; hence, no perfect measure of adherence exists, and this may lead to measurement error.

1.3 Definition of key concepts

A number of key concepts are used throughout this thesis; it is, therefore, important to define these from the outset. The concepts include: counterfactual outcome; causal effect; estimand and estimator; confounding; directed-acyclic graphs (DAGs); selection bias; censoring; and drug forgiveness. These concepts are briefly introduced in the following subsections.

1.3.1 Counterfactual outcome

The counterfactual outcome framework was originally developed by Neyman[9] and Rubin[10] for estimating the effect of time-fixed treatments (which do not vary over time), and further extended by Robins[11, 12] to time-varying treatments (which vary over time). To introduce the counterfactual outcome, suppose we want to estimate the effect of treatment A on outcome Y for individual *i* compared to no treatment. The ideal way is to compare the 'observed' outcome on individual *i* if s/he had received the treatment with the outcome that would have been observed if the same individual had not received the treatment. We cannot observe treated and untreated outcomes for an individual patient (except in a crossover trial design), and whichever one we do not observe is the 'counterfactual outcome'. Hence, the counterfactual outcome for each individual can be defined as the outcome that would have been observed under a hypothetical condition (e.g. if the individual had not received the treatment or *vice versa*).

1.3.2 Causal effect

Estimating the causal effect of treatment is the main objective of most epidemiological studies using data from RCTs. However, we cannot estimate the causal effect for an individual patient, but we can estimate the average causal effect (ACE) of treatment for people like that individual - this is known as the population causal effect.

The population causal effect is the primary interest in estimating clinical effectiveness, and I refer to this as the 'causal effect', or ACE, throughout this thesis. It is worth noting that the term 'population' can refer to different groups of individuals depending on the causal question and objectives of the study. The causal effect may involve a 'direct' effect of the exposure (e.g. residence in mouldy dwelling) on the outcome (e.g. depression). It may also involve an 'indirect' effect of the exposure on the outcome via an intermediate variable (e.g. the indirect effect of mouldy dwelling on depression via affecting physical health). The total effect combines direct and indirect causal effects. The term 'causal inference methods' is used in this thesis to refer to the methods originating in the causal inference literature and this is further clarified in Chapter 2.

1.3.3 Estimand, estimator and estimate

The 'estimand' is defined as the parameter of interest that we can use to make inferences about a population using a sample from that population.[13] In the context of RCTs, the estimand can be defined using four attributes:[14, 15] (i) the population (patients of interest targeted by the scientific question); (ii) the outcome variable or endpoint (e.g. time to incidence of graft loss); (iii) the specification of how to deal with intercurrent events (e.g. include compliers only); and (iv) the population-level summary of the outcome variable, that is, the measurement of the causal effect of the intervention (e.g. hazard ratio [HR]). Using the examples given above, a possible estimand would be the HR relating to graft survival in the compliers only. The estimand is also known as the 'target of inference'.

In contrast, the 'estimator' is the method of estimation (e.g. maximum-likelihood estimator [MLE]) that can be applied to any sample from a population to produce a numerical value for the estimand. The estimate is the numerical value of the estimand obtained from a particular sample (e.g. HR=0.78).

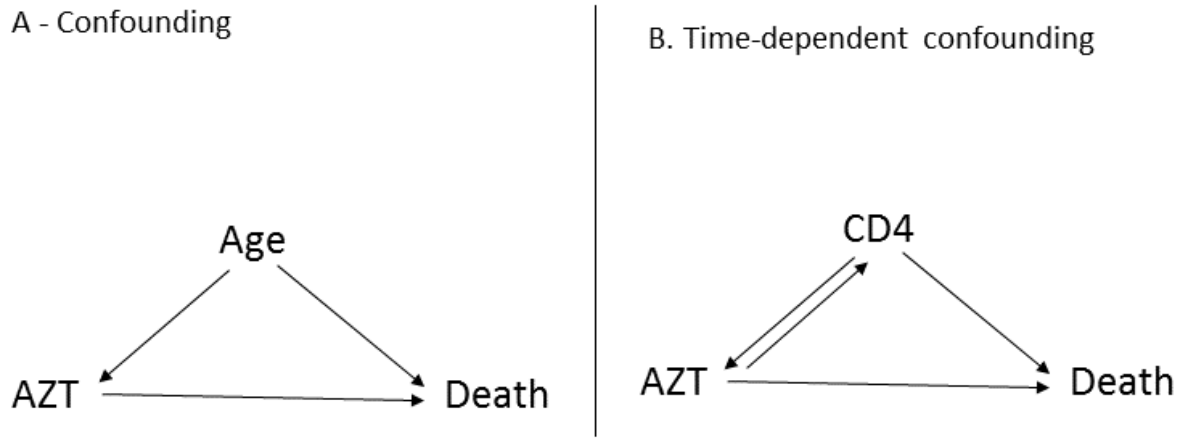
1.3.4 Confounding

Confounding is a major risk for causal inference in epidemiological studies (including RCTs), and failure to control for it may lead to a biased estimate of the ACE of treatment. To better understand the concept of confounding, I use the example in Figure 2 representing the causal effect of azidothymidine (AZT) treatment (an antiretroviral drug used for prevention and management of human immunodeficiency virus [HIV]/acquired immunodeficiency syndrome [AIDS]) on the survival of HIV-positive patients with time-independent and time-dependent confounding variables (see Figure 2). In this example, Figure 2A shows 'age' as a time-independent confounder. In this figure, AZT treatment

has a direct causal effect on mortality (as indicated by the arrow). However, this causal relationship is confounded by age which is a common cause of the exposure (adherence to AZT treatment) and the outcome (death). In this case, the analyst should adjust for age using traditional methods such as stratification or regression analysis.

Figure 2B shows time-dependent confounding where the cluster of differentiation 4 (CD4 count - a measure of the immune system functionality used in HIV/AIDS diagnosis) is considered as a time-varying confounder for the effect of AZT treatment on death. CD4 count is a time-dependent confounder because: (a) CD4 count affects AZT prescribing, and AZT treatment affects CD4 count; and (b) CD4 count predicts survival outcome (time to death). In this case, standard methods for estimating the ACE of time-varying AZT treatment on survival may produce biased estimates, and therefore, more advanced methods are needed.

Figure 2: Time-independent and time-dependent confounding



Note: In Figure 2A, confounding refers to time-independent confounding

Sterne and Tilling[16] characterised the confounding bias by CD4 count and its implications using three possible analytical strategies:

(i) Crude analysis (without controlling for CD4) will produce a biased estimate because AZT is more likely to be given to patients with higher CD4 count who tend to experience a higher death rate

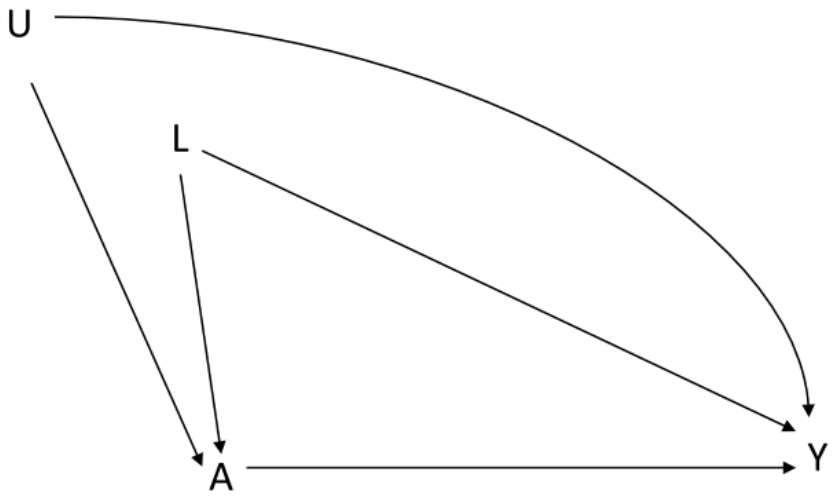
(ii) Controlling for CD4 count at baseline will also produce a biased estimate because it ignores the fact that patients who actually initiate the treatment after baseline are more likely to be those who had higher CD4 count (immunosuppressed patients).

(iii) Controlling for time-dependent CD4 count is problematic because the causal effect of AZT is partly mediated via time-updated CD4 count (Figure 2B) and controlling for it will block the indirect causal effect AZT→CD4→Death causal path resulting in an underestimate of the total treatment effect of AZT treatment.

1.3.5 Directed-acyclic graphs

DAGs, which are generally known as 'causal graphs', are a graphical representation used in causal inference to conceptualise the causal links between the treatment, outcome and other variables.[13] To illustrate the concept of DAGs, I use a generic example (see Figure 3) conceptualising the causal effect of treatment A on outcome Y, with two other random variables L and U representing measured and unmeasured confounders, respectively. The nodes in the DAG representing variables A, L, Y, and U are called '*vertices*'. The variable A (for example) is called an *'ancestor'* of variable Y because it has a '*directed*' path arrow leading to variable Y, and therefore, Y is called a '*descendant'* of A. The unmeasured confounder U (e.g. genetics) occurs prior to treatment A and the outcome Y assuming that time goes from left to right as illustrated in the DAG.

Figure 3: Directed-acyclic graph (DAG)



This causal DAG has five arrows representing directed causal relationships. For example, treatment A has a direct (non-mediated) causal relationship with outcome Y, that is the causal effect of interest. Each arrow connecting two vertices (variables) in the DAG is called an 'e*dge*' or an '*arc*'. Causal DAGs are *'acyclic'* because there are no loops which imply that a variable in the DAG cannot have a direct or

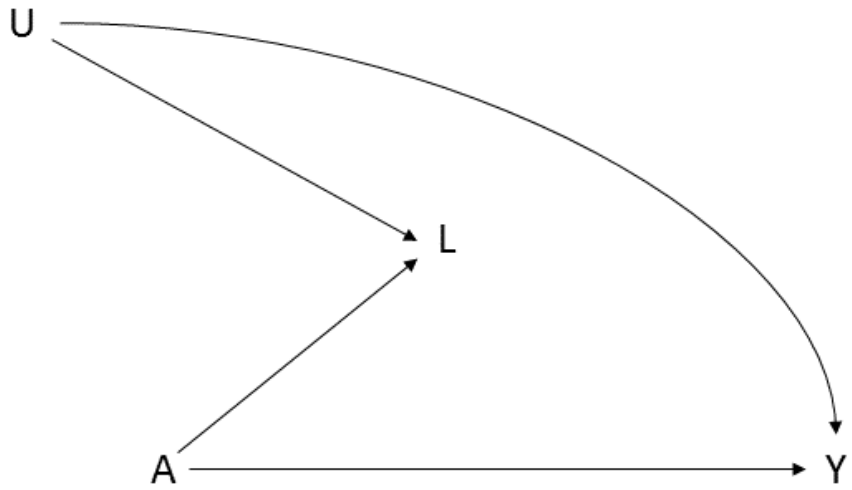
indirect causal effect on itself (i.e. no cycles). The A→Y causal path is called a *'frontdoor path',* and this causal relationship is confounded by variable L via a '*backdoor path'* A←L→Y. The backdoor path is also called a *'non-causal path'* through which an association probability can flow from A to Y via L, creating a confounding bias and the analyst can control for this by conditioning on L (see Section 1.3.4). Similarly, A←U→Y is an open backdoor path which can create confounding bias, but in this case, we cannot control for it because U is unmeasured. To address this problem, the investigator may find an intermediate variable M that mediates the U→A or U→Y causal relationships and control for M in order to control for the confounding effect of U. This indicates that investigators need to identify and collect sufficient data on all variables that have important causal relationships in the causal network as conceptualised by their particular DAG. The main disadvantage of DAGs is that they describe causal relationships based on discrete times which may not reflect the continuous process of cause-effect relationships, but DAGs can include time-varying events as an attempt to overcome this shortcoming.[17] Despite this limitation, causal DAGs can be useful in identifying important variables that need to be measured for use in estimating an unbiased causal effect of treatment.

1.3.6 Selection bias

Selection bias is another risk for causal inference caused by an association link which may be created by the process of selecting individuals included in the analysis.[13] In this section, I introduce two specific types of selection bias: (i) *'conditioning on a collider bias';* and *(ii) 'confounding by indication'*. Conditioning on a collider bias is a specific type of selection bias that is distinct and often confused with confounding bias.[18] To introduce this concept, I use the causal DAG presented in Figure 4 representing a direct causal effect of treatment A on outcome Y via the frontdoor path A→Y. The variable L is a descendant of treatment A and unmeasured variable U. Consequently, the causal effects of A and U collide on L, and in this case, L is called a 'collider'. This means that probability does not flow and there is no association between U and A in Figure.

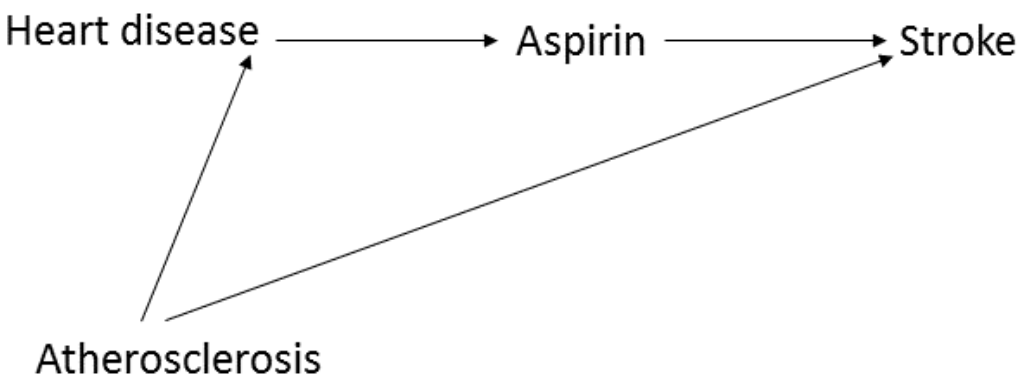
In this particular DAG, variables L and U have no confounding effect on the causal effect of A on Y because the backdoor path A→L←U→Y is already blocked by the collider variable L. In other words, conditioning on L will essentially create a backdoor path from A to Y through L←U. This indicates that the analyst should not condition on L because conditioning on it will open the closed backdoor path creating selection bias - this is called conditioning on a collider bias. Therefore, it is very important to investigate each variable in the causal network to ensure that it does not meet the conditions of being a collider before inclusion in the adjustment model.

Figure 4: A causal DAG illustrating conditioning on a collider bias



Confounding by indication is another form of selection bias which is common in the context of clinical decision analysis. This is best described using the causal DAG in Figure 5. In this DAG, the causal effect of treatment (Aspirin) on the outcome (risk of stroke) will be confounded because Aspirin treatment is commonly prescribed to patients with heart disease, which is both an indication for Aspirin treatment and a risk factor for the outcome (stroke). That is because both heart disease and stroke are affected by the common cause atherosclerosis (see Figure 5).[13] Confounding by indication is also known as *'channelling'*, a term commonly used to describe confounding bias in specific situations where the selection bias is due to patient risk factors that influence doctors to prescribe a particular treatment within a class of drugs.[13]

Figure 5: A causal DAG illustrating confounding by indication



1.3.7 Censoring

In epidemiological research, survival analysis is used to estimate the ACE of treatments on time-to-event outcomes. Censoring is an important issue in analysing RCT data with a time-to-event outcome. An individual is said to be censored at a particular time point when data on the time-to-event outcome is unknown due to reasons such as loss to follow up or non-occurrence of the outcome before the end of the study. End of study censoring is known as 'administrative censoring' which is likely to be non-informative in RCTs. Administrative censoring can be addressed using standard methods of survival analysis that allow for data from censored observations to be included in the analysis.[19] Standard parametric survival models assume that censoring is 'non-informative', that is the survival-time outcome is prognostically independent of censoring. In other words, patients censored at a particular time point have the same probability of experiencing the event of interest compared to uncensored patients at the time of censoring.[19] 'Informative censoring', on the other hand, implies that the survival outcome and time-to-censoring are dependent, for example, if a patient is censored when they stopped taking the treatment (discontinued) or switched to another treatment.[19, 20] This is also referred to as 'dependent censoring' which is likely to introduce confounding bias when standard survival analysis methods are used.[19] Because censoring happens over time, potential time-dependent confounding is an issue, and therefore, more advanced methods than standard regression adjustment may be needed to address this particular problem.

1.3.8 Drug forgiveness

Drug forgiveness is defined as the number of consecutive doses that can be missed by a patient while still maintaining the intended therapeutic effect.[21, 22] To determine the drug forgiveness, the analyst needs to define the adherence level above which the therapeutic effect will be maintained (i.e. adherence threshold). The adherence threshold can be objectively determined using the PKPD characteristics of the drug. However, there is no universal adherence threshold that applies to all drugs as this will be drug/disease-specific. For drugs with a short duration of action, there may be no adherence threshold (i.e. adherence threshold equals 100%). Formally, drug forgiveness (F) can be calculated as the difference between the postdose duration of beneficial action of the medication (D) and the prescribed dosing interval (I) using the following equation.[21, 23]

$$F = D - I \qquad [1]$$

1.4 What is the problem?

This section describes the impacts of non-adherence on clinical outcomes, treatment costs and cost-effectiveness based on existing evidence. The difference in adherence levels in the real world compared to trials and the problem faced in HTA (due to impacts on costs, effects, different adherence levels) is provided.

1.4.1 Existing evidence

Non-adherence to prescribed medications is a major and intractable problem in health care, with significant negative consequences for both clinical outcomes and health care costs.[24, 25] To identify the relevant evidence, I undertook a literature review that identified 43 papers - see Appendix A1 for details of the search strategy and sifting. This review was used to provide a summary of the existing evidence on the impact of non-adherence on clinical outcomes, costs and cost-effectiveness (see Section 1.4.2 and 1.4.3). The differences between adherence levels in clinical trials and the real world are discussed in Section 1.4.4.

1.4.2 Impact of non-adherence on clinical outcomes

Non-adherence to prescribed treatment may have a major negative impact on clinical outcomes by affecting the link between the process of healthcare provision and the outcome.[26] The clinical consequences of non-adherence may be determined from the complex interplay of three factors: (i) the type(s) of non-adherence; (ii) the nature of the disease, and (iii) the pharmacokinetics (PK) and pharmacodynamics (PD) of the drug.[27] In simple terms, PK is the study of the drug's path through the body in terms of absorption, distribution, metabolism and excretion (i.e. what the body does to the drug). PD is the study of the biochemical and physiological effect of the drug (i.e. what the drug does to the body).[28] A review of 10 cohort studies on diabetes mellitus revealed an increase in hospital admissions of 10.3% and 15% when the MPR, a measure of adherence, was lower than 80% and 40%, respectively.[29] Kidney transplantation is an important chronic condition where the implications of non-adherence to maintenance immunosuppressive therapy are significant. These may include: rejection of the transplanted kidney; return to dialysis, and potentially death.[30]

1.4.3 Impact of non-adherence on cost-effectiveness

The economic consequences of non-adherence are driven by the reduced efficacy of the drug leading to an increased probability of therapeutic failure, and consequently, impacts on cost-effectiveness.[27] In the USA, crude estimates reported the prevalence of non-adherence as 30-50% of all patients on chronic medications, with the total cost of all non-adherence estimated at $100 billion annually.[25] The cost of non-adherence to just 10 medications was estimated between $396 and $792 million per year.[25]

In England, findings from a modelling study aimed at estimating the impact of non-adherence for five conditions were striking.[31] These conditions were: asthma; hypertension; diabetes; schizophrenia and cardiovascular diseases. In this study, non-adherence was found to be associated with both increased cost and deterioration in health outcomes. For many of these conditions, the treatment cost alone associated with non-adherence was estimated to be more than £100 million per year.[31] The study suggests that investment to improve adherence appears to be a strong case in asthma and schizophrenia in particular, where the expected annual cost savings to the NHS are significant (£130 million and £190 million, respectively).[31] If there is a high level of discontinuation (non-persistence) to an expensive medicine, there may be savings that outweigh the costs of reduced benefit. Many treatments are marginal/not cost-effective, and not taking them e.g. in patients who do not perceive a benefit, could be cost-saving.

I have identified a relevant systematic review (Hughes et al.[32]) which outlines a number of methods used for modelling non-adherence in cost-effectiveness analyses. However, there are three main limitations associated with the Hughes et al.[32] review: (i) it is more than 10 years old, and newer methods may have been proposed since its publication; (ii) the literature search was run on two databases only (MEDLINE and NHS EED); and (iii) the review scope was restricted to pharmacoeconomic evaluations and therefore does not capture relevant papers from the clinical and statistical literature unless they also relate to pharmacoeconomic evaluations. These approaches and their limitations are discussed in Section 1.6.2.

1.4.4 Adherence levels in clinical trials and the real world

RCTs are considered the gold standard for providing unbiased estimates of the ACE of treatment as a key component of economic evaluations for HTA. However, a key issue with the evidence from RCTs is their generalisability to routine clinical practice.[33] A major factor which could impact on treatment

effectiveness is patient non-adherence. Evidence suggests that adherence levels in the real world are likely to differ from clinical trials[34] – this leads to uncertainty in the actual effectiveness of treatments. The term 'real world' (and its popularisation in the pharmaceutical literature) is increasingly used in the context of health care decision making and I use it interchangeably with 'normal clinical practice' in this thesis. In this context, non-adherence varies depending on the type of treatment, disease area, and health-care setting.[33, 35] For instance, in a case study of statins, adherence levels based on evidence from pivotal trials with 4 to 6 years follow-up were estimated to range from 81% to 99%, whereas in pragmatic studies in community settings, the level of adherence ranged from 21% to 87%.[33]

1.5 Why is it important?

In the context of HTA, it is also well recognised that most health economic models do not adequately address the issue of non-adherence.[33] In many cases, non-adherence observed in RCTs is either assumed to be applicable outside of a clinical trial setting or addressed using very simplistic approaches (e.g. assuming a proportionate reduction in treatment effect with non-adherence rate). The key issue is that the simplistic approaches are unlikely to produce robust and externally valid estimates of the treatment effect due to the complexities associated with patterns of non-adherence and their link to time-dependent confounding and clinical outcomes. This could lead to any associated cost-effectiveness analysis producing misleading results which may lead to the wrong decisions being made about the allocation of scarce healthcare resources.

In the HTA context, resource allocation decisions are usually made for a specified population defined by the scope for each decision problem. In this context, HTA needs to consider effectiveness under real-world conditions, where non-adherence levels typically differ from those observed in clinical trials. With this analytical requirement in mind, methods for adjusting estimates of treatment effect in the presence of patient non-adherence needs to be assessed for suitability for use in HTA. This will require defining features of HTA that will influence the choice of methods (e.g. population estimates and relevance to clinical practice in the real world). This is better defined as the 'appropriateness' of non-adherence adjustment methods for the HTA context and this concept is further developed in Chapter 3, Section 3.7.

1.6 What are the potential solutions?

1.6.1 Existing evidence: accounting for non-adherence in estimating the clinical effectiveness of treatments

ITT and Per-Protocol (PP) analyses are the most common analytical approaches used for estimating the ACE of treatment using data from RCTs. However, ITT analysis may fail to produce an unbiased estimate if the estimand of interest is the effectiveness of the treatment in the presence of non-adherence in standard clinical practice. PP is a different analytical approach as it answers the question of effectiveness according to adherence to the trial protocol (which may or may not include patient non-adherence to the assigned treatment as defined by the estimand). PP is an unbiased estimate of the effectiveness of following the trial protocol, conditional on the assumption that protocol violations are non-informative. If this assumption is violated, then the analyst should consider alternative methods to adjust for non-adherence. In the methodological literature, several papers have introduced alternative methods for correcting estimates of the causal effect of treatments in the presence of non-adherence. For instance, proposed adjustment methods include the g-computation algorithm, g-estimation, inverse probability weighting, structural nested models and marginal structural models.[13] Many of these methods have been described and/or compared empirically,[36, 37] in simulation studies and case studies.[38-40] Latimer reported a systematic review of methods for adjusting the causal effect of treatments in the presence of treatment switching which has identified a range of relevant methods.[41] However, to the best of my knowledge, there is no existing systematic review of the alternative methods explicitly used for adjusting estimates of the causal effect of treatment for non-adherence on time-to-event outcomes. This research contributes to filling the gap and provides new evidence on this area (see Chapter 2).

1.6.2 Existing evidence: accounting for non-adherence in estimating the cost-effectiveness of treatments

The health economics literature has adopted five different methodological approaches for correcting estimates of the cost-effectiveness of treatments for non-adherence, as characterised by Hughes and colleagues (see Table 1).[32] The classification of these approaches is not perfect, and there is substantial overlap between the categories. The classification was partly based on the implementation of methods in pharmacoeconomic evaluations (including trial-based and model-based analyses) rather than on adjusting for the impact of non-adherence on treatment effects. Nevertheless, this was very

helpful in identifying current practices for modelling non-adherence in the health economics literature. These methods are described in Table 1.

These existing approaches are mostly based on attempting to handle the non-adherence problem using structural assumptions in economic models, rather than by estimating the causal effect. In the health economics literature, the 'traditional' method is to apply a sensitivity analysis around treatment effect and costs. Generally, these approaches share the same characteristics and limitations. The key limitation is making strong assumptions about the causal relationship between patient non-adherence and treatment effect. For example, some approaches assigned reduced treatment effects proportional to the level of non-adherence (e.g. 20% reduction in treatment effect for 20% lower levels of adherence). This does not take into account the complexity of the relationship between medication adherence and treatment effect as discussed in Section 1.4.2. Consequently, those simplistic approaches are more likely to produce misleading cost-effectiveness results, leading to suboptimal allocation of scarce health resources for the NHS and other healthcare systems around the world. The summary of each approach is described in Table 1 including the design of the economic model, the disease area where the model was applied and how non-adherence was incorporated into the economic model. The PKPD modelling approach is more promising and the theoretical characteristics and application of this method are further discussed in Chapter 2.

Table 1: Summary of methodological approaches used to account for the impact of non-adherence on cost-effectiveness

Methodological approach	Description of the methodological approach	Disease area in which the methodological approach was used	The economic model design used	Incorporation of non-adherence into the economic model
Regression-based covariate adjustment	By modelling adherence as an interaction term, the regression coefficient could be used to estimate the incremental net benefit of treatment at different levels of adherence.	Urge incontinence [42] and Tuberculosis [43]	The net benefit regression model, [32] empirical treatment effect model [42] or structural mean model [44]	Including adherence metrics as a covariate in a regression model using individual patient-level data on costs, health outcomes and adherence. [32, 45]
Categorization of patients into clusters with different levels of adherence	Modelling relapse rates as a function of non-adherence and treatment effect; or spontaneous remission rate for those who discontinued treatment after a specified period and higher remission rates for premature discontinuation. [46]	Schizophrenia, [47] Severe depression, [46, 48] Childhood attention deficit hyperactivity disorder [49]	Decision tree model	Differentiating modelled patients to broad strata of adhered and non-adhered, or by different levels of adherence (i.e. ≥80%, 50% – 80% and <50%) based on tablet counts or other measures of adherence. [46, 47] Branches of the decision tree may be used to model different levels of adherence. This approach requires data on the relationship between adherence and the outcome based on evidence from the literature
Incorporating non-adherence as a modifier in deriving the rate of transition between health states	Adjustment of disease progression rates by level of adherence; proportionate reduction in outcomes with non-adherence rate and assuming therapeutic failure for non-adhered patients.	Urge incontinence related with overactive bladder, [50] Idiopathic epilepsy and epileptic syndromes, [51] and Tuberculosis [52]	Markov model	For each cycle in a Markov model, a proportion of patients with a specified level of non-adherence experience a higher probability of disease progression compared to those who adhered to the assigned treatment [27]. Movement between the modelled health states is determined by a set of transition probabilities with non-adherence modelled as one parameter in deriving transition rates
Specifying non-adherence as an event within time-based simulation models	The model simulates time-dependent adherence variable based on patient history where patients were considered as either 'adhered' or 'non-adhered'.[53] Patients with greater adherence were assigned a decreased probability of suffering a relapse.	Schizophrenia [53]	Discrete event simulation model (DES)	Within the DES model, patients are specified as entities and non-adherence as time-based events. [35, 54] This approach allows patient attributes to be assigned which could be altered over time based on continuous measures of non-adherence. DES also allow for the interaction between non-adherence, time and individual patient characteristics to be modelled
Characterising non-adherence effect on the Pharmacokinetics (PK) and Pharmacodynamics (PD) parameters	Quantifying variability in drug exposure according to different adherence profiles for time-averaged and trough drug concentrations (i.e. C_{avg} and C_{min}) and the percentage of days spent below C_{min} (PK); [55, 56] and quantifying adherence effect (as a covariate) on dose-response (PD).[57]	Kidney transplantation [55] and HIV/AIDS [58]	PKPD modelling	Individual doses for each patient are simulated according to a particular pattern of adherence or different adherence profiles, and the PK and PD consequences are then propagated for estimating the reduction in treatment effect. [27, 33, 58, 59]

There are limitations associated with the identified approaches summarised in Table 1. Generally, they are different ways of attempting to model the impact of non-adherence using structural models (Decision trees or Markov models) based on simplifications that might not reflect the complexity of the relationship between non-adherence levels and treatment effect. The key issue is that they make strong assumptions about the causal links between adherence levels and treatment effect rather than adjusting the treatment effect for non-adherence to obtain valid estimates for incorporation into the economic model. Moreover, guidance is needed to provide researchers with the data requirements and analytical steps to properly adjust for the impact of non-adherence in economic evaluations for HTA. This gap in the health economics literature has motivated me to undertake this doctoral study as a contribution to resolve this issue.

1.6.3 The gaps in the methodological literature

Despite the increased attention to modelling non-adherence in the literature, Muszbek and colleagues[3] suggest that the impact of non-adherence on the cost-effectiveness of medications is inconclusive, and therefore, further research is warranted to resolve the issues. Despite this, methods to account for non-adherence in economic evaluations remain underdeveloped, with no consensus on the appropriate approach to deal with the key issues.

The National Institute for Health and Care Excellence (NICE) produces evidence-based guidance to NHS England on the clinical effectiveness and cost-effectiveness of health technologies including pharmaceuticals. The NICE Guide to the Methods of Technology Appraisal (2013) recommends that the value of additional benefits from the mode of treatment delivery through its impact on adherence should be quantified.[60] However, the guide does not mention any preferred method(s) for modelling non-adherence. There seems to be a gap in the health economics literature on conceptualising and modelling the link between non-adherence and treatment effectiveness, and hence, the impact on cost-effectiveness.

1.7 Aim, objectives and research questions

1.7.1 Research questions

This doctoral research study aims to address the following research questions:

a. What are the key methodological approaches used to account for the impact of non-adherence on the effectiveness and cost-effectiveness of health technologies used in chronic conditions with time-to-event outcomes?
b. What is the relative performance of the alternative methods in estimating the impact of non-adherence on treatment effectiveness?
c. How should economic evaluations incorporate the impact of non-adherence using evidence from both RCTs and real-world data?

1.7.2 Aim

The aim of this doctoral research study is to develop a methodological framework to account for non-adherence to prescribed chronic medications with time-to-event outcomes when undertaking economic evaluations for HTA.

1.7.3 Objectives

a. To undertake a systematic review to identify relevant non-adherence adjustment methods used in analyses of clinical effectiveness and cost-effectiveness (Chapters 2-3).
b. To assess the relative performance of alternative adjustment methods using simulated RCT datasets (Chapters 4-5).
c. To adapt an economic model as a case study for estimating the adherence-adjusted cost-effectiveness of immunosuppressants used as maintenance therapy for kidney transplantation in adults (Chapter 6).
d. To develop a methodological framework outlining the appropriate methods for incorporating non-adherence into economic evaluations for HTAs (Chapter 7).

1.8 Expected contributions of the thesis

The main contribution of this research study will be the development of the methodological framework. This framework will help to improve the overall quality of economic models used for estimating the cost-effectiveness of prescribed chronic medications with time-to-event outcomes. The framework will provide guidance to academic researchers and the pharmaceutical industry for choosing the appropriate method to incorporate non-adherence into cost-effectiveness analysis, providing better evidence for healthcare decision makers such as NICE Technology Appraisal Committees. Better evidence about adherence-adjusted cost-effectiveness will lead to improvements in healthcare decision making, patients' quality of life, reduction in mortality risk and cost savings to

the NHS and other healthcare systems, and ultimately improvements in population health through better healthcare resource allocation.

Other expected contributions include: (i) systematic review evidence on methods for adjusting estimates of the causal effect of treatment in the presence of non-adherence for time-to-event outcomes; (ii) simulation evidence on the relative performance of alternative adjustment methods; and (iii) new evidence on the long-term adherence-adjusted cost-effectiveness of immunosuppressants used after kidney transplantation for adults in the UK based on data from an RCT and the real-world used within a decision-analytic model.

1.9 Thesis structure

This thesis is structured into eight chapters including this introductory chapter. Chapter 2 reports methods identified to adjust estimates of treatment effectiveness in the presence of patient non-adherence based on a systematic review of methodological papers. Chapter 3 compares the alternative non-adherence adjustment methods based on existing evidence identified by the systematic review. Chapter 4 describes the design and implementation of the simulation study. Chapter 5 presents and discusses the results of the simulation study. Chapter 6 presents the case study on kidney transplantation. Chapter 7 presents the methodological framework put forward in this thesis to account for the impact of non-adherence on the cost-effectiveness of prescribed chronic medications. Chapter 8 provides a recap of the thesis and highlights the contributions in the context of the health economics literature, and draws the overall conclusions.

Chapter 2: Systematic review of statistical methods for adjusting estimates of treatment effectiveness and cost-effectiveness for patient non-adherence

2.1 Introduction

This chapter presents a systematic review which was undertaken to identify the key methodological approaches used to account for the impact of patient non-adherence on the causal effect of treatments. The intention of the review was to systematically identify each relevant method published in the methodological literature. The review followed the 'Comprehensive Pearl Growing' (CPG) search technique[62] as described in Section 2.3.2.1. The review also followed guidance from the Centre for Review and Dissemination (CRD) on undertaking systematic reviews.[63]

This chapter is structured as follows. Section 2.1 (this section) introduces the topic and set out the aim and objectives of the review. Section 2.2 specifies the review question. The review methods are then introduced in Section 2.3, including the review protocol, search strategy, selection criteria, quality assessment, data extraction and data synthesis. The results of the review are presented in Section 2.4. These include the search results and details of included and excluded papers. A taxonomy of methods for adjusting estimates of treatment effectiveness for non-adherence in the context of time-to-event outcomes is proposed in Section 2.5. A summary of estimands and key assumptions of identified methods is presented in Section 2.6. A narrative synthesis of the identified adjustment methods is provided in Sections 2.7-2.10. The discussion and conclusions of the review are presented in Section 2.11.

In the methodological literature, a range of methods has been proposed for adjusting the causal effect of treatments[36, 37, 64] and cost-effectiveness[27, 32] in the presence of patient non-adherence. Most of these methods that have been developed attempt to estimate what treatment effectiveness would have been in the absence of non-adherence, a term that is often referred to as "adjusting" in this thesis. In the HTA context, the analyst needs to consider effectiveness under real-world conditions, where non-adherence levels typically differ from those observed in trials and I use the term

"accounting for" to refer to this in this thesis. However, it should be noted that the terms are sometimes used interchangeably.

Some methods are not making adjustments at all such as ITT analysis. Other methods use the CACE estimand by simply estimating the effect in compliers. The standard IPCW is estimating the effect in the population, had they all complied. However, in the HTA context we are interested in methods that estimate the effect in the presence of non-adherence (e.g. real-world adherence levels), so that adjusted effectiveness estimates can be used in economic models. In other words, I am interested in finding methods that give the relevant effect estimates that allow the analyst to make adjustments for different adherence levels. However, all methods identified are reviewed with this analytical requirement in mind and then the appropriateness of each method to the HTA context is assessed in the next Chapter (Chapter 3). The subset of appropriate methods is then carried forward for assessing their relative performance in the simulation study (Chapters 4-5) and further applied in a case study (Chapter 6).

Several papers have compared some of these methods empirically, in simulation studies and/or case studies. However, to the best of my knowledge, this is the first systematic review which has attempted to identify all the existing methods for adjusting the causal effect of treatments for non-adherence in the context of time-to-event outcomes.

The aim was to undertake a systematic review to identify relevant non-adherence adjustment methods used for estimating clinical effectiveness and cost-effectiveness. The focus is on adjusting for the impact of non-adherence using individual patient-level data (IPD) in the context of RCTs and cost-effectiveness analyses.

The objectives were:

- To identify potentially relevant papers used in the methodological literature to adjust the causal effects of treatments and/or cost-effectiveness in the presence of non-adherence.
- To select relevant papers to be included in the review for narrative synthesis using the selection criteria (see Section 2.3.4).
- To appraise the papers included in the review using a framework for critical appraisal of methodological papers.
- To identify relevant methods for adjusting for non-adherence and extract the key characteristics for each method identified.
- To undertake a narrative synthesis for each method identified and develop a taxonomy of non-adherence adjustment methods.

2.2 Review question

The review question was: *"What methods have been proposed in the methodological literature to account for the impact of non-adherence to treatments on clinical effectiveness and cost-effectiveness?"* Specifically, the review was focused on methodological papers reporting methods to adjust for non-adherence in estimating the causal effect of treatment and cost-effectiveness for time-to-event outcomes.

2.3 Review methods

2.3.1 Review protocol

A review protocol was developed setting out the methods used for undertaking this review. The review protocol was agreed with the supervisory team and advisors (Professor Dyfrig Hughes [DH], Professor Ian White [IW], Dr James Fotheringham [JF] and Dr Ruth Wong [RW]). The protocol was published on the CRD's international prospective register of systematic reviews (PROSPERO) database.[65] The review methods were pre-specified in the protocol to reduce the risk of bias in conducting the review. The review therefore strictly followed the published version of the protocol and no protocol amendments were made during the review process. The search strategy is summarised in Section 2.3.2, and more details are published in the protocol (available from http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018095544).

2.3.2 Search strategy

The search strategy involved a scoping search, database searching, citation searches and reference list checking and advice from methodological experts. These strategies are described in the following subsections.

2.3.2.1 Pearl growing iterative search technique

The CPG search technique was used to identify potentially relevant papers. The CPG process was described by Schlosser et al.[62] and involved the following steps:

a) The search started with a compilation of nine relevant papers ("pearls").[1, 26, 32, 36, 53, 57, 66-68] These pearls were identified from a citation search for three key papers[26, 32, 69] and a scoping search using terms extracted from these papers. The three papers were originally identified as highly cited publications which address the problem of non-adherence and further agreed

with an expert in the area of medication adherence research – Professor Dyfrig Hughes (the first author of the previous review[32]);

b) Relevant electronic databases were determined to specify where the identified pearls are indexed. This was done using the Ulrichweb tool http://ulrichsweb.serialssolutions.com/

c) I determined how each pearl is indexed (by the journal) in database 'X' including MeSH heading, author-assigned keywords and index terms. This analysis was done using the Yale MeSH Analyzer http://mesh.med.yale.edu/ and the Online-Utility Text Analyzer https://www.online-utility.org/text/analyzer.jsp;

d) Steps (c) was repeated in other databases;

e) Other relevant pearls in database X were found by undertaking two search iterations using new search terms identified from the papers identified in steps (c) and (d); and

f) The search was terminated when the point of saturation was reached (when retrieval of relevant articles diminishes). This diminishing rule was applied after discussion with the supervisory team, as specified by the protocol.

2.3.2.2 *First search iteration*

The first iteration of searching started with terms and MeSH headings identified from the analysis of pearls.[1, 26, 32, 36, 53, 57, 66-68] These included terms relating to patient adherence (compliance) to treatments. Specifically, the adherence terms used were: compliance, adherence, pharmacoadherence, persistence, persistency, concordance, initiation, implementation, noncompliance, nonadherence, nonpersistence, discontinuation, pharmionics, therapeutic alliance, patient irregularity or treatment refusal. MeSH headings and methods terms used were: "models, statistical" or "models, structural" or "models, economic" or "models, econometric" or "models, biological". The full search strategy applied in the first iteration is provided in Appendix A2.

Six databases (*Medline, Embase, Cochrane Library, EconLit, Scopus, Web of Science)* were searched for potentially relevant papers published in English from inception to an end date between 9th February to 8th May 2018 (see Appendix A2 for the exact end date used for each database). The database searches were complemented by citation searches and reference list checking for all relevant papers identified at that stage. The citation searching and reference list checking were conducted using the *Web of Science* database.

2.3.2.3 *Second search iteration*

The search strategy for the second iteration was informed by the results from the first search iteration. Specifically, the new pearls identified were analysed using the MeSH and Text Analysers[70, 71] as

outlined in Stage (c) of the CPG search process. Consequently, new search terms were identified and used in the second iteration database searching (e.g. “survival analysis”, “proportional hazards models”, “logistic models”). In addition, two methodological terms were recommended by expert advisors (“causal inference” and “pharmacometric”). All new terms incorporated in the second iteration search strategy are provided in Appendix A3.

As specified by the CPG search process (step c), I determined how each pearl identified (by the end of the first iteration of searches) was indexed in databases. Based on discussions of the indexing analysis results with the supervisory team (informed by the review protocol and expert advice), I decided to limit the second iteration searches to four databases (*Medline, Embase, Web of Science and MathSciNet*). This meant that three databases searched in the first iteration were excluded (Cochrane Library, EconLit and Scopus) and a new database *(MathSciNet)* was included. *EconLit* and *Cochrane* were excluded because no/few relevant papers were identified from these databases in the first iteration. *Scopus* was excluded to minimise duplicates as all papers identified by this database were also indexed in the *Web of Science* database. *MathSciNet* was included because many relevant papers were indexed on this database; therefore, it was added to avoid missing relevant papers. *Science Citation Index* was identified as a relevant database, but it was not included because it is a subset of the *Web of Science*.

The four databases were searched for potentially relevant papers published in English from inception to the end date of 22nd May, or 23rd May 2018 (see Appendix A3 for exact end dates). The second iteration database searches were complemented by citation searches and reference list checking for the new relevant papers identified.

2.3.2.4 Second reviewer and expert advice

My primary PhD Supervisor (Professor Simon Dixon) double-checked a set of papers which were initially marked as “excluded” based on the inclusion and exclusion criteria. This was followed by discussion and agreement on decisions to exclude or include each one of these papers.

As a final check to ensure that no relevant methods were missed by search, the list of identified papers was checked by two expert advisors. The expert advisors were: (i) Professor Ian White (Institute of Clinical Trials & Methodology, University College London); and (ii) Professor Dyfrig Hughes (Centre for Health Economics and Medicines Evaluation, Bangor University). The expert advisors recommended an additional list of potentially relevant papers which were further assessed against the inclusion and exclusion criteria.

2.3.3 Data management

EndNote bibliographic software (*version X8.2*) was used for managing references retrieved from different sources. This included removing duplicates, finding and storing full-text articles, coding records into groups and combining records identified from all sources in a single library for use in data synthesis and referencing.

2.3.4 Selection criteria

The selection of papers included for narrative synthesis was conducted in two stages: (i) records retrieved from all sources were screened by titles and abstracts; and (ii) potentially relevant full-text articles were assessed for eligibility. The PRISMA flow chart was used for reporting the selection process.[72] The inclusion and exclusion criteria (as specified by the review protocol) were applied for selecting the relevant papers included in the review (see Table 2).

Table 2: Inclusion and exclusion criteria for the selection of papers included in the review

No	Inclusion criteria	Exclusion criteria
1	Peer-reviewed methodological papers which describe the method(s) in detail such that they can be applied without the need for further assumptions;*	Non-peer reviewed reports, books or book chapters, theses, or other grey literature;
2.	Methods explicitly applied to adjust for non-adherence in estimating treatment-effects for survival-time outcomes and/or cost-effectiveness;	Papers which merely apply previously developed method(s) without any additional extension to the original method(s)+;
3.	Papers published from databases inception to date; and	Methods which are not explicitly applied to adjust for non-adherence to treatments
4.	Papers published in the English language	Methods based on aggregated data such as meta-analysis; or
5.		Theoretical papers with no application of the method.

** This criterion is not objective and required a judgement on my part informed by expert advice*

+ This implies that the first paper proposing the method is included and any paper published afterwards with the application of the method without any methodological extension is excluded. Although these papers were excluded, I retained them because they might have explained the method better than the first paper and further used a subset of these papers for comparison of identified methods presented in Chapter 3.

2.3.5 Quality assessment

To the best of my knowledge, there is no existing tool for assessing the quality of methodological papers. A framework was adapted from Latimer [41] and used to critically appraise the papers included in this review (Table 3). This framework was also used to extract the key characteristics for each method identified by the review.

Table 3: Framework for critical appraisal of the methodological papers included in the review

Domain	**Issues considered**
Origin of the method	Was the method originally developed to adjust for non-adherence? If not, what was the original context and how the method was adapted? Does the method represent an extension to another method adjusting for non-adherence?
Theoretical suitability	How does the method work? What are the key assumptions? What are the potential biases? Why might the method not be appropriate? What are the advantages and disadvantages associated with the method? What are the similarities and differences of the method compared to other methods identified?
Application	Has the method been applied to adjust for non-adherence in a case study/simulation study? What disease/condition is applied in the case/simulation study? What is/are the intervention(s) assessed in the case/simulation study? What were the results compared to traditional methods (ITT, PP, AT), if compared with these methods?

Adapted from Latimer (2012)[41]

2.3.6 Data extraction

A standardised data extraction form was developed to extract the basic information and the key characteristics for each method identified. The basic information extracted included: author; year of publication; journal; and the number of citations (Web of Science). The key characteristics included the details for each component included the framework presented in Table 3. In addition, I also extracted information about the disease area in which the method was applied (i.e. case study and/or simulation study), the interventions assessed and the primary outcome.

2.3.7 Data synthesis

A narrative data synthesis approach was followed for each relevant method identified. This included a description and discussion of the key characteristics of the method based on three domains (origin, theoretical suitability and practical application (see Sections 2.7-2.10). The extensions identified for each method are also described and discussed in Sections 2.7-2.10. A taxonomy of methods to adjust for non-adherence for time-to-event outcomes is proposed in this thesis (Section 2.5). A comparison of methods based on existing evidence is discussed in Chapter 3.

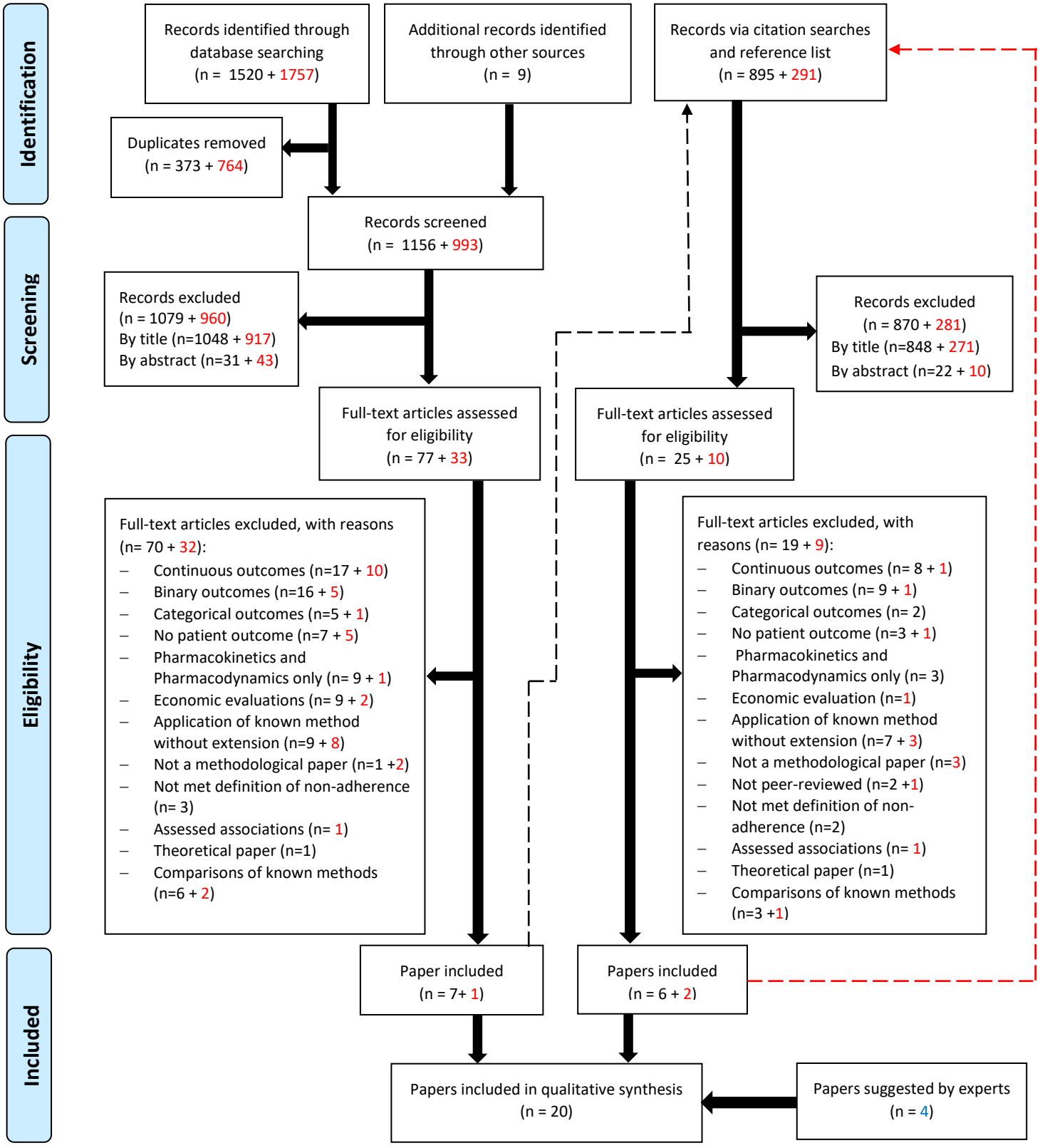
2.4 Results of the review

2.4.1 Search results

The searches identified a total of 4472 records from all rounds of the iterative search process applied in this review. The PRISMA flow diagram (Figure 6) shows the number of records identified from all sources during the first and second search iterations. The dashed lines illustrate that the citation searches and reference list checking were done for relevant papers identified from the database searching only. The only exception was papers labelled as 'comparisons' papers which were considered relevant for that purpose. The red numbers and dashed lines in Figure 6 represent records identified from the second search iteration.

The PRISMA diagram shows the number of records excluded at the title and abstract screening, full-text eligibility assessment and a final number of records included for qualitative synthesis. Details of the included papers are provided in Section 2.4.2.1. The number of full-text articles excluded is also provided in the PRISMA diagram.

Figure 6: PRISMA flow diagram



PRISMA= preferred reporting items for systematic reviews and meta-analyses. Numbers in red represent records from the 2nd stage of searches. The dashed lines show that citation searches and references lists checking were done for pearls identified from databases searching. Papers excluded for the reason of ''comparison of known methods'' are included in the citation searches and references lists checking as these were considered relevant for this purpose.

2.4.1.1 Results from the first iteration

The first search iteration of databases identified 1520 records; plus 9 records (original pearls) which were identified from the initial citation search of key papers and expert recommendations. After the removal of duplicates, 1156 unique records were screened by title and abstract resulting in 1079 records being excluded. The full-text papers for the remaining records (n=77) were retrieved and assessed for eligibility. Seven of these papers met the inclusion criteria.

Results from the citation searching and reference list checking generated 895 unique records after the removal of duplicate records. Based on the title and abstract screening, 870 records were excluded resulting in 25 records. A further 19 records were excluded because they did not meet the inclusion criteria leading to 6 new papers identified as relevant for inclusion in the review. A total of 13 relevant papers were identified from the first search iteration (databases=7; citation search and reference checking=6).

2.4.1.2 Results from the second iteration

The second search iteration on four databases identified 1757 records; 993 of these were unique records. The latter were screened by title and abstract resulting in 960 records being excluded. The full text of the 33 remaining records were assessed against the eligibility criteria leading to a further 32 exclusions, and only 1 record was included from the second iteration databases searching.

The subsequent citation searching and reference list checking generated 291 unique records of papers. After excluding 281 records at the title and abstract screening stage, full-texts for the remaining 10 records were retrieved and assessed for eligibility. This resulted in 9 further exclusions and 2 inclusions. Hence, a total of 3 records were identified from the second search iteration (databases=1; citation search and reference checking=2).

The diminished number of relevant papers identified in the second search iteration (n=3) from a large volume of records screened (n=1284) was used to inform the discussion with the supervisory team to stop further searching. The stopping rule (the point of saturation beyond which no further searching was done) was pre-specified by the review protocol.[65] By the end of this stage, a total of 16 relevant papers were identified from the first and second iterations of searches combined.

2.4.1.3 Papers suggested by expert advisors

The two expert advisors (IW and DH) and my PhD supervisor (Dr Nicholas Latimer) suggested a combined list of 42 potentially-relevant papers after checking the final list of included papers. After further investigation, 29 of these papers had already been identified by the searches and excluded as they did not meet the review inclusion criteria. Papers which were excluded at title screening were revisited with full-text retrieved and assessed for eligibility. The main reason for exclusion was that the method did not relate to the analysis of time-to-event outcomes (i.e. continuous, binary, categorical outcomes). The details of all papers excluded as well the methods discussed in these papers are provided in Appendix B, Table 32-34.

The remaining papers suggested by expert advisors which were not picked up by the searches (n=13) were retrieved and full-text assessed against the eligibility criteria. At the end of this process, 4 additional papers met the eligibility criteria and therefore were included in the review. Of those, one paper was an article in press that has not been published when the search was conducted and hence not picked up by databases search. One paper was not picked up by the databases search because the outcome was not a time-to-event but further included on the basis of assessing cost-effectiveness using the PKPD-based method with non-adherence incorporated in the analysis. The other two papers were picked up by the searches but excluded at the title screening stage as the adherence related term is not mentioned in the title.

2.4.1.4 Summary of results from all sources

A final list of 20 relevant papers was included for narrative synthesis in this review. The number of papers included by the source is summarised in Table 4.

Table 4: Summary of the number of papers included by the source of identification

Source	Number of papers included
First database search	7
First citation searches and reference list checking	6
Second-iteration database search	1
Second citation searches and reference list checking	2
Expert advice	4
Total	20

2.4.2 Details of included and excluded papers

2.4.2.1 Included papers

The basic information of included papers is shown in Figure 7 and 8. The year of publication ranged from 1992 to 2018 with one or two papers published per year over the past three decades, and some years with no papers included (Figure 7). The number of citations per paper since publication is shown in Figure 8. This is considered as a proxy measure of impact indicating the extent by which the method may have been used in practice. Robins and Finkelstein followed by Hernan et al. were the most cited papers. These two papers proposed Marginal Structural Models (MSMs) with Inverse Probability of Censoring Weighting (IPCW) and Inverse Probability of Treatment Weighting (IPTW), respectively. Pink et al., who propose a Pharmacokinetics and Pharmacodynamics (PKPD) based method had a relatively high number of citations, despite it having been recently published (see Figure 8). Three out of 20 included papers looked at adjusting for non-adherence in cost-effectiveness analysis with the remaining 17 papers focused on methods for adjusting estimates of treatment effect. This provides further evidence about the gap in the health economics literature in terms of methods for accounting for non-adherence in economic evaluations.

Figure 7: Years of publication for papers included in the review

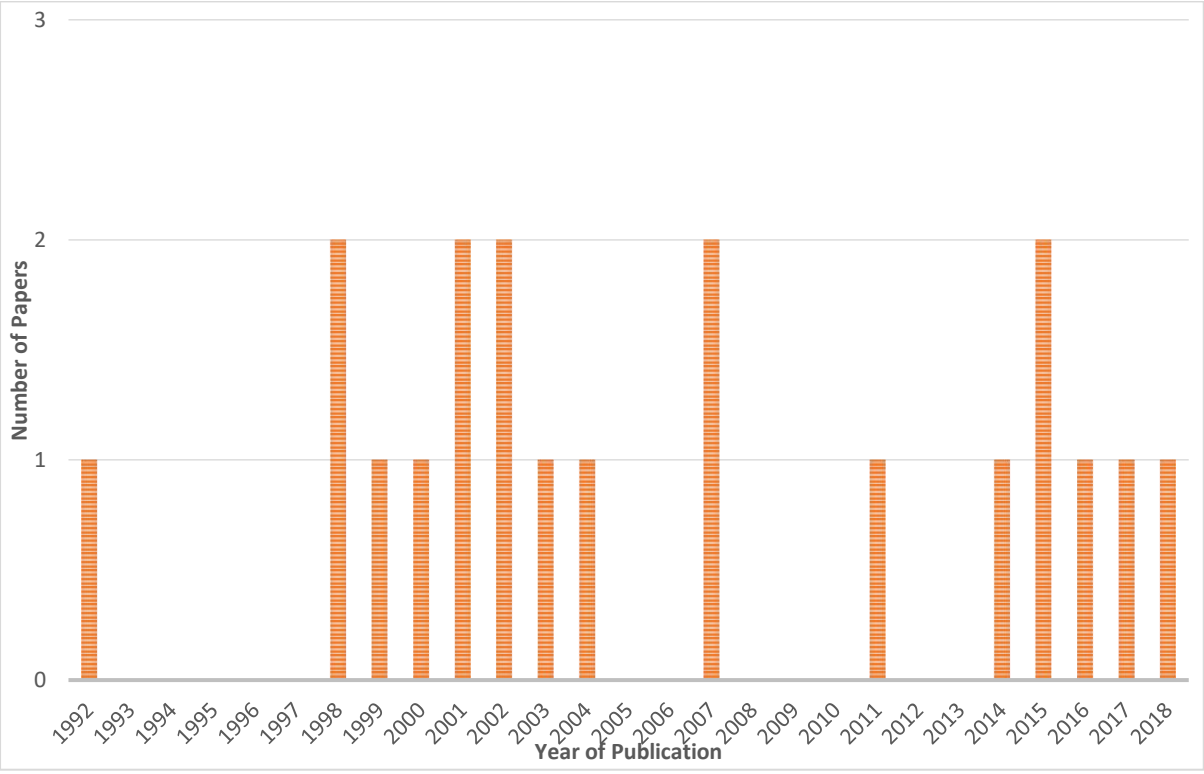
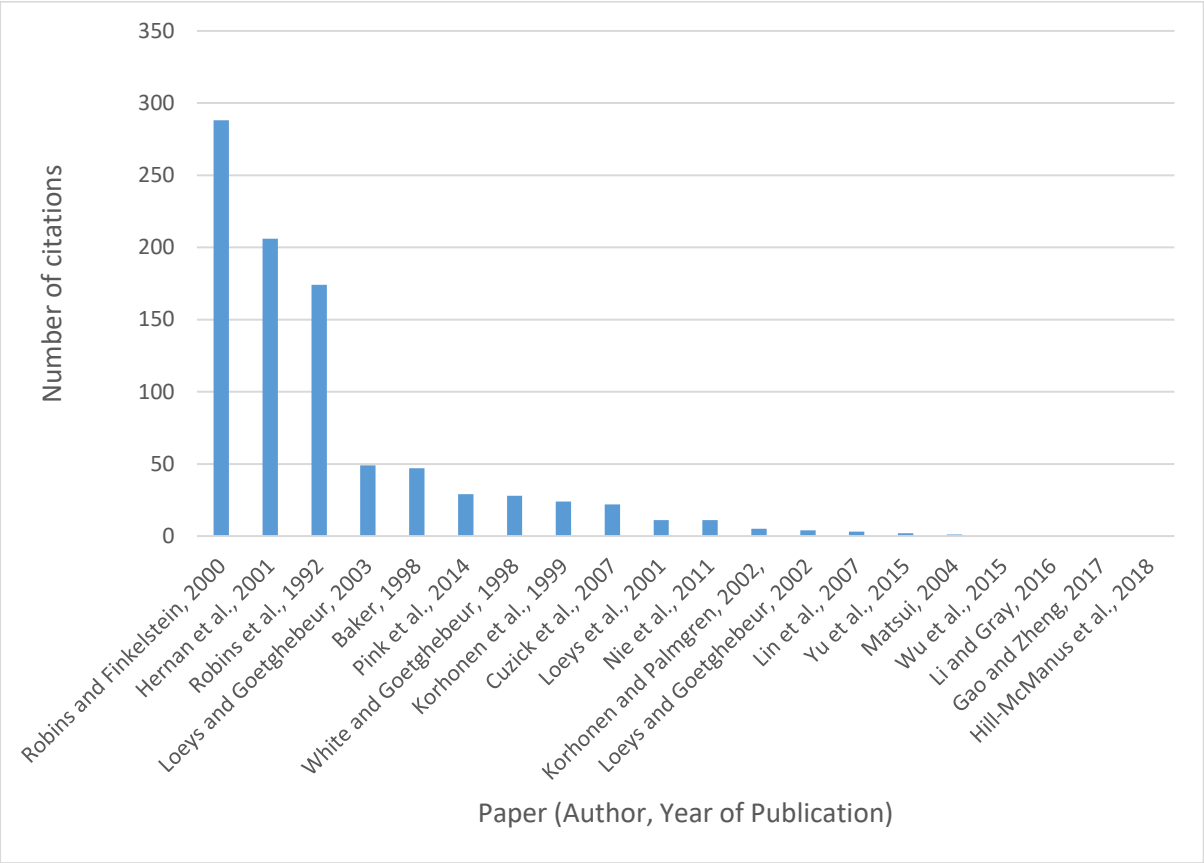


Figure 8: Number of citations per paper included since the publication date



2.4.2.2 Excluded papers

A total of 130 of the potentially relevant papers were excluded at the full-text eligibility assessment stage. The details of the excluded papers are provided in Appendix B, Table 32.

2.5 Taxonomy of methods identified to adjust for non-adherence

A taxonomy of methods for adjusting estimates of treatment effectiveness for non-adherence in the context of time-to-event outcomes is proposed in Table 5. The purpose of the taxonomy is to increase understanding of the concept behind each method and its relation to other methods in terms of estimands and estimators. The initial structure of the taxonomy was revised based on consultations with advisors (DH, JF) and an expert in causal inference methods (IW).

In the proposed taxonomy, methods are broadly classed into four groups: (1) simple methods which do not appropriately adjust for non-adherence (e.g. exclude not adhered patients from the analysis); (2) principal stratification methods for estimating the Complier Average Causal Effect [CACE] estimand;[73] (3) generalised methods (g-methods) which are based on the counterfactual outcome framework originally developed by Neyman[9] and Rubin[10] for estimating the effect of time-fixed treatments, and further extended by Robins[11, 12] for time-varying treatments; and (4) pharmacometric-based methods as a unique approach using pharmacokinetics and pharmacodynamics (PKPD) analysis commonly used in clinical trials for evaluating newly developed pharmacological interventions. The estimand and key assumptions used by each method are provided in Table 6. I provide an overview of methods in each group in the following subsections. Each method identified is described in detail in Sections 2.7-2.10. However, an overview of methods included in each group of the taxonomy is provided in the following subsections.

2.5.1 Simple methods

Simple methods include ITT, PP and As-Treated (AT) analysis. The ITT and PP represent the conventional methods applied in practice in analysing clinical trials data. The ITT analysis strategy is to include all patients randomised regardless of non-adherence, protocol deviation, switching or any other post-randomisation event. The intention is to maintain the prognostic balance between treatment arms as generated by the original randomisation. PP is another simple method which includes a subset of the study population who complied with the study protocol. The intention is to estimate the efficacy of the treatment among those who followed the protocol. The PP analysis strategy, therefore, excludes protocol non-compliers, though different definitions of protocol

deviations are used. The AT analysis estimates the ACE among individuals who actually received the treatment; therefore, it does not respect the randomisation in the trial.

2.5.2 Principal stratification methods

This group covers five methodological approaches. The methods are based on the 'principal stratification framework' proposed by Frangakis and Rubin.[73] The methods are considered an extension to traditional methods, and they are based on stratifying patients into different classes of adherence. These methods can be used to estimate the ACE within principal strata (e.g. compliers). The methods covered by this group are: (i) Cox PH model with Partial Likelihood Estimator (PLE); (ii) Markov Compliance Class (MCC) model in a Three-Stage Method (3SM); (iii) Weighted Per-Protocol (Wtd PP) analysis using a PH model with an Expectation-Maximisation (EM) Estimator; (iv) Compliers PROPortional Hazards Effect of Treatment (C-PROPHET) model; and (v) Instrumental Variable (IV) methods.

2.5.3 G-methods

G-methods are largely based on the counterfactual outcome framework and include different models compared to those introduced in the previous two categories. This includes three methods and their extensions: (i) MSM with IPCW/IPTW; (ii) Structural Nested Failure Time Models (SNFTMs) with g-estimation; and (iii) Rank-Preserving Structural Failure Time Models (RPSFTMs) with g-estimation. SNFTMs and RPSFTMs belong to a broader class of models known as Structural Nested Models (SNMs).

2.5.4 Pharmacometrics-based methods

This class includes one method, the Pharmacokinetics and Pharmacodynamics (PKPD) based method, which is a unique approach based on pharmacometrics analysis commonly used in clinical trials evaluating newly developed pharmacological interventions. This is a mechanism-based method for modelling varying adherence patterns to estimate adherence-adjusted causal effects.

Table 5: Taxonomy of methods for adjusting for non-adherence in the context of time-to-event outcomes

Methods group	Method sub-category	Method/Extension	Reference
Simple methods	ITT *	Intention-To-Treat (ITT) analysis	Yu et al., 2015[74]
	PP	Per-Protocol (PP) analysis	Wu et al., 2015[75]
	AT	As-Treated (AT) analysis	Korhonen et al., 1999[39]
Principal stratification methods	CPH with PLE	Cox Proportional Hazards (CPH) Model with Partial likelihood Estimator (PLE)	Cuzick et al., 2007[76]
	MCC	Markov Compliance Class (MCC) Model in a Three-Stage Method (3SM)	Lin et al., 2007[77]
	Wtd PP	Weighted Per-Protocol (Wtd PP) analysis using a Proportional Hazards Model with an Expectation-Maximisation (EM) Estimator	Li and Gray, 2016[78]
	C-PROPHET	Compliers PROPortional Hazards Effect of Treatment (C-PROPHET)	Loeys and Goetghebeur, 2003[79]
	IV	Instrumental variable (IV) with Likelihood Estimator	Baker, 1998[80]
		IV with Plug-in Non-Parametric Empirical Maximum Likelihood Estimation (PNEMLE)	Nie et al., 2011[81]
		Transformation Promotion Time Cure Model with MLE to estimate the Compliers Average Causal Effect (CACE) and the Compliers Effect on Survival Probability (CESP)	Gao and Zheng, 2017[82]
G-methods	MSMs	Marginal Structural Models (MSMs) with Inverse Probability of Censoring Weighting (IPCW)	Robins and Finkelstein, 2000[66]
		MSM Extension: MSMs with Inverse Probability of Treatment Weighting (IPTW)	Hernan et al., 2001[83]
	SNFTMs	Structural Nested Failure Time Models (SNFTMs) with g-estimation	Robins et al., 1992[84]
	RPSFTMs	Rank-Preserving Structural Failure Time Models (RPSFTMs) with g-estimation	Loeys et al., 2001[85]
		RPSFTM Extension: Incorporating covariates to improve the precision of estimators	Korhonen and Palmgren, 2002[86]
		RPSFTM Extension: Improving the efficiency of the estimators	Loeys and Goetghebeur, 2002[87]

		RPSFTM Extension: Allowing for dependent censoring	Matsui, 2004[88]
		RPSFTM Extension: Choice of model and impact of recensoring	White and Goetghebeur, 1998[89]
Pharmacometrics-based methods	PKPD	Pharmacokinetics and Pharmacodynamics (PKPD) based method	Pink et al., 2014[90]
		PKPD Extension: Modelling varying implementation and persistence types of non-adherence	Hill-McManus et al. 2018[91]

* ITT does not adjust for non-adherence but included in the taxonomy as a "do nothing" approach (i.e. ignoring non-adherence)

2.6 Summary of estimands and key assumptions of identified methods

The estimand for each method identified and key assumptions are summarised in Table 6. The main point is that different estimands are used making comparability across alternative methods problematic. For example, the ITT effect of treatment assignment in the entire study population cannot be compared with an estimate from the IV CACE estimand which is restricted to the subgroup of compliers.

Table 6: Estimands, causal interpretations and key assumptions for non-adherence adjustment methods

Method	**Estimand***	**Estimand Attribues**	**Causal Interpretation of the Estimate**	**Key Assumptions**
ITT	The effect of treatment assignment (not the effect of treatment itself)	Entire study population; ignoring events such as non-adherence and dropout	The average causal effect of treatment assignment on the survival outcome in a particular study (regardless of adherence, dropout, etc…)	The randomisation assumption (i.e. group membership is randomly assigned), which implies that groups are comparable or exchangeable.
PP	The effect of following the study protocol	Sub-population of the protocol compliers in the study; excluding protocol non compliers from the analysis set	The average causal effect of treatment on the survival outcome in individuals who adhered to the protocol in terms of eligibility, adherence, outcome assessment, etc...	The groups of patients who adhered to the protocol in each arm are comparable after covariate adjustment.
AT	The effect of treatment actually received	Sub-population of patients who initiated treatment; with patients who switched treatment analysed with the group they switched to regardless of randomisation	The average causal effect of treatment on the survival outcome among individuals who actually received the treatment in the experimental group (including control group patients who switched on to the experimental treatment) compared to those who actually received the standard treatment (or those actually not received the treatment in placebo-controlled trials) regardless of treatment assignment	The group of patients who received the treatment is comparable to those who did not, regardless of their treatment assignment after covariate adjustment.
CPH with PLE	The complier average causal effect (CACE)	Sub-population who adhered to the protocol; excluding patients who did not adhere to the protocol in each arm of the study	The average treatment effect on the survival outcome in the compliers sub-population (patients who adhered to the protocol)	Covariates included in the model are independent of adherence
MCC	CACE	As above	As above	- The Markov assumption - Time-varying adherence depends on the history of adherence - Latent and ignorable missing data mechanism

Wtd PP	CACE	As above	As above	The patient population consists of three (possibly latent) subgroups: '*ambivalent', 'insisters'* and *'refusers'*
C-PROPHET	CACE	As above	As above	The exclusion restriction assumption
IV	CACE	As above	As above	- The exclusion restriction assumption - Randomisation has no effect on the probability of adherence to treatment - Monotonicity assumption
MSMs with IPCW/IPTW	The effect of treatment had everyone remained adherent to the protocol	Entire study population; had everybody adhered to the protocol with perfect adherence to the prescribed dosing regimen (or had everybody adhered to the protocol at an alternative level of adherence to the prescribed dosing regimen than what was observed in the trial (e.g. real-world adherence level)	The average causal effect of treatment that would have been observed if everybody adhered to the protocol. MSMs estimate the average treatment effect in the entire population, but the causal effect in a subset of the population (defined by a combination of variables L) can also be estimated. The IPCW estimand can also be interpreted as a comparison of the potential (counterfactual) outcomes under different levels of adherence in the same group of subjects.	- No unmeasured confounders - Positivity assumption
SNFTMs with g-estimation	The effect of treatment had everyone remained adherent to the protocol	As above	The average treatment effect that would have been observed if everybody adhered to the protocol (or remained at a particular adherence level such as real-world adherence level). SNFTMs can be used to estimate the average causal effect in a subset of the population defined by a combination of factors (L), e.g. men, patients aged >60 years, etc…	- No unmeasured confounders - Survival times and treatment-free survival times are proportional by an unknown factor that depends on the exposure

RPSFTMs with g-estimation	The effect of treatment had everyone remained adherent to the protocol	As above	The average treatment effect that would have been observed if everybody adhered to the protocol compared to none treated.	- The randomisation assumption - The common treatment effect assumption - Survival times and treatment-free survival times are proportional by an unknown factor that depends on the exposure
PKPD method	The effect of following a particular adherence pattern in the study population	Entire study population; given a particular pattern of adherence to the prescribed dosing regimen	The average causal effect of treatment if individuals followed a particular adherence pattern.	- The exclusion restriction assumption - Correctly specified model (there are at least 2 components to this - the link between adherence and PKPD and the link between PKPD (surrogate) and the final endpoint

* The estimand is the parameter of interest defined using four attributes: (i) the population, (ii) the outcome variable or endpoint, (iii) the specification of how to deal with intercurrent events (e.g. include compliers only), and (iv) the population-level summary of the outcome variable. The description of the estimand in this table is focused on two attributes (the population and specification of how to deal with intercurrent events)) as the other two attributes (the outcome variable and the population-level summary of the outcome variable) are expected to be similar in the context of time-to-event outcomes.

Narrative synthesis of identified methods

The following sections provide the narrative synthesis of the identified non-adherence adjustment methods. The sections describe the identified methods (categorised based on the proposed taxonomy (Table 5) including their extensions. Details are not given for ITT and PP as they are so commonly used, but all subsequent methods are described in terms of their origins, theoretical characteristics and applications. The latter provides the application of the method in a case study, simulation studies or both. These were reported as demonstrative case studies in the included methods papers. No additional case studies or simulation studies were included to describe these methods. The methods that I subsequently chose not to take forward (including their extensions) are only described briefly, whereas the more relevant methods are described in more detail.

2.7 Simple methods

This section describes the identified non-adherence adjustment methods including their extensions. Details are not given for ITT and PP as they are so commonly used, but all subsequent methods are described in terms of their origins, theoretical characteristics and applications.

2.7.1 Intention-To-Treat analysis

In the ITT analysis approach, individuals are analysed as randomised regardless of whether they adhered to their assigned treatments or not.[74] The treatment effect estimated from an ITT analysis will reflect observed adherence levels but does not include any causal link between adherence and treatment effect. The focus is to maintain randomisation. ITT analysis produces an unbiased estimate of the observed treatment strategies; however, the question is whether these strategies represent the estimand of interest in answering the scientific question. If there is non-adherence, and the analyst needs to know what the effect is with a different level of adherence (e.g. real-world adherence level), then ITT analysis does not answer the question we are interested in and so will likely not give an accurate estimate of the effect they are interested in.

In RCTs with non-adherence, the ITT analysis is likely to mix the benefit of treatment among individuals with a high level of adherence with the absence of (or suboptimal) benefits among those with a low level of adherence.[75, 78] Therefore, in a situation where there exists non-adherence to the assigned treatment, the ITT analysis may not be the appropriate analytical approach. In RCTs with time-to-event endpoints, the conventional model used to estimate treatment effects is the Cox Proportional Hazards

(CPH) model.[92] In this context, the Kaplan-Meier (KM) non-parametric survival model is commonly used for creating survival curves using the fraction of patients surviving for a certain amount of time after receiving the treatment.

2.7.2 Per-Protocol analysis

The PP analysis approach excludes protocol non-compliers from the analysis.[75] The idea is to estimate the efficacy of treatment among patients who adhered to the study protocol. The main concern is that excluding some patients from the analysis may undermine the prognostic balance generated by the randomisation which may introduce selection bias. This is likely to be the case if non-adherence is not random, i.e. if non-adherence is influenced by other patient characteristics and prognostic factors. Therefore, estimates using unadjusted PP analysis may produce biased estimates as a result of failure to adjust for confounding by measured and unmeasured factors. Even if prognostic factors which are associated with non-adherence are correctly identified, the standard PP analysis will introduce bias if there is time-dependent confounding. The key issue with PP analysis is that the estimand is not marginalised to the entire study population, therefore, the method does not answer the question of interest.

The limitations of the PP (and ITT) analysis approach in adjusting for non-adherence is well established in a wide body of the causal inference methodological literature.[38, 39] The key limitations include difficulty in adjusting for time-dependent confounding and potential bias introduced by excluding patients from the analysis when using PP analysis to estimate treatment effect in the presence of non-adherence. As ITT and PP are widely used as conventional methods for estimating treatment effects, it is, therefore, important to know how their estimates differ from other non-adherence adjustment methods. Therefore, I will present estimates based on these methods (where reported in the reviewed papers). It is worth noting that the results estimated by the alternative methods may not be directly comparable to ITT and PP due to different estimands.

When there exists time-dependent confounding, more complex methods are needed to adjust the causal effect of treatments in the presence of non-adherence. These include g-methods such as MSMs and SNMs; these methods are introduced in Section 2.9.

2.7.3 As-Treated analysis

2.7.3.1 Origin of the method

The As-Treated analysis approach was first introduced by Kohren et al.[39] to adjust for patient non-adherence in the analysis of trial data from the Alpha-Tocopherol Beta Carotene Lung Cancer Prevention Study (ATBC study) as described in further detail in Section 2.7.3.3.

2.7.3.2 Theoretical characteristics

As-Treated (AT) analysis is a simplistic approach for estimating the ACE of treatment in the context of RCTs. The method estimates the causal effect among patients who actually received the treatment compared to those who did not receive the treatment, regardless of randomisation, therefore, it does not respect the randomisation in the trial. The main issue is that the group who actually received the treatment is unlikely to be comparable to the group who do not, making this approach prone to selection bias. The method has been used to adjust for non-adherence at the initiation stage, mostly for time-fixed treatments.

2.7.3.3 Application in a simulation study and case study

Kohren et al.[39] applied the AT approach for adjusting for non-adherence using a simulation study and further application using real data in a case study. The simulation study was designed to mimic a two-arm trial (active treatment vs placebo) with a sample size of 1000 and simulations repeated 500 times. Scenarios assessed included an outcome-dependent case where the survival time depends on patients receiving the active treatment (time on treatment) and an independent case scenario. The simulated dataset also included a case with baseline unmeasured confounders as common causes of the survival-time outcome and the exposure. The method was compared with ITT and RPSFTM with g-estimation.

The results from this simulation study found that the AT approach can produce misleading estimates when a non-adherence-dependent outcome exists in the data. In contrast, RPSFTM produced a valid causal effect, under the assumptions made by this approach, even in scenarios with outcome-dependent non-adherence. The paper concluded that RPSFTM with g-estimation would be the best alternative in these situations.

In addition, the investigators applied the three methods in a case study using real data from the ATBC study. In brief, the study compared the causal effect of beta carotene versus placebo on the survival of lung cancer patients. The level of persistence non-adherence observed in the study was estimated

as 25%. The AT estimated an HR of 2.13 (95% CI: 1.93-2.3) compared with an ITT HR of 0.92 (95% CI: 0.86-0.99) and an RPSFTM with g-estimation HR of 0.93 (95% CI: 0.87-0.99). The results from the case study are consistent with the findings from the simulation study, which concludes that the AT analysis may produce misleading results when estimating the ACE in the presence of non-adherence.

2.8 Principal stratification methods

2.8.1 Cox Proportional Hazards Model with Partial Likelihood Estimator

2.8.1.1 Origin of the method

The Cox PH model with Partial Likelihood Estimator (PLE) approach was originally proposed by Cuzick et al. to adjust for the non-initiation type of non-adherence for binary outcomes.[93] In the context of RCTs with time-to-event outcomes, the standard Cox PH model is widely used for estimating the ACE of treatment.[92] The method was extended by Cuzick et al.[76] to deal with non-adherence and the description of the method below is based on this paper.

2.8.1.2 Theoretical characteristics of the method

The CPH model with PLE is a method for adjusting for non-adherence while respecting the randomisation (randomisation-based). This is a semi-parametric model where the treatment effect on the distributions of failure times is the parametric part. A simple (non-iterative) version of the method can be used when there are no covariates to adjust for (or if the analyst decides not to adjust for them). This simple version is considered as a generalisation of the classical Mantel-Haenszel (MH) estimator.[76] The more general PLE version of the method was developed to incorporate covariates assuming that they are independent of adherence. The method is designed to incorporate time-independent covariates. An extended version of this method which can incorporate time-dependent confounders is introduced in Section 2.8.2.

The concept of this method is based on patient stratification by adherence status (i.e. binary adherence) and then incorporating covariates to adjust for confounding under a strong assumption that covariates are independent of adherence. The method can be applied to estimate the treatment effect based on the CACE estimand. The method works by stratifying the study population into different classes and then estimating the ACE by comparing the outcomes between different strata of interest (e.g. compliers). The method can also be used to adjust for non-adherence in situations where non-adherence is dependent on covariates, but this approach requires a more complex estimator. An important assumption is that covariates are independent of adherence (which might be unrealistic).

Therefore, the method can be generalised to situations where non-adherence is dependent on covariates.

To introduce the basic model, the method classifies patients (randomised in a two-arm trial) into three classes: (a) insisters (individuals who always want the new treatment; (b) ambivalent (individuals who accept any treatment offered to them); and (c) refusers (individuals who refuse the new treatment, if randomised into the intervention group). Each class is divided into two groups at random for receiving the new treatment (T) or the control (C). This is best explained in Figure 9 which was adapted from Cuzick et al.[76] to aid the explanation. Because it is not possible to determine each individual's latent adherence class, we can only observe the following four groups in the RCT data (also see Figure 9):

- CT (yellow) – insisters assigned to the control arm
- CC (green) – ambivalent and refusers assigned to the control arm
- TT (blue) – insisters and ambivalent assigned to the treatment arm
- TC (orange) – refusers assigned to the treatment arm.

Figure 9: Groups of patients by compliance class and randomisation group

	Randomised to:	
	Control arm	**T**reatment arm
Insisters	**CT**	**TT**
Ambivalent	**CC**	**TT**
Refusers (non-compliers)	**CC**	**TC**

In the basic model provided by Cuzick et al.[76] an individual with covariates (*k, z_0, z*) will have a hazard function presented in equation [2].

$$\exp(\gamma_T z_0 + \beta z + \gamma_k)\,\lambda(t) \qquad [2]$$

where λ_T is the treatment effect in compliers (CACE estimand) expressed in terms of the hazard at time *t* for a cumulative hazard function $\Lambda_k(t)$ (this is only observable for ambivalent class), γ_k is the adherence class of the kth individual, z_0 is a vector of baseline covariates, and *z* is a set of time-dependent covariates. In some cases, it should also be noted that z_0 and *k* are not observed in the RCT data.

The method proposed three different estimators to estimate the CACE: (i) no covariates adjustment using the CPH estimator or using the iterative approaches of Mantel-Haenszel weights (MH) or efficient one-step weights (EW); (ii) using covariates which are considered as independent of class membership; and (iii) using covariates correlated with class membership. The latter two methods rely on the partial likelihood estimator (PLE) or full likelihood estimator (FL). Further details about the technical properties of these estimators are reported in Cuzick et al.[76]

The method makes three assumptions:

(a) The proportional hazards assumption, which implies that cumulative hazard functions are proportional in the absence of treatment; that is the treatment effect (HR) must be constant over time;
(b) Independent (non-informative) censoring; and
(c) Covariates included in the model are independent of adherence.

In terms of potential bias, the proposed method is generally powerful in producing unbiased estimates of causal effect when there are no, or minor, deviations from its assumptions.[76] However, the method may produce biased estimates of treatment effect in extreme cases (extremely unequal proportions of insisters and refusers).[76] This was found in a simulation study designed to test this method.[76] Moreover, the EW and MH approaches are less powerful when there is a high level of non-adherence in the trial population.[76] The findings from the simulation study suggest that the EW approach has no substantial benefits compared to the MH approach.

The method has some advantages and disadvantages. The FL approach performs well across different classes of non-adherence for scenarios with both small sample size and large sample size. PLE is easy to compute and also showed good performance except in a situation where there is a high insister or refuser effect. The method can be applied for adjusting for non-adherence in the context of two-arm RCTs as well as placebo-controlled RCTs. Moreover, the method can easily handle ties between censored observations and un-censored observations by making an additional assumption that un-censoring occurred first. The main limitation of this method is the difficulty of modelling time-varying

treatments and other forms of non-adherence beyond initiation (i.e. implementation and persistence non-adherence). This is formally possible, but the causal interpretation of estimates from these models is difficult. Finally, the CACE estimand used by this approach is problematic and not appropriate for the HTA context as discussed earlier. Therefore, I do not provide detail on the more advanced applications of this method (time-dependent covariates and other forms of non-adherence) as the method does not give estimates that are useful for an HTA perspective.

2.8.1.3 Application in a simulation study

The method was applied in a simulation study for evaluating its performance.[76] This involved a detailed study for assessing the performance of five estimators proposed by this method (i.e. PH, MH, EW, PL and FL estimators). The simulation study was conducted for evaluating the ACE of hypothetical treatments on survival time outcome. The study was done for two main scenarios of sample sizes (200 and 2000 observations) for a two-arm RCT design with a dual assignment. The simulation study also considered three scenarios of covariate structures (no covariates, covariates independent of randomisation, and two covariates).

The simulation study showed that the treatment effect was underestimated by the standard CPH model. An underestimation was also reported for models with weighted estimators without adjustment for covariates. Further details about the performance of the proposed estimators in different scenarios are reported in Cuzick et al.[76]

2.8.2 Markov Compliance Class Model in a Three-Stage Method

2.8.2.1 Origin of the method

The Markov Compliance Class (MCC) model was originally developed by Lin et al.[94] using the compliance class model framework of Imbens and Rubin.[95] In Lin et al., the method was extended to adjust for time-varying non-adherence in the context of longitudinal studies where patients are randomised at baseline and randomisation is maintained over time.

2.8.2.2 Theoretical characteristics

The method is based on specifying two possible adherence classes and several follow-up time points, e.g. five-time points resulting in a total of 32 (2^5) adherence patterns. A stratification strategy can then be used to stratify adherence patterns into super-classes (low compliers, decreasing compliers and high compliers). This can then be used to estimate the ACE of treatment accounting for the non-

adherence superclass variable. CACE is the estimand used by this method which means the treatment effect estimated cannot be marginalised to the entire study population. A nested MCC model was extended to incorporate baseline covariates (patient characteristics) as predictors of time-varying adherence.[77]

The MCC model relies on the following key assumptions:

a. Time-varying adherence depends on the history of adherence
b. Missing data mechanism is latent and ignorable
c. Data are missing at random conditional on adherence class and the covariates.

The MCC method works by searching for relevant individual patient-level baseline predictors of the superclass strata that describe time-varying adherence patterns and assess the relationship of these super-classes and the outcome. The analysis follows three stages:

(a) Principal stratification of patients to describe time-varying adherence patterns;
(b) An incomplete data model to compute the "posterior predictive distributions" of the adherence superclass. In this stage, the 'latent' adherence superclasses are treated as missing data, and a multiple imputation technique is applied using the Markov Chain Monte Carlo (MCMC) simulations to compute the posterior distributions; and
(c) A complete data model to relate the imputed adherence superclass to baseline covariate and the survival outcome, which is a model that gives the CACE estimand.

As reported in Lin et al.[94], the method can be used to estimate the CACE estimand among the compliers superclass. Model [3] can then be used to account for the relationship between adherence and survival time at time *t*.

$$h(t|U_i = k) = h_0(t)\exp\left(\beta_k I(U_i = k)\right) \quad [3]$$

where β_k for one of the adherence superclasses is assumed 0 for identification (reference superclass) and U_i is individual i's latent adherence superclass for a k number of superclasses. Lin et al.[94] also provided formulae for estimating the survival function using the Kaplan-Meier model for computing the hazard rate when there are no covariates. As a limitation, the method cannot deal with time-dependent confounding.

In a two-arm RCT setting, Lin et al.[94] specified the following four possible adherence classes. These classes are considered as latent because they are not observable at baseline.

(i) Adherers who would always adhere to the treatment they were randomised to;

(ii) Always takers who want to receive the new treatment regardless of their randomisation;

(iii) Never-takers who choose to be in the control arm regardless of their randomisation; and

(iv) Defiers who always take the opposite treatment they were assigned to receive.

In terms of situations where the MCC method might not be appropriate to use, one was identified relating to the pattern of non-adherence. That is when the number of subjects experiencing a particular adherence pattern is too small; the method may produce estimates which are not clinically meaningful. In these situations, inference will be sensitive to the modelling assumptions. To avoid this, broader adherence superclasses should be used.

While the MCC method allows us to assess the relationship between baseline covariates, adherence and the survival outcome, the method does not allow the analyst to understand the reason why particular adherence patterns lead to poor health outcomes. Similar to previously described methods (principal stratification methods), the method estimates a treatment effect for each latent adherer class of patients (e.g. compliers) which would make it difficult for policymakers to make decisions based on evidence estimated by this method. This is because they are not identifiable at baseline. As a limitation, the method can not deal with time-dependent confounding.

2.8.2.3 Application in a case study

The MCC method was applied in a case study on depression. This was applied using data from the Prevention of Suicide in Primary Care Elderly Collaboration Trial (PROSPECT) study. Using this study data, three initial latent adherence classes were assumed, but one class was found to be very small leading to clinically non-meaningful estimates, and therefore, excluded. This left only two latent superclasses in the analysis (increasing compliers and high compliers) which were assessed for five follow-up time points (4, 8, 12, 18 and 24 months). The analysis showed that a beneficial effect (improved survival) among compliers compared to non-compliers (HR=0.32; CI:0.15-0.68). Further details about this case study are reported in Lin et al.[77] This case study shows that the method has been applied, but no information was provided about the performance of the method.

2.8.3 Weighted Per-Protocol using PH Model with EM Estimator

2.8.3.1 Origin of the method

Li and Gray[78] proposed the Weighted Per-Protocol (Wtd PP) by extending the approach introduced in Section 2.8.2. This method is focused on the ambivalent class with two main contributions: (i)

proposing a Wtd PP estimator by using time-varying weights in the in the PH model; and (ii) proposing an EM algorithm to maximise the FL and PL which were originally considered by Cuzick et al.[76] The method was developed to adjust for time-dependent confounders which are associated with non-adherence.

2.8.3.2 Theoretical characteristics

The theoretical characteristics of this method are similar to the CPH approach introduced in Section 2.8.2; therefore, I briefly describe the advantages and disadvantages in this section. The key advantage of the Wtd PP approach is its ability to update the baseline risk over time by using a weighted Cox model. In other words, the method attempts to estimate the ACE of treatment weighted by time-varying confounders included in the model. In addition, the method proved to be robust in modelling misspecification in both insisters and refusers classes as no distributional assumptions are imposed among these two groups in the analysis.

2.8.3.3 Application in a case study

The Wtd PP approach was tested in a simulation study which showed good performance in terms of bias; but in most situations, the method proved less efficient than the likelihood estimator. Details about the simulation study are reported in Li and Gray.[78] The method was further applied in a case study on breast cancer for evaluating combination chemotherapy, CMFP (cyclophosphamide, methotrexate, 5-fluorouracil, prednisone) versus observation. The analysis included 424 patients with analysable data with 11% non-adherence level in the CMFP arm and 16% in the observation arm. The Wtd PP estimated an HR of 0.48 compared to 0.40 and 0.41 estimated by unweighted PP and ITT analysis, respectively. Further details about this case study are reported in Li and Gray.[78]

2.8.4 Compliers PROPortional Hazards Effect of Treatment (C-PROPHET)

2.8.4.1 Origin of the method

The Compliers PROPortional Hazards Effect of Treatment (C-PROPHET) method was proposed for estimating the ACE of treatment actually received. The method was originally developed by Loeys and Goetphebeur[79] to adjust for the initiation type of non-adherence in an RCT context. The estimand which could be estimated by this method is CACE.[96] C-PROPHET is considered as a semi-parametric model with the parametric side being the effect of the exposure on the survival times distributions.

2.8.4.2 Theoretical characteristics

The concept of this method is that the analyst could identify adherent patients (compliers) at baseline and estimate the treatment effect in this group, adjusting for baseline covariates. If individual patients who actually adhered to the assigned treatment can be predicted at baseline in the intervention and control arm of an RCT, then one could fit a PH model for this subgroup of the study population to estimate the treatment effect. The ACE estimated by this method is not marginalised to the entire study population, which is a limitation.

The C-PROPHET model assumes that the hazard of survival time (T_i) is as provided in equation [4]:

$$\lambda(t|Z_i = 1, E_{1i} = 1) = \lambda(t|Z_i = 0, E_{1i} = 1)\exp(\psi_0) \quad [4]$$

where Z_i is the randomisation variable for individual *i* ($Z_i = 1$ for the intervention group, $Z_i = 0$ for the control group), E_{1i} represents the principal stratum at the treatment initiation stage. The parameter ψ_0 denotes the causal proportional hazards effect in the subpopulation of compliers. This is the parameter of interest that is called C-PROPHET.[79]

The C-PROPHET method relies on the 'jack-knife' resampling technique to produce a finite sample correction of the point estimate.[79] The jack-knife technique showed a conservative estimate of variance (with 1% standard error) and high coverage (the percentage of simulations where the 95% CI includes the true outcome) based on a simulation study used to evaluate the performance of the C-PROPHET method.

The C-PROPHET method relies on four key assumptions:[79]

(a) The randomisation assumption;
(b) No access to the new treatment in the control arm (this assumes "treatment-free" outcome when an individual is randomised to the control arm);
(c) The "exclusion restriction assumption" which is also known as the "absence of indirect effect" assumption that is randomisation affects the survival outcome only through the exposure; and
(d) The independent censoring assumption.

Regarding potential biases associated with the C-PROPHET method, the violation of assumptions (particularly the exclusion restriction assumption) is the main risk for bias. The method can incorporate baseline covariates; however, the causal effect of time-varying treatments with more

levels of adherence cannot be estimated using the C-PROPHET model. This is a limitation given the time-varying nature of the non-adherence problem, and this is considered as a situation where the method may not be appropriate to use.

Analysis using the C-PROPHET method is easy to communicate which is an advantage. However, the estimates are only valid if the exclusion restriction assumption holds. This assumption is likely to hold in double-blinded RCTs, but this may not hold in single-blinded or non-masked RCT designs. On the positive side, the C-PROPHET method does not require the application of artificial censoring as it does not construct individual counterfactual outcomes. The method rather works by fitting adherence-specific distributions of observed outcomes to estimate the treatment effect.

On the negative side, the C-PROPHET cannot be used to deal with time-dependent non-adherence. The method is restricted to the binary initiation type of non-adherence and therefore may not be applicable to adjust for 'sub-optimal implementation' or 'non-persistence' types of non-adherence. The method relies on the non-informative censoring assumption, and if this assumption does not hold, then the method may produce biased results.

2.8.4.3 Application in simulation and case studies

The C-PROPHET approach was applied in both a simulation study and a case study. The simulation study was used to evaluate the performance of C-PROPHET compared with an RPSFTM proposed by Robins and Tsiatis.[37] In the reported simulation study, 21 sequences of data were generated under three situations: (i) a situation where only the PH model specified in Section 2.8.4.2 holds; (ii) only the Robins and Tsiatis model holds; and (iii) both models hold. The simulation study used a sample size of 150 observations randomised in a 1:1 ratio between the intervention and control arms, with 400 simulations (repetitions). The data-generating mechanism used an exponential distribution for failure times and varying causal effects with adherers having a better prognosis compared to non-adherers. The generated data also included a time-varying hazard which was changed at 20, 40 and 50 months from baseline.

The simulation study showed that the C-PROPHET model and the RPSFTM model are generally similar in terms of producing unbiased results with the C-PROPHET estimator producing a smaller empirical standard error. The method showed high coverage of the 95% CI with a 1% standard error demonstrating conservative variance estimates based on the jack-knife technique. The method also showed a high reduction in mean square error (MSE) in the situation (ii) where the RPSFTM model

only holds but larger MSE in the situation (i) where the PH model only holds. Further details about the asymptotic properties of the estimator are reported in Loeys and Goetphebeur.[79]

The method was further applied in a case study on colorectal cancer using data from an RCT comparing surgical resection followed by chemotherapy administered via an atrial device (intervention arm) compared with surgical resection alone (standard treatment arm). The ITT analysis estimated a causal effect (in terms of an HR) of 1.26. The C-PROPHET estimate of the causal effect was 1.43 - this is the proportional hazards effect among compliers, i.e. the estimate from the CACE estimand).[79] Although these are two different estimates, they are not directly comparable due to different estimands. This point is relevant to the simulation study undertaken in this research and will be discussed further in subsequent chapters.

It should be noted that the RCT used in the case study was un-blinded which raises questions about the possibility of violating the exclusion restriction assumption. However, the authors reported that they have no reason to believe that has happened. The danger of violating the exclusion restriction assumption is based on an intuitive suspicion that individuals assigned to the intervention arm may have received more attention. If that is the case, then this may have had a direct effect on survival outcomes that is not caused through treatment received.

2.8.5 Instrumental Variable Methods

2.8.5.1 Origin of the method

The Instrumental Variable (IV) approach is an overarching term for a set of methods which could be used to recover the ACE of treatment in the presence of "unmeasured confounders". IV methods have traditionally been used by econometricians based on original work by Wright (1928) and Golderberger (1972) in the context of structural equation modelling (SEM). This approach has been adapted by Angrist et al.[64] with further work by Imbens and Rubin[97] to adjust for non-adherence using the counterfactual outcome framework. The IV method was developed to estimate the ACE of treatment received, rather than treatment assignment employed by the ITT analysis. The technique was originally developed in a non-time-to-event endpoint context, and further extended to time-to-event outcomes.

If the investigators are interested in estimating the causal effect but are no longer prepared to assume no unmeasured confounding because they have not collected data on potential confounders, then the IV approach provides an attractive alternative. The key characteristic of this method is that it does not

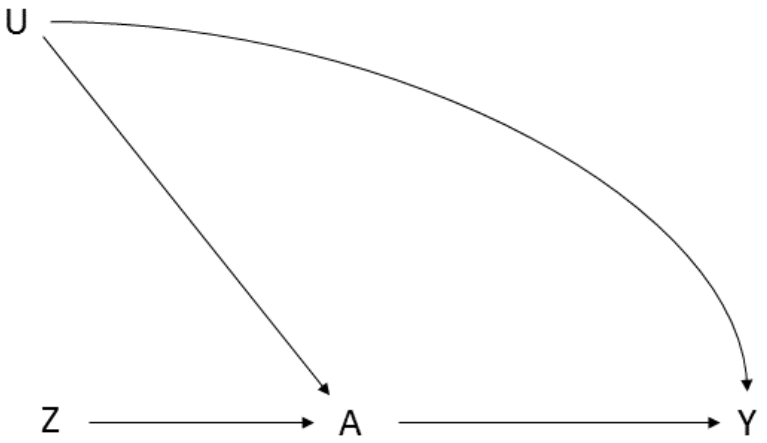
rely on the strong assumption of no unmeasured confounding. However, instead, it relies on the exclusion restriction assumption, which is described in the next section.

Three versions of the IV approach have been identified for adjusting for non-adherence in the context of time-to-event outcomes: (i) IV with Likelihood Estimator;[80] (ii) IV with Plug-in Non-Parametric Empirical Maximum Likelihood Estimator (PNEMLE);[81] and (iii) Transformation promotion time cure model with MLE to estimate the CACE and the compliers effect on survival probability (CESP) estimands.[82] I discuss method (i) in Sections 2.8.5.2and 2.8.5.3 and further discuss the two extensions in Section 2.8.5.4.

2.8.5.2 Theoretical characteristics

The concept of this method is based on SEM where the ACE can be estimated in two steps. The key assumption to all IV methods is the requirement to identify a pre-exposure 'un-confounded' instrumental variable that satisfies the "exclusion restriction assumption", that is the IV affects the survival outcome only through its effects on the exposure. This is best explained by a causal DAG, as presented in Figure 10. In this DAG, the instrumental variable Z affects the outcome Y only through the exposure A (e.g. adherence to the assigned treatment). The absence of an arrow from Z to Y indicates no direct causal effect of the IV on the outcome which satisfies the exclusion restriction assumption. The variable U represents an unmeasured confounder (e.g. genetics). The exclusion restriction assumption is likely to hold in double-blinded RCTs where randomisation is used as an instrument. The IV variable must be un-confounded; this condition implies that Z is independent of U.

Figure 10: Instrumental variable causal DAG



The IV method works by adjusting for binary non-adherence for estimating the causal effect using the CACE estimand. This can simply be done using a two-stage estimating process which can be applied easily in standard software. (a) estimate the effect of the Z on treatment initiation to predict A for each individual patient; and (b) estimate the effect of the predicted treatment initiation on the outcome Y, which will produce the causal effect of A on Y among the compliers subpopulation.

The approach of Baker[80] approach used the randomisation factor, denoted *(Z)* in the DAG. This is best explained by imagining an RCT where individuals are randomised to the intervention group (Z=1) or control group (Z=0). Individuals in the intervention group who immediately initiate treatment are denoted T_1 and those who refuse to initiate treatment denoted (T_0). Individuals randomised to the control arm are also denoted T_0. The method works by considering an experiment classifying individuals in the trial population into four groups (similar to the classification used by Lin et al. for the MCC method) using a variable (R) to indicate the subject type as outlined below:

N - Never-takers (T_0 if either Z=0 or Z=1)

C – Adherers (T_1 if Z=1 and T_0 if Z=0)

D – Defiers (T_0 if Z=1 and T_1 if Z=0)

A – Always takers (T_1 if either Z=1 or Z=0)

Because the randomisation variable R can be considered as a baseline covariate, its distribution will be independent of Z and therefore R can be used as an IV. Under specific assumptions (outlined below) the maximum likelihood estimation is formulated using cause-specific hazards. This is used to compute the probability of having the cause-specific event of interest (e.g. breast cancer death) at time *t* for each latent adherence type. Then the joint probability of receiving the treatment at time *t* across adherence types should be estimated (e.g. the sum of the joint probabilities of T_0 among N and C types conditional on Z=1). Then the treatment effect in terms of the difference in hazards or hazards ratio can be computed. It is worth noting that individuals in group D and some individuals in group A are actually switching treatment rather than exhibiting non-adherence. This issue was observed in a number of methodological papers included in this review. Group D is excluded from the analysis by the "monotonicity assumption" associated with this method (see next paragraph).

In the context of survival analysis, The IV method makes the following assumptions:

(a) The exclusion restriction assumption (discussed above)
(b) The randomisation assumption
(c) Randomisation has no effect on the probability of adherence to treatment
(d) Monotonicity assumption which excludes defiers

(e) Independence of failure time and censoring times
(f) Only compliers and never takers are included in the model

The exclusion restriction is a strong assumption which is difficult to evaluate systematically. However, evaluation of this assumption could add credibility to the adherence-adjusted causal effect estimated using the IV method. There are a number of formal tests which can be used to evaluate the IV assumptions such as the "IV inequality test" which can detect extreme violations.

Regarding potential biases, the method considers death from competing risks as a type of non-ignorable missing data, and this might be problematic. This is because, in some cases, death from the cause of interest and death from other competing risks may both depend on treatment adherence, making the reliance on the randomisation assumption more vulnerable to bias.

The IV method with the likelihood estimator is easy to apply, and if the assumptions hold, the method can produce a valid estimate of causal effect in the presence of non-adherence. However, the method estimates the CACE estimand which is not marginalised to the entire study population. The method is widely accepted in the economic literature, and its results would be easy to communicate. The MLE used in this method can produce negative estimates of hazards. This problem can be avoided by fitting a polynomial logistic model of the cause-specific hazards. The polynomial model can also smooth estimates over time. As a disadvantage, the method may be inefficient in some cases. The reason for this asymptotic inefficiency relates to the standard IV approach observed in non-survival analysis settings. Another disadvantage is that, finding an instrumental variable that meets all the criteria of a valid IV can be challenging. More importantly, the method uses the CASE estimand which might not be relevant to HTA policymakers and this is a limitation (This is discussed in further detail in Chapter 3).

2.8.5.3 Application in a case study

The IV method with likelihood estimator was applied in a case study for estimating the ACE of breast cancer screening on survival using secondary data from the Health Insurance Plan for Greater New York (HIP) study.[80] The HIP study had a sample size of 60,000 women randomised to screening (intervention group) and usual care (control group). The analysis estimated the causal effect in terms of life-years saved adjusted for initiation type of non-adherence. Baker[80] further applied a polynomial model to estimate adherence-adjusted cost-effectiveness. Further details are reported in Baker's paper.[80]

2.8.5.4 IV extensions

IV with Plug-in Non-Parametric Empirical Maximum Likelihood Estimator (PNEMLE)

The IV with Plug-in Non-Parametric Empirical Maximum Likelihood Estimator (NELME) was developed by Nie et al.[81] This is an extension to the standard IV approach for improving efficiency by making use of the mixture structure in the data without making any assumptions on outcome distributions. The mixture structure was previously used by Imbens and Rubin in a latent compliance class model. The method classifies individuals in a two-arm RCT into four adherence classes (always takers, compliers, never takers and defiers) as previously defined. This means there will be a mixture of compliers and never takers in the control arm. The method relies on the same assumptions reported in Section 2.8.5.2.

The PNEMLE approach assumes the following survival functions for compliers in the intervention group denoted as $S_{s1}(V)$ and control group denoted as $S_{c0}(V)$, while never-takers have a similar survival function in both groups denoted as $S_{nt}(V)$.

$$S_T|R = 1(V) = \pi_c S_{c1}(V) + (1 - \pi_c) S_{nt}(V) \qquad [5]$$

$$S_T|R = 0(V) = \pi_c S_{c0}(V) + (1 - \pi_c) S_{nt}(V) \qquad [6]$$

where π_c is the fraction of compliers in the intervention group.

The PNEMLE estimator works by imposing a constraint in the estimation such that the probabilities of survival among never-takers at that particular time point are similar in both arms. Based on this constraint, and assuming that the proportions of compliers are similar in the two arms, the following three analytical steps should be followed to estimate the difference in survival probability among compliers (in the two arms) at a specific time point:

(a) Estimate the survival probability among compliers in the intervention group [$S_{c1}(V)$], the survival probability for never takers $S_{nt}(V)$ in the two groups, and the observed proportion of compliers in the intervention group π_c. The first two parameters can be estimated using the Kaplan-Meier estimator.

(b) Estimate survival probability $S_{c0}(V)$ in the control arm using the non-parametric empirical likelihood estimation.

(c) Estimate the treatment effect using estimates from steps (a) and (b).

The PNEMLE method was evaluated in a simulation study and also applied in a case study showing how the analytical steps should be applied. A two-arm RCT data with a sample size of 2000 patients and 1000 simulations were used. The simulation study was restricted to include only compliers and never takers. Generally, the PNEMLE performed well compared to the standard IV method in terms of unbiasedness and efficiency. Relative bias was 11.1% in the worst situations, and the PNEMLE is proved to be at least as efficient as the standard IV approach, and in many cases more efficient (28% reduction in RMSE).[81] An advantage is that PNEMLE does not rely on parametric assumptions used in accelerated failure time models. As a disadvantage, a finite sample size bias was observed in the simulation study. Further details about the asymptotic properties of the PNEMLE estimator are reported in Nie et al.[81]

Nie et al.[81] reported that they had not investigated a situation whereby the probability of non-adherence depends on baseline covariates. This is a serious concern as adherence is more likely to depend on baseline patient characteristics such as comorbidity and it is not clear if this method would produce an unbiased estimate under these situations. This issue falls under a broader methodological question of estimating causal effect under dependence. However, in theory, IV methods have the advantage of not relying on the no unmeasured confounding assumption.

In the case study, the PNEMLE estimator was applied to the HIP study (an RCT designed for estimating the effectiveness of breast cancer screening on survival).[81] The PNEMLE method produced very similar estimates and confidence intervals compared to the standard IV method. The PNEMLE estimated a 12.3% higher probability of survival over 10 years among compliers who received treatment compared to those who adhered to receiving controls.[81] The ITT estimates are smaller than PNEMLE indicating that the level of non-adherence in the HIP study was very high. However, the estimands from the two approaches are not directly comparable to the ITT. This is because the ITT estimates the effect of treatment assignment on the survival outcome; while the PNEMLE estimates the effect of treatment actually received.

IV using transformation promotion time cure model

Gao and Zheng[82] proposed a semi-parametric causal transformation model with MLE for estimating the CACE and CESP estimand in a two-arm RCT setting with non-initiation (all-or-nothing) non-adherence. This models the individual's survival time in the intervention and control group, conditional on covariates and latent adherence type based on a key assumption that censoring time and survival time are independent.

The method makes the same IV assumptions described by Angrist et al.[64] and Nie et al.[81] – see Section 2.8.5.2. The randomisation assumption used is a weaker version conditional on covariates to allow for confounders to be taken into consideration. The method makes an additional assumption known as the "conditional non-null compliance class" assumption which implies the existence of compliers in the data. The cured individuals considered by the CACE are assumed to have infinite survival time.

For the potential outcome (event time) detonated by T^z, the conditional survival function in the transformation promotion time cure model is provided by the following formula:

$$S^z\left(t\middle|X_i = \sum_{k=1}^{3} G(v(\beta_k^c + \eta_{k1}^c z + \eta_{k2}^{cT} X_i\right) F_0(t)\, I(U_i = k) \qquad [7]$$

where $v(.)$ is an unknown link function, G(.) is an unknown transformation function, F_0 denotes a baseline distribution function and $I(.)$ represents the indicator function. The parameter of interest η_{k1}^c is the causal effect in class K, η_{k2}^c is a vector of parameters that relates the covariates to the survival function and β_k^c is the intercept in the regression output.

Based on the causal model in the above formula, the CACE estimand for cured patients is given by the following equation:

$$CACE(x) = S^1(\infty|U = 2, X = x) - S^0(\infty|U = 2, X = x) \qquad [8]$$

Considering the CESP estimand for un-cured patients, the complier survival probability is given by this equation:

$$CESP(t; x) = S^1(t|U = 2, X = x, T < \infty) - S^0(t|U = 2, X = x\, T < \infty) \qquad [9]$$

However, the estimand specified by the above formulas is the effects at the individual patient level, conditional on covariates. Once the individual effects are estimated, then a population-level causal effect can be estimated using the following unconditional versions of equations.

$$CACE(x) = E_x[S^1(\infty|X) - S^0(\infty|X)|U = 2] \qquad [10]$$

$$CESP(t) = E_x[S^1(t|X, T < \infty) - S^0(t|X, T < \infty)|U = 2] \qquad [11]$$

where E_x is an expectation parameter related to the distribution of *X*.

This method was evaluated in a simulation study and further applied to real data in a case study. The mixture structure in the data and infinite-dimensional parameters in the model resulted in computational difficulties. These difficulties were dealt with using the EM algorithm proposed by this method to estimate the parameters. The CACE and CESP estimators used by this method were proved

to be consistent as well as asymptotically normal. The method showed reasonable performance and resulted in estimates closer to the "true" values and coverage was close to nominal levels. More details about the asymptotic properties of CACE and CESP estimators are reported in Gao and Zheng.[82]

The model was further demonstrated by the application in the HIP study data with a sample size of 60,696 women randomised to receive screening (intervention group, n= 30,131) or usual care (control group, n= 30,565). The analysis using the causal transformation model showed that the intervention (screening) has a beneficial effect on survival and cure rate, conditional on covariates. The CACE estimated that compliers (women who adhered to the screening intervention) have a 17.97% higher probability of being cured compared to women who adhered to usual care (no screening). For CESP, the causal effect among non-cured women who adhered to the assigned screening intervention has a 3.82% higher probability of survival over 3 years follow up compared to women who received usual care (conditional on being uncured).

2.9 G-methods

2.9.1 Marginal Structural Models with Inverse Probability of Censoring Weighting

2.9.1.1 Origin of the method

The inverse probability of censoring weighting (IPCW) is a method that can be used to address informative censoring. It can be used to address non-adherence by censoring non-adherers and then using IPCW for estimating the ACE of treatment using MSMs. Exchangeability in an RCT context implies that individuals in one randomised group are comparable to the other group (i.e. they have the same baseline risk had they received treatment assigned to the other group). The IPCW approach relies on 'conditional exchangeability', meaning that the outcome is independent of everything except the treatment, conditional on variables included in the model. The MSM estimates the marginal (unconditional) effect, that is the ACE of treatment across the study population, had everyone adhered to the assigned treatment.

The IPCW is a form of a generic method known as inverse probability weighting (IPW). IPCW was originally developed by Robins and Rotnitzky[98] to adjust for 'dependent censoring' by incorporating data on time-dependent confounders. The Robins and Rotnitzky[98] paper was excluded because it was focused on continuous and binary outcomes, which are both outside the scope of this review. Therefore, the description of the IPCW approach reported here is mainly based on the Robins and

Finkelstein paper, as it focused on time-to-event outcomes.[66] The IPCW method focuses on two key aspects: (i) the individual's weight is inversely proportional to an estimate of the conditional probability of remaining uncensored up to time *t*; and (ii) this estimate is obtained by fitting a time-dependent Cox PH model for censoring incorporating time-dependent prognostic factors. In other words, if we are concerned about informative censoring, we weight patients to avoid the related selection bias. Patients with similar characteristics to people who are censored are up-weighted to account for themselves and for the censored patients, thereby avoiding the selection bias associated with the censoring. In order for the method to work, the analyst needs data on baseline and time-dependent variables so that we can appropriately weight people who have not been censored according to their similarity to people who have been censored.

2.9.1.2 Theoretical characteristics

In the context of non-adherence, the IPCW method can be used to obtain a valid treatment effect by adjusting for baseline and time-dependent prognostic factors that predict both non-adherence and the survival outcome (i.e. confounders). To adjust for confounding in causal analysis, the analyst should adjust for all baseline and time-dependent confounders. The issue is that theoretically baseline confounders can be adjusted for using simple regression; however, time-dependent confounders cannot be adjusted for using simple methods. The IPCW is a method that can attempt to remove selection bias caused by such time-dependent confounders.

The IPCW makes the key assumption of 'no unmeasured confounders' which is also known as the assumption of 'ignorability of censoring'. Under this assumption and the exchangeability condition, the IPCW method can produce an adherence-adjusted treatment effect. The IPCW adjustment is based on the counterfactual-outcome framework (introduced in Section 1.3.1) for estimating weighted survival times in people who have not been censored. The analysis involves censoring patients at the first time they become non-adherent, which is likely to introduce informative censoring. The application of IPCW weighting attempts to remove this informative censoring by up-weighting individuals with characteristics similar to those who have been censored based on the exchangeability assumption. IPCW also requires a positivity assumption which implies that the probability of non-adherence is non-zero. The method assumes that adherence is binary which could be considered as a limitation of accounting for implementation non-adherence.

The IPCW estimation involves four steps:

(i) Censor observations at the time of non-adherence;

(ii) Model the probability of censoring (non-adherence) for each individual patient at each time interval *k* adjusting for baseline and time-dependent confounders;

(iii) Compute the probability of remaining uncensored for each individual patient up to time *t*; and

(iv) Use the inverse probabilities of remaining uncensored as weights in a weighted analysis which can be applied to any survival analysis (e.g. MSM) to estimate the ACE of treatment. In step (ii), it is very important to fit non-adherence models (e.g. logistic models) separately for each randomisation arm, as the reasons for non-adherence (or censoring) may differ by treatment assigned to each arm.

The IPCW weights produced from the inverse probabilities will be 'unstabilised' which are valid but not efficient. According to Robins and Finkelstein, 'stabilised' weights should be used instead because weighting can be inefficient compared to straightforward regression adjustment if only baseline variables are important. To estimate the stabilised weights, the analyst should first construct the un-stabilised weight ($\widehat{w}_{it}$) for each individual *i* in time interval *t* by multiplying all the probabilities of remaining uncensored (adhered) up to time *t* using the following equation.

$$\widehat{w}_{it} = \prod_{k=0}^{t} \frac{1}{1-\hat{p}_{ik}} \qquad [12]$$

where $\hat{p}_{ik}$ is the predicted probability of non-adherence in time interval *k* given the randomisation group and adjusting for baseline covariate and time-dependent covariates. The stabilised weights for each individual ($\widehat{w}_{it}^{stab}$) can then be estimated using equation [13].

$$\widehat{w}_{it}^{stab} = \prod_{k=0}^{t} \frac{1}{1-\hat{p}_{ik}} / \prod_{k=0}^{t} \frac{1}{1-\hat{p}_{0ik}} = \prod_{k=0}^{t} \frac{1-\hat{p}_{0ik}}{1-\hat{p}_{ik}} \qquad [13]$$

where $\hat{p}_{0ik}$ is the probability of non-adherence given the randomisation group and adjusting for baseline covariates only. $\hat{p}_{0ik}$ should be estimated using the same model applied to obtain $\hat{p}_{ik}$ in the denominator of the stabilised weights equation, but without including time-dependent covariates. In the stabilised model [13], baseline variables cancel out as they are in the numerator and denominator, so if time-dependent variables are actually not important, the weights will just be equal to 1. Then, because the baseline variables are included in the outcomes model, the method collapses to just a straightforward regression adjustment.

A pseudo-population should then be created by weighting each individual in the study population using the stabilised weights obtained from equation [13]. The last step is fitting the outcome model adjusting for baseline covariates that have been used in the numerator of the stabilised weights in

equation [13]. The IPCW Kaplan-Meier estimator and Cox partial likelihood estimators (or any other survival analysis) can be fitted to obtain survival curves and estimates of HR weighted by the inverse of the conditional probability of remaining uncensored using stabilised weights. This estimation will produce a valid treatment effect adjusted for non-adherence.

One problem with applying this approach is that it is impossible to formally test whether there are no unmeasured confounders, however, there are some strategies which can be followed to assess whether it is a feasible assumption given the data. Reviewing the literature and developing a causal DAG can help understand the relationships between covariates in order to identify known important variables which are potential confounders. To identify the prognostic factors of interest, one could fit a model (e.g. time-dependent Cox model of failure) which includes all potential confounders and then keep only those which are statistically significant. There are some limitations to this approach – for example, a prognostic factor can be important without being statistically significant. In other words, the analyst should just include anything believed to be prognostic irrespective of significance. Another approach is to run tests of independence between the potential confounders, the exposure and the outcome to determine if there are common causes.

In terms of advantages, the IPCW method (similar to the other g-methods) theoretically has a greater power to detect the effect of treatment in the presence of non-adherence compared with the standard ITT-analysis approach. The method may allow us to estimate a valid causal effect at different counterfactual (non-observed) levels of adherence if the assumptions hold. The main limitation of IPCW is the assumption of no unmeasured confounders, which cannot be proven empirically.

2.9.1.3 Application in a case study

Robins and Finkelstein[66] applied the IPCW approach to adjust for non-adherence using data from the AIDS Clinical Trial Group 021 study to estimate the treatment effect of Aerosolised Pentamidine (AP) (versus Bactrim) on overall survival. The study enrolled 310 patients with time to *pneumocystis carinii pneumonia* (PCP) recurrence used as a primary outcome and survival as a secondary outcome. The IPCW analysis used stabilised weights and both the IPCW Kaplan-Meier and IPCW Cox partial likelihood estimators were used in two separate analyses. The analysis included three time-dependent prognostic factors: haemoglobin levels, Karnovsky score (a measure of functional status), and asthenia score (a measure of weight loss and lean body mass). The three prognostic factors that met the conditions of being confounders were selected by fitting a stratified time-dependent Cox model for failure - the investigators only kept these three factors which were significant at the p=0.12 level. As reported, it seems that this was the sole criterion used to assess whether the confounding conditions

were not met and the cut-off point seems rather arbitrary. After checking the data, the investigators eliminated three other potential prognostic factors – white cell count, PCP episodes and CD4 count from the adjustment as they did not meet the conditions of being confounders. This means they were not common causes of the exposure and the outcome.

Robins and Finkelstein[66] undertook four causal analyses: (a) standard ITT analysis with censoring applied at the point of the patient being lost to follow-up or end of follow-up; (b) applying dependent censoring at the point of the patient being lost to follow-up and treatment switching; (c) applying dependent censoring at the point of the patient being lost to follow-up, switching and discontinuation (for non-medical reasons); and (d) applying dependent censoring at the point of patient being lost to follow-up, switching and discontinuation (for any reason).

Robins and Finkelstein[66] applied the IPCW log-rank test for adjusting for dependent censoring which takes into account data on the included time-dependent confounders. The IPCW partial likelihood estimate of mortality HR (AP compared to Bactrim) was 0.84 (IPCW 95% confidence interval [CI]: 0.21-1.57, z-value of IPCW log-rank test=2.55).[66] The ITT estimate of HR was 0.63 (95% CI: -0.1-1.5, z-value of log-rank test=1.7).[66] The IPCW estimate shows that the log-rank z-score is greater than 1.96 and the 95% CI excludes zero which implies that there is strong evidence that Bactrim has a beneficial effect on survival compared to AP therapy. This was not the case in the ITT estimates where the 95% CI includes zero, although the estimated IPCW HR is not expected to be numerically higher than the ITT estimate. If the assumptions hold, the IPCW estimates imply that Bactrim therapy has a beneficial causal effect on survival, but that the standard ITT analysis failed to detect it. If the estimand of interest is the causal effect of treatment in the presence of patient non-adherence, then the ITT estimate is not comparable to the IPCW estimate. The key point is that IPCW has a different estimand, that is the causal effect that would have been observed had everyone adhered to the assigned treatment; whereas ITT estimates the effect of treatment assignment in that particular study.

An additional issue that is raised by the Robins and Finkelstein study is the importance of the way in which censoring is specified within the analysis. It should be noted that the different ways of specifying censoring in analyses (b), (c) and (d) produced different estimates of the causal effect; and therefore, it is crucial to specify a censoring mechanism that meets the definition of non-adherence introduced in Section 1.2.

2.9.1.4 MSMs Extensions: Inverse Probability of Treatment Weighting (IPTW)

Origin of the IPTW method

Hernan et al.[83] introduced an MSM with the inverse-probability of treatment weighting (IPTW) as another class of estimators. This method relates to IPCW as they both belong to the broader IPW approach. IPTW was originally developed for estimating the ACE of treatments from observational data with time-dependent confounders. Similar to IPCW, the IPTW approach is based on the counterfactual-outcome framework and is mainly focused on time-varying treatments. The method also relies on conditional exchangeability. The key feature of this method is that it allows for modelling longitudinal adherence patterns where patients follow erratic adherence behaviours in terms of implementing the prescribed dosing regimen (i.e. on/off adherence patterns).

Hernan et al.[83] argued that even if the no unmeasured confounding and model misspecification assumptions hold, standard ITT analysis estimates of the ACE of time-varying treatments would be biased if two conditions are met: (a) there exist time-dependent prognostic factors that predict the survival outcome and subsequent treatment, simultaneously; and (b) treatment history is a predictor of subsequent risk level. If these two conditions are met, then the proposed IPTW estimator can be used instead of standard (unweighted) analysis for estimating an unbiased causal effect of treatments in the presence of patient non-adherence.

IPTW theoretical characteristics

In the IPCW approach (introduced in Section 2.9.1), individuals were censored after they first become non-adherent to their assigned treatment and they remain censored for the rest of the study follow-up. This approach might not be appropriate in situations time-varying non-adherence to prescribed chronic medications. As an alternative approach, one could allow individuals to become adherent again following a period of being recorded as non-adherent - this can be modelled using the IPTW approach. The time-varying adjustment procedure offered by the IPTW approach could be used to adjust for non-adherence at the implementation stage. In other words, IPTW works in the same way as IPCW except that non-adherers are not censored, but their non-adherence indicator just switches to “1” when they are non-adherent and then back to “0” when they become adherent again over time.

Another characteristic relating to IPTW (and IPCW) is the concept of treatment ‘causal exogeneity’ which is important for understanding the interpretation of causal parameters obtained from a correctly specified MSM. Assessing whether the treatment is causally exogenous requires ‘statistical exogeneity’ to be met, that is the probability of initiating the treatment at time *t* is independent of the history of time-dependent confounders up to time *t,* conditional on the history of treatment (i.e. adherence to the treatment before time *t*).

IPTW makes the no unmeasured confounders assumption as used by IPCW (see Section 2.9.1.2). IPTW assumes non-informative censoring and no model misspecification given covariates measured in the past and treatment history.[83] As an aside, if we have varying adherence over time, and informative censoring, one could potentially combine IPCW and IPTW to adjust for both problems in the causal analysis.

The analytical steps are similar to those described for IPCW, and the main difference is the censoring mechanism as described above. To introduce the MSM with IPTW model, I use an example from Hernan et al. for estimating the joint causal effect of AZT and PCP prophylaxis therapy on mortality of HIV-positive patients. The marginal structural Cox model specified by Hernan et al. is provided in equation [14]: [83]

$$\lambda_{T_{\bar{a},\bar{c}\equiv 0}}(t|V) = \lambda_0(t)\exp\left[\beta_1 a_1(t) + \beta_2 a_2(t) + \beta_3' V\right] \qquad [14]$$

where, $\lambda_{T_{\bar{a},\bar{c}\equiv 0}}$ is individual's failure time if he had received treatment regimen $\bar{a}$ under the ignorable (non-informative) censoring assumption. The ignorable censoring assumption implies that the conditional hazard of censoring at time *k* is independent of failure times given the past AZT treatment $\bar{A}_1(k-1)$, prophylaxis history $\bar{A}_2(k-1)$ and the history of time-dependent prognostic factors before *k*. *V* is a vector of pre-treatment (baseline) covariates and $\exp(\beta_1)$ and $\exp(\beta_2)$ are the causal parameters of interest for AZT treatment and PCP prophylaxis therapy, respectively. The IPTW estimate of the parameter β can be obtained using a weighted-logistic-regression model with weights calculated using the following formula:[83]

$$W(t) = \prod_{k=0}^{t} \frac{f[A(k)|\bar{A}(k-1),V]}{f[A(k)|\bar{A}(k-1),\bar{L}(k)]} \qquad [15]$$

where A indicates whether the patient is on treatment or prophylaxis, $\bar{A}(k-1)$ is the treatment history and $\bar{L}(k)$ is the history of measured time-dependent confounders up to and including time *k*. It is worth noting that V was not included in the denominator because it is a subset of covariates $\bar{L}(k)$, so it is already included. Under the assumptions, a consistent IPTW estimator of the causal parameter can be obtained by multiplying the individual's probability of remaining alive and uncensored at time *t* by $W(t)$ x $W^{\dagger}(t)$, where[83]

$$W^{\dagger}(t) = \prod_{k=0}^{t} \frac{Pr[C(k)=0|\bar{C}(k-1)=0,\bar{A}(k-1),V,T>k]}{Pr[C(k)=0|\bar{C}(k-1)=0,\bar{A}(k-1),L(k),T>k]} \qquad [16]$$

$W^{\dagger}(t)$ is the inverse probability of un-censoring up to time *t* divided by the probability estimated without including time-dependent predictors of censoring, but only including treatment history and

baseline covariates. The calculated weights should then be used to create a pseudo-population in which the treatment and prophylaxis therapy are statistically and causally exogenous. The causal effects (estimate of parameter β) should then be estimated using the MSM specified in equation [14].

MSMs with IPCW/IPTW estimators have the advantage that they are considered as a natural extension to standard models (e.g. standard logistic models, time-dependent Cox models) which makes it convenient to communicate their causal effect estimations.[83] This is because once a pseudo-population is created with the weights, any standard survival analysis can then be performed to estimate a valid causal effect. IPTW can be used to adjust for non-adherence in situations where individuals are allowed to become adherent again following a period of non-adherence. However, the additional information gained by modelling re-adherence could be offset by the additional variability introduced in the model, which is a disadvantage.[66] Additional assumptions may also be introduced when modelling re-adherence.

IPTW application of in a case study

Hernan et al.[83] applied the IPTW approach in a case study using data from the Multicentre AIDS Cohort Study (MACS) to estimate the joint causal effect of AZT and PCP prophylaxis therapy on mortality of HIV-positive men. The MACS study enrolled more than 5000 participants with a median follow-up of 67 months. However, Hernan and colleagues restricted their causal analysis to 2168 participants who had no diagnosis of AIDS and had not initiated AZT treatment or PCP prophylaxis therapy at the time of starting their follow-up.

As a first step in the analysis, the investigators assessed if conditions (a) and (b), introduced above, are true given the data. As a reminder, (a) existence of time-dependent prognostic factors that predict the outcome and subsequent treatment; and (b) treatment history is a predictor of subsequent risk level. To determine whether condition (a) holds, they fitted an unweighted time-dependent Cox model that included baseline and time-dependent covariates (CD4 count and PCP), AZT and prophylaxis variables. The model was used to assess whether CD4 count and PCP were independent predictors of the survival outcome (death). The model estimated an HR of 3.77 ($p<0.001$) for PCP before time *t*.[83] For low CD4 count and moderate CD4 count (with respective values of <200 and 200-500 relative to a normal value of $\geq$500), the corresponding estimates of HRs were 16.5 and 3.48 ($p<0.001$ for both estimates), respectively.[83] To assess if the current CD4 count and PCP were predictors of subsequent treatment, they fitted two models for AZT treatment initiation and prophylaxis therapy (pooled logistic models) including baseline covariates (V), PCP and CD4 count as time-dependent covariates. The treatment initiation model resulted in estimated HRs of 2.18, 3.38

and 2.56 (p<0.001 for all) corresponding to PCP incidence, low CD4 count and moderate CD4 count covariates, respectively.[83] In the prophylaxis initiation model, the corresponding HRs obtained were 1.87, 4.94 and 2.82 (p<0.001 for all estimates). These results imply that condition (a), that there exist time-dependent prognostic factors that predict the survival outcome and subsequent treatment, holds in the MACS dataset.

To identify confounders, they fitted a logistic model to estimate the probability of developing PCP in time *t* given baseline and time-dependent covariates, AZT treatment and prophylaxis history at time t-1. The model estimated HRs of 1.03 (p=0.64) and 0.77 (p<0.001) for AZT treatment and prophylaxis therapy, respectively.[83] This means that prophylaxis is a predictor of subsequent treatment, that is a protective factor for PCP which informs subsequent treatment initiation. This result implies that condition (b) also holds in the MACS dataset. Subsequently, the authors concluded that the standard ITT analysis could not be used for estimating a valid causal effect. The problem associated with condition (a) is that standard ITT analysis does not allow for adjusting for time-dependent covariates (CD count and PCP) which implies confounding bias. Regarding condition (b), the incorporation of current CD4 count and PCP as covariates in standard analysis implies bias because these variables are affected by previous treatment.

As both conditions (a) and (b) hold in the dataset, Hernan et al.[83] have applied an MSM with IPTW estimator as introduced above. The IPTW weighted analysis produced a mortality HR of 0.67 (95% CI: 0.46-0.98) for individuals who initiated the AZT treatment compared to those who did not initiate the treatment. For prophylaxis therapy, the estimated HR was 1.14 (95% CI: 0.79-1.64) which represents an adjusted treatment effect.[83] The causal interpretation of the result is that AZT treatment has a beneficial effect in reducing mortality of HIV-positive men. In contrast, the HRs estimated by the unweighted (ITT) analysis was 1.85 (95% CI: 1.49-2.30) and 1.58 (95% CI: 1.31-1.89) for AZT treatment and prophylaxis therapy.[83] This result clearly shows that failure to adjust for time-dependent prognostic factors (i.e. CD count and PCP) in standard ITT analysis has resulted in failure to detect the beneficial effect of AZT treatment.

2.9.2 Structural Nested Failure Time Models with G-estimation

2.9.2.1 Origin of the method

Structural Nested Failure Time Models (SNFTMs) were originally developed by Robins et al.[84] SNFTMs belong to the wider class of SNMs. The method is based on the counterfactual outcome framework (introduced in Section 1.3.1). SNFTMs were originally proposed to adjust the causal effect of treatment

for time-dependent risk factors, in general. The method was further applied to adjust for non-adherence for time-to-event outcomes in the context of observational studies and RCTs. The method uses the g-estimation technique (described in the next sub-section) to estimate the parameters of interest in the SNFTM.

2.9.2.2 Theoretical characteristics

The SNFTM relates the individual's observed time of the event (e.g. time of death) and observed treatment history to the counterfactual outcome (the time at which the individual would have died) if the treatment had been withheld. This approach will be essential to control for potential bias in which there exists a confounder (prognostic factor) for the survival time outcome that: (i) influences subsequent treatment exposure; and (ii) itself can be predicted by previous treatment received (i.e. time-dependent confounder). Adjusting for such confounders using traditional methods such as Cox PH models will produce biased estimates.[13]

The method makes the assumption of "no unmeasured confounders" to estimate the causal effect of treatment conditional on confounding variables. According to Robins et al.,[84] SNFTMs makes two further assumptions: (a) the causal model for the effect of the treatment on survival outcome is correctly specified; and (b) the recorded treatment data are accurate (no measurement errors). It may be difficult to verify or test if these assumptions hold. This is used in an AFT mode assuming treatment effect can be summarised as a time ratio (as opposed to an HR). It also assumes that adherence is binary which could be considered as a limitation of accounting for implementation non-adherence.

To explain how the method works for adjusting for non-adherence, I use a simple SNFTM. The simple treatment effect model using the SNFTM can be written as follows:

$$Y^{(0)} \sim \int_0^Y e^{\psi A(t)}\, dt \qquad [17]$$

where $Y^{(t)}$ denotes the counterfactual outcome at time t (survival time), $\sim$ means *"has the same distribution as"*, multiplied by a factor e^{ψ} if treatment is withheld, and the observed treatment variable $A(t)$ =1 if the individual initiated treatment and $A(t)$ =0 if not.

If the individual is alive at time t, the counterfactual survival time is provided by equation [18] below.

$$Y^{(t)} - t \sim \int_t^Y e^{\psi A(t)}\, dt\ given\ \left(\overline{L}_t, \overline{A}_t\right) \qquad [18]$$

where $\overline{L}_t$, is the recorded history of time-dependent covariates up to time t, $\overline{A}_t$ is the history of treatment received, e^{ψ} is an unknown parameter representing the factor by which the individual's remaining survival time is expanded or contracted by initiating treatment (adhered) at time t. The unknown parameter e^{ψ} is the "acceleration factor [AF]" with negative value means a beneficial effect of the treatment. This is the parameter of interest that should be estimated using the g-estimation procedure.

The g-estimation technique uses a class of estimators known as the "g-estimators". The g-estimation is considered as a generalisation and improvement of the g-formula (g-computation algorithm) and the associated G-null test previously proposed by Robins (1986) and Robins and Greenland (1992). In the SNFTM framework, g-estimation is used to search for ψ value which adds the least to the prediction model of treatment initiation. This means we search for a value of $\hat{\psi}$ that results in a ψ term having a coefficient of zero in the model for treatment initiation (i.e. treatment initiation is independent of counterfactual outcomes).

The no unmeasured confounding assumption implies that $Y^{(t)}$ (the potential outcome at time t) does not add to the prediction model for treatment initiation at time t. To formally explain the g-estimation procedure, let us assume the treatment effect model in equation [19].[99] The analyst could fit a logistic regression model to obtain the coefficients in equation [20].

$$Y^t \sim X_{\psi}(t) \;\; given \left(\overline{A}_t, \overline{L}_t\right) \qquad [19]$$

$$P[A(t)] \; = \beta_0(t) + \beta_1 A(t-1) + \beta_2 L(t) + \beta_3 X_{\psi} \qquad [20]$$

where Y^t is the observed survival time, $\sim$ means has the same distribution as, $X_{\psi}(t)$ is the counterfactual outcome, $\overline{A}_t$ is the past treatment, $\overline{L}_t$ is the history of covariates and $P[A(t)]$ is the probability of initiating the treatment at time t.

G-estimation is used to search for ψ value which adds the least to the prediction model (i.e. treatment initiation is independent of counterfactual outcomes). This means searching for a value of $\hat{\psi}$ that results in a X_{ψ} term having a coefficient $\beta_3 =$ zero in model [20]. That value of ψ provides the best estimates of counterfactual survival times adjusted for non-adherence.

To help further explain the g-estimation technique, I use Figure 11, which was adapted from Latimer et al.[100] to aid the explanation. This example from a hypothetical RCT illustrates the g-estimation process using the experimental group. The g-estimation process involves the following three steps, as

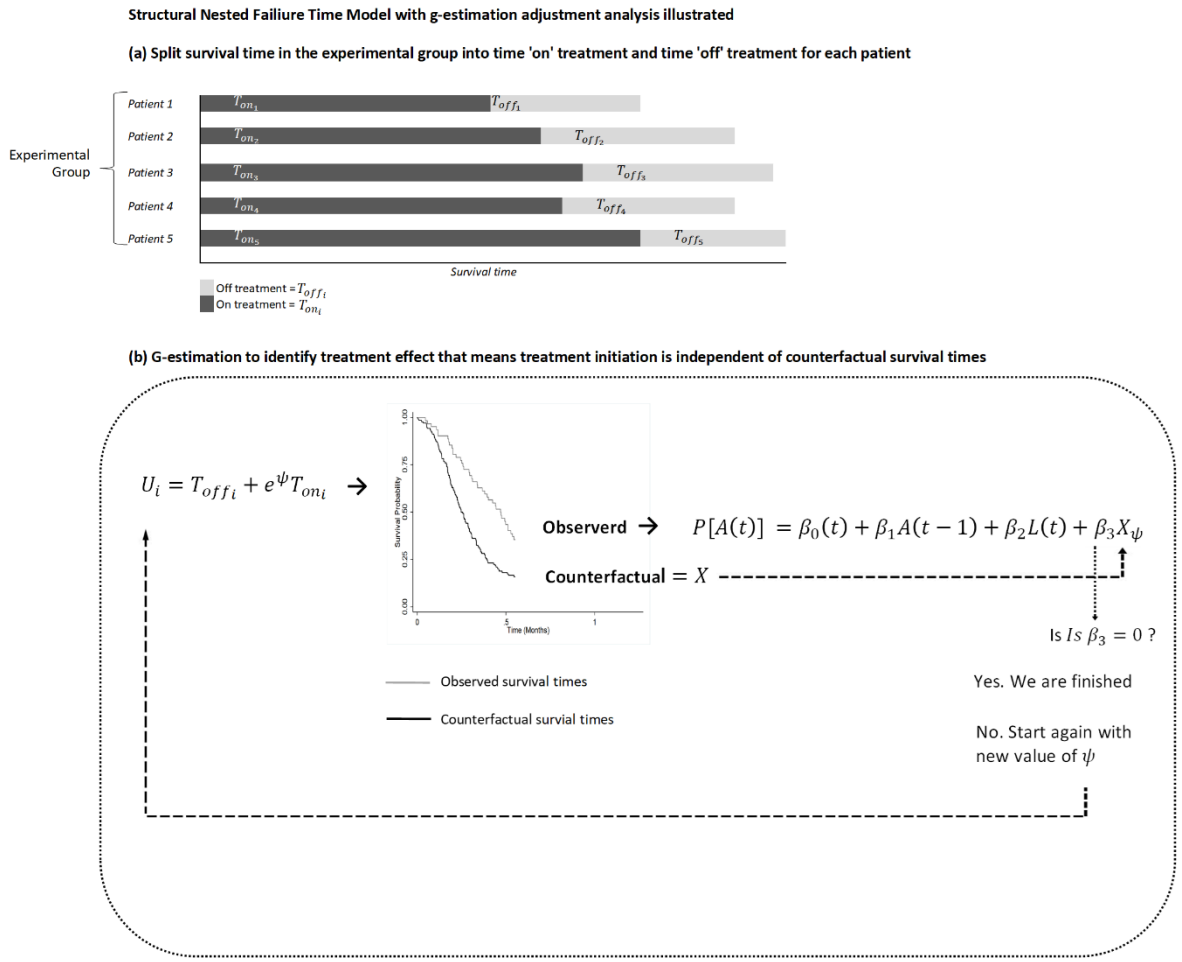
described by Latimer et al.[100] in the context of treatment switching and further adapted to explain the application in the context of adjusting for non-adherence using a SNFTM:

a) In the first step, for each individual patient in the experimental group, survival time is split into two parts: time 'on' the treatment when the patient adheres to the assigned treatment (denoted as T_{on_i}), and time 'off' the treatment (denoted as T_{off_i}) when the patient did not adhere to the treatment (see Figure 11a).

b) In the second step, assuming the abovementioned method assumptions hold, survival time is calculated as a function of time on and time off the treatment given the history of covariates (baseline and time-dependent confounders). In this step, g-estimation technique is used to estimate the 'true' treatment effect adjusted for patient non-adherence. G-estimation involves finding the value of ψ that results in treatment initiation being independent of counterfactual survival times. This is done through a 'grid search' of a range of possible values of the treatment effect, plugging each value of ψ into the counterfactual survival model, then plugging the resulting counterfactual survival times into a prediction model for treatment initiation (which includes all other prognostic covariates) and finding the value that adds the least to the prediction model (see Figure 11b). That value represents the best estimate of treatment effect which could be used to obtain e^{ψ} as an acceleration factor (AF) - the parameter of interest to be used to adjust for non-adherence in the next step. In practice, the g-estimation could be applied using standard software (e.g. Stata).[16, 101]

c) In this final step, the best value of the AF that gives the true treatment effect could then be used to adjust survival time for non-adherence for each patient in the experimental group. A similar approach could be applied to the control group (in case of two active treatments) to obtain the average treatment effect in terms of contrast, adjusted for patients' non-adherence (i.e. adjusted survival times in the experimental group compared to adjusted survival times in the control group had there been no non-adherence in both groups).

Figure 11: SNFTM with g-estimation adjustment analysis illustrated



Source: Adapted from Latimer et al.[100] This is an open access article distributed under the terms of the CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium.

SNFTMs are powerful in estimating the ACE of treatment in the presence of non-adherence taking into account how treatment effects depend on pre-treatment patient characteristics. It should be noted that the estimand used by SNFTM is for the entire study population, not just the compliers. The main limitation of SNFTMs is the potential biases related to the no unmeasured confounding assumption, which cannot be formally tested.

2.9.2.3 Application in a case study

Robins et al.[84] applied the SNFTM to a case study to assess the effect of prophylaxis therapy [high- versus low-dose AZT treatment for PCP on survival (time-to-death) of HIV-positive patients. The analysis used data from the AIDS Clinical Trial Group 002 RCT, which involved an embedded observational study designed to estimate the ACE of PCP prophylaxis on survival outcome. The analysis

was restricted to the low-dose AZT arm. Patients who remained alive and at risk at 10 months from randomisation were split into two groups depending development of PCP at 8 months. This was interesting because PCP will be a confounder for the outcome between 8-10 months. That means patients who develop PCP at 8 months have a higher probability of dying at 10 months and at the same time, they are more likely to receive treatment between 8 and 10 months because PCP is a predictor of treatment initiation. An SNFTM with g-estimation was applied to estimate the causal effect using the analytical steps described in Section 2.9.2.2.

The causal effect of prophylaxis on survival obtained using the SNFTM was a fractional change in life expectancy with a 95% CI of -0.18 to 0.16, indicating that there is no evidence of beneficial effect for PCP prophylaxis therapy. As the confidence interval includes zero, this finding means that PCP prophylaxis increases survival by 18% or decreases survival by 18% at the 0.05 confidence level.[84] Robins et al.[84] did not provide comparative data with no adjustment using traditional methods such as ITT analysis. However, Robins[36] provided empirical comparisons focused on robustness, plausibility, and strength of the assumptions used by this method compared to alternative methods for adjusting for non-adherence (see Chapter 3, Section 3.3).

2.9.3 Rank-Preserving Structural Failure Time Models with G-estimation

2.9.3.1 Origin of the method

Rank-Preserving Structural Failure Time Models (RPSFTMs) were originally developed by Robins and Tsiatis [37] to adjust for non-adherence. This is a randomisation-based method for adjusting for non-adherence using the counterfactual outcome framework. The RPSFTMs are semi-parametric, structural (or strong) versions of the accelerated failure time models (AFTMs) with time-dependent covariates. They are called “structural” because they directly model the counterfactual survival outcome from observed data. RPSFTMs belong to the broader class of SNMs which also include SNFTMs introduced in Section 2.9.2.

Loeys et al.[85] extended the RPSFTMs for adjusting for non-adherence in the context of cluster randomised trials (CRTs) and time-to-event outcomes. The argument is that cluster randomisation design could potentially involve different selective adherence levels on top of differences observed at the individual level. The description of this method is mainly based on Loeys et al.[85], and the extensions to this method are introduced in Section 2.9.3.4.

2.9.3.2 Theoretical characteristics

The RPSFTM works by developing a large sample for the rank estimators, deriving the optimal estimator, and then constructing some partially-adaptive estimators with efficiency approaching that of the optimal estimator. Loeys et al.[85] proposed two models to account for clustering: (a) structural failure time model; and (b) structural model using "Wald" test in three working models. The three working models proposed in the second approach are: (i) standard estimation ignoring clustering; (ii) robust approach with the same model in (i); and (iii) a marginal model with a gamma distribution.

RPSFTMs requires recensoring, which implies that the treatment effect may not represent the entire follow-up period as it is weighted towards an early stage of follow-up. Recensoring is used to address bias via informative censoring. However, this may be problematic as it involves discarding more information, especially when the treatment effect size is large. This is a key limitation of this method.

In an RCT context, the RPSFTM works to adjust for non-adherence using the randomisation factor and the observed survival time and treatment history. The method splits the observed event time for patient *i* into two parts: time on treatment *($T_{i,on}$)* and time off treatment *($T_{i,off}$)*. The observed time-to-event T_i can be computed using equation [21]:

$$T_i = T_{i,on} + T_{i,off} \qquad [21]$$

To estimate the treatment effect, the simple model multiplies only the $T_{i,on}$ part of the observed time, and therefore, the simplest outcome model becomes as follows:

$$T_i^0 = T_{i,off} + \exp(\psi) \times T_{i,on} \qquad [22]$$

where $\exp(\psi)$ is the acceleration factor introduced in Section 2.9.2.2. The g-estimation procedure can then be used to find the value of ψ using a grid search which can be applied using standard software (e.g. Stata).

The RPSFTMs approach makes the following assumptions for estimating the causal effect:

(a) The randomisation assumption;
(b) Equal treatment effect regardless of timing for receiving the treatment, but relative to the time treatment was taken. This is known as the assumption of "common treatment effect";
(c) The treatment of one individual has no influence on the outcome of another individual; and

(d) Survival times and treatment-free survival times are proportional by an unknown factor that depends on the exposure.

A simple RPSFTM (equation 23) can be constructed to estimate the counterfactual survival time (T_i^0).[37, 87]

$$T_i^0 = \int_0^{T_i} \exp\left[-\psi Z_i A_i(t)\right] dt \qquad [23]$$

where Z_i is the randomisation variable, A_i is a binary adherence variable which equals 1 when a patient initiated the treatment and 0 otherwise, T_i is the observed survival time and the factor $\exp(\psi)$ is the causal effect (the value by which survival time is shrunk or expanded as an effect of the treatment). At the "true" value of the parameter ψ (which we can find using g-estimation), the counterfactual survival between randomised groups will be equal and that value of ψ would be the point estimate of the treatment effect.

In the context of RPSFTM, g-estimation searches for the value of psi that gives equal average untreated survival times between randomised groups. It is designed specifically for use in a randomised context and uses randomisation as an instrument which means that it does not require the no unmeasured confounding assumption, but is reliant on exclusion restriction (randomisation assumption) and common treatment effect. These are the key differentiators compared to SNFTM. Moreover, in common with other g-methods, RPSFTM assumes binary non-adherence which might not be the best approach for implementation non-adherence.

In terms of disadvantages, the "equal treatment effect" assumption is particularly problematic as it might not hold in many situations. In addition, misspecified models may produce biased results, and in some cases, the analyst may face convergence problems when the g-estimation search for the best value of psi. In a situation where there is no clustering effect, the models proposed by this method may underestimate the treatment effect. Failure to include relevant baseline covariates relating to either clustering or individuals may produce estimates biased toward zero. However, based on the simulation study reported by Loeys et al.,[85] this bias is likely to be small. On the other hand, and under certain conditions, incorporating cluster-specific covariates may overestimate the variability producing conservative estimates. A key limitation of the RPSFTM is that it can only work for adjusting for non-adherence in a placebo-controlled trial making it inappropriate to account for non-adherence in RCTs evaluating two active treatments. This is not the case for the other g-methods (IPCW and SNFTM).

2.9.3.3 Application in a simulation study and a case study

The RPSFTM was evaluated using a simulation study and further applied to a case study. The case study used data from a placebo-controlled CRT to assess the causal effect of Vitamin A supplements on the short-term survival of infants (aged <6 months) with Vitamin A deficiency. The study recruited 9178 infants from 261 wards (villages in Nepal) used as clusters with 6 4-monthly visits. The causal analysis was adjusted for treatment initiation type of non-adherence. The application in the case study showed how the RPSFTM could be used to account for clustering with differential levels of non-adherence with a marginal approach (ignore clustering) and frailty approaches (adjust for clustering) used in the analysis. ITT analysis was also performed for comparison.

The non-adjusted ITT results using KM estimator showed no beneficial effect on 4-months, although the results show a significant effect in the longer term (24 months). The RPSFTM results from the different causal analysis performed produced slightly different estimates compared to non-adjusted ITT analysis. However, only one causal analysis where the point estimate $\hat{\psi}$ had a lower bound of the 95% CI greater than zero was found, indicating beneficial effect. Further details about the case study are reported in Loeys et al.[85]

The simulation study was designed to assess two approaches: (i) a marginal approach (ignoring clustering or robust approach); and (ii) a frailty approach (adjusting for cluster effect on adherence). The main findings from the simulation study suggest a larger (but non-significant) effect of clustering using the RPSFTM compared to unadjusted ITT analysis. The study showed that the RPSFTM performed well under controlled conditions. Further details about the findings of the simulation study are reported in Loeys et al.[85] The paper concluded that researchers should adjust for clustering and non-adherence when estimating treatment effect in CRT, which can be done using an RPSFTM without affecting the causal interpretation or adding any further complications to the analyses.

2.9.3.4 RPSFTMs Extensions

Incorporating baseline covariates to improve the precision of estimators

The RPSFTM approach was extended by Korhonen and Palmgren [86] to include baseline covariates for improving the precision of estimators of the structural effect. Based on the counterfactual outcome framework, the extended method works by estimating the causal effect via a parametric expression incorporating adherence-related covariates back-transformed to the original variables. From a range of estimated values from the procedure, the parameter of interest which best achieves equal distributions between the two arms is the point estimate. The extension was proposed to address the

issue of reduced power caused by non-adherence, and this approach attempts to recover power in the structural analysis by incorporating baseline covariates.

In terms of potential biases using this extended approach, there is a trade-off between more bias and less variability as a result of adjusting for prognostic factors. Korhonen and Palmgren [86] propose rules of thumb for situations where efficiency may be gained by adjusting for prognostic covariates. However, if there is a strong association between adherence and prognosis (i.e. worse prognosis leads to less adherence), then this version of the RPSFTM may not be appropriate. Another situation where the method might not be appropriate is related to the role of several structural parameters (e.g. the effect of nuisance estimation from the control [NEC]) in the presence of an interaction effect between baseline covariates and adherence. The latter situation is an area that may need further research. Furthermore, the analyst may face complications because not all back-transformed datasets retain the proportional HR between the study arms.

This extended version of the RPSFTM approach was assessed in a simulation study. The key evidence from the assessment proved that this extension could substantially improve efficiency compared to g-estimation with no baseline covariate adjustment. The method was further applied in a case study on leukaemia. This was done using data from the alpha-tocopherol beta-carotene (ATBC) double-blinded RCT which was conducted in Finland. The study evaluated the causal effect of bone marrow transplantation versus conventional chemotherapy. The causal effect using this approach was reported as a 7.4% reduction in survival compared to 5.9% obtained from ITT analysis.

Improving the efficiency of the estimators

Loeys and Goethgebeur[87] introduced a further extension in the same line proposed by Korhonen and Palmgren [86] as an attempt to improve efficiency by adjusting for covariates. They proposed estimating equations that combine the effect of covariates in one arm (control arm) with log-likelihood estimates of causal effect using data from a two-arm trial. The authors further assessed the proposed extension in a simulation study. They concluded that the efficacy gains would depend on the selective nature of non-adherence. Further details and the results are reported in Loeys and Goethgebeur.[87]

Allowing for dependent censoring

Matsui[88] proposed a further extension to the RPSFTM approach to allow for censoring to depend on the underlying event processes that would have been observed if patients adhered to the assigned treatment. The extended approach allows the analyst to distinguish dependent censoring from

censoring due to random loss to follow up (random non-adherence) for improving adjustment for non-random non-adherence. In this approach, the idea is to adjust for baseline covariates (that are good predictors of non-adherence) to capture the non-random underlying mechanism of non-adherence and censoring. However, the proposed approach might not be appropriate in a situation with non-null parameter values where the un-censored event time and uncensored dependent-censoring time are re-censored. At that point, the method will lose efficiency as compensation to gain unbiasedness. To gain more efficiency in this situation, an MSM with IPCW would be an alternative approach.

This extended approach was applied in a simulation study and a case study on acute myeloid leukaemia. This was undertaken using data from the acute myeloid leukaemia (AML) two-arm RCT. The study evaluated macrophages colony-stimulating factor (M-CSF) versus placebo for time to blood count recovery outcome. The estimated HR using the proposed approach was 0.65 (95% CI: 0.57-0.81) – this was similar to estimates produced by the Robins and Tsiatis standard RPSFTM. In contrast, the ITT produced an HR of 0.68 (95% CI: 0.58-0.81), which is likely to be an underestimate of the treatment effect if the causal link between non-adherence and the outcome is taken into account.

Choice of model and impact of re-censoring

White and Goetghebeur [89] extended the RPSFTMs to estimate the causal effect relating to several treatments, focusing on the choice of model and dealing with censoring and re-censoring. The method used a similar approach described in the previous papers (including covariates); and therefore, the description here is limited to the issue of model choice and re-censoring. The paper presented three models: (i) univariate model with no censoring, (ii) considered censoring at the next stage, and (iii) multivariate model (incorporating treatment-related covariates). The method chooses a re-censoring time for all individuals such that the censoring becomes non-informative. The multivariate model can be used to incorporate covariates related to non-adherence such as side effects. The paper presented a case study and provided some guidance for the choice of covariates to go into the model.

Regarding potential biases, if the effect of the randomised group is not constant, then re-censoring may produce a biased causal effect even if there is a balance between study arms. Model misspecification is another risk for producing misleading results. The univariate model (with no covariates) may not produce a valid causal effect when non-adherence is taken into account, and the results may be similar to ITT analysis. Further details about the results estimated from the case study are reported in White and Goetghebeur.[89]

2.10 Pharmacometrics-based methods

2.10.1 Pharmacokinetics and Pharmacodynamics-based Method

2.10.1.1 Origin of the method

This is a novel method which was designed to simulate a clinical trial based on pharmacokinetics-pharmacodynamics (PKPD) modelling to estimate the effectiveness of various dosing algorithms and incorporating non-adherence. The method is based on pharmacometrics analysis of the dose-exposure-response relationship, which was extended to incorporate adherence data for estimating adherence-adjusted effectiveness.

The PKPD method was originally adapted by Pink et al.[90] to estimate time within the therapeutic range (C_{min} - C_{max} range of drug concentration levels in the body) for different dosing algorithms of the assigned treatment. The method was further extended by Hill-McManus et al.[91] to adjust for the impact of varying non-adherence (implementation and persistence) on the effectiveness of treatments. The description of this method is based on Pink et al.[90] with the Hill-McManus et al.[91] extension discussed in Section 2.10.1.4.

2.10.1.2 Theoretical characteristics

The PKPD–based method could be used to model all types of non-adherence for estimating treatment effectiveness. The methods require model development and fitting using appropriate data, typically collected during each phase of clinical drug development, as well as a simulation based on different patterns of adherence, dosing schedules, and patient characteristics where covariate effects are relevant. The method works by linking the output from the PK model to the PD model for estimating the outcome in terms of a biomarker. Pink et al.[90] described the PKPD linked model in a two-stage process: (a) simulate clinical trial data of different dosing algorithms (or adherence levels) based on a PKPD model that generates an output parameter (e.g. time within anticoagulant international normalised ratio (INR) therapeutic range); (b) estimate the link between INR-range and risk of cardiovascular events using evidence from the literature. The PKPD simulation reported by Pink et al. was based on published population single-compartment pharmacokinetics (PK) model and a kinetic-pharmacodynamics (PD) model with two 3-state transit compartment chain, and the last state of each chain models the effect site.[90] Further details of the PKPD model are reported in Hamberg et al.[102]

The PKPD method works by incorporating variable dose implementation as an input function in the PKPD model. To reflect the real-world patterns of warfarin use, non-adherence was incorporated by

making two assumptions: (i) a fixed proportion of prescribed doses were missed at random (random non-adherence); and (ii) assumed a normal distribution for the timing of dosing with a standard deviation of 2 hours.

The PKPD method relies on the "exclusion restriction" assumption, that is treatment assignment (randomisation) affects the outcome only through its effects on the exposure. The PKPD approach proposed by Pink et al. also assumed that patients who discontinued treatment have switched to the control arm. This latter assumption is less relevant to my definition of non-adherence, but this needs to be dealt with in the analysis. The key advantage of this method is that it provides an alternative to RCTs for estimating effectiveness for counterfactual levels of adherence to medications using evidence from the real world and the literature. In other words, the method could be based on simulations informed by data from existing trial estimates. Another advantage is that the method can be used to adjust for implementation and persistence types of adherence. A further advantage is that the method could potentially be used to account for non-binary implementation non-adherence.

In terms of disadvantages, the method relies heavily on PKPD data which might not be routinely available in RCTs or observational studies. Another challenge is that calibrating a PKPD model is based mostly on phase I and II trial data with findings from the phase III trial and all the required data might not be available. While I have access to PK data in the context of immunosuppressive therapy after kidney transplantation (the area of my case study in Stage 3 of this research), this is unlikely to be the case for other disease areas, and this may limit the generalisability of this method to other chronic conditions. It might be difficult to apply the PKPD method in disease areas where there is no valid surrogate, or where that surrogate relationship might not hold for a treatment with a different mechanism of action. If the analyst has any reason to believe that the exclusion restriction assumption is violated, then a sensitivity analysis might help to explore potential bias. In addition, the PKPD method also requires an accurate/unbiased estimate of the link between drug concentration and patient outcomes.

Other potential biases could come from indirect comparisons needed to include all treatment options in the model using evidence (e.g. distributions) from different studies. Also, the short-term nature of PKPD data tends to create an even greater reliance on extrapolation methods to estimate long-term effectiveness than for conventional outcomes. A further limitation relates to model misspecification and the two-stage process as potential sources for added uncertainty, which is very difficult to assess using a single measure of uncertainty.

2.10.1.3 Application in case studies

Pink et al.[90] applied this method in estimating the effectiveness of pharmacogenetics-guided warfarin compared with clinically-guided warfarin and other treatments for patients with AF. The PKPD output parameter used was the anticoagulant INR. Data were simulated for 5000 hypothetical patients and INR measurements were generated to produce proportions of time (below, within and above) the therapeutic range for 3-months follow-up. Both initial doses and maintenance doses were simulated for each patient based on demographic and clinical variables. The result showed that time within range from the PKPD simulations produced a relative risk RR of 1.00047 of thromboembolic events and RR of 0.94100 of bleeds for pharmacogenetics-guided warfarin versus with clinically-guided warfarin. Further details are reported in Pink et al.[90]

2.10.1.4 PKPD Extension: Modelling varying implementation and persistence non-adherence

The PKPD method was extended by Hill-McManus et al.[91] for modelling varying non-adherence (implementation and persistence). The theoretical characteristics are similar to those described above, and the main feature of this extension is the incorporation of both implementation and persistence measures of adherence in the analysis. As such, it is compatible with the adherence taxonomy described in Section 1.2.

This version of the method used a published PKPD model which was extended to simulate the time course of serum Uric Acid (sUA) concentration for patients with varying levels of adherence to allopurinol for the treatment of gout, using NONMEM® PKPD software. The investigators modelled the plasma concentration of oxypurinol (an active metabolite of allopurinol) using a one-compartment PK model with first-order absorption and elimination kinetics. The PD model was based on a simple direct effect sigmoid E_{max} model provided in equation [24].[91]

$$E_{urate} = E_0 - E_{max} * \frac{{C_{oxy}}^{\lambda}}{{C_{50}}^{\lambda} + {C_{oxy}}^{\lambda}} \qquad [24]$$

where E_{urate} is the serum urate level, E_0 is the baseline sUA concentration, E_{max} is the maximum possible reduction in sUA concentration, C_{oxy} is the oxypurinol plasma concentration, C_{50} is the oxypurinol plasma concentration at half the maximum reduction, and λ is a shape parameter known as the Hill coefficient.

Hill-McManus et al.[91] simulated PKPD data for 500 patients using 4 options of dual urate-lowering therapy (ULT). These included Allopurinol 300mg, Allopurinol 300mg + optional lesinurad 200mg, febuxostat 80mg and febuxostat 80 mg plus lesinurad 200mg. Three patterns of medication adherence (100%, 80% and 50% levels of dose implementation) were modelled. The PKPD data was simulated for 120 days including the first 30 days used to reach steady-state concentration. The main finding from the study is that the percentage of patients achieving target sUA concentrations decreases with lower levels of adherence. At an adherence level of 80%, the percentages for allopurinol and febuxostat were 35.7% and 71.3%, respectively. At a 50% adherence level, the percentages fell to 12.7 and 25.1%, respectively. Further details are reported in Hill-McManus et al.[91]

2.11 Discussion and conclusions

A comprehensive pearl growing and iterative search technique were used across seven electronic databases to identify relevant methodological papers for adjusting the causal effect of treatment in the presence of non-adherence for time-to-event outcomes. Citation searching and reference list checking for each "pearl" identified were used to complement the databases searching. In total, 20 relevant papers covering 12 methods and 8 extensions to those methods were identified and included in the narrative synthesis reported in Sections 2.7-2.10 of this chapter.

A taxonomy is proposed broadly categorising the identified methods into four different classes: (a) traditional methods: ITT, PP and AT; (b) principal stratification methods: Cox PH model with PLE, MCC in a three-stage method and Wtd PP with EM estimator; C-PROPHET, IV; (c) g-methods: MSMs with IPCW/IPTW, SNFTMs and RPSFTMs with g-estimation; and (d) Pharmacometrics-based methods (PKPD). Each of these methods was described in terms of its origin (if originally developed to adjust for non-adherence or represents an extension to another method); theoretical characteristics (including how it works to adjust for non-adherence, key assumptions, advantages and disadvantages and potential biases); and application in a simulation study and/or a case study.

Each method makes specific assumptions and has associated limitations, and many of these assumptions are non-testable. The "no unmeasured confounding" assumption is a key assumption used by MSMs and SNFTMs. It should be noted that even if the authors of the identified papers do not talk about no unmeasured confounding if the method relies upon using covariates to control for differences between groups, the analyst would need no unmeasured confounding assumption. If this assumption seems implausible for a particular dataset, then the analyst may choose an alternative method (such as IV, C-PROPHET or PKPD) which does not rely on this assumption. However, IV and C-PROPHET rely on another key assumption, the exclusion restriction assumption which is likely to hold

in a double-blinded RCT design when randomisation is used as an instrument. The choice of the non-adherence adjustment method will partly depend on the trade-offs based on the assumptions used by the alternative methods.

Existing evidence from simulation studies showed that the g-methods (SNM, MSM, and RPSFTM) plus C-PROPHET and IV methods have generally performed well in terms of asymptotic properties of the estimators. However, violation of any of the assumptions they make or failure to incorporate relevant time-fixed and time-varying confounders may produce biased results. While it is important to think about which covariates should be included, the analyst should also think carefully about which covariates should not be included. The latter is particularly important to avoid confounding by indication bias, or "conditioning on a collider" bias when using simple methods. Many tools were proposed for identifying and evaluating relevant covariates including tests of independence and causal DAGs informed by evidence from data, or the literature might help. The PKPD-based method is also promising, but the data requirements might be an issue for application in research practice. Problems of comparability across alternative methods, due to different estimands, was also identified as a key issue. The type of estimand is an important aspect when selecting methods for estimating causal effects for use in an HTA context. This is further discussed in Chapter 3.

One of the limitations of this review relates to the screening process at the title stage. A higher number of papers were excluded at that stage because the title was not relevant. While this might be an issue, the final list of included papers was checked by two expert advisors. In addition, the review has been published in a leading academic journal and presented at an international conference, with neither prompting any communications relating to additional relevant methods.[61]

In conclusion, a range of statistical methods is available for adjusting the causal effect of treatments in the presence of non-adherence for time-to-event outcomes. Each method makes specific assumptions and has associated limitations. G-methods and PKPD methods are promising to adjust for non-adherence in estimating the real-world effectiveness of treatments. G-methods rely on adherence being binary which might not be the best approach for implementation non-adherence.

Chapter 3: Comparison of non-adherence adjustment methods and appropriateness to the context of health technology assessment

3.1 Introduction

This chapter compares the identified non-adherence adjustment methods based on existing evidence from the literature. Section 3.2 summarises the set of papers used for comparing methods. Section 3.3 discusses methods compared using empirical evidence. Section 3.4 discusses methods compared in simulation studies. Section 3.5 discusses methods compared using real data in case studies. Section 3.6 discusses comparisons based on both simulation and read-data case studies reported in the same paper. Section 3.7 assesses the appropriateness of non-adherence adjustment methods for the context of HTA. Provides the selection of methods for assessment in the Simulation study. Section 3.8 presents the selection of methods for assessment in the Simulation study. Section 3.9 provides conclusions of the comparisons.

3.2 Existing evidence comparing non-adherence adjustment methods

A set of nine papers that compared some of the identified methods in a simulation study, a case study or both, specifically in the context of non-adherence, was identified from the systematic review reported in Chapter 2. These papers were excluded from the set of 20 papers identified by the review on the basis that they report method(s) already known without any extension. However, the papers provide a better understanding of the performance of each method compared to alternative methods based on existing evidence. Table 7 shows the methods compared by each paper including the type of study, disease area, interventions compared and the outcome.

Table 7: Comparison of methods in a simulation study or case study, disease area, interventions compared and the outcome

No	Reference	Method	Type of study	Disease area	Interventions compared	Outcome
1	Robins, 1998[36]	MSMs with IPCW, SNFTMs with g-estimation	Empirical	None	None	None
2	Odondi and McNamee, 2010[103]	ITT, IV, C-PROPHET	Simulation study	Hypothetical	Active treatment vs control	Time to event
3	Lee et al., 1991[104]	ITT, AT	Case study	Epilepsy	Phenobarbital vs placebo	Time to seizure recurrence
4	Mark and Robins, 1993[105]	CPH, RPSFTMs	Case study	CHD	Special intervention involves stepped care protocol vs usual care by community physicians	Time to death (all-cause deaths, all CHD deaths)
5	Robins and Greenland, 1994[106]	RPSFTMs	Case study	AIDS	High-dose AZT vs Low-dose AZT	Time to death
6	Yamaguchi and Ohashi (2004)[107]	SNFTMs, MSMs	Case study	Lung cancer	CDDP+CPT-11 (CPT-P) vs CDDP+VDS (VDS-P)	Time to death
7	Kubo et al., 2015[38]	ITT, MSMs with IPCW, RPSFTMs	Case study	CVD	Cinacalcet versus Placebo	Time to primary composite endpoint CVD event
8	Cain and Cole, 2009[108]	ITT, MSMs with IPCW	Simulation study, Case study	AIDS	Highly Active Anti-Retroviral Therapy (HAART) vs combination ART	Time to death
9	Zhang et al., 2011[40]	ITT, MSMs with IPCW	Simulation study, Case study	Non-ST-segment elevation acute coronary syndromes	Enoxaparin versus Unfractionated heparin (UFH)	Time to all-cause death or myocardial infarction

3.3 Methods compared empirically

In 1998, Robins[36] published a comprehensive paper describing and comparing alternative g-methods for adjusting for non-adherence in the context of equivalence trials. The paper discussed SNMs and MSMs (among other methods) using empirical examples. The paper was focused on comparing the methods in terms of plausibility and robustness in correcting the causal effect for non-adherence. The paper also assessed the strength of assumptions used by each method. The paper provided an

overview of programming and computational burden for each method and suggested that this should be considered when selecting the appropriate analytical approach. The assumption for correcting for non-adherence relates to whether non-adherence is random (ignorable) or not, conditional on the history of baseline covariates and post-randomisation time-dependent prognostic factors. The choice of the appropriate model will depend, largely, on whether the conditions are met in the data. The paper was focused on random (ignorable) non-adherence; but, since this paper was published, many of the identified methods were further developed to adjust for "non-random" non-adherence, which is of primary interest within this thesis.

3.4 Methods compared in simulation studies

In 2010, Odondi et al.[103] published findings from a simulation study that assessed a wide range of methods for adjusting for non-adherence in the context of survival analysis. Six variants of methods were compared: (i) ITT Cox PH model ignoring non-adherence; (ii) Cox PH model adjusting for non-adherence using simple regression adjustment; (iii) Cox time-dependent PH model incorporating time-dependent covariates; (iv) C-PROPHET; (v) RPSFTM; (vi) and the IV method. The study assumed that once a patient becomes non-adherent, they remain so for the rest of the study follow-up; this may not be realistic in the context of time-varying chronic medications. In addition, the treatment effect was assumed to be homogenous in the simulation models (i.e. the true HR was assumed to be either 0.5 or 1).

The simulation study design involved a two-arm RCT with active treatment and control using a sample size of 1000 patients and a follow-up of 24 months. Simulated scenarios included 'random' and 'non-random' non-adherence as well as an alternative adherence measure factor using all-or-nothing non-adherence (binary variable) or partial non-adherence (time-dependent variable). The causal parameter of interest was the HR and three performance measures were used: (i) bias, (ii) 95% CI coverage (the proportion of times that 95% CI contains the true parameter value), and (iii) root mean square errors (RMSE), which is a measure of overall accuracy incorporating bias and variability measures.

The main results of the simulation study showed that the simple regression adjustment produced a small bias of -0.007 on the HR scale when adjusting for random time-dependent non-adherence.[103] However, the bias increased to -0.057 in the non-random non-adherence scenario.[103] In contrast, the results showed no important bias produced by C-PROPHET and RPSFTM under both random and non-random non-adherence scenarios. However, the good performance in unbiasedness came at a price in terms of RMSE and coverage. The paper concluded that the RPSFTM performed the best in terms

of unbiasedness and coverage; however, the method had the largest RMSE. In contrast, C-PROPHET coverage was the poorest. The key findings of this simulation study support the importance of incorporating relevant prognostic factors (confounders) when adjusting for non-adherence.

3.5 Methods compared in case studies

Lee et al.[104] compared AT with ITT using data from a two-arm double-blind RCT for estimating the causal effect of phenobarbital vs placebo on recurrence of febrile seizures in young children. The study had a sample size of 217 patients with clinic visits at 6 weeks followed by 6-monthly visits up 30 months follow up from baseline. The study considered three definitions of actual treatment received: (i) full adherence to the assigned treatment; (ii) adequate drug load estimated as average daily drug level over six-month time intervals; and (iii) adequate drug load defined as daily phenobarbital load of ≥10 mg/ml for one year prior to the two-year clinic visit.

Mark and Robins [105] and Robins and Greenland [106] both compared the RPSFTM with the standard Cox PH model. Mark and Robins[105] used data from the Multiple Risk Factor Intervention Trial (MRFIT) for assessing the effectiveness of a cigarette cessation intervention on survival while controlling for time-dependent confounders (e.g. angina). Robins and Greenland [106] used data from AIDS Clinical Group 002 RCT (introduced in Section 2.9.2.2). Both papers concluded that the RPSFTM performs well while the standard Cox model produced biased estimates when there exists non-adherence to the assigned treatment. Further details about these comparisons are published elsewhere.[105, 106]

Yamaguchi and Ohashi [107] compared SNFTM, MSM and ITT methods in a case study on non-small-cell lung cancer using data from a superiority trial. The analysis was used to estimate the counterfactual causal effect (survival differences) that would have been observed had every randomised individual received the radiotherapy treatment (second-line treatment). The HR from the ITT analysis was 0.97, while SNFTM (assuming a Weibull distribution) produced an HR of 0.50. It is clear that the ITT provides a different estimate of the causal effect; however, this is not directly comparable to the SNFTM or MSM estimates due to different estimands. Further details about this case study are reported in Yamaguchi and Ohashi.[107]

Finally, Kubo et al.[38] have recently published a paper comparing a range of methods for estimating the causal effect of cinacalcet using real data from a double-blinded placebo-controlled RCT (EVOLVE trial). The EVOLVE trial had a sample size of 3883 participants with extensive discontinuation non-adherence levels of 67% and 71% in the intervention arm and the control arm, respectively. Methods compared and their corresponding estimates of the HR for death were: ITT=0.93, IPCW=0.85, and RPSFTM=0.85.[38] Although the ITT analysis produced slighted higher estimate indicating a bias towards

the null, it was not clear if the investigators have fully explored and adjusted for relevant baseline and time-dependent covariates. Further details are reported in Kubo et al.[38]

3.6 Methods compared in both simulation and case studies

Cain and Cole[108] reported the first simulation evidence that evaluated the IPCW compared with ITT analysis in the context of time-to-event outcomes and non-adherence. The simulation study design involved a two-arm RCT comparing highly-active antiretroviral therapy (HAART) to placebo using a sample size of 1000 men and 2000 simulations draws. Scenarios assessed included 100%, 80% and 60% levels of non-random adherence to the assigned treatment. The findings from the simulation study showed good performance of the IPCW estimator (in terms of unbiasedness and RMSE) compared to ITT in scenarios with imperfect adherence. However, bias and imprecision appear to increase as the level of non-adherence increases. This indicates that in situations where the level of non-adherence is high, the method might not work properly. In other words, bias and impression appear to increase as adherence level decreases.

Cain and Cole[108] also reported the application of IPCW and ITT in a case study using data from the AIDS Clinical Trial Group 320 comparing HAART with a regimen of combination ART. The trial recruited 1156 patients aged at least 16 years with an HIV positive test and immunosuppressed with a maximum of 52 weeks follow-up. The authors reported that they adjusted for two time-varying covariates; CD4 cell count and a summary measure of HIV-related symptoms. They also reported adjustment for time-fixed covariates in the model; particularly, age, gender, race, ethnicity, CD4 count at baseline, and whether the patient was on *Zidovudine* for more than one year before randomisation. IPCW was applied where patients were censored at the first time they become non-adherent or when they were lost to follow-up. The IPCW adjusted HR estimates was 0.46 (95% CI: 0.25-0.85) compared the ITT unadjusted HR estimates of 0.75 (95% CI: 0.43-1.31).[108] The adjusted analysis shows the treatment effect had everyone in the trial adhered to the assigned treatment. The finding suggests that the IPCW adjusted estimates were 63-13% farther from the ITT estimates. These results are only valid if the IPCW assumptions hold and the method has worked. The results clearly show that the ITT approach has underestimated the actual treatment effect as the method is designed to estimate the effect of treatment assignment rather than the effect of treatment itself. The authors concluded that IPCW could be used to help in adjusting for non-adherence in treatment effect estimates.

In 2011, Zhang et al[40] published a paper that compared the same methods (IPCW and ITT) that led to similar conclusions. The methods were applied in the SYNERGY trial, an open-label multicentre RCT of 9,487 patients compared two anticoagulant drugs (enoxaparin versus unfractionated heparin) with 12

months follow-up. The primary time-to-event outcome was a composite endpoint defined as time to all-cause death or myocardial infarction within 30 days of randomisation. The IPCW adjusted HR estimate was 1.08 (95% CI: 0.92-1.22) compared to ITT estimates of 1.06 (95% CI: 0.92-1.26). In addition, the authors investigated a modified version of IPCW designed to incorporate "augmentation" (proposed by Robins 1994) for improving the precision of the estimator. The result of the modified IPCW adjusted HR was 1.07 (95% CI: 0.91-1.25).[40] The authors concluded that the efficiency gain is not worth the increased complexity. Further details are reported in Zhang et al.[40]

3.7 Appropriateness of non-adherence adjustment methods to the HTA context

The concept of the appropriateness of non-adherence adjustment methods to the HTA context was introduced in Chapter 1, Section 1.5. However, the concept needs to be developed further in order to make a more robust assessment of 'appropriateness'. In this regard, I have identified three criteria for assessing the appropriateness of the alternative adjustment methods. The criteria were: (a) the suitability of the estimand (as described below); (b) the types of non-adherence the method is capable of dealing with; and (c) whether it is possible to use the method to account for real-world non-adherence levels.

In the HTA context, resource allocation decisions are usually made for a specified population defined by the scope for each decision problem. The estimands of interest are those covering the entire study population (as specified by the RCT eligibility criteria), and this should be identifiable at baseline for resource allocation decision making. Therefore, estimands focused on latent subgroups of patients (e.g., compliers) may not be appropriate for the HTA context. Often there are issues in NICE appraisals because the RCT populations do not reflect NHS populations. This is another reason that the HTA needs to consider probably different adherence levels in the real world compared to trials. In addition, HTA probably prefers a treatment effect from a full selected trial population compared to a treatment effect from a select group within a selected trial population. Therefore, presenting the cost-effectiveness results with real-world adherence levels accounted for alongside the cost-effectiveness results based on standard ITT unadjusted analysis will provide useful information for HTA bodies (such as NICE) and Technology Appraisal committees.

The appropriateness for HTA is provided in Table 8. The table shows how each identified method is assessed against the selection criterion with more details provided in the notes column. The results based on the criteria applied for assessing appropriateness (suitability of the estimand, type of non-adherence and possibility to account for real-world non-adherence levels) for each of the identified

adjustment methods is provided in Table 8. Five methods (ITT, MSMs, SNFTMs, RPSFTMs and PKPD) generate the estimand that is appropriate for HTA (covering the entire study population), with only three of these being capable of accounting for all types of non-adherence (MSMs, SNFTMs and PKPD). Five methods are thought to be capable of re-estimating effectiveness for real-world levels of non-adherence. When looking across all three facets of estimating effectiveness for HTA, g-methods and PKPD appear to be more appropriate.

Table 8: Appropriateness of estimand for the HTA context, types of non-adherence, possibility to account for real-world adherence levels and suitability of the effectiveness estimates for HTA using the alternative adjustment methods

Method	**Appropriateness of estimand for HTA context** *	**Type of non-adherence which can be adjusted for using the method**		**Possibility to account for real-world non-adherence levels**‡	**Suitability of the method for use in HTA**	**Notes**
		Initiation, Implementation, Persistence	**Random, Explainable non-random, No-random**†			
ITT	Yes	None	None	No	No	- The estimand is marginalised to the entire population. - Cannot estimate counterfactual estimands (i.e. treatment effectiveness given adherence levels in the real world).
PP	No	Initiation, implementation, persistence	Random	No	No	- The estimand is not marginalised to the entire population. - Excluding the protocol non-compliers may break the randomisation balance leading to selection bias if protocol non-compliance is related to the underlying prognosis.
AT	No	Initiation	Random	No	No	- Does not respect the randomisation balance which may lead to selection bias. - Cannot estimate counterfactual estimands

CPH with PLE	No	Initiation	Random, Explainable non-random	No	No	- The CACE estimand used by all five methods is not marginalised to the entire population. - The compliers class is a latent group of patients which is not identifiable at baseline, making it difficult for policymakers to make resource allocation decisions based on CACE estimand. - IV can estimate effectiveness given real-world adherence level based on the counterfactual outcome framework
MCC	No	Initiation, Implementation	Random	No	No	
Wtd PP	No	Initiation	Explainable non-random	No	No	
C-PROPHET	No	Initiation	Non-random	No	No	
IV	No	Initiation, Implementation, persistence	Non-random	Yes	No	
MSMs	Yes	Initiation, implementation, persistence	Explainable non-random	Yes	Yes	- Effectiveness estimates are marginalised to the entire study population. - Can be used to account for real-world adherence levels - RPSFTM only estimates the "all treated" vs "non-treated" estimand making it applicable to adjust for "initiation" type of adherence only.
SNFTMs	Yes	Initiation, implementation, persistence	Explainable non-random	Yes	Yes	
RPSFTMs	Yes	Initiation	Non-random	Yes	Yes	
PKPD	Yes	Initiation, implementation, persistence	Explainable non-random	Yes	Yes	- The estimand is marginalised to the entire population. - Can estimate effectiveness given different adherence patterns.

* In the HTA context, the estimand of interest includes the entire study population and this should be identifiable at baseline for resource allocation decision making.

† This column specifies the type of non-adherence that each adjustment method is capable of dealing with in terms of random (non-selective) non-adherence, explainable non-random (selective) non-adherence (i.e. non-adherence explainable by observed covariates), or no-random (selective) non-adherence.

‡ In the HTA context, methods for adjusting trial data for non-adherence needs to be capable of re-estimating treatment effectiveness for any given level of adherence (e.g. real-world adherence levels).

The main differences between the four classes of methods (as categorised by the proposed taxonomy) are the estimands, assumptions, and the types of non-adherence that each method is capable of dealing with. Simple methods are only valid in the presence of random (non-selective) non-adherence. Principal stratification methods are capable of adjusting for some types of non-adherence, but their estimands seem inappropriate for the HTA context based on the criteria I set out in the first paragraph of this section. Both g-methods and PKPD can deal with real-world non-adherence and their estimands are appropriate for HTA. G-methods are similar in terms of their capability for adjusting effectiveness estimates for counterfactual non-adherence levels. In other words, can be used to re-estimate the treatment effect taking into account a predicted adherence within the RCT dataset and this is explained further in Chapter 6. However, PKPD is a unique method that uses a different approach compared to g-methods.

In practice, the analyst could apply g-methods to individual patient-level data from an RCT to re-estimate treatment effectiveness (adjusted for real-world predicted non-adherence) for populating cost-effectiveness models. The analytical approach is described further in Chapter 6. Real-world adherence levels could be estimated from registry data or observational studies. All g-methods could be applied using standard software (e.g. SAS, Stata or R).[16, 109-111] While g-methods could be applied to real RCT datasets, the PKPD approach relies on simulating an RCT dataset based on a specified pattern of non-adherence (e.g. real-world adherence), and then uses the simulated data for generating the adjusted estimates. This would require data (including PKPD data) collected at different phases of clinical drug development. PKPD method can be applied using specialist software (e.g. NONMEM) or standard software (e.g. R) for simulating the dataset.[112]

The PKPD approach also, typically, require additional modelling in order to link the Pharmacometrics measures to the clinical and patient-based outcomes necessary for HTA. It is also about the specification of the surrogate-final endpoint relationship which may be difficult. This adds an additional layer of complexity and uncertainty. It would also require a shift away from the preference in HTA of using Phase III trial data. Consequently, there is sufficient uncertainty around its appropriateness for HTA to exclude it from the simulation study and this has been identified as a key area for further research.

3.8 Selection of methods for assessment in the simulation study

A subset of the identified methods was carried forward for assessment of performance in the simulation study (Stage 3). The selection was based on the appropriateness of the non-adherence adjustment to the HTA context (see Section 3.7 for greater detail). The application of the assessment

criteria on the adjustment methods is provided in Table 9. One of the key requirements for selecting methods to be carried forward is the ability to use the methods to adjust effect estimates accounting for real world adherence levels.

Based on this selection criterion, five methods were considered for assessment in the simulation study (ITT, MSMs, SNFTMs, RPSFTMs and PKPD). RPSFTM and PKPD were further excluded from further assessment in the simulation study. The RPSFTM only works for adjusting for initiation non-adherence in placebo-control RCTs. The method cannot deal with implementation or persistence non-adherence. However, the planned simulation study was focused on comparing two active treatments informed by the design of an RCT assessing maintenance immunosuppressive therapy in kidney transplantation (tacrolimus versus cyclosporine regimens). Therefore, the RPSFTM model was excluded as it was not possible to directly compare it with the alternative adjustment methods in this particular simulation study. Attempting to develop a simulation study that links PKPD, to covariates and clinical outcomes, such that it can be directly compared to the other adjustment methods was deemed to be too complex for the timescales of a doctoral study. However, the application of the PKPD method for adjusting cost-effectiveness for different levels of non-adherence was assessed in a recent doctoral thesis by Hill-McManus et al (see Section 2.10.1.4).[91]

A final list of four methods was selected to be carried forward for assessment in the simulation study comprising ITT, PP, MSMs, and SNFTMs. Based on the selection criterion, PP analysis was considered as not appropriate for the HTA context; however, it was included in the simulation study as it was thought to provide a useful benchmark given its widespread use.

Table 9: Selection of methods for assessment in the simulation study based on the appropriateness to HTA context

Method(s)	Appropriateness for the HTA context	Selected for assessment in the simulation study	Notes
ITT	Yes	Yes	The standard ITT method was selected for inclusion as a do-nothing strategy (comparator against the alternative adjustment methods.
PP	No	Yes	Although there are some concerns about the relevance of PP analysis for HTA, I decided to include this method in the simulation study as it is commonly used as a conventional method of analysis in RCTs. The main concerns are: - The estimand is not marginalised to the entire population. - Excluding the protocol non-compliers may break the randomisation balance leading to selection bias if protocol non-compliance is related to the underlying prognosis.
AT	No	No	Does not respect the randomisation balance which may lead to selection bias.
CPH, MCC, Wtd PP, IV C-PROPHET	No	No	- The estimand is not marginalised to the entire population. - The compliers class is a latent group of patients which is not identifiable at baseline, making it difficult for policymakers to make resource allocation decisions based on CACE estimand.
MSM with IPCW, SNFTM with g-estimation	Yes	Yes	This is a subset of g-methods in the taxonomy which meets the three criteria regarding the appropriateness for the HTA context
RPSFTM with g-estimation	Yes	No	RPSFTM only estimates the "all treated" vs "non-treated" estimand making it applicable to adjust for "initiation type of adherence only. The causal effect estimated by RPSFTM is marginalised to the entire population, but the method only works for a placebo-controlled RCT design.
PKPD	Yes	No	The method requires a different simulation study design and DGMs and it is not possible to directly compare it with the alternative adjustment methods within the planned simulation study. The method has been assessed for adjusting for different levels of non-adherence in a recent doctoral thesis by Hill-McManus et al.[91]

3.9 Conclusions

In summary, this chapter compared the adjustment methods based on a review of nine studies and these identified several important issues, including different performance in situations with high levels of non-adherence, computation burden and convergence problems in simulations. I then assessed the appropriateness of the methods for HTA based on a set of three criteria. Using the evidence from Chapter 2, the comparative evidence in this chapter and the assessment of appropriateness for HTA, I identified four methods to examine in a simulation study (Chapters 4-5).

As a limitation, I recognise that the choice of criteria and the assessment of methods against them involve an element of subjectivity and others may not agree with them.

Chapter 4: Simulation study assessing the performance of non-adherence adjustment methods: Study design and methods

4.1 Introduction

Chapter 3 compared non-adherence adjustment methods based on existing empirical evidence, simulation studies and case studies. A subset of the identified methods was selected for assessment of performance in a simulation study based on their appropriateness to the HTA context.[61] This chapter presents the design and methodology of the simulation study to assess the performance of these methods across a range of realistic scenarios. Sections 4.2-4.7 describe the design of the simulation study using the ADEMP (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) structural approach for planning simulation studies.[113] Section 4.8 outlines the steps undertaken to perform the simulation study. The results of the simulation study are presented and discussed in Chapter 5.

The simulation study was based on evaluating the treatment effect of hypothetical maintenance immunosuppressive drugs on graft survival for kidney transplantation in adults. A series of patient-level RCT datasets were simulated and non-adherence was applied within these datasets based on different profiles of non-adherence identified from the literature.[55, 114] Simulation studies are performed using computer-intensive procedures for different purposes including assessing the performance of statistical methods in estimating outcomes relative to a known truth.[113, 115] Whilst not perfect, simulation studies have been widely used to test the performance of a variety of statistical methods.[113] We can "assess" statistical methods by applying them to real-world data (RWD), for example, by seeing if they converge, but we do not know if they give us a correct result or not, because in real-world data we do not know the "truth".

4.2 Simulation study design overview

The simulation study plan is based on guidelines published by Burton et al.[115] and Morris et al.[113] The pre-specified protocol for the simulation study is presented in the subsequent sections describing the design of the simulation study. This protocol was developed using the ADEMP structural approach for planning simulation studies.[113]

In the following sub-sections, I state the aim of the simulation study, then describe the data-generating process through the development of a directed acyclic graph (DAG), followed by the

identification of covariates and specification of adherence and time-to-event models. I then describe the selection of a series of scenarios over which the performance of the different methods is assessed.

4.3 Aim of the simulation study

The simulation study aimed to assess the performance of alternative non-adherence adjustment methods in estimating the treatment effect using simulated RCT datasets with a time-to-event outcome. The simulation study was designed to answer the following research question: *"What is the relative performance of the alternative methods in estimating the impact of non-adherence on treatment effectiveness?".* The simulation study provides evidence on the relative performance of methods across a range of realistic scenarios with different types and levels of patient non-adherence.

4.4 Data-generating mechanisms

The data-generating mechanisms (DGMs) refer to how the simulated data were created. These involve the specification of the pattern of non-adherence, prognostic variables, distributions, covariate correlation structures and random number generators. It was important to simulate baseline and time-dependent variables that were strongly prognostic and strongly related to adherence so that I can properly try to test the alternative adjustment methods in such circumstances (irrespective of what the variables are). However, in reality, there will be more variables in the dataset, things would be more complex, and these assumptions and simplification were necessary to test the performance of methods.

Complex DGMs were used to allow the assessment of alternative non-adherence adjustment methods across a range of pre-specified scenarios. The scenarios covered alternative representations of factors thought to influence estimated efficacy, including its associated uncertainty. The factors are; type of non-adherence, level of non-adherence, sample size, and the pattern of hazards (as represented by an underlying graft survival model, effect size, relationship between treatment effect and adherence level and the existence of any time-dependent treatment effect). The choice of these factors and the levels chosen for each are described in Section 4.4.6. A parametric simulation approach was used to compare the performance of each method against a known "truth", as specified in Section 4.7.

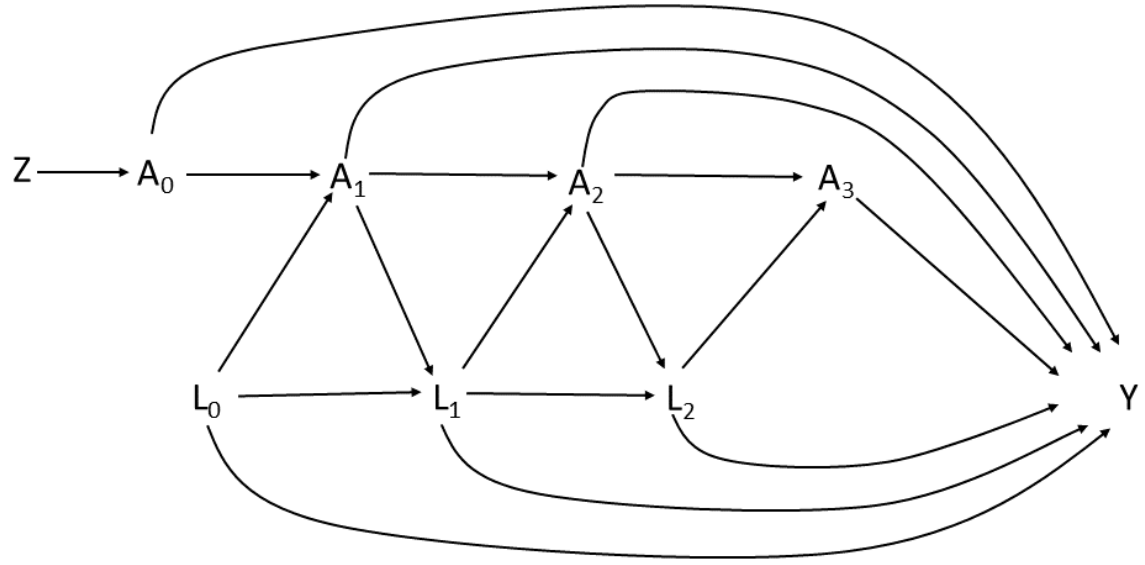
4.4.1 Directed acyclic graph (DAG)

The DAG presented in Figure 12 shows the relationships between covariates incorporated in the data-generating model, including randomisation, baseline and time-varying covariates, non-adherence and

graft survival outcome. The DAG was used to conceptualise and guide the process of simulating the datasets. The variables included in simulating the datasets are described in this section and defined below the figure. The DAG comprises nodes denoted Z, A, L and Y, representing the abovementioned variables at different time points. The DAG also include edges (arrows) representing the relationship between these variables. In this Figure, Z is the randomisation variable, A, L and Y represent non-adherence, baseline and time-dependent confounders and graft loss outcome, respectively. The non-adherence variable A also represents treatment, as essentially treatment is the inverse of non-adherence. In this DAG, time is assumed to flow from left to right and therefore the process starts with the randomisation variable (Z) followed by As and Ls and each one of them affecting the Y outcome at the end (see Figure 12).

In the DAG figure, the randomisation variable Z is assumed to affect the initiation of treatment and non-adherence only occurs after that. Ls affect subsequent As and the outcome and the Ls themselves are affected by previous As, representing time-dependent confounding. For example, time-varying confounder L_1 influences both the probability of non-adherence A_2 and graft survival outcome Y and was also influenced by previous treatment A_1. Baseline covariates and the values of time-dependent confounders at baseline, both denoted as L_0, are common causes which means they influence both graft survival outcome and non-adherence between baseline and the first follow-up time point A_1. The primary outcome Y is time to graft loss with administrative censoring at the end of the study (12 months) and this is influenced by the values of baseline and time-dependent covariates. The variable relationships are modelled such that they satisfy the conditions of time-dependent confounding (see Chapter 1, Section 1.3.4). This is important because time-dependent confounding is expected to be present in reality.

Figure 12: A DAG representing variable relationships in the data-generating model for implementation non-adherence



Z= is randomisation
L_0= is a vector of baseline covariates that includes age. L_0 also include values of time-varying covariates measured at baseline.
L_1, L_2 = updated time-varying covariate (BMI) at 4 and 8 months, respectively.
A_0, A_1, A_2, A_3= time-varying non-adherence at baseline, between baseline and 4 months, 4- 8 months and 8-12 months, respectively.

4.4.2 The pattern of non-adherence and simulation factors

The pattern of non-adherence to medications is an important aspect of the simulation study's design. The simulation study needed to be capable of reflecting real-world non-adherence; I attempted to achieve this by simulating different levels of non-adherence (e.g. high/low implementation) that differ by treatment arm as non-adherence is likely to be influenced by treatment and dosing regimen.

To consider the pattern of non-adherence for the simulation study design, I held a one-hour meeting with two clinicians (WM and JF, Consultant Nephrologists, Northern General Hospital, Sheffield Teaching Hospitals NHS Foundation Trust). The main topics discussed included: the important patterns of non-adherence to immunosuppressants in adult kidney transplant patients; patient characteristics which may predict non-adherence patterns; biomarkers and prognostic factors associated with non-adherence and graft loss; and measurement of adherence in clinical trials and the real world.

The meeting highlighted the following issues:

- Sub-optimal implementation was identified as the most important type of non-adherence followed by non-persistence (treatment discontinuation).
- Treatment initiation was considered less important as patients initiate their immunosuppression therapy under clinical supervision within the hospital.
- Age, graft rejection, comorbidity, time since transplantation, and donor-specific antibodies (DSA) which are associated with the quality of matching were identified as important prognostic factors. Patients aged 18-24 years old were known as having the worst levels of adherence to immunosuppressive medications.
- Comorbidity may also influence non-adherence due to psychological issues and practical issues (e.g. polypharmacy).
- Patient weight (or high BMI) was identified as an important time-varying factor that may affect both adherence and graft survival.

Based on the discussion with the clinicians complemented with evidence from the literature, a subset of the identified factors was chosen for incorporation as covariates in the simulated datasets. These were age as a baseline covariate and BMI as a time-dependent confounder. The main reason for including two baseline and time-dependent confounders was simplicity. Including more variables requires ever more assumptions about relationships between variables, which could distract from the focus of the study, which is investigating the performance of the methods in the presence of time-dependent confounding. Based on the DGMs applied in this simulation study, it is possible to simulate time-dependent confounding with just one time-dependent variable. The correlation between these factors and graft loss were based on evidence from the literature and assumptions. All three types of non-adherence (initiation, implementation, persistence) were considered and included in this simulation study. The coefficient of variation (CV%) for drug concentration levels was considered as reasonable for measuring implementation non-adherence as it is widely used and accepted in the area of kidney transplantation.

In the simulation study design, patients non-adhering at the previous time point were assumed to remain non-adherent for the rest of the study follow-up. This rules out MSM with IPTW from the simulation study as patients were not allowed to reinitiate the treatment after the first instance of implementation non-adherence. This assumption does not affect the other types of non-adherence (initiation and persistence).

4.4.3 Simulating baseline covariates and randomisation assignment

I simulated datasets for a two-arm, RCT with 1:1 random allocation, a time-to-event outcome and non-adherence metrics. The data were simulated to mimic an RCT evaluating the effectiveness of maintenance immunosuppressive drugs for adult kidney transplant patients (incidence cases) with 12 months of follow up.[116]

Data generation in this particular simulation study starts with specifying the number of observations (sample size) followed by creating a baseline covariate (age) which was simulated using coefficient values based on distributions from the literature and further assumptions. Then, the values of the time-dependent covariate (BMI) at baseline was generated. This is followed by randomisation to assign observations to the experimental group or control group using *"randomize"*, a user-written command in Stata for performing the randomisation procedure checking for the balance of baseline covariates between the two groups.[117] Age and BMI at baseline variables were used to test for balance between the two arms in the randomisation procedure.

4.4.4 Simulating non-adherence and time-varying covariates

Patient non-adherence metrics in the simulated datasets followed the ABC taxonomy.[1] Non-adherence was simulated for three-time intervals (baseline to 4 months, 4 - 8 months, and 8 – 12 months). The time intervals were assumed to mimic follow-up time points in a clinical trial. Time-dependent covariates were measured at baseline and two follow-up time points (4 months and 8 months). A key problem with non-adherence is that it is associated with post-randomisation events – an experience of adverse drug reactions (ADRs), difficulties of pill burden, and behavioural characteristics that change over time, and might not be readily predicted at baseline. This means it is crucial to account for time-dependent predictors of time-varying non-adherence in the analysis.

Time-varying covariates (L's) and non-adherence variables (A's, implementation, persistence, or initiation; depending on the scenario) were generated sequentially for each follow-up time point. For each time point, the non-adherence variable by treatment group was simulated first using the general equation [25]. This was dependent on the history of time-varying covariates and non-adherence at the previous time point (t-1). The "*rbinomial(n,p)*" function in Stata was used for generating binary time-varying non-adherence where n is the number of trials and p is the probability of non-adherence.

$$A_t = rbinomial(n,p)\ if\ (L_0 = x\ \&\ L_{t-1} = m\ \&\ Z = g) \qquad [25]$$

where A_t is non-adherence between the previous follow-up time point up to time t (e.g. 4-8 months), L_0 is baseline covariate (Age) with x=1 if age ≤ 24 years and 0 otherwise, L_{t-1} is the history of time-varying confounder (BMI) measured at the previous time point t-1 (e.g. 4 months) with m= 1 if BMI>30 and 0 otherwise, and Z is the randomisation variable with g=1 for the experimental group and g=0 for the control group.

As an example, for generating non-adherence between 4 and 8 months (A_2) in the experimental group, I used the following five lines of code in Stata. This simulates non-adherence assuming people with age ≤ 24 years at baseline and high BMI (>30) at 4 months have a 45% chance of non-adherence during the time interval 4 to 8 months. This risk is reduced to 30% if they have a high BMI but aged > 24 years and 15% risk if aged ≤ 24 years but normal BMI. For people with normal BMI and aged > 24, the probability of non-adherence is 7.5%.

```
gen A2=rbinomial(1,0.45) if (Age==1 & BMI1==1 & Z==1)
replace A2=rbinomial(1,0.30) if (Age==0 & BMI1==1 & Z==1)
replace A2=rbinomial(1,0.15) if (Age==1 & BMI1==0 & Z==1)
replace A2=rbinomial(1,0.075) if (Age==0 & BMI1==0 & Z==1)
replace A2=1 if A1==1
```

A_2 was set equal to 1 if A_1=1 assuming patients non-adhering at the previous time point remained non-adherent for the rest of the study follow-up.

Time-dependent confounding variables were simulated using the general equation [26], which incorporates the history of time-dependent confounders at the previous time point and non-adherence, assuming similar risk across treatment groups. The latter assumption allowed for time-dependent confounding to be influenced by previous non-adherence and the value of time-dependent confounders at the previous time point as illustrated in the DAG (Figure 12).

$$L_t = rbinomial(n,p)\ if\ (A_t = n\ \&\ L_{t-1} = m) \qquad [26]$$

where L_t is time-updated covariate at time t, A_t represents non-adherence during the interval *t-1* up to time *t* with *m*=1 means non-adhering patient and 0 otherwise, L_{t-1} is the history of time-varying confounder (BMI) at the previous time point t-1 with *m= 1* if BMI>30 and 0 otherwise.

To generate BMI at 8 months (L_2) for example, the following lines of code were used in Stata:

```
gen L2=rbinomial(1,0.90) if (A2==1 & BMI1==1)
replace L2=rbinomial(1,0.30) if (A2==0 & BMI1==1)
```

replace L2=rbinomial(1,0.60) if (A2==1 & BMI1==0)

replace L2=rbinomial(1,0.20) if (A2==0 & BMI1==0)

As can be seen, A_2 was used in the above equations to simulate L_2 as it represents non-adherence history for the time interval from month 4 up to month 8, which is likely to influence BMI at 8 months, as illustrated by the DAG (Figure 12). Whereas for time-varying covariate BMI, I used values from the previous time point BMI_1 (i.e. BMI at month 4) as they represent the history of the confounding covariate at the previous follow-up time point.

4.4.5 Graft survival time data-generating models

Crowther and Lambert (2013)[118] proposed a framework for simulating survival data under exponential, Gompertz and Weibull distributions. This framework was used as a basis for specifying the data-generating models simulating the survival data in this study. The generated datasets were checked to ensure their resemblance to realistic situations before using the simulated datasets for assessing non-adherence adjustment methods. This was achieved by using summary statistics, Kaplan-Meier survival curves and model fitting statistics.

Two graft survival-time data-generating models were chosen to simulate the RCT datasets: (i) a standard parametric survival model with Weibull distribution; and (ii) a two-component parametric survival model with a mixture Weibull-Weibull distribution. The two models were chosen to improve transferability beyond kidney transplantation where the shape of the survival curves differ by disease area and intervention. Both a standard parametric survival model and a two-component model were used in the simulation as specified by each scenario (Appendix C). For the two-component model, a visual examination of fit suggested that a Weibull-Weibull mixture model is most appropriate for mimicking the case study trial data used in Chapter 6.

To generate the truth, graft survival time outcomes were simulated using the *survsim* Stata command by incorporating all baseline and updated time-varying covariates at all time-points except without non-adherence variables (i.e. the values of the L's are based on perfect adherence). This parametric survival model was used to generate the "true" graft survival outcomes. The model required the specification of shape and scale parameters which were specified using evidence from the literature such that the generated survival times mimic graft survival times observed in trials in the context of immunosuppression after kidney transplantation. The hazard function was transformed onto the survival time scale based on a user-defined function within the survsim model.[118]

$$H(t) = h_0(t)\exp\left[\beta_0 L_0 + (\beta_1 L_1) \times (0 \leq t < 0.33) + (\beta_2 L_2) \times (0.33 \leq t < 0.66)\right] \quad [27]$$

where L_0 is a vector of baseline covariates and of time-dependent covariate at baseline, L_1 is time-dependent covariate updated at 4 months, and L_2 is time-dependent covariate updated at 8 months of the study follow-up. β_0 represents the coefficients for baseline covariates, and $\beta_1 \; and \; \beta_2$ are the coefficients for time-dependent covariates at 4 and 8 months, respectively. The values of time-dependent covariates at 12 months were not included in the model as these will not influence graft survival at 12 months (the study end date), although these would influence graft survival beyond the trial follow-up. Model [27] was used to generate the true estimates by simulating very large datasets using one million iterations for each scenario.

Then, time-varying non-adherence (A_1, A_2 and A_3) were incorporated as covariates into the survsim model to generate RCT datasets for testing the alternative non-adherence adjustment methods in the causal analysis – Model [28].

$$H(t) = h_0(t)\exp\left[\beta_0 L_0 + (\beta_1 L_1) \times (0 \leq t < 0.33) + (\beta_{1a1} A_1) \times (0 \leq t < 0.33) + (\beta_2 L_2) \times (0.33 \leq t < 0.66) + (\beta_{2a2} A_2) \times (0.33 \leq t < 0.66) + (\beta_{3a3} A_3) \times (0.66 \leq t < 1)\right] \quad [28]$$

where β_{1a1}, β_{2a2} and β_{3a3} are the coefficients for time-varying non-adherence with values influencing the relationship between non-adherence and graft survival as they interact with baseline and time-dependent covariates within the graft survival time data-generating model.

The only difference between Model [28] and Model [27] is the incorporation of non-adherence variables (A_1, A_2 and A_3) corresponding to measurements at the time up to 4, 8 and 12 months as these are measures of non-adherence from the previous time point. For example, A_3 represents non-adherence during the time interval between month 8 and month 12. The idea is to assess how each method performs in estimating the true value of the outcome and treatment as obtained from Model [27] (i.e. the outcome that would have been seen with no non-adherence) by applying the alternative non-adherence adjustment methods to the datasets affected by non-adherence generated by Model [28]. Time-dependent effects were incorporated (in some scenarios) by specifying the "tde" option within the *survsim* Stata command.

To implement Model [28] in Stata for generating graft survival times, I used the *survsim* command, incorporating baseline and time-dependent covariates, time-varying non-adherence, and the treatment effect in a delayed entry model.[118, 119] In this model, delayed entry times at which the impact of particular variable start to affect the time-to-event outcome were specified. These were specified using a common time (#) for all observations in the dataset based on follow-up time points. For example, the impact BMI at 4 months (hBMI1) affects graft loss between 4 and 8 months, and this was

specified using the command time (0.3333333:<=#t:<0.6666667) expressed in years. These were implemented to produce biologically plausible graft survival data using the following code:

```
survsim stime event, loghazard(-1.2 :+ 0.2:*#t :- 0.03:*#t:^0.5 :+ 0.05:*#t:^-0.5 :* (hBMI0:* 0.35) :* ///
(0:<=#t:<0.3333333) :+ (MNA0:* 0.40) :* (0:<=#t:<0.3333333) :+ (hBMI1:* 0.35) :*  ///
(0.3333333:<=#t:<0.6666667) :+ (MNA1:* 0.40) :* (0.3333333:<=#t:<0.6666667) :+ ///
(hBMI2:* 0.35) :* (0.6666667:<=#t:<=1) :+ (MNA2:* 0.40) :* /// (0.6666667:<=#t:<=1)) ///
cov(trt -0.75 age 0.25) tde(trt 0.15) maxt(1)                                    [29]
```

where MNA represent medication non-adherence and hBMI represents high body mass index, (i.e. BMI ≥30). The "tde" involves the time-dependency of the treatment effect. In some scenarios, the treatment effect was constant over time, whilst in others, the 0.15 parameter value was used, which results in a 15% reduction in treatment effect over time. Parameter values in Model 29 were varied to produce datasets across a range of realistic scenarios specified by the simulation study protocol (see Section 4.4.6 for parameter values). The simulated datasets were used for assessing the performance of alternative non-adherence adjustment methods.

Before applying the methods, for patients who experienced graft loss in the trial, the values of time-dependent covariates were set to missing after the event time. This was undertaken to mimic the practice in many clinical studies of ending data collection upon the occurrence of the event related to the primary outcome measure.

4.4.6 Parameter values and distributions

To generate biologically plausible RCT datasets, the parameters' values for the DGMs (described in Sections 4.1-4.4) were specified. The following factors were varied to specify the scenarios simulated: sample size, non-adherence metrics (implementation, persistence, and initiation), baseline hazard function for the survival-time data-generating model, the relationship between adherence level and graft survival, time-dependent treatment effects and treatment effect size. These were chosen based on discussions with two clinicians (WM and JF) and complemented with evidence from the literature. The seven factors were varied in a partly factorial design to specify a range of realistic scenarios in the simulation study. The rationale for specifying the sample size for the simulations and baseline hazard function is provided in the subsequent subsections. The parameter values for the levels of non-adherence, the correlation between non-adherence and outcome and the other factors in the simulations were assumed to test the methods in arrange of scenarios based on varying the values of these factors.

The levels specified for each factor are provided in Table 10 and the full listing of scenarios is given in Appendix C. A partly factorial design was used resulting in 90 across the three types of non-adherence (Initiation= 38, implementation=34, persistence=18) scenarios assessed in the simulation study. More scenarios were specified for initiation and implementation non-adherence as these are considered the most important types across chronic disease and this will aid the transferability of findings from the simulation study. The specification of scenarios is described in more detail in Section 4.4.7.

Table 10: Factors included in the simulation study scenarios

Factor level	Sample size	Type of non-adherence	Non-adherence (implementation/ persistence / initiation	Graft survival time data-generating model (DGM)	Relationship between treatment effect and adherence level	Time-dependent treatment effect	Treatment effect size
1	Large	Implementation	High	Standard parametric survival model (PSM) – Weibull distribution	Strong	Yes	Large
2	Small	Persistence	Low	Two-component parametric survival model - Weibull-Weibull (Mixture) distribution	Weak	No	Small
3	-	Initiation	-	-	-	-	-

The parameter values for each factor that were used to simulate the dataset are specified in the following subsections (4.4.6.1 to 4.4.6.6).

4.4.6.1 Sample size

To specify the number of observations in the simulated datasets, the sample size for large studies (n=450) was assumed based on the 75th percentile sample size among 40 trials (Appendix D, Table 36). These were clinical trials conducted in the area of maintenance immunosuppression after kidney transplantation, as identified by a published systematic review and included in a network meta-analysis (NMA).[120]

A small sample size (n=120) was assumed for some scenarios which were based on the 25th percentile sample size among the same list of trials (Appendix D, Table 36). Initially, I planned to use a sample size of 90 but based on testing, that resulted in a very high number of failed simulations. This was largely due to non-convergence associated with the nature of how some non-adherence adjustment methods work (i.e. g-methods). This issue is discussed later in this thesis.

4.4.6.2 Non-adherence metrics

Non-adherence was simulated using binary covariates for implementation, persistence or initiation (depending on the scenario specifications). Initiation and persistence use a binary variable that reflects the non-adherence event at a particular time point; whilst implementation reflects non-adherence to the prescribed regimen during a particular time interval (e.g. 4 to 8 months).

For implementation, non-adherence is often measured by the coefficient of variation (CV%), which is a validated measure of adherence in kidney transplant recipients that has been demonstrated to be associated with graft loss.[121, 122] A higher CV% indicates a more erratic level of medication-taking behaviour, and therefore, a higher level of non-adherence. At least three data points of drug concentration levels are needed to calculate the CV% for each individual patient for each time interval, which then needs to be combined with a cut-off point for CV% above which the patient will be categorised as non-adherent. Implementation non-adherence was simulated as a binary variable without simulating concentration levels and/or CV% and then converting it to a binary variable. Implementation non-adherence is expected to be more prevalent, but to have less impact on the graft survival outcome. Therefore, that is what I simulated in the implementation scenarios, by changing the numbers used for non-adherence probability and the effect of non-adherence on graft survival times in the DGMs.

The probabilities of non-adherence at each time interval (% of non-adherence) were simulated such that the overall non-adherence patterns are classified as high/low. These were numerically defined as relative values depending on the type of non-adherence (e.g. 10% low implementation non-adherence and 40% high implementation non-adherence). The probability of non-adherence in the control arm was different to the experimental arm. This mimics the usual pattern of non-adherence seen in clinical trials. The values for the probability of non-adherence were assumed with values varied depending on follow-up time points and the type of non-adherence evaluated in each set of scenarios (See Appendix C for an example of these parameter values).

4.4.6.3 Baseline hazard function

Two DGMs (Standard parametric survival model (PSM) with Weibull distribution and Two-component parametric survival model with Weibull-Weibull (Mixture) distribution) were used to generate graft survival times (Figures 13-14). For each model, a user-defined log hazard function with polynomial fractions and delayed entry was specified.[118] The shape of the KM survival curves generated by the standard PSM was sharply decreasing in hazards as shown from the analysis of a simulated dataset with 1000 observations. The shape of the KM survival curves generated by the mixture Weibull-Weibull model with similar patient characteristics is shown in Figure 14. Despite the similar patient characteristics, the mixture model produces survivor functions that drop a lot more slowly, so may be considered to represent less severe disease. The mixture model mimics graft survival curves observed in clinical trials conducted in kidney transplantation. The two models were used in this simulation to boost the transferability of findings beyond kidney transplantation as shape parameters for survival data varies across disease areas.

Figure 13: KM curves using standard parametric survival model with Weibull distribution

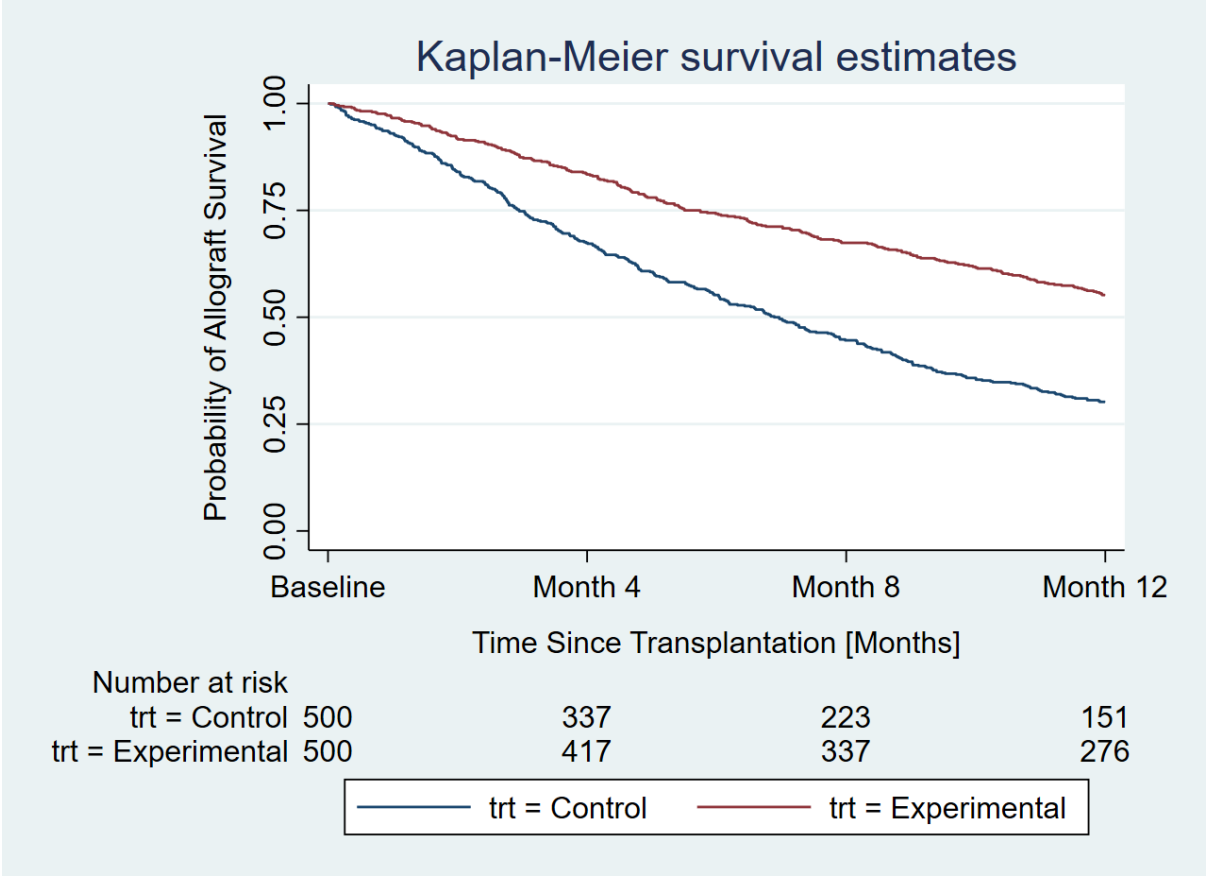
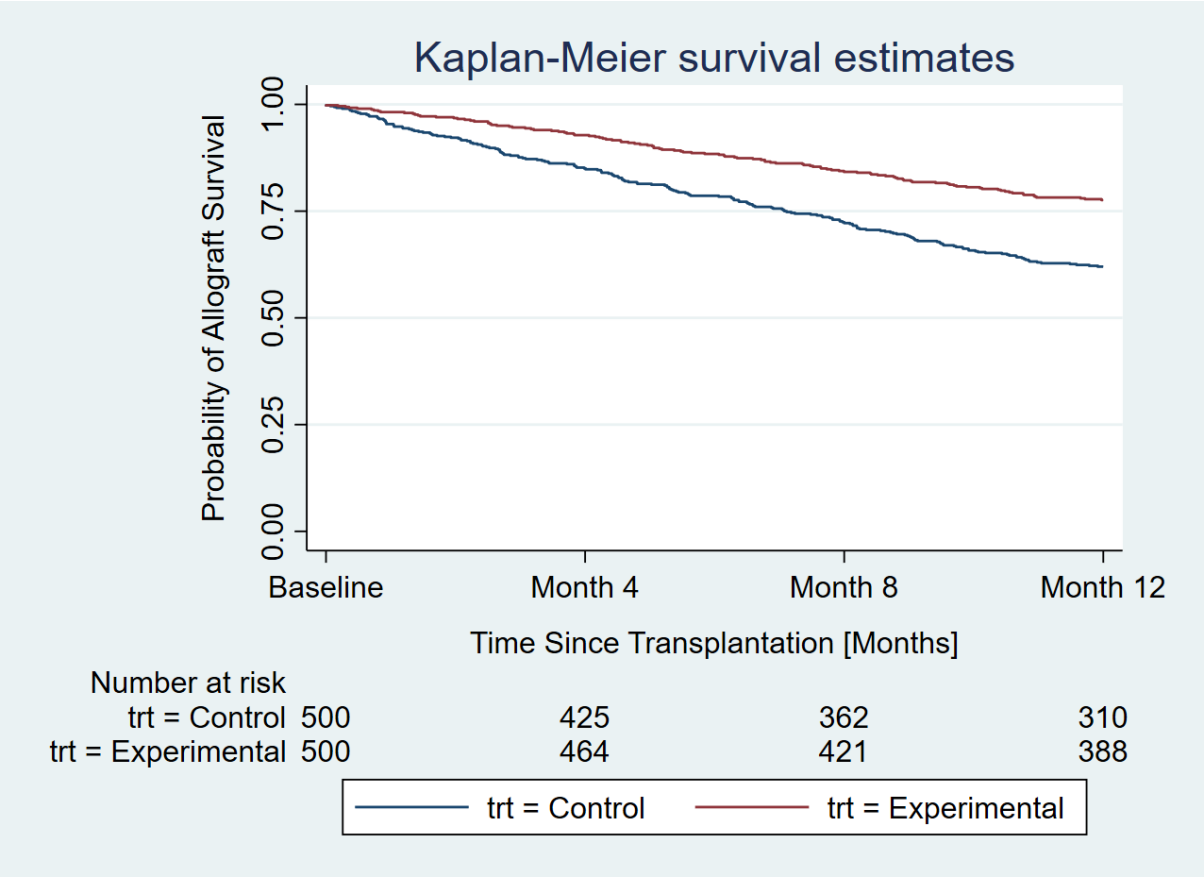


Figure 14: KM curves using two-component parametric survival model with Weibull-Weibull (Mixture) distribution



For the standard PSM, the parameter values used in the user-defined log hazard function were: *loghazard(-1.2:*0.2:*#t:^(0.2:-1)*. For the two-component mixture model, the parameters values for the hazard function were: *loghazard(-1.2 :+ 0.2:*#t :- 0.03:*#t:^0.5 :+ 0.05:*#t:^-0.5).* To implement these complex hazard functions, the Mata code was combined with the *survsim* command in Stata in a way that allowed for incorporating time-dependent covariates and non-adherence. The *"moremata"* Stata package was used to apply the models. The technical details and the full code used to simulate the datasets are provided in Appendix (G).

4.4.6.4 Relationship between non-adherence and graft survival

A parameter was used to represent the correlation coefficient as a measure of how strong the relationship between non-adherence and the time-to-event outcome (time-to-graft loss) was. The value of this parameter was specified as a coefficient within the survsim model such that the relationship is classed as "strong" or "weak" depending on the scenario. For a strong relationship, a value of 0.40 was used across implementation non-adherence scenarios. For persistence and initiation

non-adherence, higher values were used ranging between 0.42 to 0.55 to simulate a stronger impact on graft survival. For the weak relationship, a value of 0.22 was used in implementation non-adherence scenarios with alternative values ranging between 0.22 to 0.38 used in persistence and initiation non-adherence scenarios. These parameter values were assumed to achieve the desired Kaplan-Meier (KM) survival curves and hazard ratios (HRs). The strength values (strong/weak) are relative and these were determined using simple regression analysis on the simulated datasets. Different values of coefficients were used depending on the type of non-adherence to reflect different impacts on the survival outcome as discussed in Section 4.4.6.2.

4.4.6.5 Time-dependent treatment effect

The time-dependent treatment effect was incorporated in some scenarios as specified in Appendix C. In these scenarios, the parameter value was 0.15 allowing for a 15% time-dependent linear reduction in the effect of the treatment. This value was assumed to achieve the desired KM curves and HRs. Therefore, scenarios with no time-dependent treatment effect assumed a 0% time-dependent reduction in treatment effect, and constant HR.

4.4.6.6 Treatment effect size

The treatment effect size was specified based on HRs. Graft survival times were simulated such that the generated HR is around 0.55 indicating a beneficial treatment effect reducing the graft loss event rate by 45%. This large treatment effect is representative of comparing a very effective drug to a less effective drug (e.g. tacrolimus versus standard-dose cyclosporine regimens in some scenarios). In scenarios where a small/moderate treatment effect size was simulated, an HR of around 0.70 was generated representing other comparisons such as low-dose cyclosporine versus standard-dose cyclosporine.

4.4.6.7 Informative censoring

The simulation assumed no “non-administrative” censoring due to loss of follow-up and no missing data. Although these are common issues in real RCTs, there are established methods in the methodological literature to handle them. This assumption allowed the simulation to focus on addressing the issue of non-adherence without the need to simultaneously address other inter-current events. However, in practice, the analyst should consider these issues and apply the appropriate methods alongside the best performing non-adherence adjustment methods.

4.4.6.8 Application of coefficient values within scenarios

To simulate the datasets, a range of other parameters were specified in the form of coefficient values incorporated into the simulation program. These include coefficients for generating baseline and time-dependant covariates, time-varying non-adherence and graft survival times. In the simulated datasets, both baseline and time-dependent confounders (age and BMI) were included as covariates. The coefficient values used within the simulation program for generating time-varying non-adherence, baseline covariates, time-dependent confounders and graft survival times were kept constant across simulations. The rationale is to focus on varying the values of the key factors (specified in Table 10) to evaluate their influence on methods performance.

To explain how the simulation program is implemented in Stata Software, let us take Scenario 2 as an example. The parameter values and distributions specified for generating the datasets in this scenario are presented in Table 11.

Table 11: Parameter values for simulated RCT datasets - Scenario 2

Parameter	Value for Scenario 2	Distribution/function	Source of value
Sample size	450	-	Analysis based on Table 36 (Appendix D)
Age	18-24 (55%), 25-75 (45%)	Conditional random variable	Assumed
Treatment group	0= Control, 1= Experimental	Randomly assigned in 1:1 ratio	-
Non-adherence - implementation : 0-4 months	Control: 30% if (age ≤24 & hBMI0=1) 20% if (age >24 & hBMI0=1) 10% if (age ≤24 & hBMI0=0) 5% if (age >24 & hBMI0=0) Experimental: 22.5% if (age ≤24 & hBMI0=1) 15% if (age >24 & hBMI0=1) 10% if (age ≤24 & hBMI0=0) 5% if (age >24 & hBMI0=0)	Binomial random variable	-
Non-adherence - implementation: 4-8 months	Control: 60% if (age ≤24 & hBMI0=1) 40% if (age >24 & hBMI0=1) 20% if (age ≤24 & hBMI0=0) 10% if (age >24 & hBMI0=0) Experimental: 45% if (age ≤24 & hBMI0=1) 30% if (age >24 & hBMI0=1) 15% if (age ≤24 & hBMI0=0) 7.5% if (age >24 & hBMI0=0)	Binomial random variable	-
Non-adherence - implementation: 8-12 months	Control: 60% if (age ≤24 & hBMI0=1) 40% if (age >24 & hBMI0=1) 20% if (age ≤24 & hBMI0=0) 10% if (age >24 & hBMI0=0)	Binomial random variable	-

	Experimental: 45% if (age ≤24 & hBMI0=1) 30% if (age >24 & hBMI0=1) 15% if (age ≤24 & hBMI0=0) 7.5% if (age >24 & hBMI0=0)		
Probability of high BMI (≥30) at baseline	(1,0.60)	Binomial random variable	-
Probability of high BMI (≥30) at Month 4	Control & Experimental: 90% if (age ≤24 & hBMI0=1) 30% if (age >24 & hBMI0=1) 60% if (age ≤24 & hBMI0=0) 20% if (age >24 & hBMI0=0)	Binomial random variable	-
Probability of high BMI (≥30) at Month 8	Control & Experimental: 90% if (age ≤24 & hBMI0=1) 30% if (age >24 & hBMI0=1) 60% if (age ≤24 & hBMI0=0) 20% if (age >24 & hBMI0=0)	Binomial random variable	-
Baseline hazard function	loghazard(-1.2 :+ 0.2:*#t :- 0.03:*#t:^0.5 :+ 0.05:*#t:^-0.5)	User-written hazard function using a two-component parametric survival model - Weibull-Weibull (Mixture) distribution	
Coefficients for generating graft survival time	$\beta_0 L_0$ (Age) = 0.25 $\beta_0 L_0$ (hBMI0) = 0.35 ($\beta_1 L_1$(hBMI1) = 0.35 ($\beta_2 L_2$(hBMI2) = 0.35 (β_{1a1} (A1) = 0.40 (β_{2a2} (A2) = 0.40 (β_{3a3} (A3) = 0.40 Treatment effect= -0.75 Time-dependent effect= 0	Implemented within the *"survsim"* model in Stata	
Administrative censoring (End of study) in Years	1.0		

β_0, the coefficient for baseline covariates and the value of time-dependent covariates at baseline (L_0); β_1 , the coefficient for time-dependent covariates at 4 months (L_1); β_2 , the coefficient for time-dependent covariates at 8 months (L_2); β_{1a1}, β_{2a2} and β_{3a3}, the coefficients for implementation non-adherence between baseline at 4 months (A1), 4 to 8 months (A2) and 8 to 12 months (A3); hBMI0, high Body Mass Index at baseline; hBMI1, high Body Mass Index at Month 4; hBMI1, high Body Mass Index at Month 8. A1, Implementation non-adherence between baseline and 4 months; A2, Implementation non-adherence between 4 and 8 months; A3, Implementation non-adherence between 8 and 12 months.

The parameter values presented in Table 11 were incorporated into the simulation program and *survsim* model to produce graft survival times in the absence of non-adherence. The Kaplan-Meier survival curves produced from one simulated dataset with a sample size of 2000 and perfect adherence using the relevant parameter values (Table 11) are presented in Figure 15. This step was run 1 million times to produce the truth for one scenario in the full simulation program. Figure 16 shows KM survival curves with non-adherence incorporated.

Figure 15: Graft survival curves from the simulated dataset in the absence of non-adherence - Scenario 2

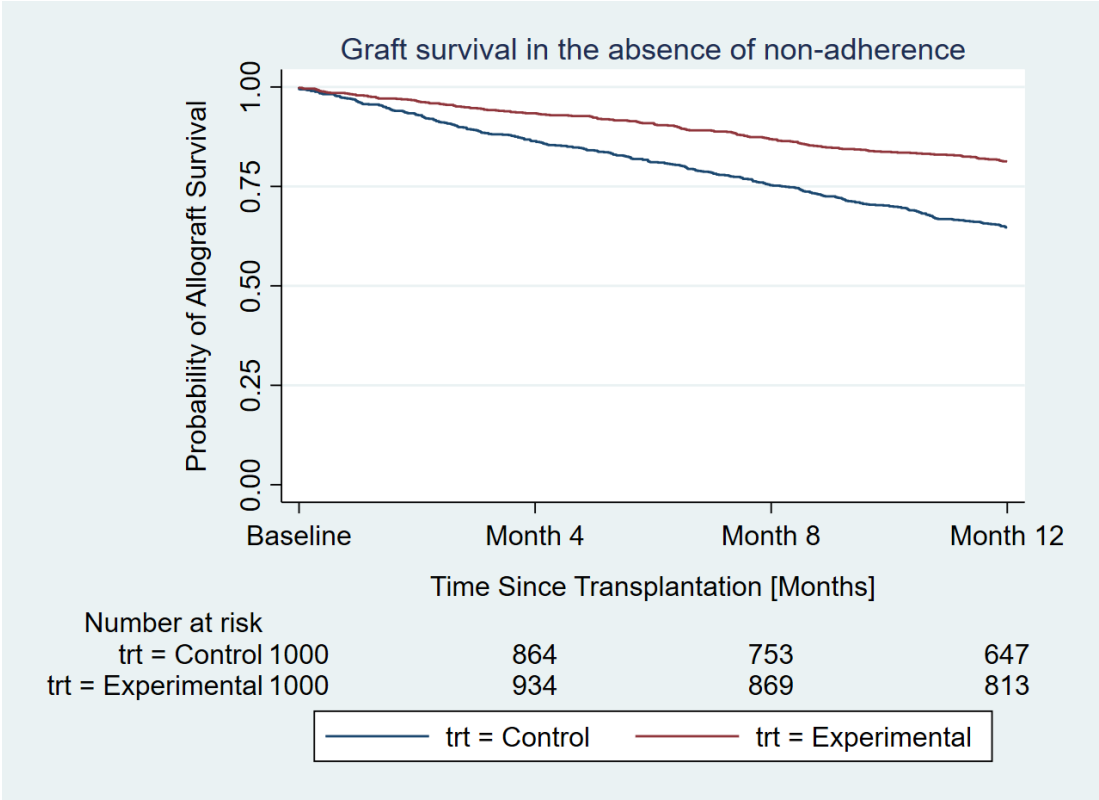
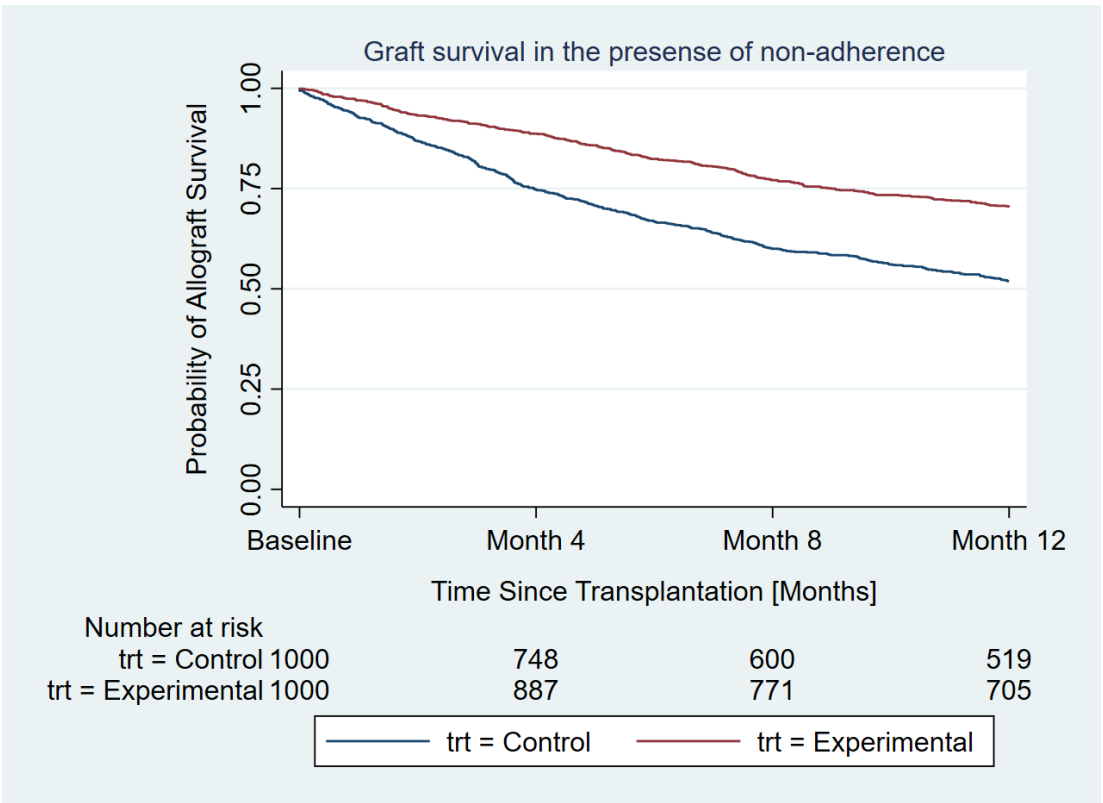


Figure 16: Graft survival curves from the simulated dataset in the presence of non-adherence - Scenario 2



To illustrate the impact of non-adherence and prognostic characteristics on graft survival over time, KM survival curves were generated based on analysis of the above-mentioned dataset. These include the impact of age (Figure 17), BMI (Figures 18-20), and non-adherence (Figures 21-23).

Figure 17: Impact of age on simulated graft survival time

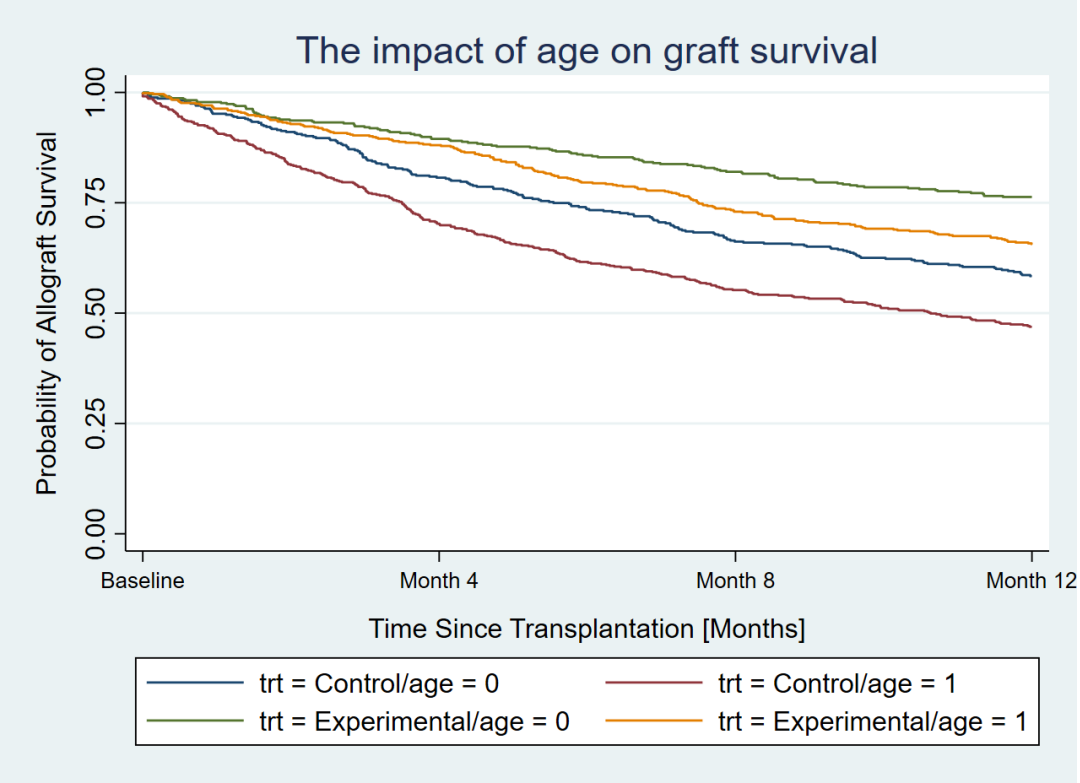


Figure 18: Impact of baseline BMI on simulated graft survival time

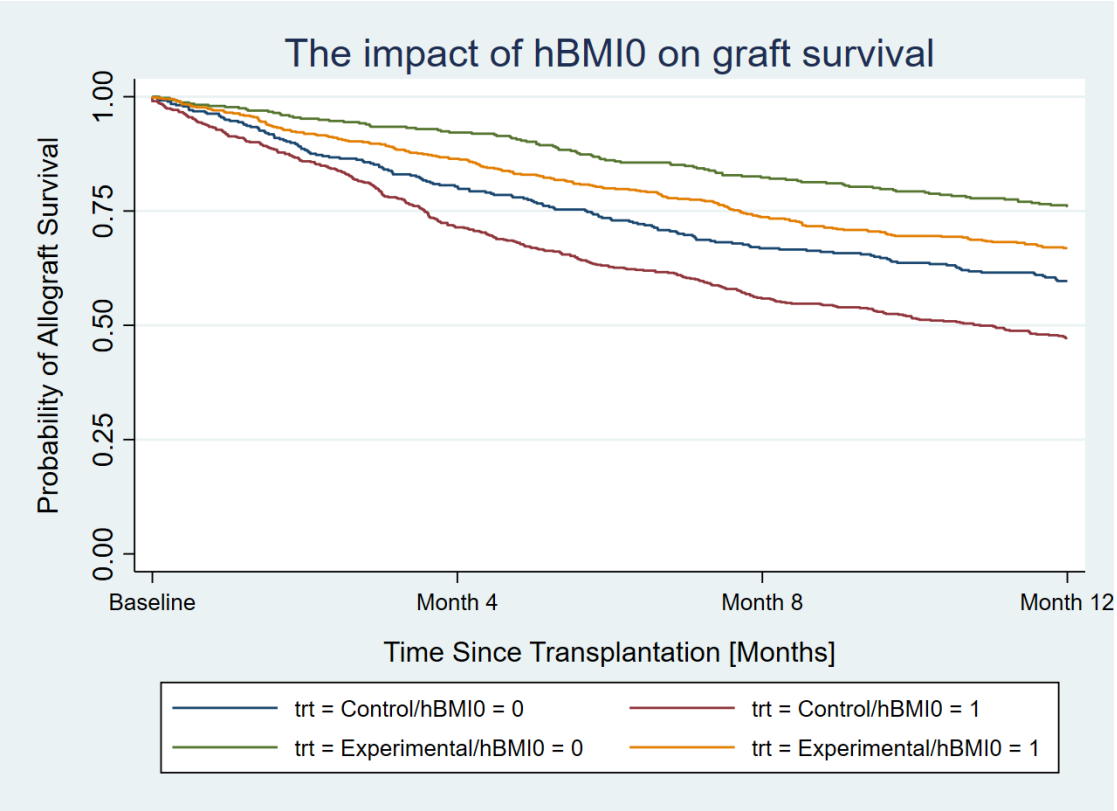


Figure 19: Impact of BMI at 4 months on simulated graft survival time

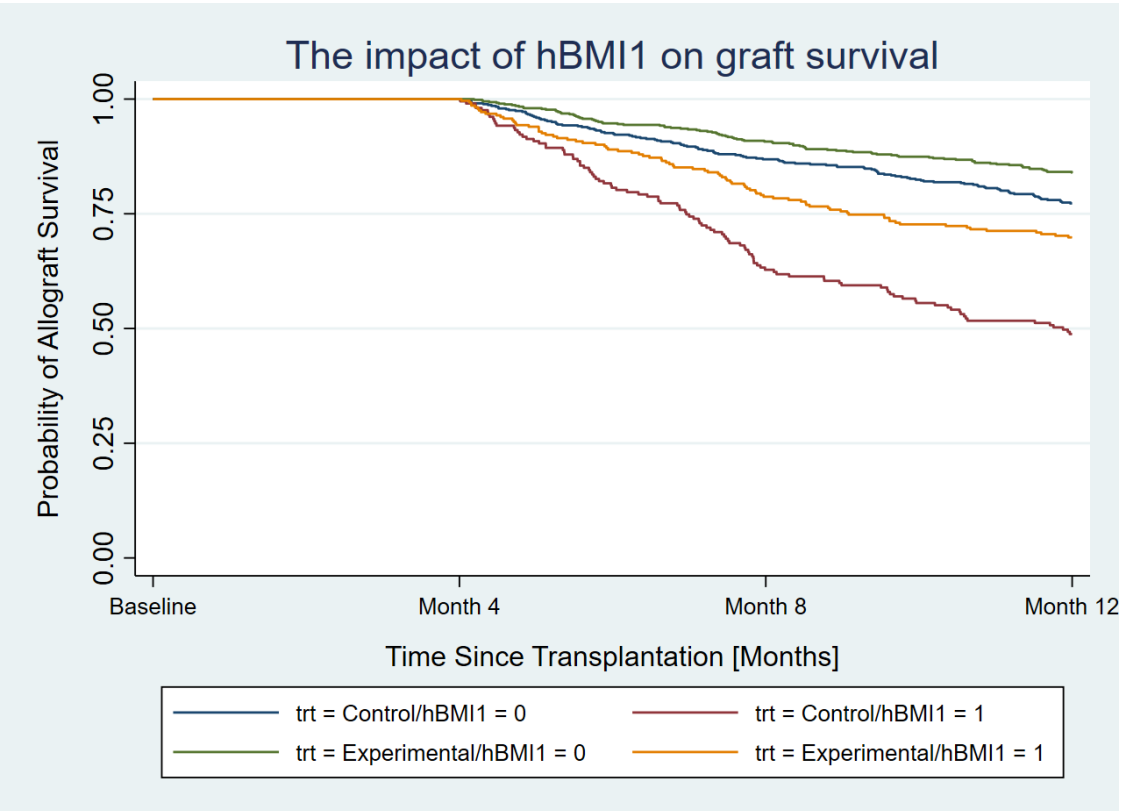


Figure 20: Impact of BMI at 8 months on simulated graft survival time

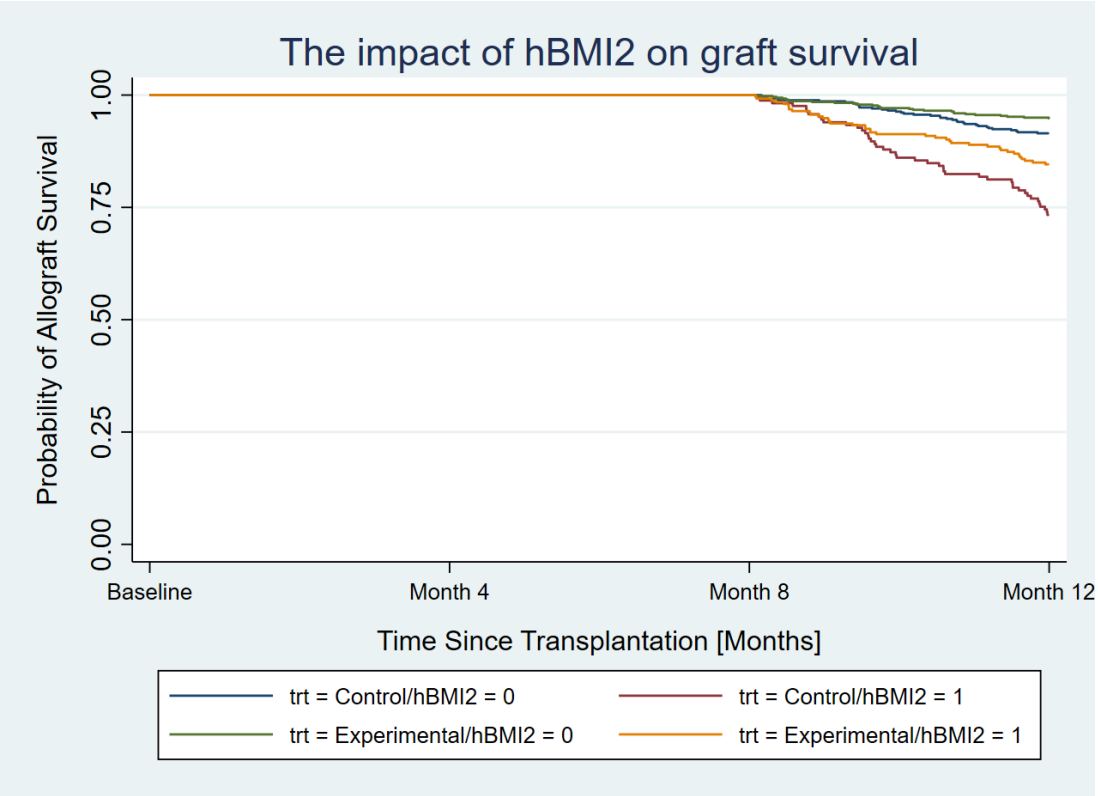


Figure 21: Impact of 0-4 month's implementation non-adherence on graft survival

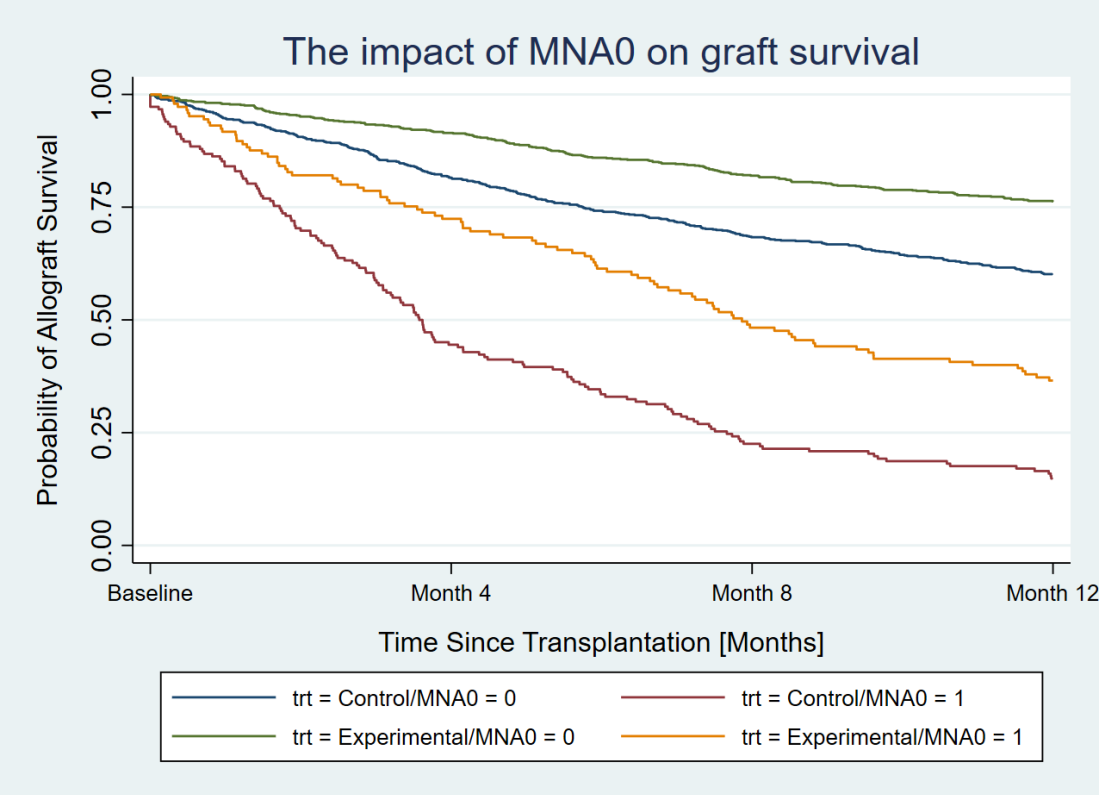


Figure 22: Impact of 4-8 month's implementation non-adherence on graft survival

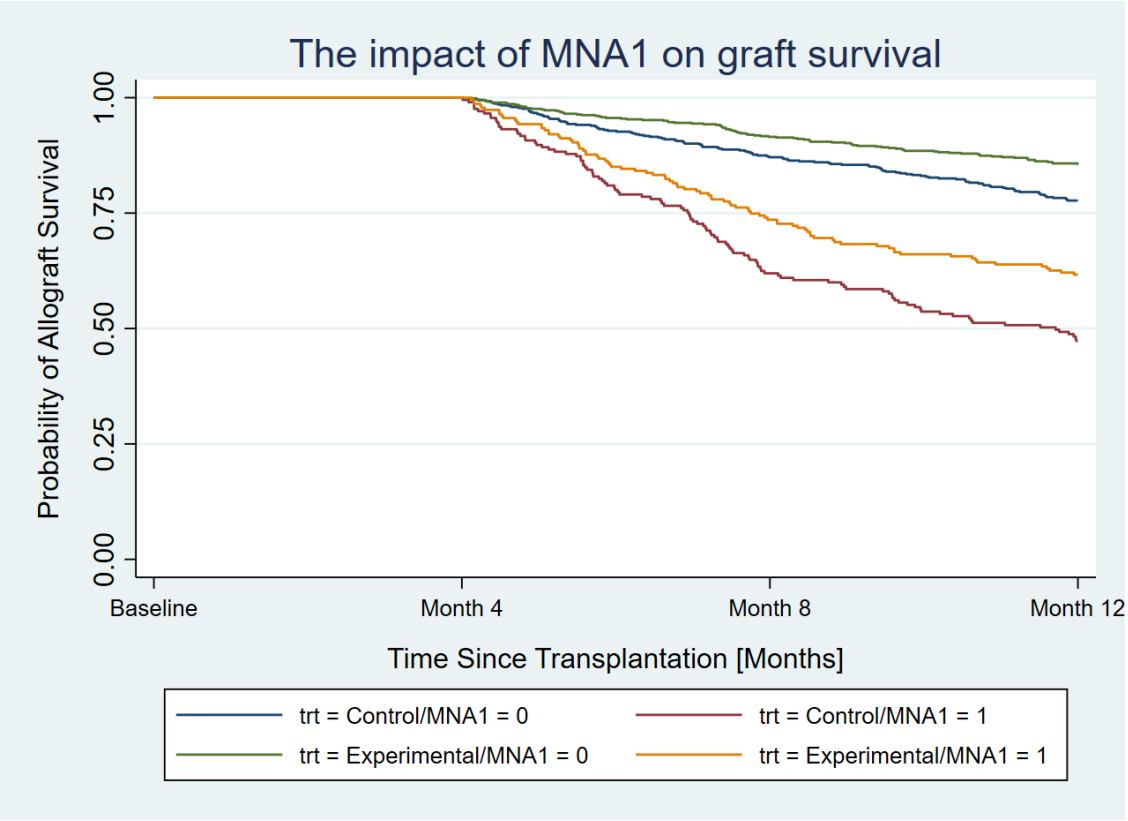
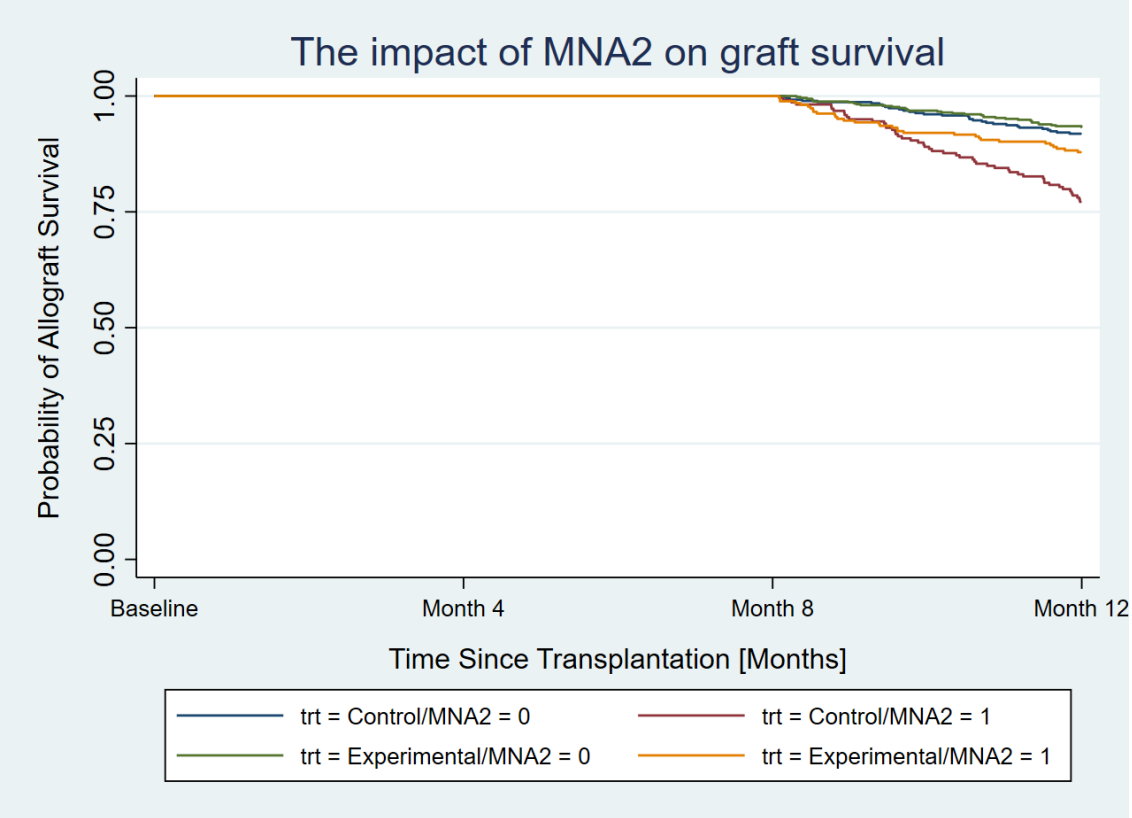


Figure 23: Impact of 8-12 month's implementation non-adherence on graft survival



4.4.7 Simulation program

Data were simulated using Stata MP 15 (64-bit) and the '*Mersenne twister'* random-number generator (mt64s) was used to set a 'stream' of random numbers for parallel runs. The input seed number for each stream of random numbers was '13183'. Parallel simulations were run on different cores of High Performing Computers (HPC) for more efficiency. This was performed using the Sheffield Advanced Research Computer (ShARC) facilities. The *"simulate"* approach in Stata was used to perform a Monte Carlo simulation.

The Shared-Memory Parallelism (SMP) approach with eight processor cores per scenario was used within ShARC to run 16 scenario sets simultaneously for 1 million iterations each. These scenarios represent RCT datasets in the absence of non-adherence to generate the truth. For scenarios with non-adherence, the SMP with four processor cores was used to run 90 scenarios for applying non-adherence adjustment methods with 1900 simulations each. These scenarios were divided into three sets based on the type of non-adherence (implementation=38, persistence=18, initiation=34). Each scenario set was run as a single job with multiple threads (each represents a single scenario) using the SMP parallel computing environment. This meant the full set of simulations were run in four separate jobs, including one job for estimating the truth.

The parallel computing approach has allowed the simulation program to run within 96 hours for truth scenarios and 6-10 hours for scenarios applying the non-adherence adjustment methods. This approach saved significant computation time and allowed the simulation program to be run in an efficient way given the large number of scenarios evaluated.

The "*survsim"* command in Stata was used for simulating biologically plausible graft survival data.[118] This was complemented by additional programming to simulate non-adherence data which was associated with covariates and the graft survival outcome.[119] Then, I used the "*simsum"* command in Stata for reporting the results of the simulation study in a more efficient way.[123] This involved using tables and graphs (e.g. nested loop plots for presenting results from the performance measures).[124]

The simulation program involves the following datasets which were generated and stored in a pre-specified directory within the University of Sheffield HPC facilities:

(a) Simulated truth datasets: these are datasets with n_{obs} generated by different DGMs with 1 million iterations, which were used to generate the truth for each set of scenarios.

(b) Simulated datasets with non-adherence: these are datasets with n_{obs} generated by different DGMs, and to which the alternative adherence-adjustment methods were applied to estimate the treatment effect (θ).
(c) Estimates: these are datasets with n_{sim} summaries (e.g. $\hat{\theta}$) from repetitions for each combination of DGM, scenario and method.
(d) Performance measures: these are datasets that contain the estimated performance measures for each combination of scenario, DGM and method.
(e) Monte Carlo Standard Errors (SEs) of estimates: These include Monte Carlo SE estimates of the estimands and performance measures.

In summary, the simulation program was run for a total of 90 scenarios (Appendix C) that were defined by a partial factorial design using sample size (2 levels), type on non-adherence (3 types), non-adherence levels (2 levels), graft survival time data-generating model (2 types), the strength of the relationship between non-adherence level and graft survival outcomes (2 levels), the existence of a time-dependent treatment effect (yes/no) (2 levels) and treatment effect size (2 levels). The simulation considered each type of non-adherence for each scenario, and therefore, the three types of non-adherence (initiation, implementation, persistence) were not combined for assessment in a single dataset.

4.5 Estimands

The primary estimand of interest is treatment effect using the difference in restricted mean survival time [RMST][125] between treatment groups, had there been no non-adherence. I also recorded experimental group RMST and control group RMST and include them as estimates for secondary estimands. In addition, the population-level HRs were estimated and reported, but these were not used for assessing methods performance. This is mainly because there is no guarantee that the proportional hazards assumption will hold in each scenario, which means the HR is a potentially misleading summary statistic. In practice, investigators and analysts are interested in different estimands depending on the purpose of each particular study. Therefore, the secondary estimands were included in this simulation to ensure that the non-adherence adjustment methods could be used to produce estimates associated with these estimands.

The difference in RMST was chosen as the primary estimand because unlike the HR it is not affected by the proportional hazards assumption. This was calculated using the *stpm2* Stata command, with the *standsurv* post estimation command. This involves fitting flexible parametric models (FPM) to the graft survival data estimated by each method under the hypothetical assumption of zero non-

adherence and calculating the RMST using these models, restricted to the 1-year end-of-follow-up timepoint. The FPMs also produced the RMSTs in the control and intervention groups alongside their standard errors. HRs were generated using a Cox proportional hazards model using the "*stcox"* command in Stata. The specification of the models is discussed in more detail in Section 4.6.

4.6 Methods assessed

Each simulated RCT dataset was analysed using each of the following methods:

(a) Intention-To-Treat (ITT) analysis
(b) Per-Protocol (PP) analysis
(c) Marginal Structural Model (MSM) with IPCW Estimator
(d) Structural Nested Failure Time Model (SNFTM) with G-estimation.

The application of four alternative methods to datasets simulated in 90 scenarios meant that 360 different causal analyses were performed (four methods x 90 scenarios). The use of the University of Sheffield HPC facilities for running parallel simulations, as described before, helped efficiently undertake the simulation study.

The non-adherence adjustment methods assessed were described in more detail in Chapter 2 including the estimands, estimators and key assumptions. However, for the sake of clarity, the application of each method to adjusting for non-adherence in the simulation study is briefly described in the following subsections.

4.6.1 ITT

The ITT method does not attempt to adjust for non-adherence. In this particular simulation, the ITT method estimates RMSTs in the presence of the estimated level of non-adherence in the simulated data, whilst the PP and g-methods (described in the subsequent sections) reflect an adjustment to non-adherence. All methods are compared against their abilities to estimate the same truth (difference in RMST in the absence of non-adherence) using the performance measures described in Section 4.7.

The ITT analysis involves fitting a flexible parametric survival model (FPM) to both treatment groups combined (i.e. with treatment group as a covariate) using the *"stpm2"* user-written command in Stata. Fitting one FPM to the treatment groups combined (rather than fitting a separate model to each treatment arm) assuming a proportional treatment effect. The FPM model was adjusted for age as a

baseline covariate using 2 degrees of freedom in the model specification. This was based on testing a range of values for the degrees of freedom and the appropriate option was selected based on model fit criteria using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The *stpm2* model specification was similar across alternative methods for a fair comparison.

The *vce(robust)* option was included in the model to produce robust standard errors. Then, the *"standsurv"* post estimation command was used to produce the difference in RMSTs between treatment groups alongside the standard error and confidence intervals around the estimates. The *standsurv* command was simultaneously used to produce estimates for the secondary estimands in terms of RMST by treatment arm alongside standard error and confidence interval around these estimates.

In addition, the ITT analysis involved applying a CPH model using the *"stcox"* command. The model was adjusted for baseline covariates to produce HRs alongside the standard error and 95% confidence intervals.

4.6.2 PP

The PP analysis strategy was applied by excluding non-adherent patients from the analysis dataset. This was applied by censoring all non-adherent patients at the first time of non-adherence. Then the *stpm2, standsurv* and *stcox* commands were applied in the same way as described in the ITT analysis (Section 4.6.1).

4.6.3 MSM with IPCW estimator

The application of MSM with IPCW estimator starts with creating a time-dependent non-adherence indicator for each time interval within the dataset (i.e. 0-4 months, 4-8 months, and 8-12 months). Then, a time-dependent outcome for graft loss was created using the same time intervals. To derive the IPCWs, four non-adherence logistic models were fitted (two models per treatment arm). The impact of BMI (as a time-dependent confounder) dependent on time is an important predictor so the interaction of confounders with time was incorporated into the IPCW weighting models.

In these non-adherence models, I used logistic regression to predict non-adherence given baseline covariate and time in the control arm (Non-adherence Model 1). Non-adherence Model 2 was then fitted on the control group with both baseline and time-dependent covariates with interaction terms

included in the model specification. Non-adherence Model 3 and Model 4 were fitted in the same way on the experimental group. Following the application of each logistic regression (non-adherence models), the "*predict"* command was used to estimate the probability of non-adherence for each patient observation included in the regression. These were then used to calculate the probability of remaining uncensored (i.e. the probability of adherence) by subtracting the probability of non-adherence from 1. This was undertaken at the individual patient-level using the estimates obtained from each of the non-adherence models.

To generate stabilised weights, I divided the probability of remaining uncensored (adherent) obtained from Model 1 by the probability of adherence generated from Model 2. This produced the IPCW stabilised weights for the control group. Similarly, I divided the probability of remaining uncensored generated from Model 3 by the one obtained from Model 4 to derive the IPCW stabilised weights for the experimental group. Then, I declared the data as survival data by using the *"stset"* command incorporating the stabilised weights. A pseudo-population dataset was created using the stabilised IPCW weights and this represents a population in which there was zero non-adherence. Then, I applied the *stpm2* parametric survival model (including the *standsurv* post-estimation command) and Cox proportional hazards model, as described above, to obtain the required estimates. The model used robust standard errors to get 95% confidence intervals that take into account the weighting of the data and the clustering of individuals.

4.6.4 SNFTM with G-estimation

The application of the g-estimation method was implemented as follows. First, I specified baseline and time-varying confounders for inclusion in the g-estimation model. This involved specifying age as a baseline covariate and BMI as a time-dependent confounder. The time-to-event indicator (time-dependent graft loss) was made time-dependent in the dataset. Then, time-lagged non-adherence variables were generated using individual patient-level adherence data for each time interval. Subsequently, I declare the dataset as survival data using *stset* command clustering observations by patient ID.

Second, I estimated the Acceleration Factor (AF) as the effect of time-dependent non-adherence on graft survival time outcome. This was done for each treatment arm separately using the *"stgest3"* command, which is a Stata program for implementing g-estimation in an SNFTM.[101] The *stgest3* model specification included baseline and time-dependent confounders using the *model(all)* option which includes all observations and the *outcome(mgale)* option for handling how the potential outcome is

entered into the g-estimation model. The latter is a new technique using the martingale potential outcome that brings together the time to event (time-to-graft loss) and censoring indicators together to improve performance. The *stgest3* model incorporated the interaction of confounders with time in the same way applied to the IPCW weighting models.

Third, I used the estimated AF to adjust graft survival times by treatment arm allowing for recensoring. This results in estimated graft survival times that would have been observed if there had been non non-adherence.

Finally, the *stset* command was applied again and the *stpm2, standsurv and stcox* models were used for generating the required estimates, as explained in previous subsections. The CIs obtained do not factor in the fact the data have been adjusted, and therefore, are likely to underestimate uncertainty, with implications for coverage.

4.7 Performance measures

The performance assessment was focused on the following key properties of the estimators: bias, accuracy, coverage, and both empirical and model-based standard errors. In addition, to quantify simulation uncertainty over n_{sim}, I have estimated the Monte Carlo standard errors (MCSE) of the estimated performance measures. The formulae for computing the performance measures are provided below.

- Bias is a measure of accuracy and was evaluated using absolute and percentage bias. The absolute bias was calculated using the following formula.

$$Bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta \qquad [30]$$

- Mean Squared Error (MSE) is a measure of overall accuracy because it includes both bias and variability measures. This is presented as a percentage of the true value. Formally, it is the sum of the squared bias and variance of $\hat{\theta}$ which can be computed using the following formula.

$$MSE = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)^2 \qquad [31]$$

- Coverage is defined as the probability that a CI contains θ (i.e. the proportion of times that 95% CI contains the true value of the estimated parameter). For a two-sided interval, coverage was computed using the following formula.

$$Covergae = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{low,i} \leq \theta \leq \hat{\theta}_{upp,i}) \qquad [32]$$

- Empirical standard error (EmpSE) for $\hat{\theta}$ is an estimate of the long-run standard deviation of $\hat{\theta}$ over the n_{sim}. The EmpSE was computed using the following formula.

$$EmpSE = \sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}(\hat{\theta}-\bar{\theta})^2} \qquad [33]$$

- Average model-based standard error (ModSE) for $\hat{\theta}$ is the average of the estimated SEs which targets the estimated empirical standard error. The ModSE was computed using the following formula.

$$Average\ ModSE = \sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}\widehat{Var}\left(\hat{\theta}_i\right)} \qquad [34]$$

Both EmpSE and ModSE were expressed as a percentage of the true value of the treatment effect (i.e. Difference in Restricted Mean Survival Times).

The number of simulations or repetitions (n_{sim}) used to simulate datasets, reported previously in Section 4.4.7, was 1900. The central issue when considering the optimal n_{sim} is the MCSE where the main performance measures should be estimated to a satisfactory degree of precision. As one of the key performance measures of interest is coverage, the following formula was used to calculate the optimal number of simulations.

$$n_{sim} = \frac{E(Coverage)x\ ((1-E(Covergae)}{(Monte\ Carlo\ SE_{req})^2} \qquad [35]$$

I assumed 0.5 % MCSE to be satisfactory for 95% coverage, which produces an n_{sim} value of 1900. This was used as the number of simulations across all scenarios. The minimum and maximum MCSE for each scenario across all methods is presented alongside the detailed results in Chapter 5.

The analysis of the estimates datasets starts by generating the number of successful iterations out of 1900 simulations. This is followed by calculating successful estimations relating to model convergence for each method. For some methods, multiple models need to converge in order to achieve successful estimation for one iteration. For IPCW, convergence for each of the four non-adherence models was captured (see Section 4.6.3). In addition, two further models (*stpm2* and *stcox*) need to converge for successful estimation of the difference in RMST and HRs to be achieved. For SNFTM with g-estimation, successful estimation involves the convergence of the "*stgest3"* model in each treatment arm, followed by the convergence of the *stpm2* and *stcox* models. All these convergence events were

captured and reported in the estimates datasets and summarised as successful simulation and successful estimation in the results.

Then, the estimates datasets were analysed using the *simsum* Stata command to generate the performance results for comparing non-adherence adjustment methods across scenarios. The results from all scenarios were compiled and saved in separate datasets (depending on the type of non-adherence) for data visualisation and reporting. These are presented using nested loop plots for comparing methods performance using the seven factors specified in the study design as descriptors. The ranking of methods based on best performance was assessed and presented using tables. The latter covered the five performance measures specified above. The analysis also involved generating the mean estimates (by method) for each estimand across the successful estimations and these are presented using tables in this chapter.

4.8 Steps of the simulation study

In summary, the simulation study involves the following steps:

a. RCT datasets were simulated using 1 million iterations with prognostic baseline and time-dependent confounding and graft survival outcomes. The average treatment effect generated from the analysis of these datasets represents the "true" treatment effect. The true treatment effect was estimated using the difference in restricted means survival time [RMST][125] as a primary estimand. The true values were also obtained for secondary estimands (HR, RMST in the control and experimental groups).

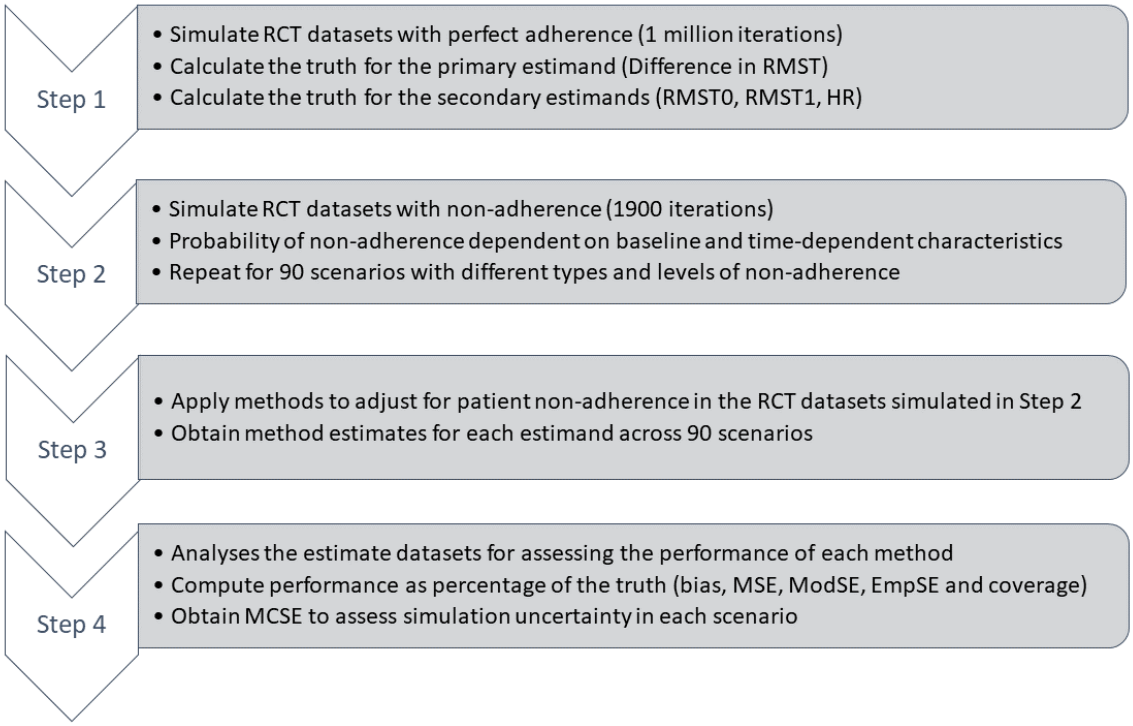
b. RCT datasets were simulated using 1900 iterations and a similar DGM specified in (a) with the only difference being the presence of non-adherence (implementation, persistence or initiation variables). In these datasets, the probability of non-adherence was associated with the prognostic baseline (age) and time-dependent (BMI) characteristics.

c. The alternative four methods were applied to adjust for patient non-adherence in the dataset simulated in (b) to estimate the adherence-adjusted treatment effect. The key estimates generated from this step include the difference in RMSTs, standard errors and confidence intervals for the difference in RMSTs, and indicators for model convergence (including four non-adherence models used by IPCW and two g-estimation models used by the SNFTM). The estimates datasets also included similar estimates for the secondary estimands (HR, RMST in the experimental and control groups).

d. Each estimates dataset was analysed using the *"simsum"* Stata package for assessing the performance of each method. The performance of each method was assessed based on five measures: (i) bias, by comparing the predicted with the true mean difference in RMST; (ii) mean square error to assess variability; (iii) coverage based on the number of simulations where the 95% confidence interval includes the true mean outcome; (iv); EmpSE to assess the precision of the estimator; and (v) average ModSE. In addition, the Monte Carlo standard error was produced to assess simulation uncertainty in each scenario.

A flow diagram showing the main steps and operations of the simulation study is presented in Figure 24.

Figure 24: Simulation steps flow diagram



Chapter 5: Simulation study assessing the performance of non-adherence adjustment methods: Results, discussion and conclusions

5.1 Introduction

Chapter 4 presented the design and methodology of the simulation study for assessing the performance of the alternative methods across all types of non-adherence scenarios. This chapter presents the results of the simulation study, discusses the findings and provides the conclusions. Section 5.2 presents the results of the simulation study and provides new evidence for using the best-performing methods to estimate treatment effects adjusted for patient non-adherence. Section 5.3 discusses the results of the simulation study and presents the conclusions.

5.2 Results of the simulation study

5.2.1 Overview of the results

The reporting adheres to published international guidelines for reporting simulation studies and medication adherence research.[113, 126] Firstly, the true values of the estimates are reported (Section 5.2.2), which are then followed by the performance results based on the primary estimand (difference in RMST expressed in years) across the 90 scenarios by the three types of non-adherence (Sections 5.2.3-5.2.5). Section 5.2.6 presents the results for the secondary estimands.

To summarise and interpret the findings of the simulation study, I use tables and graphs to illustrate the pattern across each group of scenarios using the specification factors as descriptors. These factors include sample size, graft survival time data-generation model, level of non-adherence, the relationship between non-adherence and graft survival outcome, time-dependent treatment effect and treatment effect size.

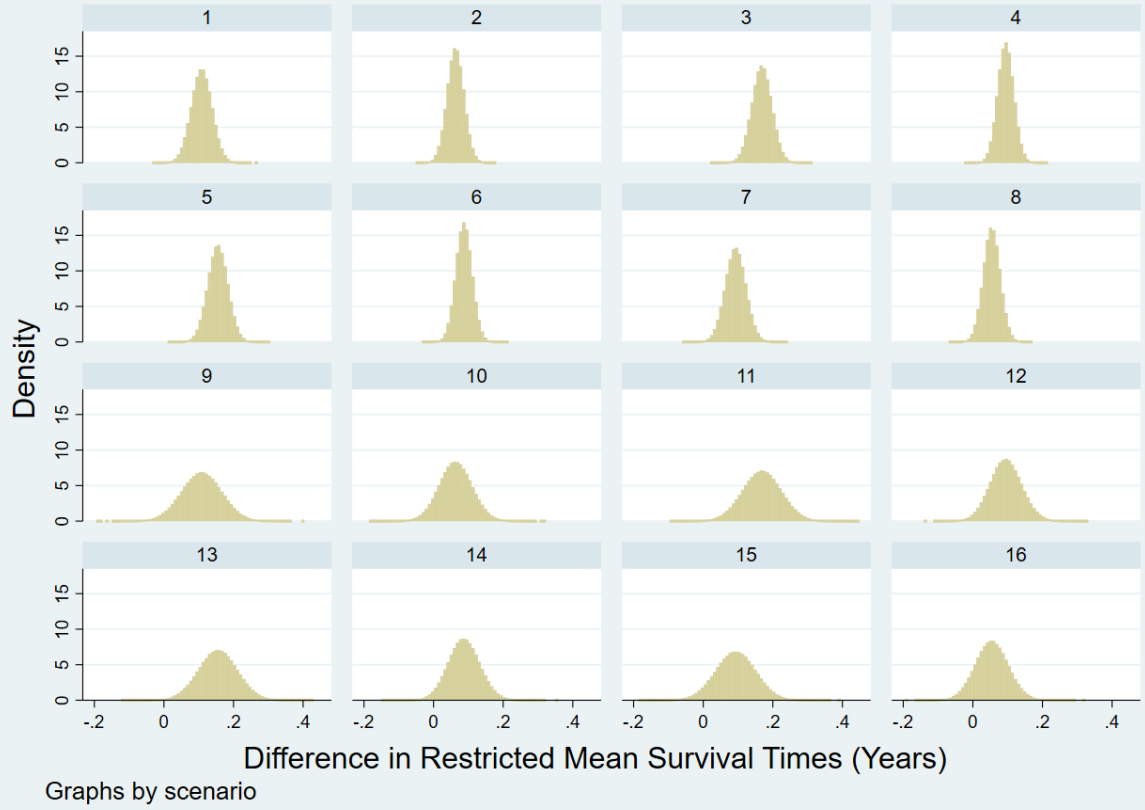
I use nested loop plots to present the performance of all methods and to illustrate how the parameter values influenced the estimates across all scenarios.[124, 127] I use tables to summarise the number of times each method performed the best across each set of scenarios (defined by the type of non-adherence) for each performance measure. These included bias, percent bias, MSE, ModSE, EmpSE and coverage as specified in the simulation study protocol (Chapter 4, Section 4.7).

5.2.2 The truth

The truth represents what would be observed with zero non-adherence, and that is what each of the adjustment analyses is designed to estimate. The true values of the estimates were generated from the analysis of large datasets with 1 million iterations for each scenario. These values included the truth for the primary estimand (the difference in RMSTs) and the secondary estimands (HR and RMST in the control and experimental groups) across 16 scenarios. Each scenario generates the truth for a set of scenarios for the application of adjustment methods as these scenarios varied by the type and level of non-adherence while the truth remains similar among each set. For example, Scenarios 1, 3 and 5 have a similar sample size, DGM, treatment effect size and they only differ in the level of implementation non-adherence and the strength of the relationship between treatment effect and non-adherence. Therefore, the three scenarios have a similar value of truth (See Appendix C details across all scenarios).

Figure 25 presents the distributions within the datasets used to generate the truth for the difference in RMST. In this figure, scenarios 1-8 illustrate the distribution of the true survival times obtained from datasets with a large sample size (n=450). Scenarios 9-12 represent small size datasets (n=120).

Figure 25: Distribution of the true difference in RMSTs (the primary estimand) across all scenarios

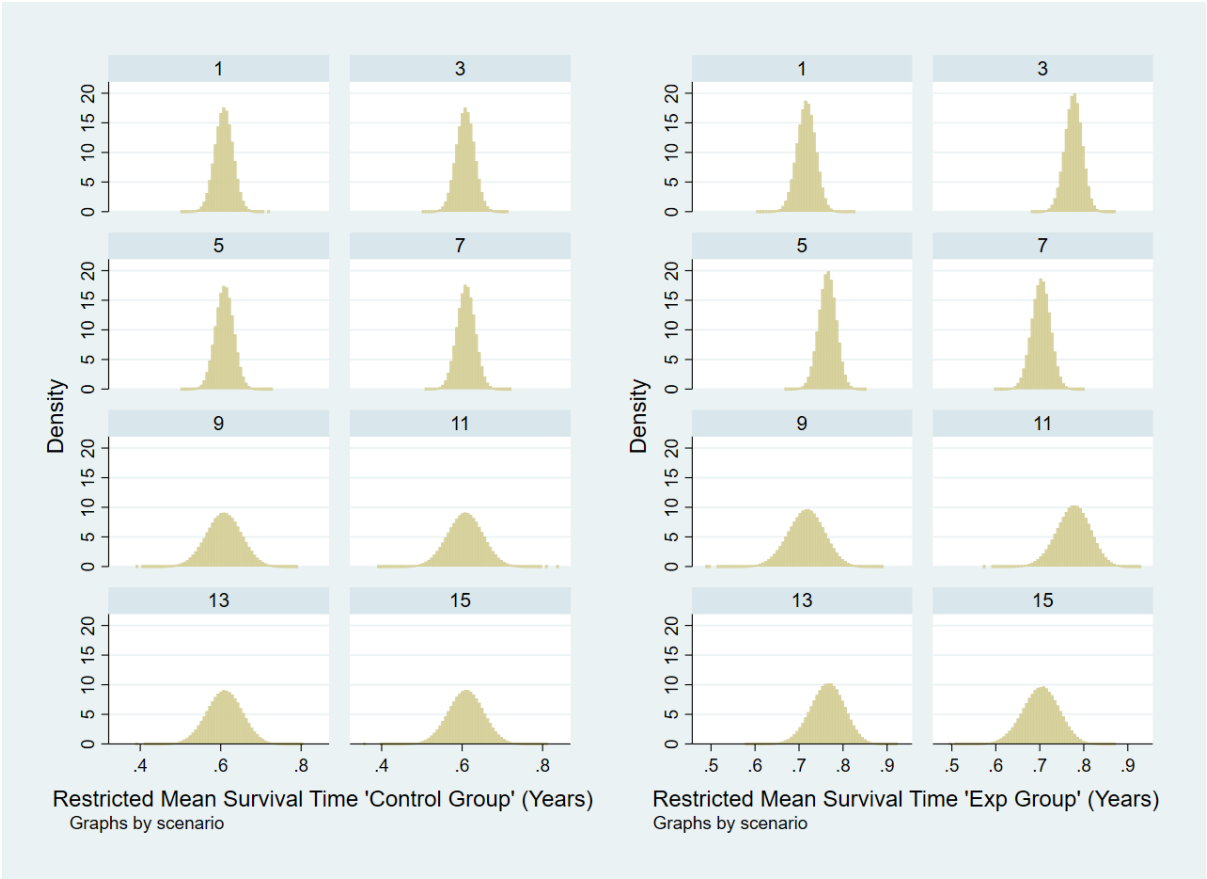


Key:			
1= n=450, standard PSM – Weibull DGM, no tde, large TE size	*2= n=450, two-component Weibull Mixture DGM, no tde, large TE size*	*3= n=450, two-component Weibull Mixture, no tde, small TE size*	*4= n=450, two-component Weibull Mixture, no tde, small TE size*
5= n=450, Standard PSM – Weibull, tde=0.15, Small TE size	*6= n= 450, Two-component Weibull Mixture, tde=0.15, Small TE size*	*7= n=450, standard PSM – Weibull, tde=0.15, large TE size*	*8= n=450, two-component Weibull, Mixture, tde=0.15, large TE size*
9= n=120, standard PSM – Weibull, no tde, large TE size	*10= n=120, two-component Weibull Mixture, no tde, large TE size*	*11= n=120, standard PSM – Weibull, no tde, Small TE size*	*12= n=120, two-component Weibull Mixture, no tde, small TE size*
13= n=120, standard PSM – Weibull, tde=0.15, small TE size	*14= n=120, two-component Weibull Mixture, tde=0.15, small TE size*	*15= n=120, standard PSM - Weibull, tde=0.15, large TE size*	*16= n=120, two-component Weibull Mixture, tde=0.15, large TE size*

TE= treatment effect, tde= time-dependent effect

Figure 26 and 27 show the distribution of the data which were used to produce the true RMSTs by treatment group.

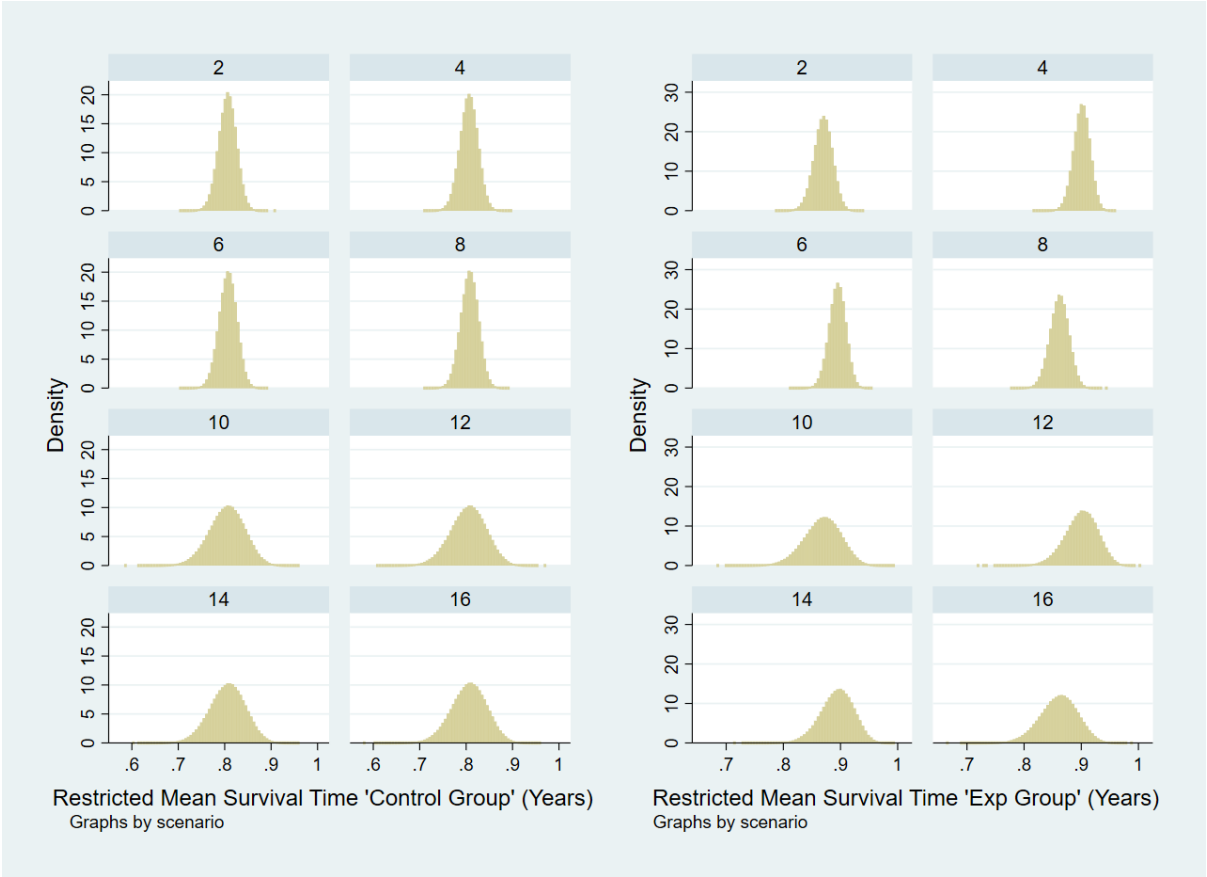
Figure 26: Distribution of the true RMST by treatment arm across scenarios - Standard parametric survival model with Weibull distribution



Key:	
1= *n=450, standard PSM – Weibull DGM, no tde, large TE size*	*3=* *n=450, standard PSM – Weibull, no tde, small TE size*
5= *n=450, Standard PSM – Weibull, tde=0.15, Small TE size*	*7=* *n=450, standard PSM – Weibull, tde=0.15, large TE size*
9= *n=120, standard PSM – Weibull, no tde, large TE size*	*11=* *n=120, standard PSM – Weibull, no tde, Small TE size*
13= *n=120, standard PSM – Weibull, tde=0.15, small TE size*	*15=* *n=120, standard PSM - Weibull, tde=0.15, large TE size*

TE= treatment effect, tde= time-dependent effect

Figure 27: Distribution of the true RMST by treatment arm across scenarios - Two-component Weibull Mixture model



Key	
2= n=450, two-component Weibull Mixture DGM, no tde, large TE size	*4= n=450, two-component Weibull Mixture, no tde, small TE size*
6= n= 450, Two-component Weibull Mixture, tde=0.15, Small TE size	*8= n=450, two-component Weibull, Mixture, tde=0.15, large TE size*
10= n=120, two-component Weibull Mixture, no tde, large TE size	*12= n=120, two-component Weibull Mixture, no tde, small TE size*
14= n=120, two-component Weibull Mixture, tde=0.15, small TE size	*16= n=120, two-component Weibull Mixture, tde=0.15, large TE size*

TE= treatment effect, tde= time-dependent effect

The true parameters values for each scenario alongside their standard errors are presented in Table 12. For the difference in RMST, the true treatment effect ranged between 0.05 and 0.21 years across 16 scenario sets. For the HRs, the true treatment effect ranged between 0.45 and 0.74.

Table 12: True values of parameters for primary and secondary estimands assuming perfect adherence

Scenario Set	Difference in RMST	SE of the difference in RMST	RMST 0	SE of RMST0	RMST1	SE of RMST1	HR	SE of HR
1	0.11	0.030	0.61	0.023	0.72	0.021	0.65	0.080
2	0.06	0.024	0.81	0.020	0.87	0.017	0.65	0.113
3	0.17	0.029	0.61	0.023	0.78	0.020	0.48	0.063
4	0.09	0.023	0.81	0.020	0.90	0.015	0.48	0.091
5	0.16	0.029	0.61	0.022	0.76	0.020	0.52	0.067
6	0.09	0.024	0.81	0.020	0.89	0.015	0.52	0.096
7	0.09	0.030	0.61	0.022	0.70	0.021	0.69	0.084
8	0.05	0.025	0.81	0.019	0.86	0.017	0.70	0.119
9	0.11	0.057	0.61	0.043	0.72	0.041	0.66	0.158
10	0.06	0.047	0.80	0.038	0.87	0.032	0.67	0.229
11	0.17	0.056	0.61	0.043	0.78	0.038	0.49	0.125
12	0.10	0.045	0.80	0.038	0.90	0.028	0.50	0.184
13	0.16	0.056	0.61	0.043	0.76	0.039	0.53	0.132
14	0.09	0.046	0.81	0.038	0.89	0.029	0.54	0.194
15	0.09	0.057	0.61	0.043	0.70	0.041	0.71	0.166
16	0.05	0.048	0.81	0.038	0.86	0.033	0.72	0.241

RMST0: restricted mean survival time in the control group; RMST1: restricted mean survival time in the experimental group; HR: hazard ratio; SE: standard error.

Key: similar to the key provided under Figure 25.

5.2.3 Performance of methods across implementation non-adherence scenarios

This section presents the results of methods performance across the 38 scenarios adjusting for implementation non-adherence in the control and experimental arms of the simulated RCT datasets (Scenarios 1-38). The detailed results across all implementation scenarios including the minimum and maximum MCSE are provided in Appendix E, Table 37.

Table 13 summarises the performance results for each method in terms of five performance measures specified in the study protocol (i.e., bias, MSE, ModSE, EmpSE and coverage). The performance results were computed as percentages of the truth for the comparison.

The results show that SNFTM with g-estimation was the best performing method in terms of producing the least bias (expressed as percent bias) in 21 out of 38 scenarios. MSM with IPCW performed best in 9 scenarios followed by PP which performed best in the remaining 8 scenarios. ITT performed the worst in all scenarios in terms of bias. In terms of MSE, the performance results favoured PP and ITT methods.

SNFTM performed best in all 38 scenarios in terms of ModSE. The EmpSE favoured ITT in all scenarios, whereas MSE favoured PP. However, this should be interpreted with caution because ITT performance lagged behind IPCW and SNFTM in terms of bias percent and ModSE. Coverage percent represents the proportion of times that the 95% confidence interval contains the true value of the estimated parameter (i.e. the difference in RMST). Although a robust standard error around the difference in RMST was estimated using the *vce(robust)* option within the FPM model, the coverage data generated seems less reliable. Coverage percent favoured IPCW and PP in 19 out of 38 scenarios each. ITT performed the worst in terms of bias, ModSE and coverage across all implementation scenarios. The detailed performance results alongside the mean estimates, standard errors and 95% confidence intervals across all implementation scenarios are provided in Appendix E. The subsequent subsections describe the performance of the alternative methods for each measure across scenarios using nested loop plots.

Table 13: Best-performing methods by performance measure across implementation non-adherence scenarios (1-38)

Method	Bias	MSE	Model-based SE	Empirical SE	Coverage
ITT	0	13	0	38	0
PP	8	22	0	0	19
IPCW	9	2	0	0	19
SNFTM	21	1	38	0	0

5.2.3.1 Bias

Bias was evaluated and presented in terms of absolute bias and percent bias (i.e. a percentage of the true value of the estimated parameter).

Figure 28 shows the nested loop plot illustrating the performance of methods in terms of percent bias across the implementation non-adherence scenarios. The parameterisation of each scenario is summarised at the bottom of the figure. For instance, the sample size used in the simulated datasets was 450 for scenarios 1-18, and a small sample size of 120 was implemented in scenarios 19-38.

SNFTM and IPCW performed the best across most scenarios although in scenarios with large treatment effect size (7-14 and 27-34) there is a modest increase in its percent bias. The level of non-adherence has some influence on percent bias, but g-methods (SNFTM and IPCW) generally handled the variation in non-adherence better than other methods. PP performed best in 8 out of 38 scenarios and generally produced results closer to g-methods in terms of bias. As can be seen in the graph, it should be noted that the levels of bias between IPCW, SNFTM and PP were often very similar. The MCSE for the three methods are also close to each other (detailed MCSEs are provided in Appendix E). When looking at each of these methods, the MCSE ranged between 0.07-0.17% for PP, 0.07-0.22% for IPCW and 0.07-0.17% for SNFTM across all implementation scenarios.

ITT was always the worst-performing method as it produced a higher bias percent (as an overestimate) reaching more than 50% in some cases (see Figure 28). ITT bias was amplified by a larger treatment effect size. Generally, levels of percent bias ranged between 7-38% for the IPCW, PP, SNFTM (20% on average), but between 20-75% for ITT (40% on average).

The relationship between non-adherence and graft survival outcome had a small but noticeable influence on IPCW and SNFTM performance with a stronger relationship increasing bias, although these methods handled this factor better than all other methods.

Looking across all factors, it is clear that treatment effect size is one of the factors contributing to higher bias percentages for ITT with a stronger relationship increasing bias. The level of non-adherence impact on bias is demonstrated by the nature of the lines in the nested loop plot. Other factors contributing to bias include the type of survival time DGM (see Figure 28).

Figure 28: Percentage bias in the estimation of the difference in RMSTs across implementation non-adherence scenarios

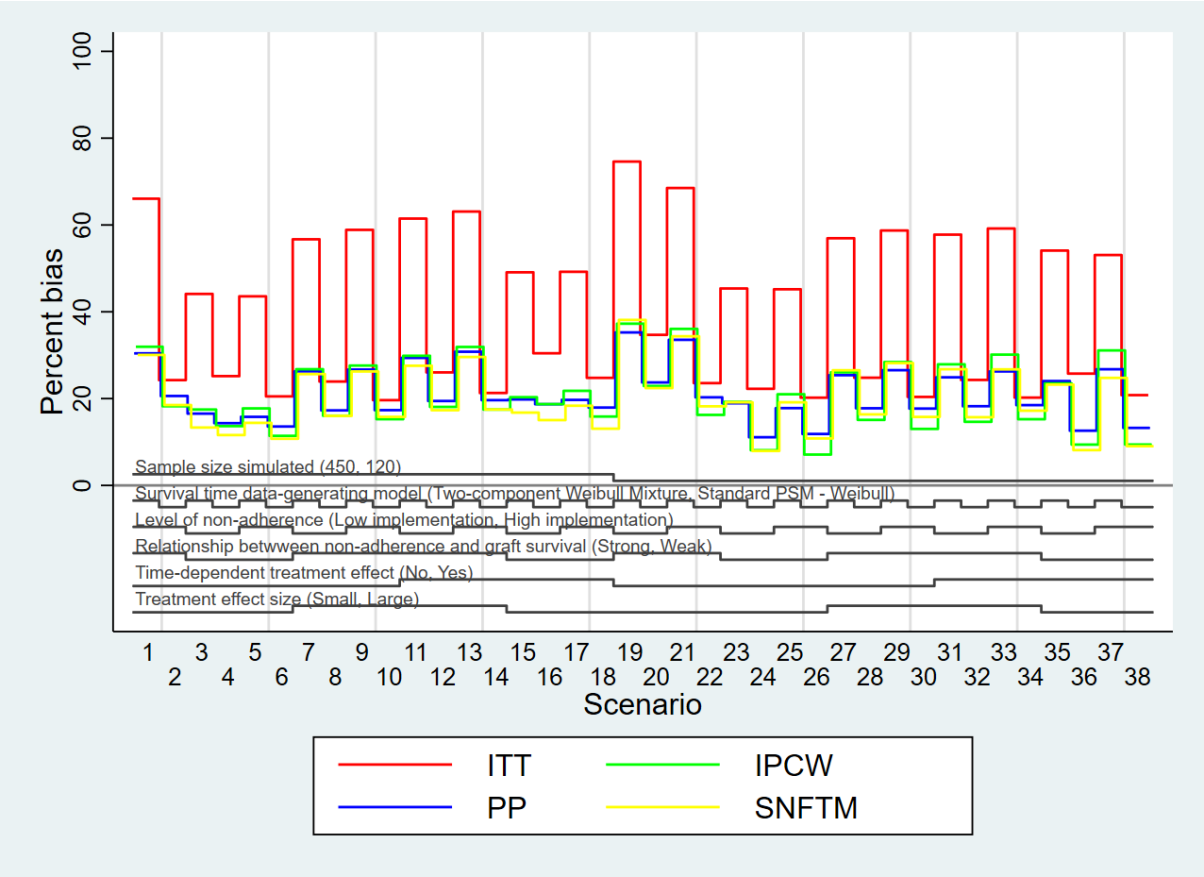
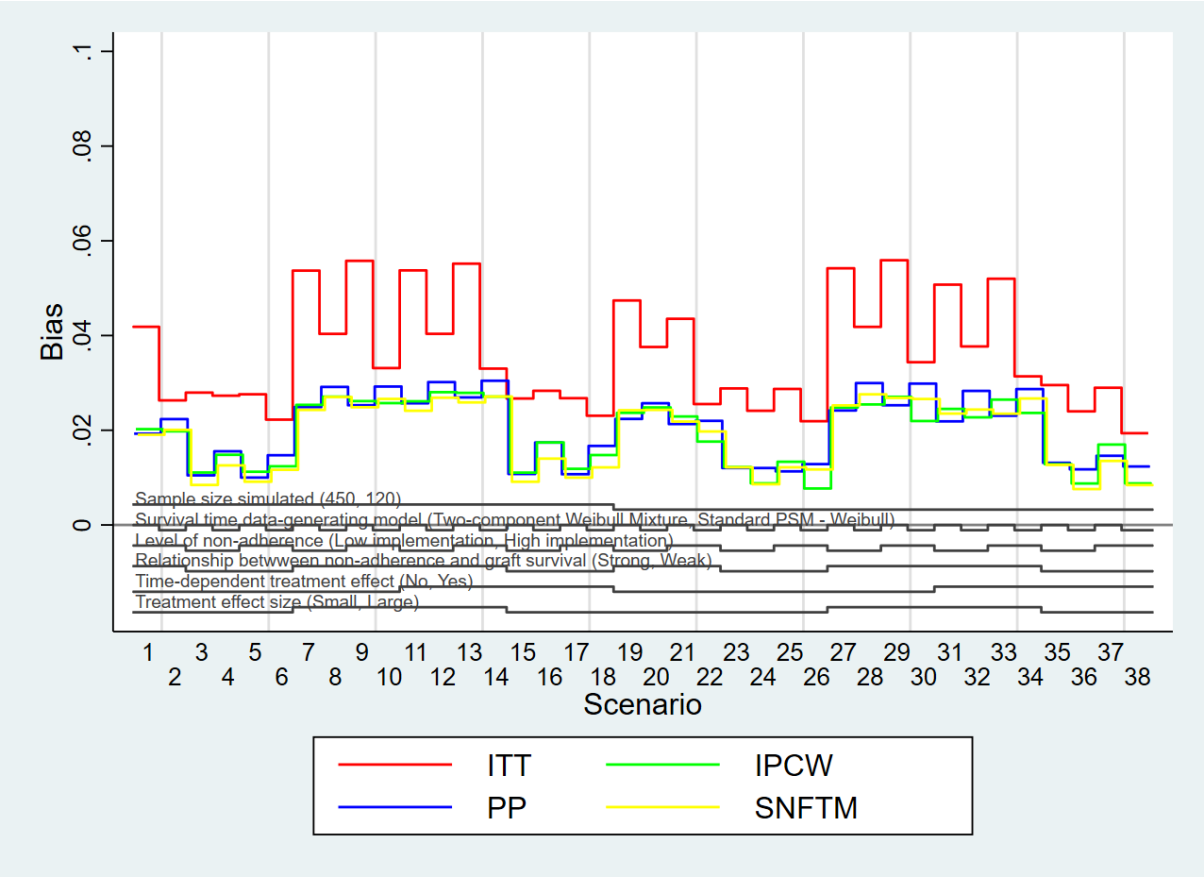


Figure 29 illustrates methods performance in terms of absolute bias. Generally, bias was very small for g-methods and PP; however, in scenarios with a small treatment effect, this has resulted in relatively higher bias percentages as shown in Figure 28.

G-methods produced a small bias of 0.019 years (7 days) on average (12 months follow-up) across all 38 implementation non-adherence scenarios, followed by PP that resulted in it an average bias of 0.020 (7.4 days). In contrast, ITT resulted in a higher bias of 0.036 years (13.1 days) across the same implementation non-adherence scenarios.

Figure 29: Bias in the estimation of the difference in RMSTs across implementation non-adherence scenarios

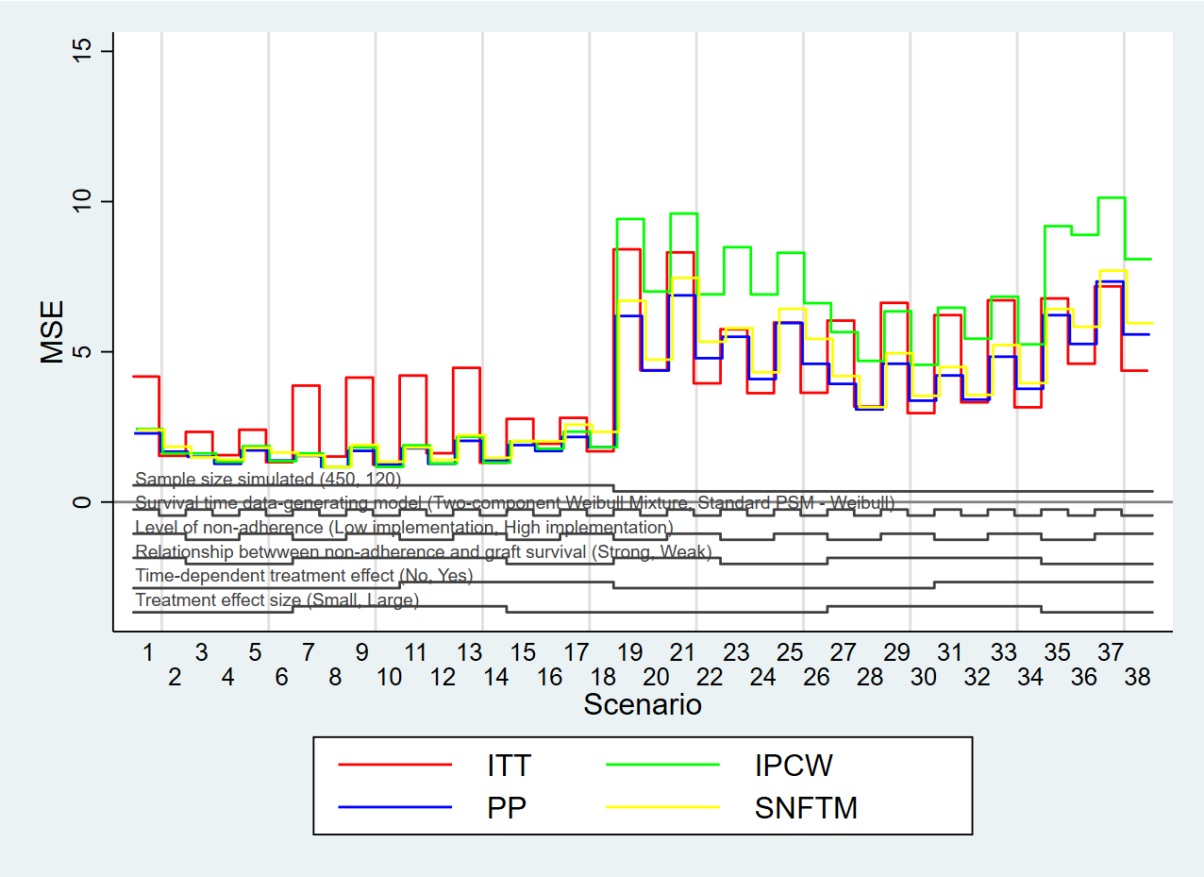


5.2.3.2 *Mean square error*

The performance of methods based on MSE is presented in Figure 30. MSE is a useful performance measure because it combines both bias and variability in a single measure. The methods' performance in terms of MSE is presented as a percent of the truth (Figure 30). PP and g-methods produced the lowest MSE across scenarios with an average MSE of 1.7% for large sample size and this remained below 3% in most of these scenarios. In contrast, ITT produced much higher MSE with higher values up to 4.5% in some cases.

Interestingly, ITT showed improved MSE performance in scenarios with a smaller sample size (Scenarios 19-38) with g-methods and PP lagging behind in these scenarios. There is some fluctuation in MSE performance resulting from a combination of treatment effect size with other factors such as the level of non-adherence and the relationship between non-adherence and graft survival outcome (see Figure 30). PP analysis produced good MSE performance compared to the alternative methods, although the method struggled to cope in scenarios with a small sample size, as did g-methods.

Figure 30: MSE in the estimation of the difference in RMSTs across implementation non-adherence scenarios

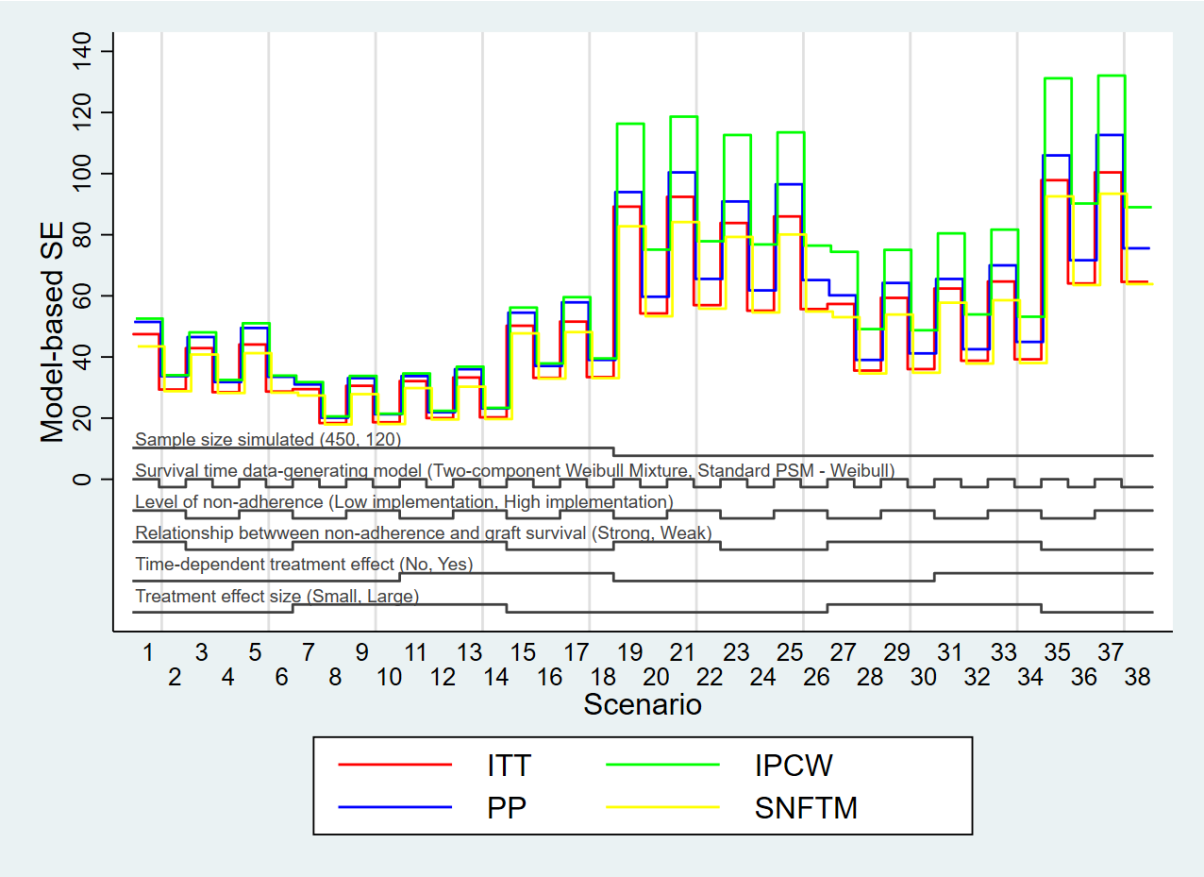


5.2.3.3 *Model-based standard error*

The performance of adjustment methods in terms of ModSE across the implementation non-adherence scenarios is presented in Figure 31. These figures are reported as a percentage of the truth to aid the comparisons of methods across the range of scenarios. As illustrated in the nested loop plot (Figure 31), scenarios with a smaller sample size (19-38) produced higher ModSE.

For ModSE, SNFTM performed the best in all the implementation non-adherence scenarios with IPCW and PP showing lower performance across scenarios. ITT always produced ModSE higher than SNFTM but better than IPCW and PP in most cases with higher ModSE % in scenarios with a small sample size. In addition to sample size, treatment effect size, time-dependent treatment effect and the relationship between non-adherence and graft survival outcome are the factors that had a noticeable influence on ModSE performance. A larger treatment effect size (Scenario 7-14 and 27-34) contributed to generating higher ModSE across methods, although SNFTM remained the best performing method in all of these scenarios.

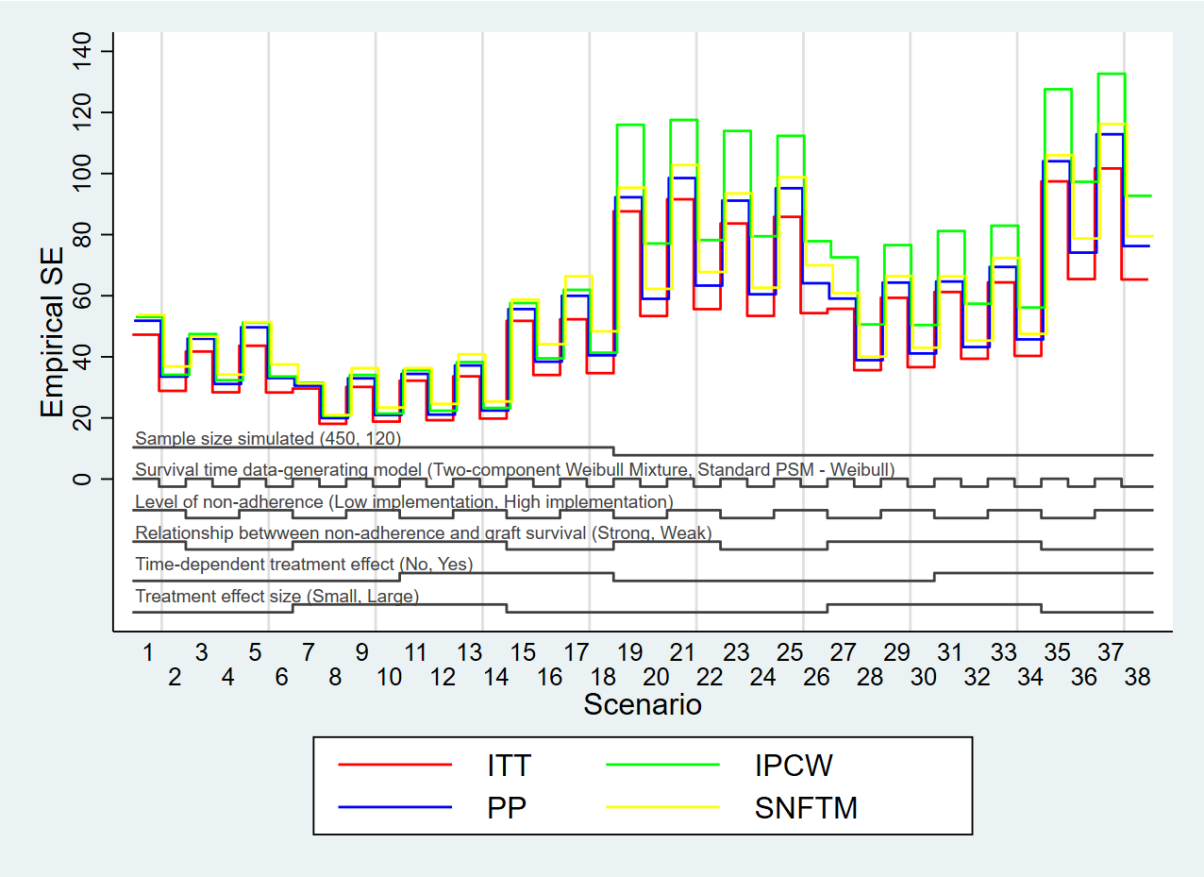
Figure 31: Model-based SE in the estimation of the difference in RMSTs across implementation non-adherence scenarios



5.2.3.4 *Empirical standard error*

The EmpSE represents the standard deviation of the estimates across the successful simulations (out of 1900 iterations). The performance of methods in terms of EmpSE across the implementation scenarios is illustrated in Figure 32. The results favoured ITT across scenarios, however, this should be interpreted with caution because ITT has generated the highest bias and produced the worst ModSE performance across all implementation non-adherence scenarios. Nevertheless, EmpSE performance is reported to provide the full picture in terms of the simulation results and to comply with the study protocol.

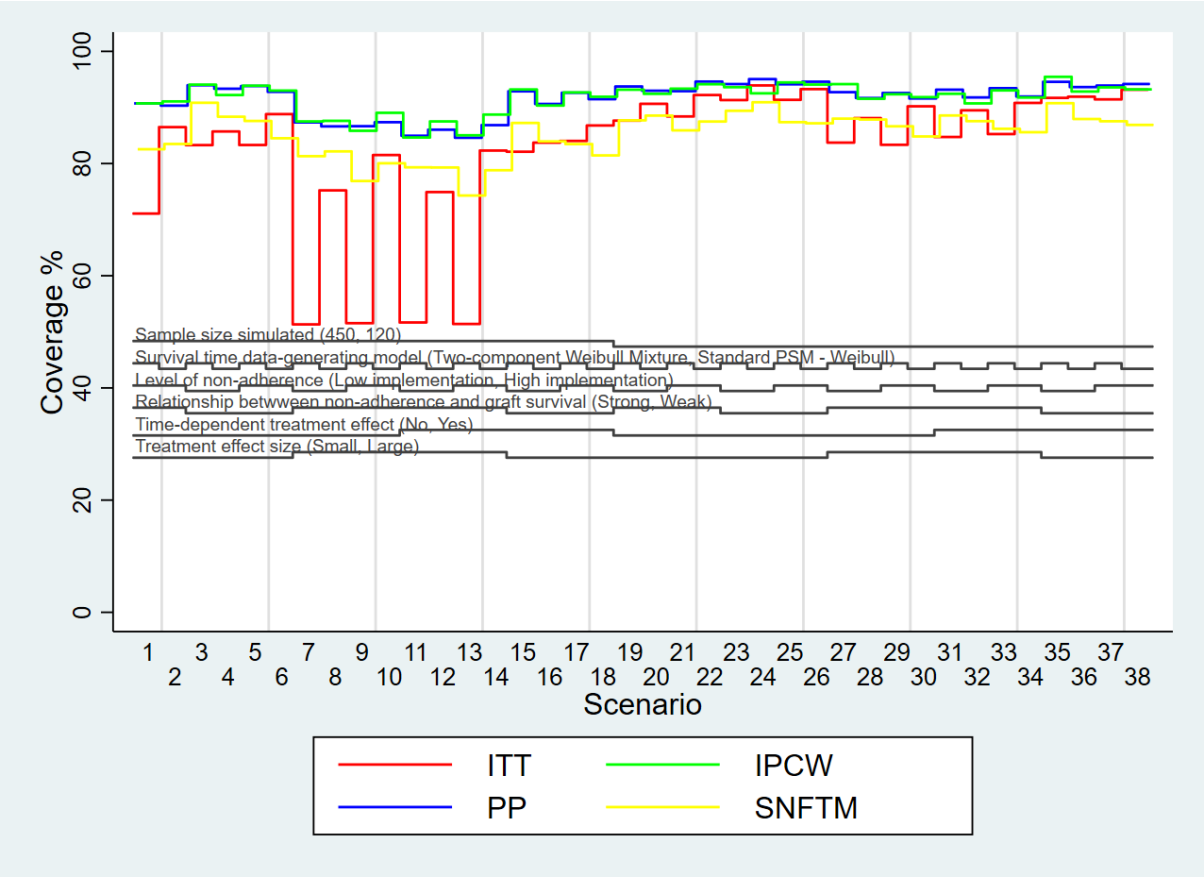
Figure 32: Empirical SE in the estimation of the difference in RMSTs across implementation non-adherence scenarios



5.2.3.5 *Coverage*

Figure 33 reports coverage percent for each method across 38 implementation non-adherence scenarios. Coverage percent represents the proportion of times that the 95% confidence interval contains the true value of the estimated parameter (i.e. the difference in RMST). Coverage performance data favoured IPCW and PP across scenarios with coverage percent reaching more than 90% in most cases. In contrast, ITT produced lower coverage especially in scenarios with large sample size. SNFTM struggled in terms of coverage in scenarios with small sample size.

Figure 33: Coverage in the estimation of the difference in RMSTs across implementation non-adherence scenarios



5.2.4 Performance of methods across persistence non-adherence scenarios

This section presents the results of methods performance across 18 scenarios adjusting for persistence non-adherence (Scenarios 39-56) as specified in the simulation study protocol. First, I present a summary table highlighting the results in terms of the number of times each method performed the best across the 18 scenarios. Then, the results by performance measure are presented and interpreted using nested loop plots in the subsequent subsections. The detailed results are provided in Appendix E.

Table 14 shows that SNFTM performed best across 7 of the 18 scenarios evaluated in terms of bias followed by IPCW (6 scenarios) and PP (5 scenarios). Often the results were very close between g-methods and PP. ITT was the worst-performing method across all persistence non-adherence scenarios. MSE favoured PP and ModSE showed SNFTM as the best performing method across all scenarios. EmpSE favoured ITT in 17 out of 18 scenarios with PP performing best in the remaining one scenario. Results using EmpSE as a performance measure should be interpreted with caution by

looking into the results from all other measures at the same time. In terms of coverage, IPCW was the best performing method in 11 out of 18 scenarios with PP performed better in the remaining 7 scenarios. ITT and SNFTM lagged behind in terms of coverage. The subsequent subsections illustrate methods performance across scenarios and the influence of the seven factors incorporated in the design of the simulation study across scenarios.

Table 14: Best-performing methods by performance measure across persistence non-adherence scenarios (39-56)

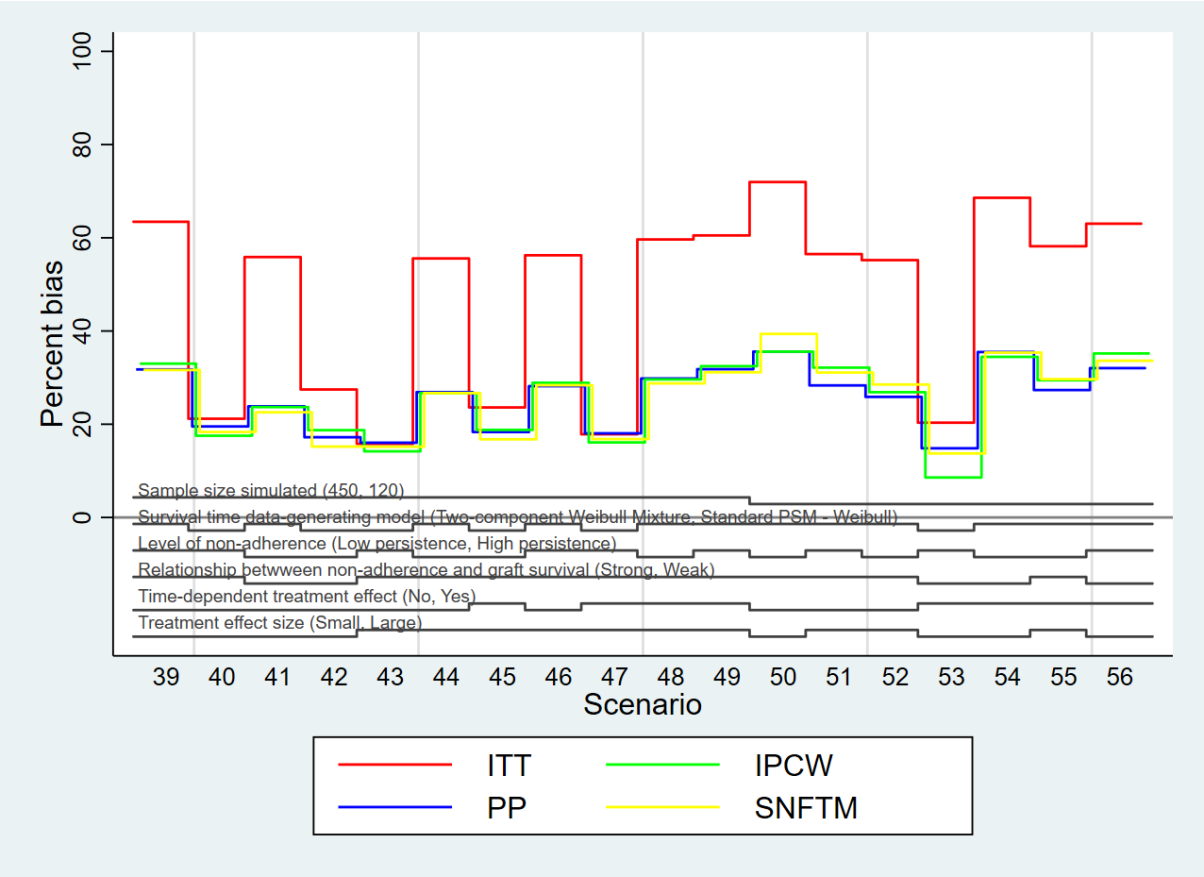
Method	Bias rank	MSE rank	Model-based SE rank	Empirical SE rank	Coverage rank
ITT	0	4	0	17	0
PP	5	12	0	1	7
IPCW	6	0	0	0	11
SNFTM	7	2	18	0	0

5.2.4.1 Bias

Figure 34 illustrates methods performance in terms of percent bias across the persistence non-adherence scenarios in both the control and experimental arms. G-methods (SNFTM and IPCW) performed well across 13 out of the 18 scenarios with PP producing better results in 5 scenarios with a small sample size. In all scenarios, the difference between g-methods and PP method is very small. When looking at MCSE (reported in Appendix E), the MCSE are also close with values ranging from 0.07 to 0.17% for PP, 0.07-0.23% for IPCW and 0.07-0.18% for SNFTM across all persistence scenarios.

Figure 34 shows that small sample size has a negative impact on the performance of g-methods leading to higher bias compared to the simple censoring method (PP). SNFTM, IPCW and PP produced lower bias in direction of overestimation with bias percent around 20% although this is slightly higher (around 30%) in scenarios with a small sample size. ITT generated substantially high bias, in most scenarios reaching up to 60% of the true value. Generally, levels of percent bias ranged between 9-39% for the IPCW, PP, SNFTM (26% on average), but between 16-72% for ITT (47% on average).

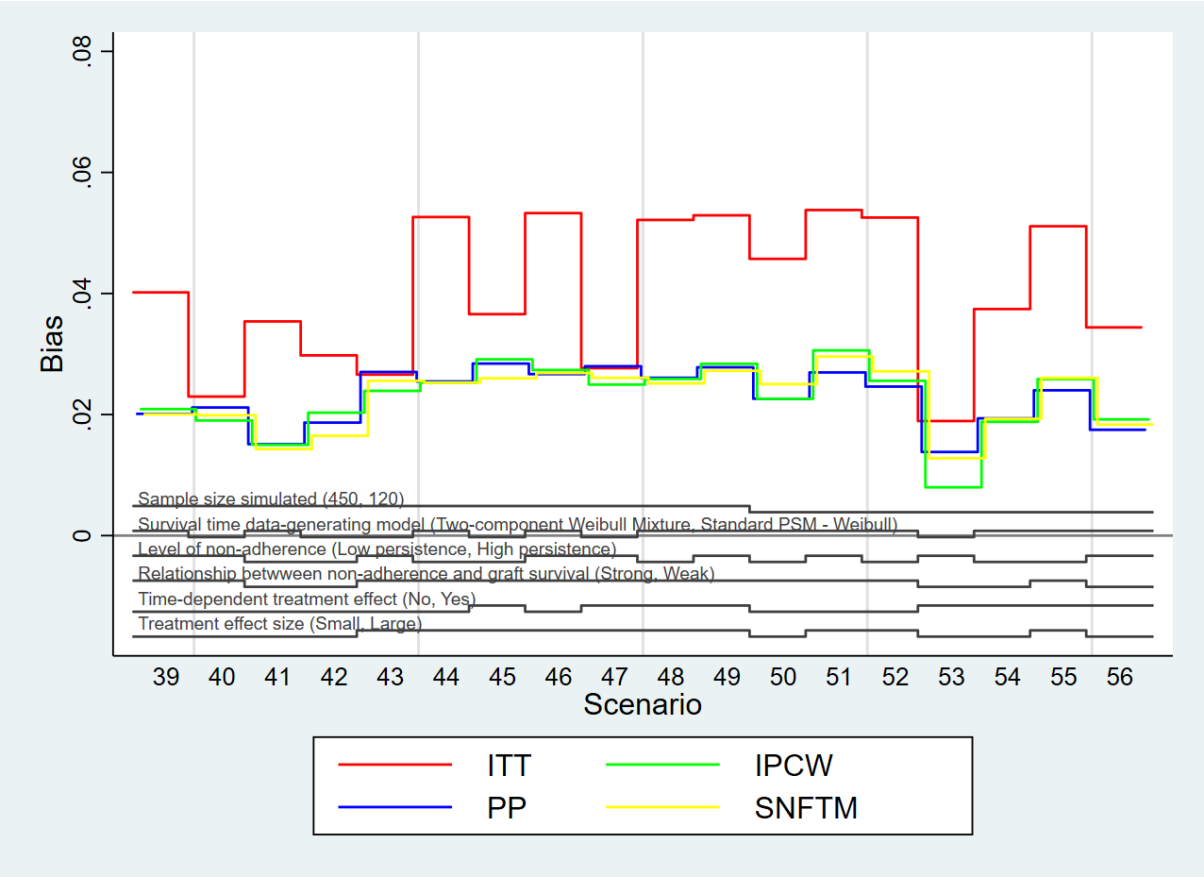
Figure 34: Percentage bias in the estimation of the difference in RMSTs across persistence non-adherence scenarios



In terms of absolute bias, g-methods produced a small bias of 0.023 years (8.3 days) on average (12 months follow-up) across all 18 persistence non-adherence scenarios. PP produced results that were very close to the g-method with an average absolute bias of 0.023 (8.4 days). In contrast to a higher bias of 0.04 years (14.7 days) produced by ITT across the same scenarios. The nested loop plot (Figure 35) clearly shows that ITT was the worst-performing method across all scenarios in terms of bias in the direction of over-estimation. As in percent bias, the type of baseline hazard function specified by the DGM has an impact on the size of bias produced by each method. The combined effect of the simulation specification factors (shown as descriptors) on bias is illustrated.

In summary, g-methods (SNFTM and IPCW) were the best performing method in terms of percent bias (compared to alternative methods) in adjusting treatment effect for persistence non-adherence, although it produced higher bias in some scenarios when the results are compared with implementation non-adherence. PP did well in most scenarios as it produced performance results closer to g-methods (better in some cases) with ITT always produced the worst performance in terms of bias.

Figure 35: Bias in the estimation of the difference in RMSTs across persistence non-adherence scenarios

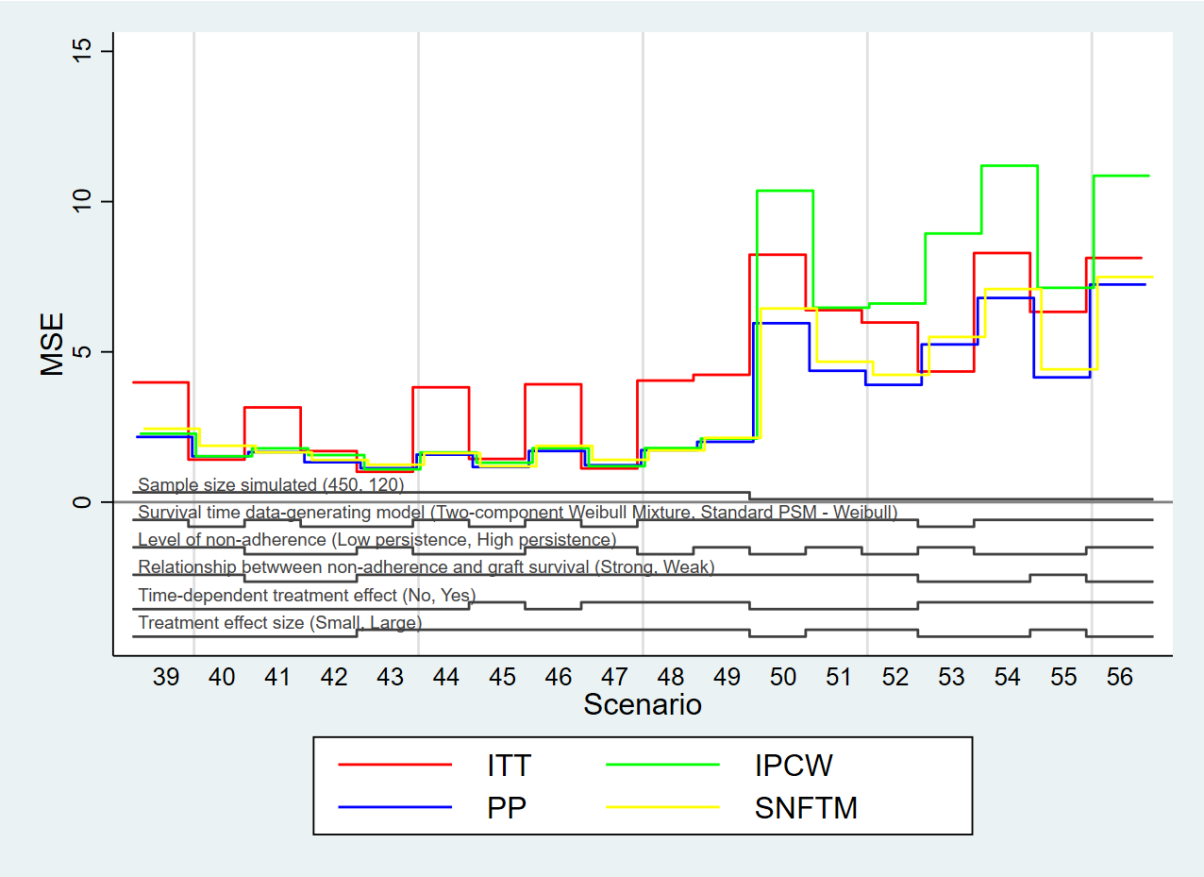


5.2.4.2 *Mean square error*

Figure 36 illustrates the results of methods performance in terms of MSE. The values are expressed as a percentage of the truth to aid comparison using the nested loop plots. As shown in the figure, PP permed best especially in scenarios with a small sample size. The IPCW performance was influenced by the combination of sample size and treatment effect size with a small sample size and low persistence leading to a higher MSE percentage.

SNFTM, IPCW and PP generally did well in scenarios with a large sample size (n=450) with MSE below 2.5 % in most scenarios. In these scenarios, ITT produced higher MSE with clear fluctuation influenced by survival time DGM and level of non-adherence. MSE performance results were influenced by sample size, with IPCW performing worst in scenarios with a small sample size with MSE values reaching up to 10 % of the truth. PP performed better than ITT in most scenarios (See Figure 36)

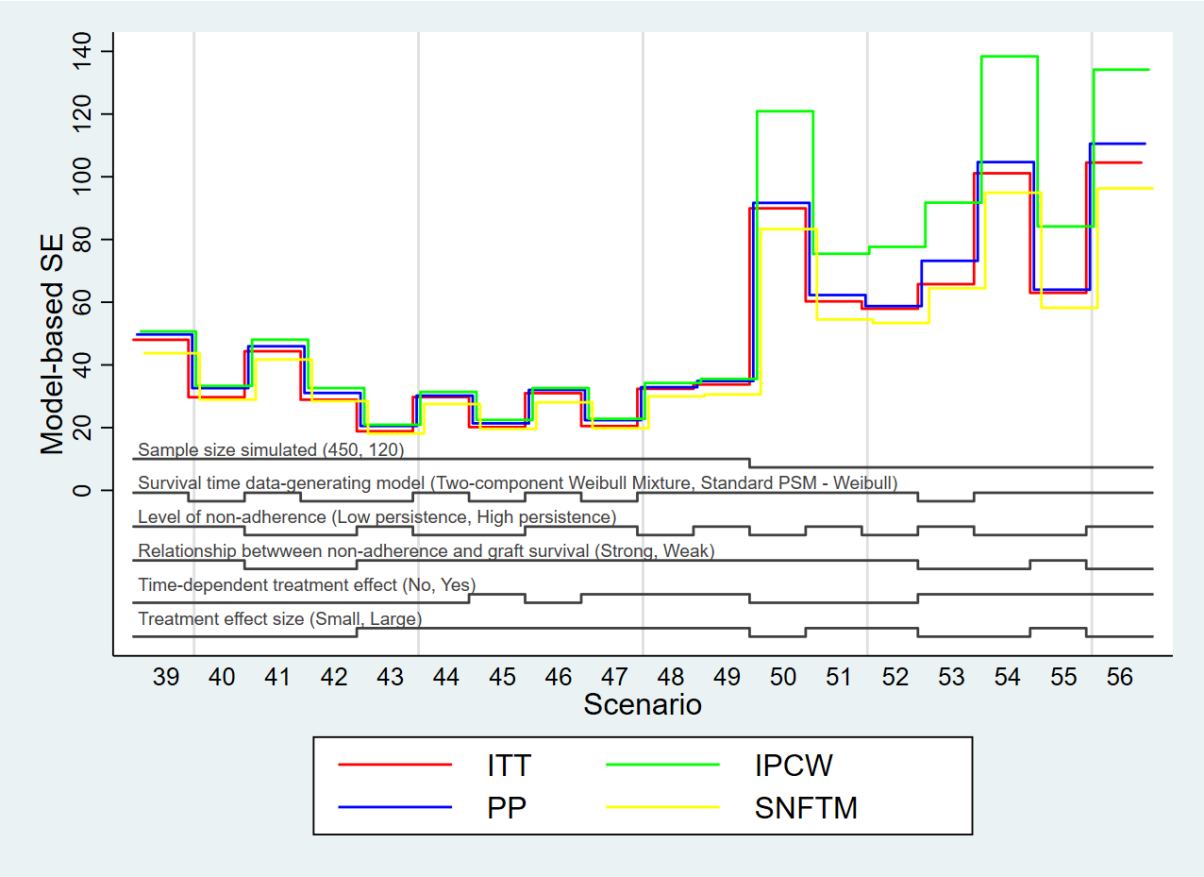
Figure 36: MSE in the estimation of the difference in RMSTs across persistence non-adherence scenarios



5.2.4.3 *Model-based standard error*

Figure 37 shows the performance results in adjusting for persistence non-adherence using ModSE. SNFTM with g-estimation performed the best across all persistence non-adherence scenarios with the most noticeable trend relating to sample size. In scenarios with a large sample size of 450 observations (39-49), SNFTM performed better than scenarios with a small size of 120 (50-56). The impact of a small sample size on ModSE was much bigger for IPCW with substantially higher values in scenarios with a small sample size. Large treatment effect size (in combination with small sample size) led to an even higher ModSE percentage for IPCW.

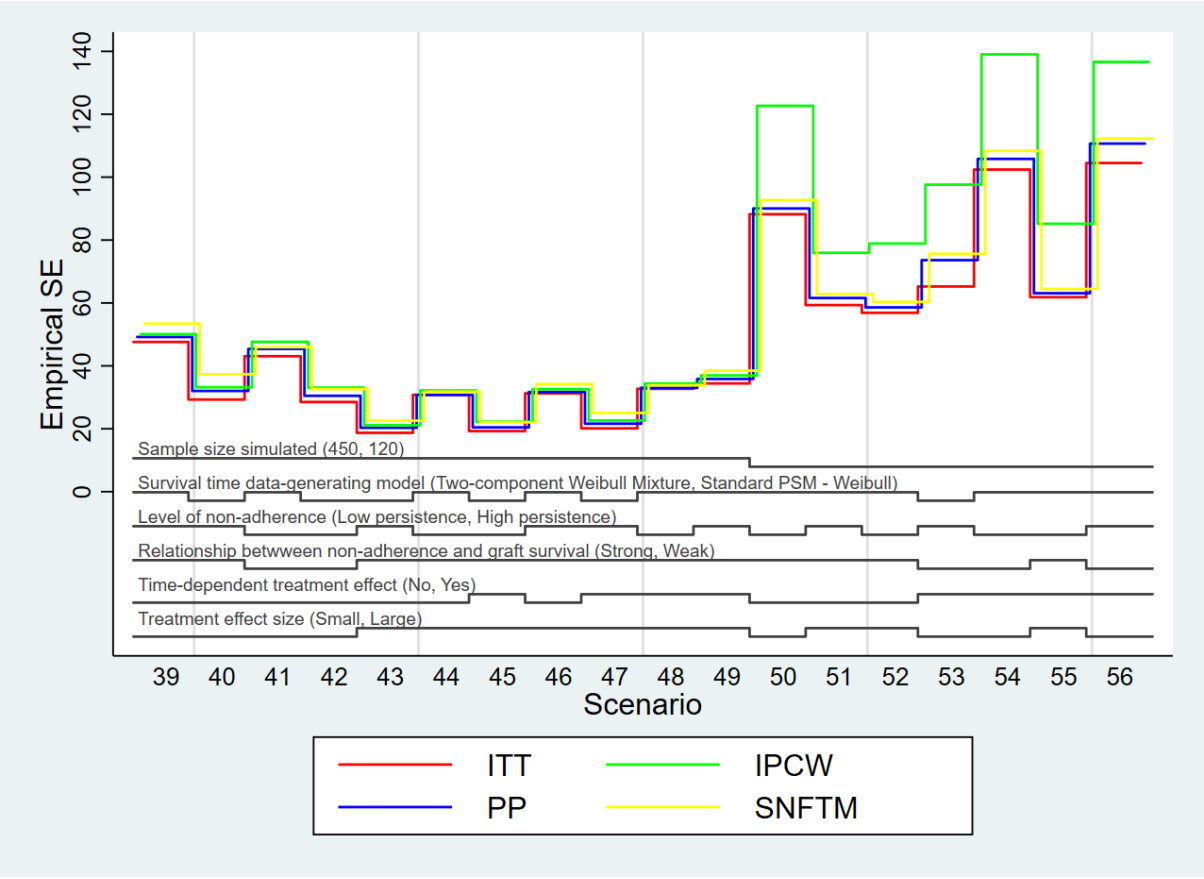
Figure 37: Model-based SE in the estimation of the difference in RMSTs across persistence non-adherence scenarios



5.2.4.4 Empirical standard error

Figure 38 illustrates methods performance using EmpSE across the 18 persistence non-adherence scenarios (39-56). These results show ITT as the best performing method followed by PP with SNFTM and IPCW produced higher EmpSE. These results should be interpreted with caution as discussed in Section 4.9.3.4.

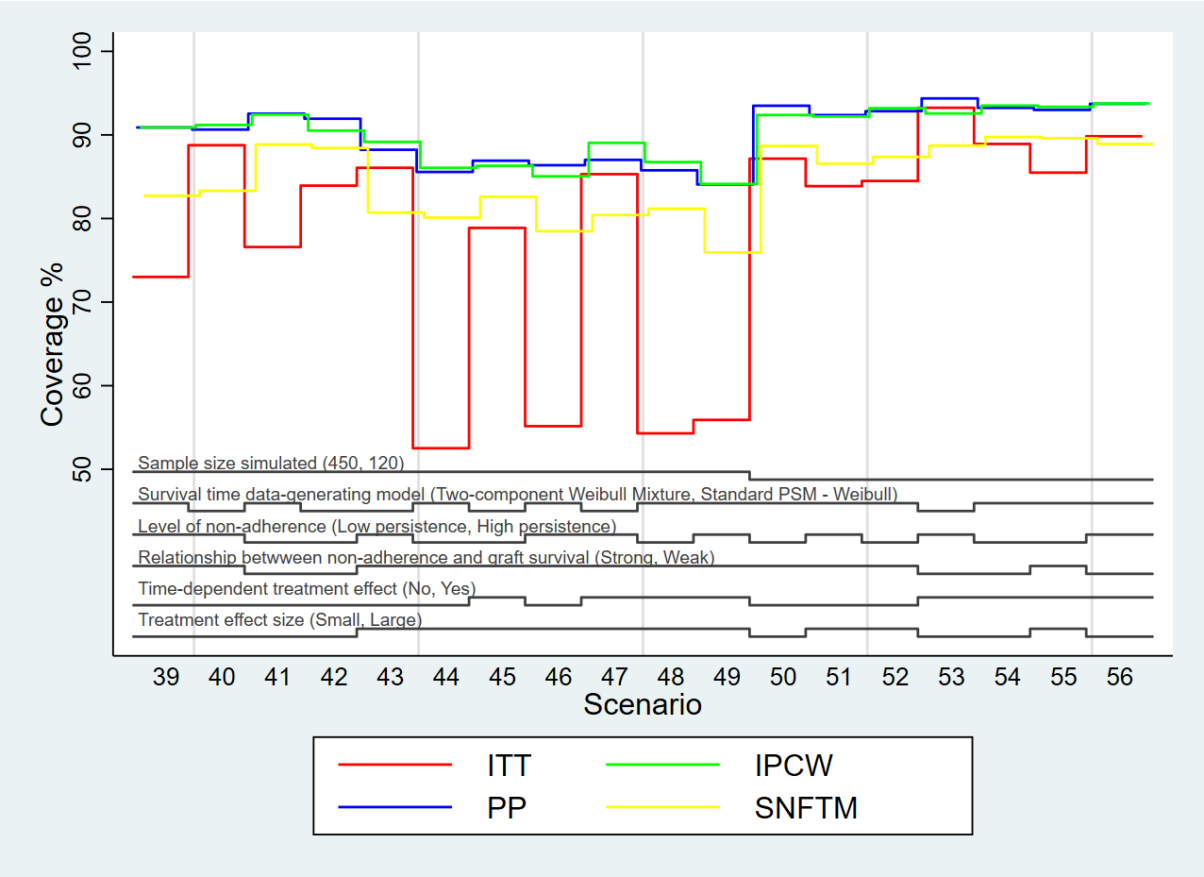
Figure 38: Empirical SE in the estimation of the difference in RMSTs across persistence non-adherence scenarios



5.2.4.5 *Coverage*

Figure 39 presents coverage percent showing the probability that the 95% confidence interval contains the true value of the estimated parameter (i.e. the difference in RMST) across the successful simulations. IPCW performed best in most scenarios but PP did better in scenarios with large treatment effect size. IPCW and PP coverage was generally very high reaching more than 90% in most cases. SNFTM lagged behind IPCW and PP in terms of coverage, but the method still produced better performance compared to ITT, which was the worst across all scenarios. Coverage was affected by the DGM with ITT resulted in coverage lower than 60% in some cases.

Figure 39: Coverage in the estimation of the difference in RMSTs across persistence non-adherence scenarios



5.2.5 Performance of methods across initiation non-adherence scenarios

The performance results of adjustment methods across 34 initiation non-adherence scenarios (57-90) are presented in this section. An overview of the results summarising the number of times each method ranked first on each performance measure is provided. This is followed by highlighting the results in the subsequent subsections in terms of performance measures (bias, MSE, ModSE, EmpSE and coverage) using nested loop plots. The detailed results for the performance of adjustment methods across all initiation non-adherence scenarios are presented in Appendix E.

Table 15 summarises the results based on the number of times each method performed the best compared to the alternative methods assessed across initiation scenarios. SNFTM performed the best in terms of bias and ModSE compared to simple methods and IPCW in adjusting estimates of treatment effect for initiation non-adherence across scenarios, although PP produced better MSE in most scenarios. In terms of bias percent, SNFTM performed the best in 19 out of 34 scenarios with PP performed the best in 11 scenarios. EmpSE error favoured simple methods with ITT performing best;

however, this should be interpreted with caution because ITT produced higher bias in most scenarios. Coverage also favours SNFTM in 21 out of 34 scenarios with PP performed best in 10 scenarios and ITT in three scenarios. Often there was very little to choose between IPCW, SNFTM and PP as the performance are very close.

Table 15: Best-performing methods by performance measure across initiation non-adherence scenarios (57-90)

Method	Bias rank	MSE rank	Model-based SE rank	Empirical SE rank	Coverage rank
ITT	3	3	0	29	3
PP	11	26	0	5	10
IPCW	1	0	0	0	21
SNFTM	19	5	34	0	0

5.2.5.1 Bias

Figure 40 illustrates bias expressed as a percentage of the true difference in RMST across 34 initiation non-adherence scenarios. SNFTM with g-estimation produced the smallest bias compared to the alternative methods in most scenarios. In contrast, to the performance results in implementation and persistence non-adherence, PP did better than IPCW in terms of bias percentage. Similar to the other types of non-adherence, the difference between IPCW, SNM and PP is very small. MCSE data also show close results for the three methods with numbers ranging between 0.06-0.18 for PP, 0.07-0.27 for IPCW and 0.7-0.20% across all initiation scenarios. Bias performance fluctuated, largely dependent on the data-generating model, sample size and treatment effect size.

In contrast, ITT performance was worse than g-methods and PP in terms of bias across all but three scenarios. The direction of bias produced by all methods was in the positive region compared suggesting an overestimation, although the direction of bias changed to negative (suggesting an under-estimation) in some scenarios (83-86). This change was influenced by a combination of small sample size, large treatment effect size and time-dependent treatment effect. In scenarios with a large sample size (57-74), the average bias percentage produced by SNFTM was 14.7% in contrast with an average bias of 52.1% generated by the ITT analysis. PP resulted in a low bias percentage (closer to g-methods) in the positive region of the nested loop plot indicating over-estimation in most scenarios.

SNFTM with g-estimation produced better performance in scenarios with a large sample size with bias percent closer to zero in most cases (see Figure 40). In addition to the small sample size, the other main influencer of bias seems to be the treatment effect size with a larger effect size leading to higher

bias, although g-methods and PP handled treatment effect size better than ITT in all scenarios. Generally, levels of percent bias ranged between -34 to 110% for the IPCW, PP, SNFTM (35% on average), but between -18 to 200% for ITT (75% on average).

Figure 40: Percentage bias in the estimation of the difference in RMSTs across initiation non-adherence scenarios

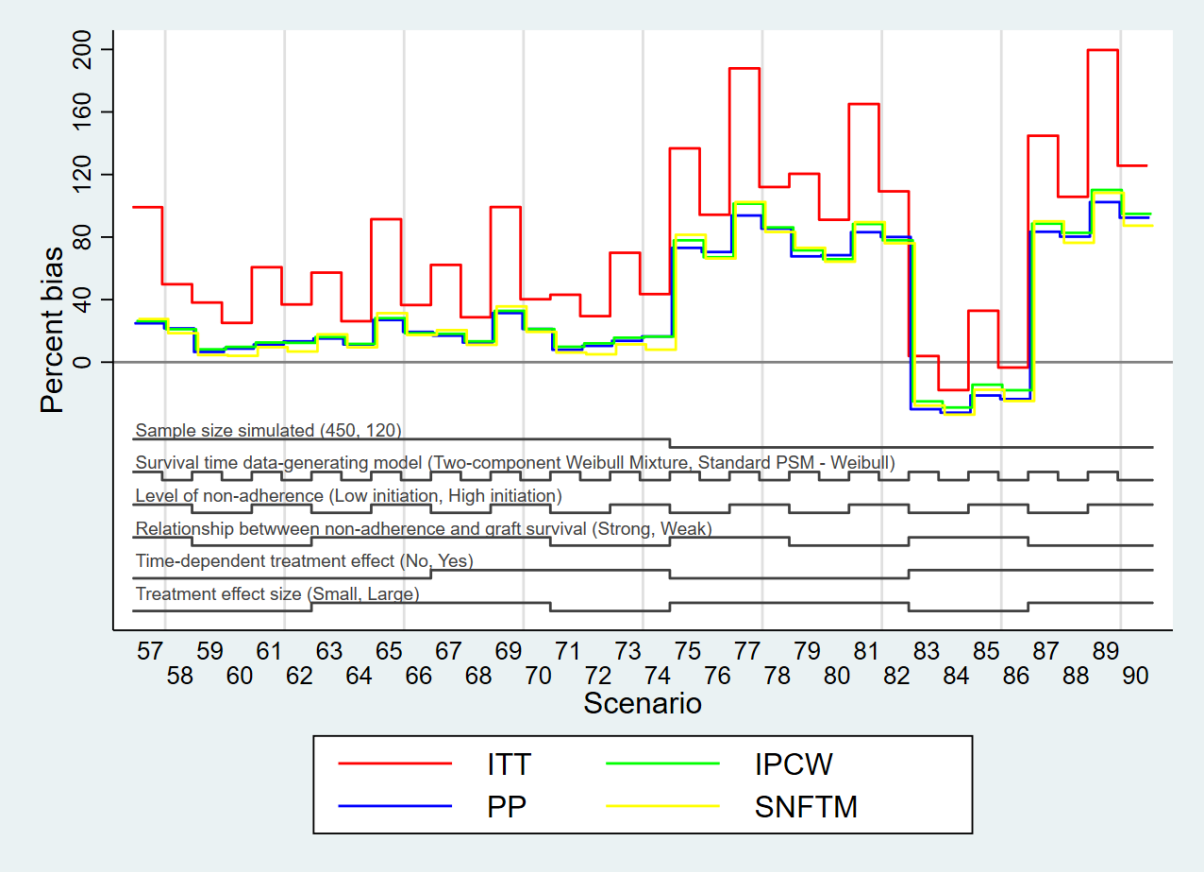
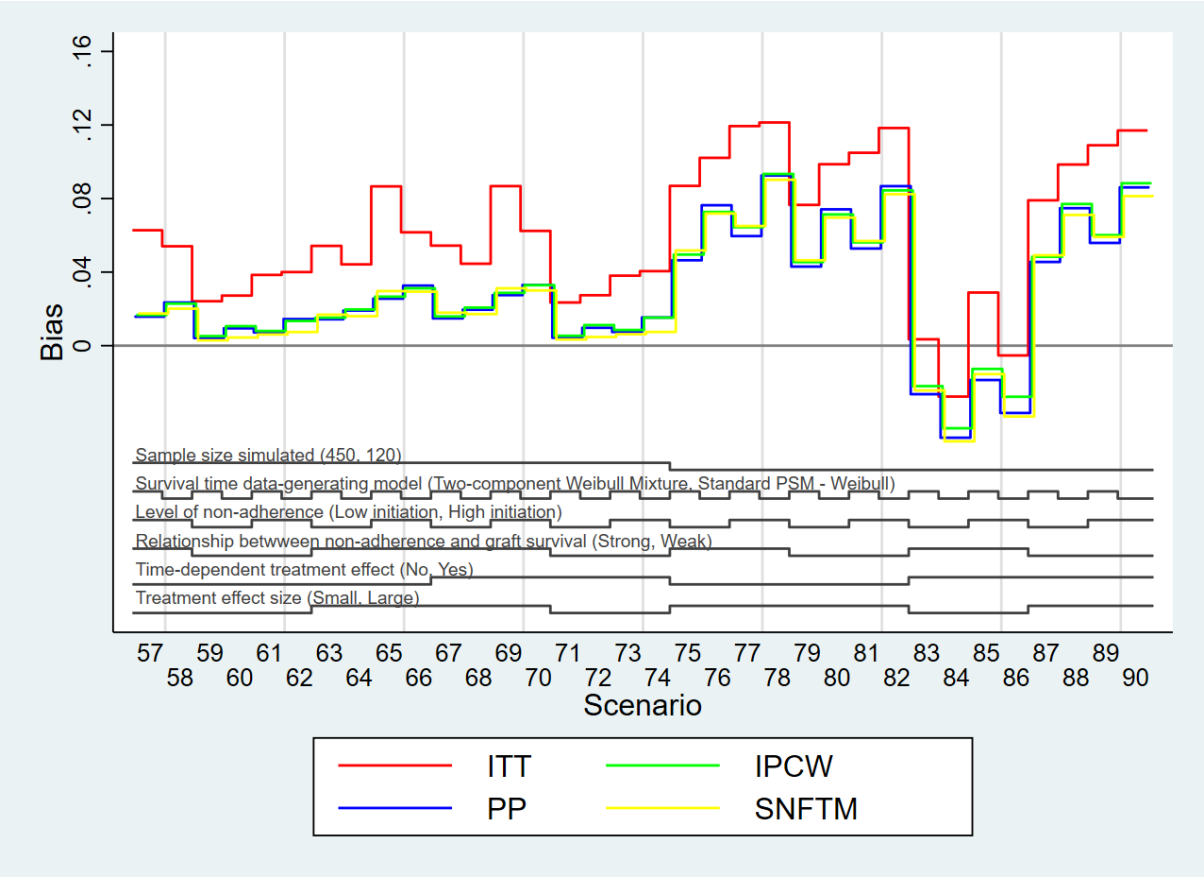


Figure 41 presents performance as absolute bias with results largely similar to percent bias as discussed above. SNFTM produced the smallest bias of 0.035 years (12.8 days) on average compared to the alternative methods across the 34 initiation non-adherence scenarios with even smaller bias in scenarios with a large sample size (0.015 years [5.4 days]). This is followed by PP and IPCW that generated absolute bias of 0.036 (13.1) and 0.036 (13.2), respectively. In contrast, ITT was the worst-performing method as it resulted in a higher bias of up to 0.064 years (23.3 days) across the same scenarios. That is 82% higher bias compared with bias produced by SNFTM with g-estimation.

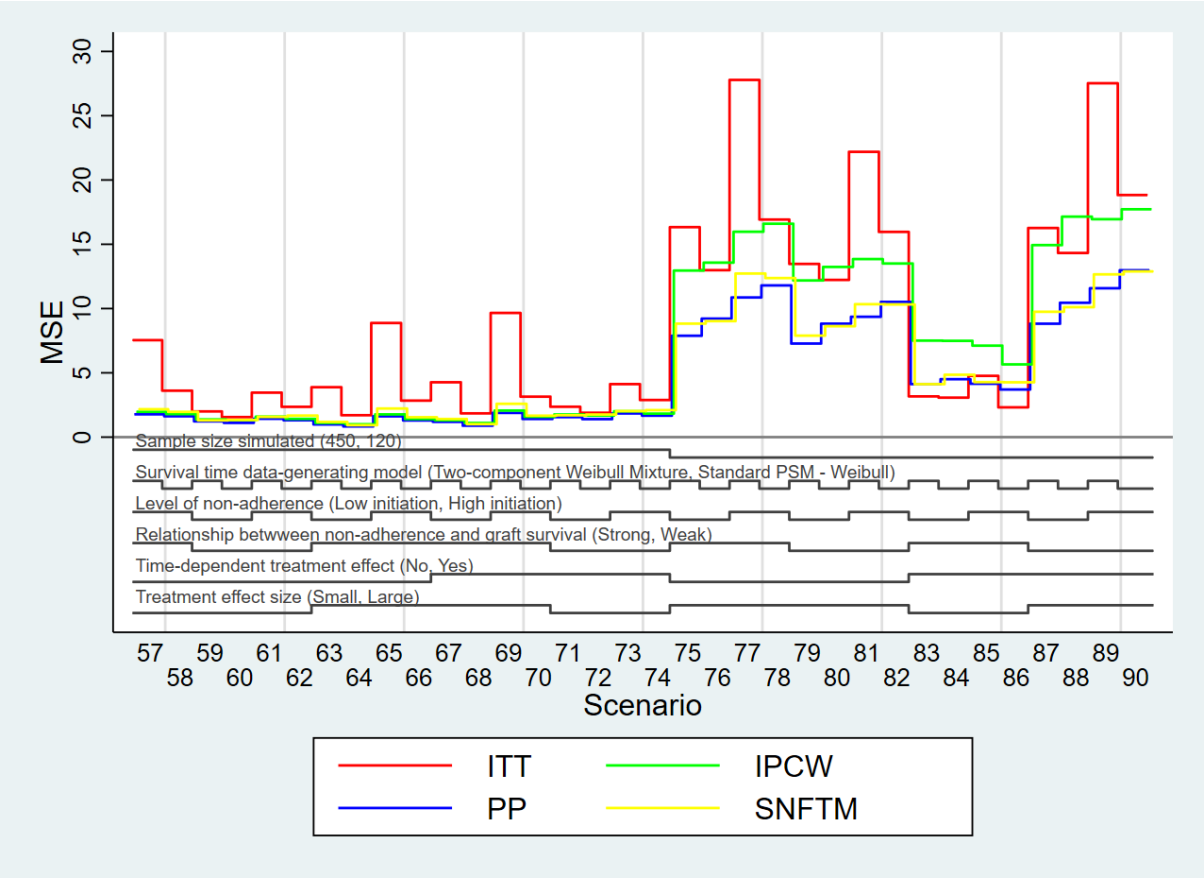
Figure 41: Bias in the estimation of the difference in RMSTs across initiation non-adherence scenarios



5.2.5.2 *Mean square error*

Figure 42 shows the results using MSE as a performance measure. G-methods (SNFTM and IPCW) resulted in very low MSE (expressed as parentage of the true difference in RMST) compared to ITT across scenarios with a large sample size. For these methods, MSE percent was less than 9.7% with an average of 2.1% across compared with an average MSE of 3.7% for ITT across the same scenarios. G-methods struggled in terms of MSE performance in scenarios with a small sample size combined with high treatment effect size as a key contributing factor. ITT was the worst-performing method across most scenarios with an MSE of more than 25% in some scenarios with a small sample size. ITT did better in scenarios with a small sample size, small treatment effect size and time-dependent treatment effect (Scenarios 83-86). The influence of each factor combined with other factors specified in the simulation study design is illustrated by the descriptors in the nested loop plots (see Figure 42).

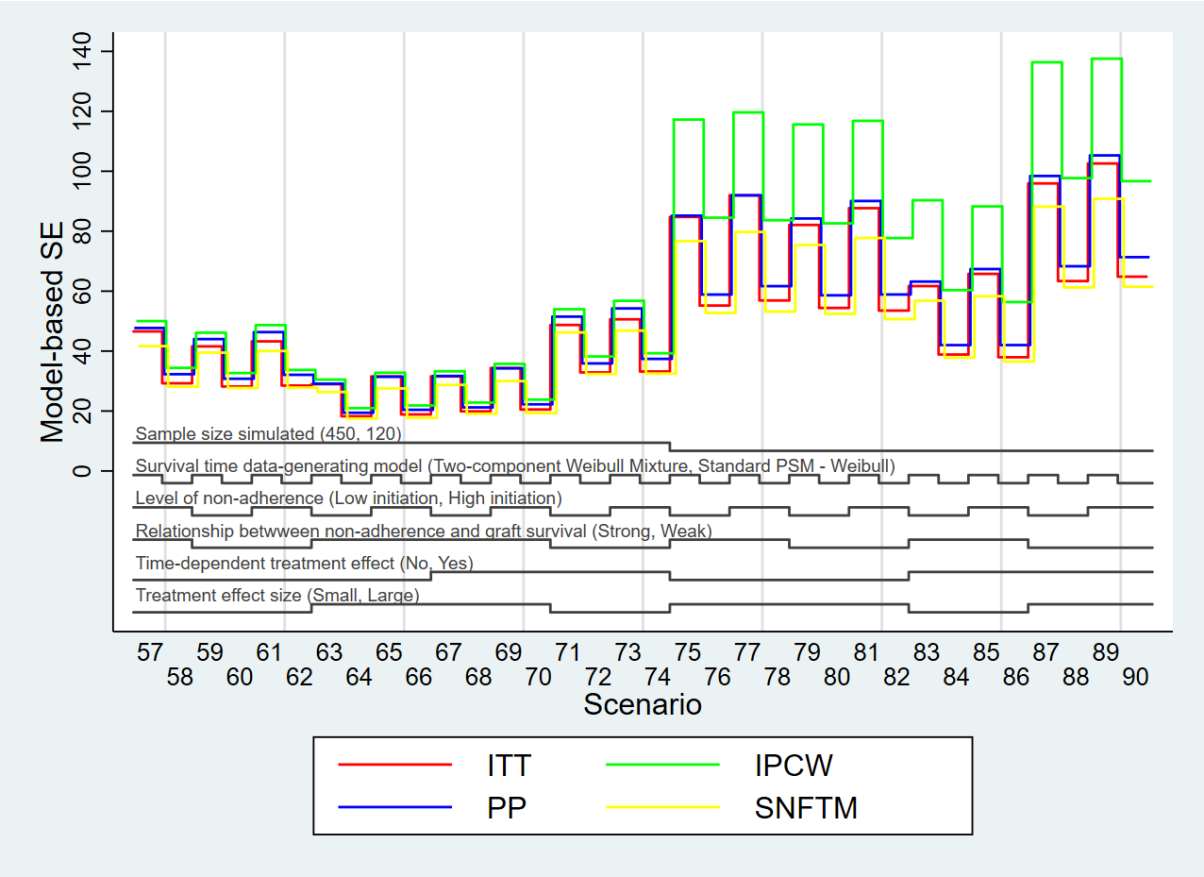
Figure 42: MSE in the estimation of the difference in RMSTs across initiation non-adherence scenarios



5.2.5.3 *Model-based standard error*

Methods performance based on ModSE (as a percentage of the truth) is presented in Figure 43. SNFTM with g-estimation showed better performance compared to the alternative methods across all 34 initiation non-adherence scenarios. ModSE in scenarios with a large sample size (n=450) is better than scenarios with a small sample size as illustrated in the nested loop plot (Figure 43). IPCW produced higher ModSE in scenarios with a small sample size (Scenarios 75-90). Similar to implementation and persistence non-adherence, treatment effect size had a large influence on methods performance when it comes to adjusting treatment effect for initiation non-adherence.

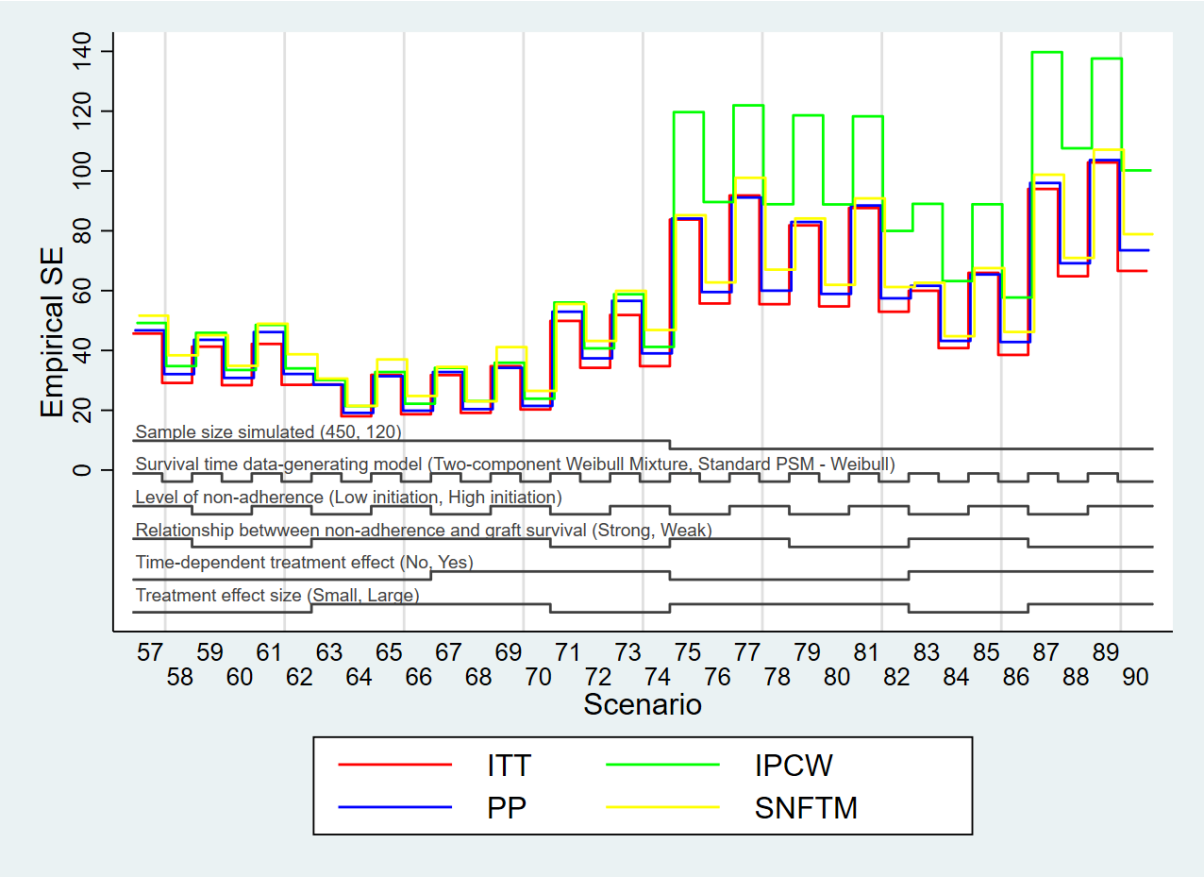
Figure 43: Model-based SE in the estimation of the difference in RMSTs across initiation non-adherence scenarios



5.2.5.4 *Empirical standard error*

Figure 44 illustrates the performance of methods using EmpSE across 34 initiation non-adherence scenarios. As shown, ITT and PP produced better results compared to g-methods with IPCW generated higher EmpSE in scenarios with a small sample size. As discussed earlier, the EmpSE results should be interpreted with caution and read alongside the results of the other performance measures.

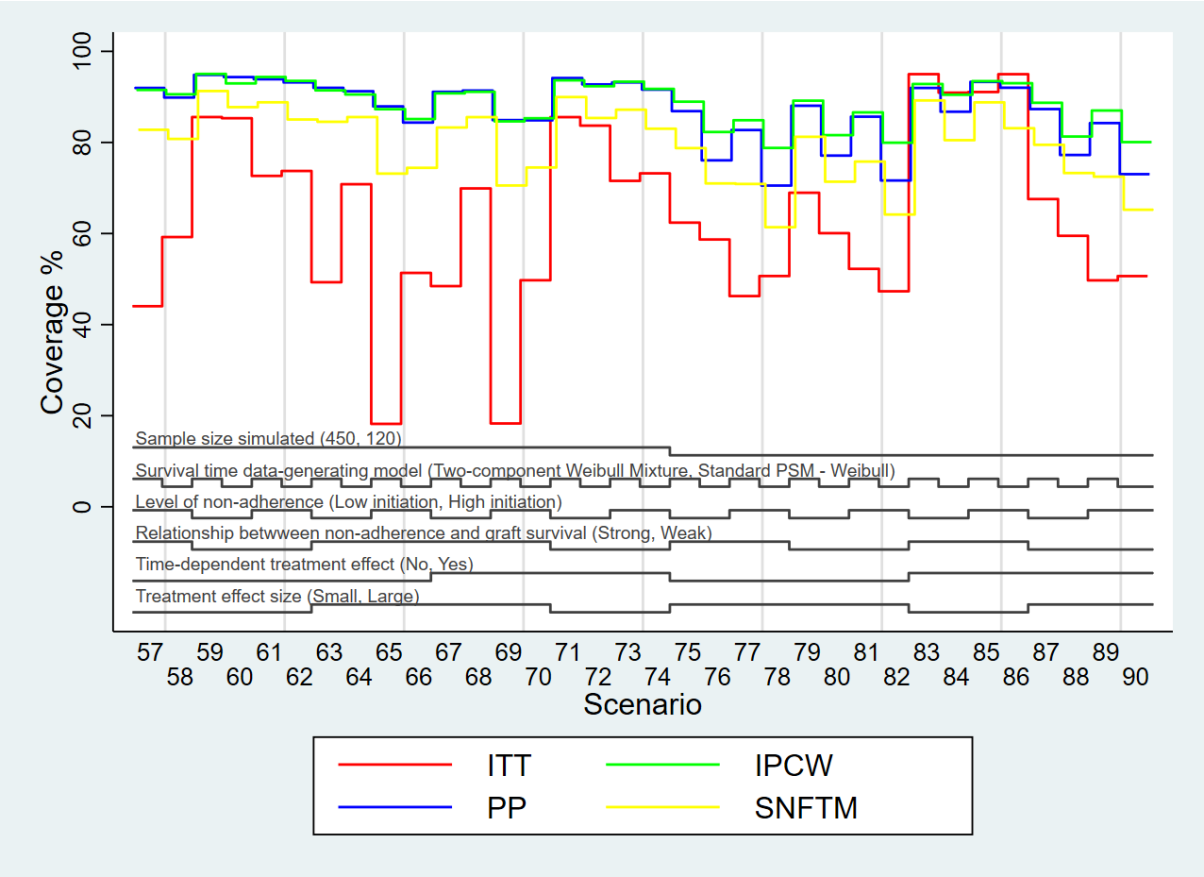
Figure 44: Empirical SE in the estimation of the difference in RMSTs across initiation non-adherence scenarios



5.2.5.5 *Coverage*

Figure 45 reports coverage percent generated by each method across the 34 initiation non-adherence scenarios. IPCW performed best across 21 out of 34 scenarios with a very high coverage percentage up to 95% (87.2 % on average). This was followed by PP and then SNFTM with ITT produced the worst coverage across scenarios. The level of non-adherence was a noticeable contribution to the magnitude of coverage for ITT in particular lower levels of initiation non-adherence leading to lower coverage percentage.

Figure 45: Coverage in the estimation of the difference in RMSTs across initiation non-adherence scenarios



5.2.6 Results for secondary estimands

The results in terms of estimates for the secondary estimands are reported in Appendix F (Tables 39-41). These include RMST in the control arm, RMST in the experimental arm and HRs alongside their standard errors across 90 scenarios. As discussed in Chapter 4 (Section 4.7), the results from secondary estimands were not used for assessing methods performance or comparing non-adherence adjustment methods. However, these estimates are reported alongside their associated standard errors to show that alternative non-adherence methods could be used to produce these estimates. Another reason for generating these estimates was to see if they highlighted any differences in interpretation of the results – e.g. if one method did well for the difference in RMST, but that was as a result of over-predicting RMST for control and experimental separately, then this would suggest that although the method seemed good based on the primary estimand, actually it was not good because it is predicting worse absolute survival. Looking at the results from the secondary estimands, there is no evidence anything like that was apparent from the simulations.

5.3 Discussion and conclusions

5.3.1 Summary

The findings of this simulation study demonstrated the importance of adjusting the treatment effect for patient non-adherence in the context of survival analysis and HTA. In this study, RCT datasets with non-adherence, prognostic characteristics and a time-to-event outcome were simulated to assess the performance of alternative methods. The study included all types of non-adherence (implementation, persistence and initiation) across a range of 90 realistic scenarios. Scenarios represent different types and levels of non-adherence, sample size, the pattern of hazards, treatment effect size, the relationship between treatment effect and non-adherence and the existence of any time-dependent treatment effect. The simulation assumed no informative censoring (other than that related to non-adherence) and no missing data. Non-adherence adjustment methods were assessed using five performance measures (bias, MSE, ModSE, EmpSE and coverage). Performance was presented using these measures as percentages of the truth for each scenario assessed. Both absolute and percentage bias were presented in the results section. Presenting the results in terms of absolute bias shows the scale in term values but the comparison of methods performance is better when using percent bias. The study assessed four non-adherence adjustment methods with three of these methods identified as appropriate for the HTA context.[61] The fours methods include two g-methods (SNFTM and IPCW) and two simple methods (ITT and PP), with the PP analysis strategy included as a benchmark comparator.

The main results show that g-methods performed better than the ITT method in terms of bias across most implementation and persistence non-adherence scenarios with marginal differences compared to the PP method. The simulation results show that IPCW produced better coverage in most scenarios across the three types of non-adherence (implementation, persistence and initiation). This was followed by PP and SNFTM with ITT producing the worst coverage performance in all but three of all scenarios assessed across all types of non-adherence. EmpSE favoured simple methods across scenarios; however, this should be interpreted with caution when considering performance results from other measures such as bias and coverage.

5.3.2 Key findings

G-methods were found to be the best performing method for adjusting treatment effectiveness in the presence of implementation non-adherence in terms of bias and ModSE. For bias, g-methods and PP had very similar patterns in the performance results and the differences between them were very marginal. The findings show slight changes with higher bias results associated with small sample sizes. The better performance of the PP method is more likely to be related to the strength of the relationship between prognostic characteristics, non-adherence and the survival outcome in the simulated datasets (see below for further discussion). Coverage performance was better for IPCW in 50% of implementation scenarios (mostly in scenarios with a large sample size) followed by PP. However, coverage did not favour SNFTM, although the method has performed the best in terms of bias and ModSE.

In scenarios with implementation non-adherence, bias ranged between 7-38% for the IPCW, PP, SNFTM (20% on average), but between 20-75% for ITT (40% on average) suggesting that all adjustment methods reduced percentage bias by about half compared to ITT, but some reasonable level of percentage bias did remain. SNFTM produced the lowest bias in most scenarios (21 out of 38) followed by IPCW (9 out of 38), although bias was higher in scenarios with large treatment effect size. Methods performance varied with variation in the values of other factors such as sample size and treatment effect size, with small sample size and large treatment effect size both leading to higher bias. The level of non-adherence and the relationship between non-adherence and graft survival outcomes were found to influence the performance of methods in terms of MSE, ModSE and EmpSE across implementation scenarios. For these performance measures, the sample size was the most noticeable influencing factor with a small sample size leading to a higher percentage of these measures (i.e. lower performance compared to scenarios with a large sample size). PP analysis produced a good performance in terms of MSE and coverage, although the method lagged behind g-methods in terms of bias and ModSE across implementation non-adherence scenarios. On the other hand, the findings show that ITT was the worst-performing method in terms of bias, ModSE and coverage across all implementation non-adherence scenarios. The findings show that ITT produced 47% higher bias (on average across all implementation non-adherence scenarios) compared with g-methods.

In scenarios with persistence non-adherence, the performance of methods was largely comparable to implementation, although the results show higher percent bias in some persistence scenarios. G-methods were the best performing methods in terms of bias, ModSE and coverage. Bias ranged between 9-39% for the IPCW, PP, SNFTM (26% on average), but between 16-72% for ITT (47% on

average) suggesting that all adjustment methods reduced percentage bias by about half compared to ITT, but some reasonable level of percentage bias did remain. ITT was always the worst-performing method when these measures were used to assess performance across 18 persistence non-adherence scenarios. The findings show that ITT led to 76.2% higher bias (on average across all persistence non-adherence scenarios) compared to g-methods. Although bias was the key performance measure in this simulation study, MSE is also important because it combines bias with variability, with the latter considered as particularly important to address uncertainty in an HTA context. MSE favoured PP; however, the performance results are very close to those produced by g-methods in most persistence scenarios with large sample size, with ITT produced a significantly higher MSE percentage across the board. In scenarios with small sample size, IPCW produced a higher MSE percentage with even higher percentages in scenarios with a large treatment effect size. EmpSE favoured simple methods in persistence scenarios; however, this should be interpreted with caution, as discussed in Section 5.3.1.

In scenarios with initiation non-adherence, SNFTM with g-estimation was the best performing method in terms of bias in most scenarios, with the best performance in terms of ModSE across all initiation non-adherence scenarios. The next best performing method found was PP followed by MSM with IPCW estimator. Bias ranged between -34 to 110% for the IPCW, PP, SNFTM (35% on average), but between -18 to 200% for ITT (75% on average) suggesting that all adjustment methods reduced percentage bias by about half compared to ITT, but some reasonable level of percentage bias did remain. PP was found to produce better performance in terms of MSE, although the results were very close to g-methods in scenarios with a large sample size. The study demonstrated that SNFTM with g-estimation outperformed the alternative methods in 19 out of 34 initiation non-adherence scenarios (mostly in scenarios with a large sample size). The treatment effect size was the next contributing factor in influencing bias produced by g-methods as a large treatment effect size led to larger bias, particularly in combination with a small sample size. However, SNFTM struggled in terms of coverage performance in which IPCW was the best-performing method in 21 out of 34 scenarios. The level of non-adherence was a key contributing factor for coverage performance with very high levels of non-adherence leading to poor coverage performance. PP performed well in terms of MSE and bias, although the method lagged behind SNFTM in terms of bias across initiation non-adherence scenarios. ITT produced higher bias, higher MSE and lower coverage across all initiation scenarios. The findings show that ITT led to 82% higher bias compared to SNFTM with g-estimation.

Another important research finding relates to PP analysis as the method performed well in many scenarios producing good performance closer to g-methods in terms of MSE. The method also did very well in terms of coverage and was found to be the best-performing method in 50% of implementation

scenarios (similar to IPCW). The latter finding was mostly associated with scenarios with a small sample size where IPCW struggled in coverage performance. Based on further investigations carried out in this study, this finding seems to relate to the impact of the time-dependent confounder on graft survival times. In the simulated scenarios where patients with worse prognostic characteristics (i.e. high BMI and aged below 24 years) are more likely to non-adhere to the dosing regimen, but over time more of these patients are likely to experience the event. Therefore, fewer patients are left to non-adhere, and ultimately in the end I get a mixture of poor and good prognosis patients non-adhering, and the relationship between non-adherence and prognostic characteristics is weakened over time. This issue may result in the IPCW having weights that are all close to 1, so it does not do much on top of the simple censoring (PP) analysis in these particular scenarios. This means that because there is not a strong relationship between prognosis and adherence, the PP approach is not very biased. This issue may require more complex DGMs and this could be an important area for future research.

A potential reason might be related to the DGM as only three time-points were simulated, and so maybe it was difficult for consistent relationships to show up in each simulated dataset, given variability. The impact of the number of data points is another important consideration for the design of future simulation studies. The abovementioned reasons might potentially explain why similar results between PP and the g-methods were generated in many scenarios, which perhaps would not have been expected based upon the advantages of the g-methods and based on findings in previous simulation studies. In circumstances such as those simulated in this study, when the pool of "at-risk of non-adhering" patients changes over time, the confounding might even out (at first worse prognostic patients non-adhering, then over time better prognostic patients non-adhering) might lead to a case where the bias in a simple censoring approach is low. In this case, methods applying a simple censoring mechanism such as PP might work. Future research is required to further investigate this issue.

Although complex simulation methods were used in this simulation study to generate RCT datasets for testing the alternative methods, the DGMs in some scenarios may not sufficient in generating substantial time-dependent confounding to test the limitations of the PP method in this context. This means the relationship between prognosis and non-adherence was not strong enough for the g-methods to generate superior performance compared to the PP method. The g-methods did not improve much compared to PP and this could be because of the change in the patient mix over time, perhaps because there were only three-time intervals, or perhaps a mixture of these reasons.

However, the simulation study was complex enough to assess the performance of g-methods compared to ITT analysis across all types of non-adherence. The study provides clear evidence in

favour of g-methods against ITT analysis. The latter analysis is not possible to perform using the PP simple method. The implications of this finding are discussed further in the methodological framework developed in this thesis (Chapter 7). It should be noted that, unlike g-methods and ITT, the suitability of the estimand used by PP is not theoretically appropriate for HTA because the estimand is not marginalized to the entire study population. In other words, it is not comparing like with like in terms of estimates generated by the methods.

The overall pattern of performance results across all the scenarios is that ITT is the worst performing method, with a quite high percentage bias; and IPCW, SNFTM, PP all reduce this bias by about half (or more) and produce very similar results to one another.

5.3.3 Comparison with other studies

Existing evidence from relevant simulation studies showed comparable results, although the design and non-adherence metrics used in these simulation studies were different. For instance, Cain and Cole (2009)[108] published the first simulation evidence in the methodological literature that compared IPCW and ITT analysis in adjusting for non-adherence in the context of survival analysis. The study aimed to correct for the effect of time-varying non-adherence on estimating the effect of a hypothetical highly active antiretroviral therapy on time to the incidence of AIDS or death. The study assessed methods performance in three scenarios defined by different levels of non-adherence (0% assuming perfect adherence, 20% and 40% non-adherence) by performing 2000 iterations for each scenario.[108] The sample size used was 1000 with a standard Weibull distribution used for generating survival times in the simulated datasets. Bias and MSE were reported as performance measures with findings showed that the g-method (IPCW) was the best-performing method in terms of unbiasedness. The paper reported that bias and imprecision increase as the level of non-adherence increased. This is similar to findings from my current simulation study, although the magnitude of bias differs, potentially due to the different estimands, DGMs, non-adherence metrics and other parameter values. In their study, HR was used as a primary estimand whereas; I used difference in RMSTs for assessing methods performance. For example, their study reported bias and RMSE as absolute numbers; whereas, I used bias percent and MSE as percentage of the truth.

Other existing simulation evidence includes a study reported by Zhang et al. in which they compared the same methods (IPCW vs ITT).[40] The study aimed to assess methods performance on adjusting for treatment discontinuation (i.e. persistence type of non-adherence). The study assessed hypothetical treatments comparing two anticoagulants with a time-to-event outcome. The study design assumed

a large dataset (n=2000) and simulated 2000 iterations with time to discontinuation as a measure of time-varying non-adherence. The simulated datasets included baseline covariates, time-dependent confounders, censoring of event time, and time-varying non-adherence, with simple Bernoulli distributions, used to generate randomisation assignment. The paper reported an average non-administrative censoring of 32% over the study follow-up. The key findings are similar to those reported by Cain and Cloe with IPCW generated the best performance in terms of bias and coverage. IPCW produced an average estimate of -0.492 in contrast with an ITT estimate of -0.334 compared with the true parameter value of -0.500 in terms of the log HR. The reported coverage performance was 96% for IPCW and 14.1% for ITT analysis.[40] Although, a different study design was used (including the estimand), the findings of this study are similar to my simulation study with IPCW outperforms ITT analysis in adjusting treatment effect for persistence non-adherence. However, the magnitude of coverage differs and this is likely to be due to differences in DGMs and sample size. My study findings show that a smaller sample size leads to poor performance, but the advantage is that it is more generalisable than findings arbitrarily chosen sample size as in the two existing studies discussed in this section. It was not possible to compare the findings related to PP method because the two relevant existing studies (discussed above) did not include PP analysis in the alternative methods assessed.

The g-methods evaluated here have also been tested in several simulation studies for adjusting effectiveness in the presence of treatment switching.[101, 128, 129] There are some arguments in the published literature considering treatment switching as a type of persistence non-adherence. However, the counter-argument is that switching prescribed medication is a different type of change in therapy, as it must be initiated by the prescriber (e.g. the medical practitioner). Whereas persistence non-adherence, which happens before the end of prescribing by the patient's own behaviour, would be considered as non-adherence. Latimer and colleagues published several papers providing evidence on the performance of these methods in the context of treatment switching and survival analysis.

The findings from these studies show that g-methods were superior to simple methods in terms of performance using a range of performance measures including bias and coverage. For instance, Latimer et al.[101] published findings from a simulation study aimed to assess the performance of IPCW, RPSFTM and ITT (among other methods). The authors simulated RCT datasets in the presence of treatment switching (from the control onto the experimental group) with a time-to-vent outcome and time-dependent confounding. The primary estimand was RMST in the control group that would have been observed in absence of treatment switching. The findings from the simulations demonstrated

that both g-methods (IPCW and RPSFTM) performed well in terms of handling the effects of time-dependent confounding. However, IPCW struggled in situations with high percentages of switching and a modest sample size.

5.3.4 Strengths and limitations

This simulation study is superior to the comparable existing studies in many aspects. The study design was based on international guidelines for planning simulation studies and followed a pre-specified study protocol.[1, 113, 115] These include using a new taxonomy of medication non-adherence using the initiation, implementation and persistence framework (i.e. ABC taxonomy). The study included the simple censoring PP method, which was not considered in the existing comparable studies, and this has produced interesting findings as discussed further down this section. Moreover, a range of five performance measures and four estimands were included in this study, which is more than used in many previous studies. The study also used a nested loop plot to assess performance patterns using different performance measures and study factors across all scenarios assessed.

Robust data-generating mechanisms were used to simulate biologically plausible survival data with baseline covariates, time-dependent confounders, time-varying non-adherence and a robust randomisation assignment procedure. The relationships between prognostic variables, non-adherence and the survival outcome followed a DAG which was developed in this study based on discussions with clinicians, external advisors and evidence from the literature. The user-written baseline hazard functions used for simulating survival data in a delayed entry model allowed for simulating a range of RCT datasets (with reasonable levels of complexity) for testing the alternative non-adherence adjustment methods.

Other strengths associated with this simulation study include the incorporation of seven important factors in a partly factorial simulation study design that has contributed to widening the range of scenarios assessed. These factors covered sample size, type of non-adherence, level of non-adherence, baseline hazard function, the relationship between patient non-adherence level and survival outcome, time-dependent treatment effect and treatment effect size. Using two types of baseline hazard functions (standard FPM and two-component mixture FPM) contributes to improving the generalisability of findings to other disease areas beyond kidney transplantation. Using five key performance measures within the best available tools (e.g. the *simsum* Stata Package) with the nested loop plots have facilitated the presentation and interpretation of the results. The simulation reporting adheres to published international guidelines for reporting simulation studies and medication

adherence research.[118, 126] Presenting the results in terms of absolute bias shows the scale in term values but the comparison of methods performance is better when using percent bias.

There are limitations associated with this study, and these are discussed here. It is difficult to accurately simulate the complexities of real-world disease pathways and biomarkers; therefore, simplification is inevitable, and this simulation study is not an exception. Key simplifications include the number of baseline and time-dependent confounders simulated in the datasets. In this study, only one baseline covariate (age) and one time-dependent confounder (BMI) were included as a simplification to simulate meaningful relationships between covariates, non-adherence and graft loss. However, in real RCT datasets, there are likely to be more confounders. The interaction between multiple confounders, non-adherence and survival outcomes might lead to different outcomes to those generated in simulated datasets and this might influence performance. In the simulated datasets, I assumed no “non-administrative censoring” and no missing data; therefore, the findings do not take these inter-current events into account. However, these problems were properly investigated in the methodological literature; and it could be argued that accounting for non-administrative censoring and missing data in real practice should be considered alongside the non-adherence adjustment methods assessed in this simulation study.

The Rank-Preserving Structural Failure Time Model (RPSFTM) with g-estimation was identified as appropriate for the HTA context. However, the method can only be applied for RCT designs with a placebo control arm (see Chapter 2 for more details about how the method works). I have discussed this with Professor Ian White and considered designing a special set of scenarios for testing this method as all the 90 scenarios simulated datasets compared two active treatments with non-adherence applied on each arm. Based on further considerations, I decided to exclude this method from assessment in the current simulation study.

Unsuccessful data simulation and non-convergence problems were captured by the simulation program with the latter possibly being related to the degrees of freedom specified in the flexible parametric survival model (FPM), although this has been rare in this particular simulation study. I considered using a different approach by allowing the FPM to choose a smaller degree of freedom if the originally specified one results in non-convergence. However, the concern is that the modified model is likely to produce different results influenced by the model specification rather than non-adherence which is not desirable. Therefore, I decided to record results from the failed simulations as missing and move to the next iteration. The argument is that if the "preferred" FPM model does not converge, the results are less comparable to other simulations where the preferred model does

converge. Nevertheless, the degree of freedom used in the FPM specification was chosen based on assessing a range of values using AIC and BIC criteria for model fitting. In trials with very small sample sizes, the application of g-methods to adjust for non-adherence might lead to non-convergence problems and therefore the analyst needs to consider this as a potential issue when designing their analysis plan. The non-convergence issue was also associated with the data-generation methods including the randomisation and survsim commands that have resulted in several unsuccessful simulations and this should be considered as a limitation. However, enough successful simulations were achieved to assess the alternative non-adherence adjustment method for each scenario assessed in this simulation study.

Another limitation relates to the coverage performance results. Although a robust SE around the difference in RMST was generated, the coverage data seems less reliable. As an alternative approach, bootstrapping might be needed on top of robust SE to improve coverage performance estimates. While this approach has not been specified in the study protocol; and therefore not applied, it might be worth considering it in the design of future simulations if better coverage performance is desirable. For ITT and PP no adjustments are made, and the SEs and CIs generated are robust. For IPCW weights are used to get a pseudo-population, but also use robust SE which account for these weights. For SNFTM robust SEs deal with the clustering of the data but do not recognise that the FPM is applied to an adjusted dataset, therefore, the uncertainty associated with the adjustment is not recognised. For SNFTM, adequate confidence intervals (CIs) and coverage could be obtained bootstrapping the entire adjustment analysis, which in a simulation study is very computationally intensive. In this particular simulation, I would need to sample the simulated dataset 1000 times (for example), apply the SNFTM and FPM in each sample, record the RMST in each sample, then take CIs across the 1000 samples to get the CI for 1 simulation. Then do that 1900 times for 1 scenario and this will be equivalent to 1900*1000 SNFTM analyses for 1 scenario. This computational burden was the main reason for not applying bootstrapping in this simulation study.

It is important to assess the sensitivity of each method performance to their departure from key assumptions. The “no-unmeasured confounding” is a key assumption used by g-methods (SNFTM and IPCW). As a limitation, the assumption of non-unmeasured confounding was not assessed in this simulation study. This is because only one time-dependent confounder was simulated, thereby making it impossible to run analyses with fewer covariates. However, the analyst should consider this to ensure that the assumption of no-unmeasured confounding is met when applying g-methods in their studies in practice. PP analysis could be considered as a baseline for this. The method does not correct for differences between non-adherers and adherers, and so, represents IPCW had IPCW not included

any covariates or if the relationship between prognosis and non-adherence is not being very strong. Future research needs to consider including more than one baseline and time-dependent confounders to allow for testing the assumption of non-unmeasured confounding.

Other limitations include the values of parameters and coefficients used for simulating the datasets. While these were based on the literature with values calibrated to achieve KM survival curves that mimic data from RCT evaluated immunosuppression in kidney transplantation, the performance of methods in scenarios with different parameter values is unknown. I tried to vary the values of certain parameters and factors including sample size, treatment effect size, the relationship between non-adherence and graft survival to improve generalisability. However, for practical reasons covering all possible options was not possible as it requires a fully factorial study design and more resources which are beyond the scope of this thesis. Nevertheless, the findings of this simulation study provide new evidence for choosing the appropriate method across a range of realistic scenarios. This also opens a range of possibilities for future research (simulation studies) for testing non-adherence adjustment methods in other scenarios with different DGMs, longer follow-up and different parameters values. Special attention should be given to the robustness of methods for generating substantial time-dependent confounding if the simple censoring method (PP) is considered among the alternative methods to be assessed. Finally, assuming binary adherence used by the adjustment methods is another limitation which is more relevant to adjusting for implementation non-adherence.

5.3.5 Conclusions

In conclusion, the findings of the simulation study demonstrated that g-methods (MSM with IPCW and SNFTM with g-estimation) are the best-performing methods in terms of unbiasedness and ModSE for adjusting estimates of treatment effect in the presence of implementation and persistence non-adherence in RCTs with time-to-event outcomes. For initiation non-adherence, SNFTM is the best-performing method in terms of unbiasedness and ModSE. The findings demonstrated that g-methods produce higher coverage compared to ITT in most scenarios across all types of non-adherence. The study provides new evidence comparing four statistical methods covering two g-methods (IPCW and SNFTM) and two simple methods (ITT and PP) across a range of implementation, persistence and initiation non-adherence scenarios. The simulation study provided evidence on nuances on relationships between prognostic variables, patient mix, and adherence to medication over time, and the ability of g-methods to model these relationships. The study findings also demonstrated that the PP analysis method performed well in many scenarios with good performance in terms of MSE and coverage, although the PP estimand is different from the ITT and g-methods estimands.

The findings favour SNFTM in most scenarios across all types of non-adherence, but it should be noted that the method might produce high bias in scenarios with high levels of non-adherence and/or large treatment effect size, although this is more likely to be lower than the potential bias produced by the alternative methods. The findings demonstrated that IPCW is the best performing method in terms of producing higher coverage in most scenarios across all types of non-adherence.

Chapter 6: A case study on the application of non-adherence adjustment methods in kidney transplantation

6.1 Introduction

This chapter presents the case study to show how the adjustment methods can be applied using real RCT data combined with real-world non-adherence levels to generate adherence-adjusted cost-effectiveness estimates of immunosuppressive regimens in kidney transplantation. Section 6.2 presents the aim of the case study. Sections 6.3 provides an overview of the methods used in the case study including the overall study design, interventions compared, study population, directed-acyclic graph (DAG) and analytical steps. Section 6.4 describes the SYMPHONY trial dataset used in analysis including patients' characteristics, baseline and time-dependent confounding and adherence to medications in the trial. Section 6.5 presents a review of implementation non-adherence to immunosuppressive therapy in the real world based on existing evidence from the literature. Section 6.6 presents the analysis of the SYMPHONY dataset using the standard ITT (unadjusted analysis) and the analyses using g-methods to adjust the treatment effectiveness for real-world implementation non-adherence levels. Section 6.7 describes the economic model and cost-effectiveness analysis. The section describes the economic analysis undertaken for estimating adherence-adjusted cost-effectiveness of immunosuppressants using an adapted economic model. Section 6.8 presents the results of the case study. Section 6.9 discusses the findings and provides the conclusions of the case study.

The case study built on the work undertaken in Stage 1 (Systematic review)[61] and Stage 2 (Simulation study) of this PhD research project, as applied to maintenance immunosuppressants for kidney transplantation in adults. This disease area was chosen for three main reasons: (1) significant implications of non-adherence for patients (i.e. graft loss, return to dialysis and potentially death); (2) significant cost implications to the NHS; and (3) availability of both RCT and real-world data with adherence to medications metrics.

The original idea behind this case study was to investigate the implications of differential adherence levels to a new once-daily modified-release tacrolimus formulation compared to a twice-daily immediate-release tacrolimus formulation among kidney transplant recipients.[130] The tacrolimus once-daily formulation was rejected in the recent update of NICE Technology Appraisal (TA481).[120] The original plan was to obtain individual-patient level data (IPD) from the UK Renal Registry (UKRR) in terms of drug concentration levels to estimate real-world non-adherence for the two tacrolimus

formulations. And then use the adherence estimates from the UKRR database within an RCT dataset from the OSAKA study (a trial that compared once-daily versus twice-daily tacrolimus formations).[131] However, my requests for both the UKRR and the OSAKA datasets have been rejected by the data owners; and therefore, a change to the original plan was made. Based on consultations with two clinicians (WM and JF), the plan was changed to use published estimates of real-world non-adherence to the twice-daily tacrolimus regimen compared to the standard and low-dose cyclosporine regimens within the SYMPHONY trial dataset (see Section 6.4 for more detail about this trial).[116]

Given the potential weakness of the UKRR dataset, the new plan may not be inferior to the original plan. Among others, the UKRR data limitations are: (a) centres started submitting data on drug concentration levels as part of their 2016 data (from January 2017) but it is likely to be a number of years before all centres do; (b) the data is collected as part of the PatientView extract (an online portal that takes data from renal unit's records at least once a day and links to useful information about the patient's kidney condition and its treatment); and (c) PatientView data is only available on people who have signed up to PatientView (i.e. around half of the transplant patients) and these patients may not be representative of all kidney transplant patients as they are signing up to check their own blood results online.

The selection of the SYMPHONY trial was also informed by advice from the clinicians. The regimens compared in the case study (and SYMPHONY trial) represent the current standard maintenance immunosuppressive regimens in the NHS and other health care systems around the world. SYMPHONY is the largest RCT that evaluated calcineurin inhibitors (CNIs; cyclosporine and tacrolimus). A third trial dataset (the 3C study)[132] was also pursued, but this was abandoned due to contractual problems encountered by the investigators due to their dataset also including data from NHS Digital (and for which they did not have the rights to share with a third party).

Non-adherence in patients after having kidney transplantation can be so serious that the transplanted kidney may be rejected or lost, the patient may need to return to intensive treatment (kidney dialysis) or they may even die. Well-designed and conducted RCTs are used as the best way of assessing treatment effects.[133] However, there is a debate around whether evidence from trials can reflect normal healthcare practice. A major factor that could impact on treatment effect is patient non-adherence to the prescribed dosing regimen. The key point is that adherence in RCTs may not reflect normal adherence in the real world, and therefore, estimates of effectiveness from trials may not be externally valid. This is particularly important when the medications are prescribed for a longer period (i.e. lifetime) as in the case of maintenance immunosuppressive therapy after kidney transplantation.

6.2 Aim of the case study

The case study aimed to assess the cost-effectiveness of maintenance immunosuppressive therapy for kidney transplantation in adults, adjusting for patient non-adherence to the prescribed dosing regimens. The analysis used adherence data from the real world to adjust effectiveness data from an RCT to produce estimates of real-world effectiveness, which were then applied within an existing decision-analytic model. The purpose of the case study is to show how the best-performing methods from the simulation study could be applied using real data and to identify any issues associated with applying the methods to unsimulated data.

Given the nature of the patient population and intervention, only implementation non-adherence is relevant. Furthermore, based on the results of the simulation study this meant that only g-methods were considered relevant to the HTA context of the study. ITT values were estimated as they represent the current approach adopted by NICE. IPCW was chosen to re-estimate effectiveness in the base-case analysis with SNFTM applied in a secondary analysis. There was not much to choose between the two methods based on performance in the simulations so I did the adjusted analysis using them both.

6.3 Methods overview

To apply non-adherence adjustment methods, the required RCT data should include demographic information, adherence metrics, and relevant prognostic confounders. The SYMPHONY trial is a study that meets these criteria. The details of this trial are published elsewhere,[116] but these are briefly described in Section 6.4.

6.3.1 Case study design

The overall design of the case study is a cost-utility analysis where the results are expressed in terms of total discounted costs, total discounted QALYs, incremental costs and QALYs, and net health benefits (NHBs). Actual adherence to maintenance immunosuppressive therapy in the real world was obtained from a review of the literature (see Section 6.5). This information was used to re-estimate the clinical effectiveness based on the assumption that adherence in the trial population was the same as observed in the real world. The clinical effectiveness estimates were then used within an adapted decision-analytic model to estimate the adherence-adjusted cost-effectiveness of these drugs over a patient lifetime horizon. I obtained NHS Research Ethics Approval for this case study (REC reference: 19/LO/0847).

An existing health-economic model was adapted to estimate adherence-adjusted cost-effectiveness comparing standard-dose cyclosporine, low-dose cyclosporine and low-dose tacrolimus (referred to as 'tacrolimus') as maintenance immunosuppressants for adult kidney transplant patients in the UK. The economic model was originally built by the Peninsula Technology Assessment Group (PenTAG) which underpinned the recent update of NICE Technology Appraisal guidance (TA481).[120] The model has been adapted in this study to incorporate treatment effectiveness estimates produced from my analysis of the SYMPHONY data. The economic model is described in greater detail in Section 6.7.

6.3.2 Interventions

The interventions assessed in this case study are the following maintenance immunosuppressive regimens as used in the SYMPHONY study and the adapted economic model:

- Group A: Standard-dose cyclosporine (Neoral or Sandimmune, Novartis) plus mycophenolate mofetil (CellCept, Roche) and corticosteroids (CsA+MMF+ST).
- Group B: Basiliximab induction and low-dose cyclosporine plus mycophenolate mofetil and corticosteroids (Bas+CsA+MMF+ST).
- Group C: Basiliximab induction and low-dose immediate-release tacrolimus (Prograf, Astellas Pharma) plus mycophenolate mofetil and corticosteroids (Bas+Tac+MMF+ST).

All three alternative treatment regimens included the maintenance immunosuppressive drugs, cyclosporine (standard-dose or low-dose) or tacrolimus, in combination with mycophenolate mofetil and corticosteroids. An induction agent, basiliximab, was also used for groups B and C. In the adapted economic model, basiliximab was used as an induction agent for the low-dose cyclosporine and tacrolimus regimens; whereas, in the SYMPHONY trial, daclizumab was used as an induction agent. Although these are two different drugs, a recent meta-analysis of six RCTs concluded that "the safety and efficacy of daclizumab and basiliximab are similar in kidney transplant recipients".[134] The study also indicated that basiliximab is more cost-effective than daclizumab.[134] Therefore, basiliximab was considered as the induction agent of interest in the analysis as used in the original economic model. The new induction agent was also accepted by NICE in their updated Technology Appraisal guidance (TA481).

Although the treatment regimens include multiple drugs, non-adherence data were only collected for the maintenance immunosuppressive agents within each regimen; and therefore, the analysis did not consider the impact of multiple adherence to all medications within each regimen. Measuring multiple medication adherence (MMA) has been identified as an issue in a recent report of the ISPOR

(International Society for Pharmacoeconomics and Outcomes Research) Medication Adherence and Persistence Special Interest Group.[135] Therefore, adjusting for MMA in cost-effectiveness estimates could be an interesting area for future research. The maintenance immunosuppressive therapies of interest (tacrolimus and cyclosporine) are described in more detail in the following subsections:

6.3.2.1 Tacrolimus

Immediate-release tacrolimus is a maintenance immunosuppressant used for the prophylaxis of graft rejection in adult transplant patients.[120] It is administered orally as capsules, twice a day, with a recommended initial dose of 0.1 to 0.3 mg/kg/day which is usually reduced over time to maintain the target levels.[116, 120] The common brand names are Prograf, Adoport, Capexion, Perxis, Tacni and Vivadex.

6.3.2.2 Cyclosporine

Cyclosporine is an immunosuppressant that may be used alone or in combination with other medication, as maintenance therapy after kidney transplantation in adults.[120] The recommended dose is 2-6 mg/kg/day administered as twice-daily capsules and reduced gradually to maintenance with a lower dose when used concomitantly with other immunosuppressive therapy (e.g. corticosteroids).[120]

6.3.3 Study population

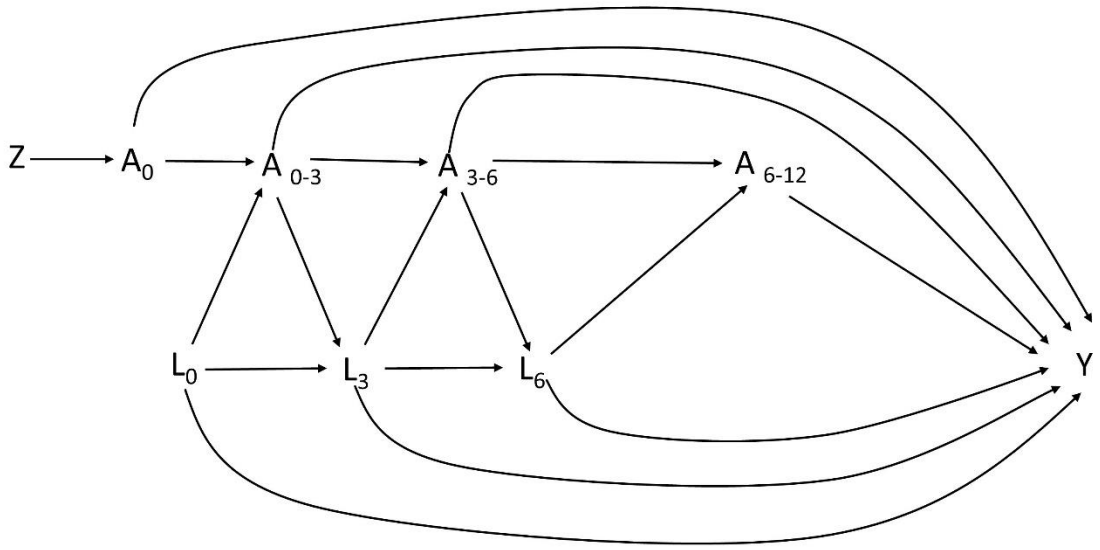
The study population included adult patients aged 18 and over receiving maintenance immunosuppressive therapy (tacrolimus or cyclosporine) after incident kidney transplantation. The study population was selected based on the decision problem assessed in the latest NICE Technology Appraisal model (TA481). No subgroup analysis was undertaken in this case study.

6.3.4 Directed-acyclic graph (DAG)

A DAG for the case study was drawn based on evidence from the literature and discussion with clinicians (Figure 46). In the DAG, time flows from left to right and therefore the sequence of variable measurements are based on this convention.[13] The DAG illustrates the randomisation variable (Z) to the left-hand side assigning patients to each of the three immunosuppressive regimens assuming perfect initiation of the treatments. The implementation non-adherence only occurs after that for a follow-up of 365 days post-transplantation. The confounders Ls affect subsequent adherence (As) and the time to graft loss outcome (Y); and the Ls themselves are affected by previous As, representing time-dependent confounding. Baseline covariates and the values of time-dependent confounders at

baseline (both denoted as L_0) are common causes which means they influence both graft survival outcome and non-adherence between baseline and the first follow-up time point A_1, as illustrated in the DAG. The justification for the identification of baseline and time-dependent confounders was discussed earlier in Chapter 4, Section 4.4. Figure 46 shows the relationships between all the variables used in the analysis. No potential collider variables were identified (for more detail see Chapter 1, Section 1.3.6).

Figure 46: Directed-acyclic graph (DAG) for SYMPHONY dataset



Z= is randomisation variable; L_0= is a vector of baseline covariates that includes age and gender and also include values of time-varying covariates measured at baseline; L_3, L_6 = updated time-varying covariate (BMI and acute rejection) at 3 and 6 months, respectively; A_0, A_{0-3}, A_{3-6}, A_{6-12}= time-varying non-adherence at baseline, between baseline and 3 months, 3-6 months and 6-12 months, respectively.

6.3.5 Analytical steps

The case study involved the following analytical steps:

1) Identification of estimates of medication adherence levels in the real world using data from the literature. This was focused on implementation non-adherence measured using CV% relating to blood concentration levels for immunosuppressants (i.e. tacrolimus and cyclosporine).
2) The predicted non-adherence levels were obtained by adjusting the CV% cut-off used for determining the presence of non-adherence, such that the overall adherence level in the trial matched real-world estimates.

3) Re-estimation of the relative clinical effectiveness of immunosuppressive regimens (tacrolimus versus low-dose cyclosporine versus standard-dose cyclosporine) using data from the SYMPHONY trial, adjusted according to the predicted real-world adherence levels (this is described in more detail in the next paragraph). The adjusted analysis was performed using g-methods (MSM with IPCW applied in the base-case analysis and SNFTM with g-estimation used as a secondary analysis). The clinical effectiveness estimates included graft survivor functions censored for death with a functioning graft (DWFG). A standard ITT unadjusted analysis was also performed and the results are presented alongside estimates from the adjusted analysis.
4) Adaptation of the health economic model for estimating long-term adherence-adjusted cost-effectiveness of immunosuppressive therapy for adult kidney transplant recipients in the UK. The economic model was originally developed for the latest NICE Technology Appraisal (TA481). Unadjusted cost-effectiveness results using effectiveness estimates from the standard ITT analysis was also performed and presented alongside the adjusted analysis for comparisons.

The predicted non-adherence levels were obtained using adherence levels observed in the SYMPHONY trial in terms of CV% and a new cut-off point such that the overall proportion of non-adhered patients matches real-world estimates. People with higher CV% based on the observed trial records of drug concentration levels have a higher probability of being non-adhered in the predicted dataset. The analysis for re-estimating treatment effectiveness conducted in this case study was adjusted for patient characteristics and prognostic factors that may cause confounding bias. These were variables that had a common cause on the exposures and the outcome of interest. Baseline and time-dependent confounders were identified using the process described in Chapter 4 (Section 4.4.2).

In brief, the process was based on the DAG which was drawn based on evidence from the literature and discussions with clinicians (see Figure 46). Simple regression was used to check the relationships between confounders and adherence to medications in the dataset. These will be subject to potential bias themselves, hence they are just a check and the main choices were made based on the DAG. Statistical significance was not used to determine the inclusion of the potential confounders as the assessment was mainly based on the values of the coefficients from the regression output required to show some degree of associations. This confirmatory step was used to satisfy the conditions of baseline or time-dependent confounding as described in Chapter 1 (Section 1.3.4).

6.4 The SYMPHONY trial data

The SYMPHONY study was a prospective, randomised, open-label, multicentre study of 1,645 kidney-transplant recipients with a 12-month follow-up. The study evaluated the clinical effectiveness of four immunosuppressive regimens involving standard-dose cyclosporine, low-dose cyclosporine, tacrolimus and low-dose sirolimus. The analysis undertaken in this case study included three arms with the sirolimus regimen excluded. The main reason for excluding the sirolimus regimen was that the treatment is rarely used in the NHS as a maintenance immunosuppressive therapy in kidney transplantation. In addition, unlike tacrolimus and cyclosporine, evidence suggests that the intra-patient variability (IPV) obtained from drug concentration levels for sirolimus is unreliable and may not be appropriate for use as a proxy measure of adherence. Moreover, sirolimus has a much longer half-life, so level variation will inevitably be less even in non-adherent patients.

In this context, IPV is defined as the amount of fluctuation in drug blood concentration within an individual patient over a particular period during which the dose of the drug has not changed.[136] Patients with higher IPV indicates a higher level of non-adherence. As the sirolimus arm was excluded, data from 1,190 patients was included in the analysis (standard-dose cyclosporine [n= 390]; low-dose cyclosporine [n= 399]; tacrolimus [n= 401]). The SYMPHONY trial was undertaken across 83 centres in 15 countries (Australia, Austria, Belgium, Brazil, Canada, Czech Republic, Germany, Greece, Israel, Mexico, Poland, Spain, Sweden, Turkey, and the United Kingdom).[116]

The SYMPHONY dataset was provided by F. Hofmann-La Roche Ltd. based on a data-sharing agreement (DSA) signed between Roche and the University of Sheffield. Written permissions for using the SYMPHONY dataset in this case study were also obtained from Professor Philip Halloran, University of Alberta, Canada (Co-sponsor of the SYMPHONY trial) and Ulf Malmqvist, Region Skane, Sweden (Sponsor representative of the SYMPHONY trial).

The SYMPHONY dataset was provided in an anonymised individual patient-level format including the raw data. The key study documentation was also provided in a redacted format including the study protocol, data definition document, case report forms (CRFs), statistical analysis plan (SAP), clinical study report (CSR), and the anonymisation orientation document. The individual patient-level datasets were checked for consistency against the CSR and the published report before the case study analysis was conducted.[116] Minor data queries were checked with the SYMPHONY trial statistician and resolved based on their advice. The dataset matches the CSR and published reports apart from minor deviations resulting from the data anonymisation process performed by the Roche Global Patient-level Data Sharing team. The trial dataset is described in the following sub-sections.

6.4.1 Baseline characteristics

Table 16 provides information on the key baseline characteristics from the analysis of the SYMPHONY dataset. The values for each variable were checked against the published data and CSR. These data checks were done to ensure that the adherence-adjusted analysis was performed on a dataset that is consistent with the published trial data (see Table 16, "Notes" column).

Table 16: Key baseline characteristics of the study population

	Group A: Standard-dose cyclosporine (n=390)	**Group B: Low-dose cyclosporine (n=399)**	**Group C: Tacrolimus (n=401)**	**Notes**
Age (Years) Mean (SD)	46.7(13.2)	47.9(12.8)	46.8(13.7)	Slightly different from Ekberg et al.[116] due to anonymization [i.e. patients <21 or >89 years old were aggregated]
Male (%)	62.3	66.4	65.8	Consistent with Ekberg et al.
Race (%)	-	-	-	Variable dropped as part of the anonymisation
Type of donor (%)				
Deceased	65.6	64.3	63.0	Consistent with Ekberg et al.
Living related	28.5	26.9	31.8	Consistent with Ekberg et al.
Living unrelated	5.9	8.8	5.3	Consistent with Ekberg et al.
Donor age (years) Mean (SD)	47.8(13.2)	48.5(13.0)	48.4(12.6)	Slightly different due to anonymisation
Donor with expanded criteria (%)	16.9	18.0	17.7	Consistent with Ekberg et al

Note: Group D (low-dose sirolimus) was dropped from the original dataset as this regimen is not included in the case study

6.4.2 Baseline and time-dependent confounding

The identified baseline confounders were age and gender. Age (in years) was recorded as a continuous variable and gender as a dichotomous variable in the trial dataset. Time-dependent confounders were BMI and acute rejection, with BMI included in the analysis as a continuous variable and acute rejection as a binary variable. BMI was calculated from the height and weight records in the SYMPHONY dataset.

6.4.3 Adherence to medications in the trial

6.4.3.1 Description of adherence to medication data

Adherence to medications in the SYMPHONY trial was based on the drug concentration level data for tacrolimus and cyclosporine. The drug concentration trough levels were measured at 10 planned visits (baseline; 1, 2, 4, 6 and 8 weeks; and 3, 6, 9 and 12 months post-transplantation). In addition, extra data on drug concentration levels were measured at additional visits to the clinics over the 12-months follow-up. The analysis of adherence data in the case study was based on a combined dataset that included measurements at protocol visits plus records from additional visits.

A total of 18,873 trough level records were available across the three arms of the SYMPHONY trial. Trough levels data were used for calculating the coefficient of variation (CV%) as a proxy measure of implementation non-adherence. Figure 47-50 show the mean trough levels over 12 months post-transplantation for cyclosporine and tacrolimus, respectively.

Figure 47: Mean cyclosporine trough levels over the study follow-up (Protocol visits data)

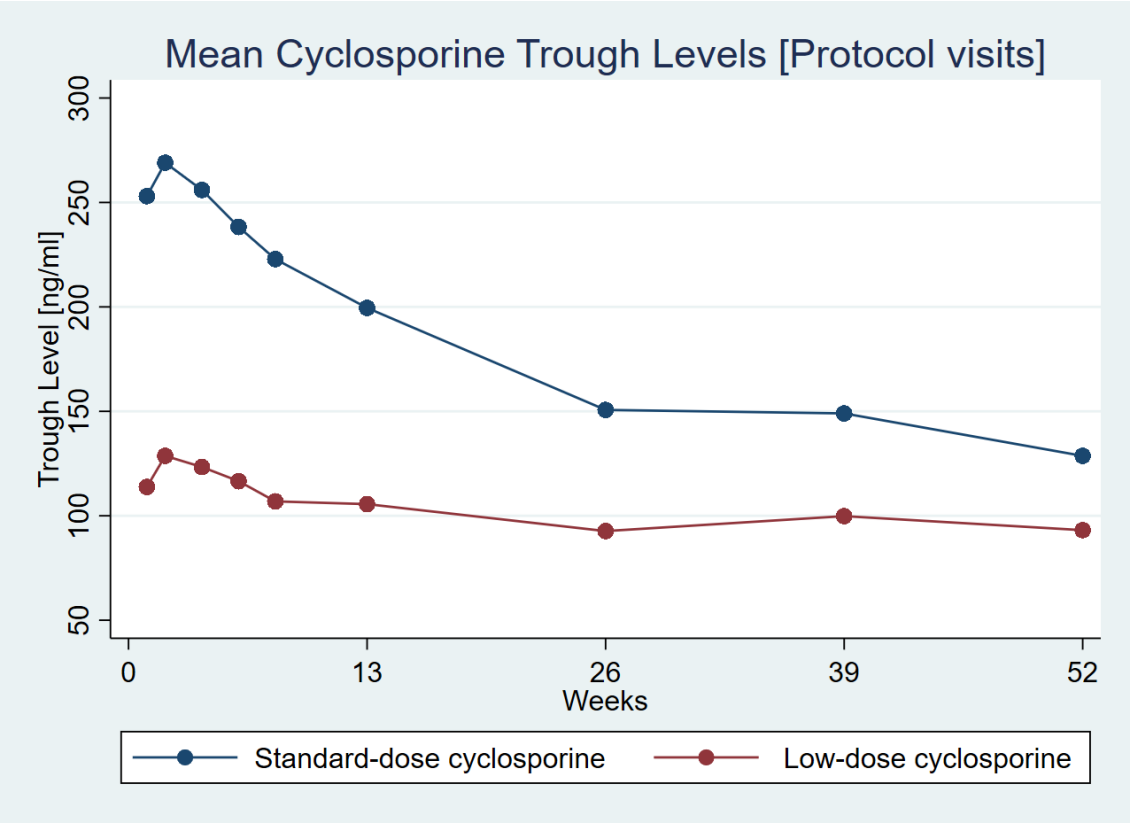


Figure 48: Mean tacrolimus trough levels over the study follow-up (Protocol visits data)

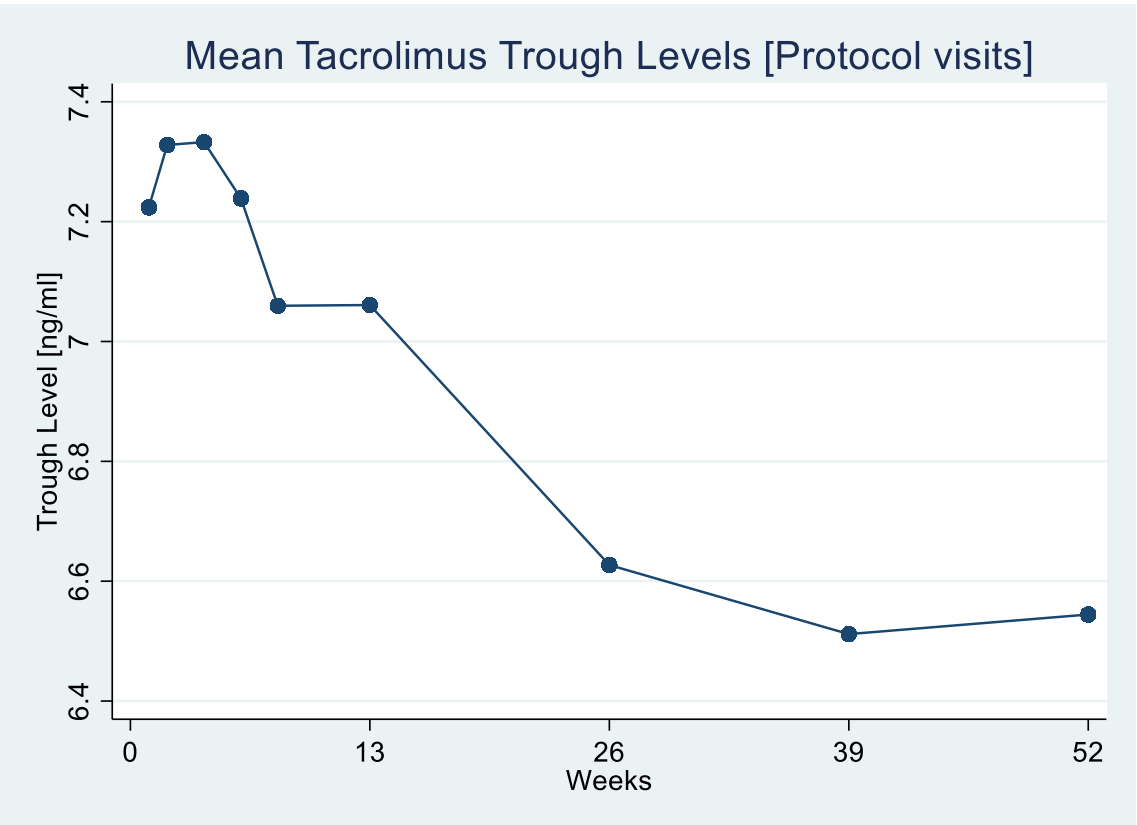
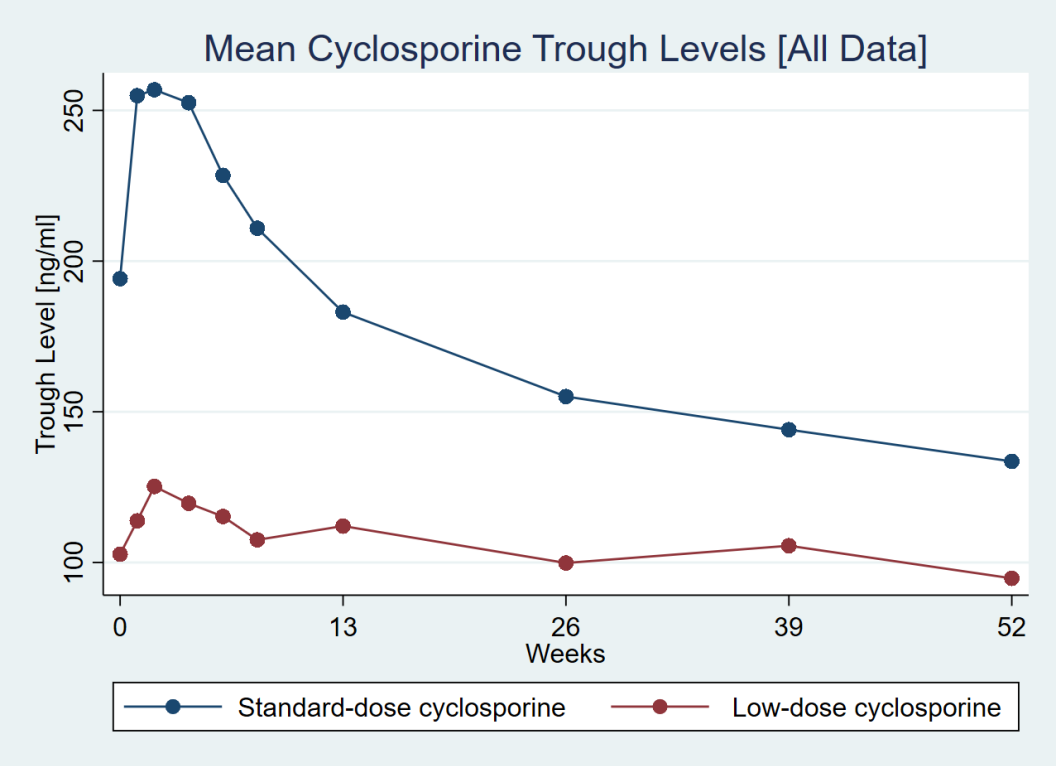
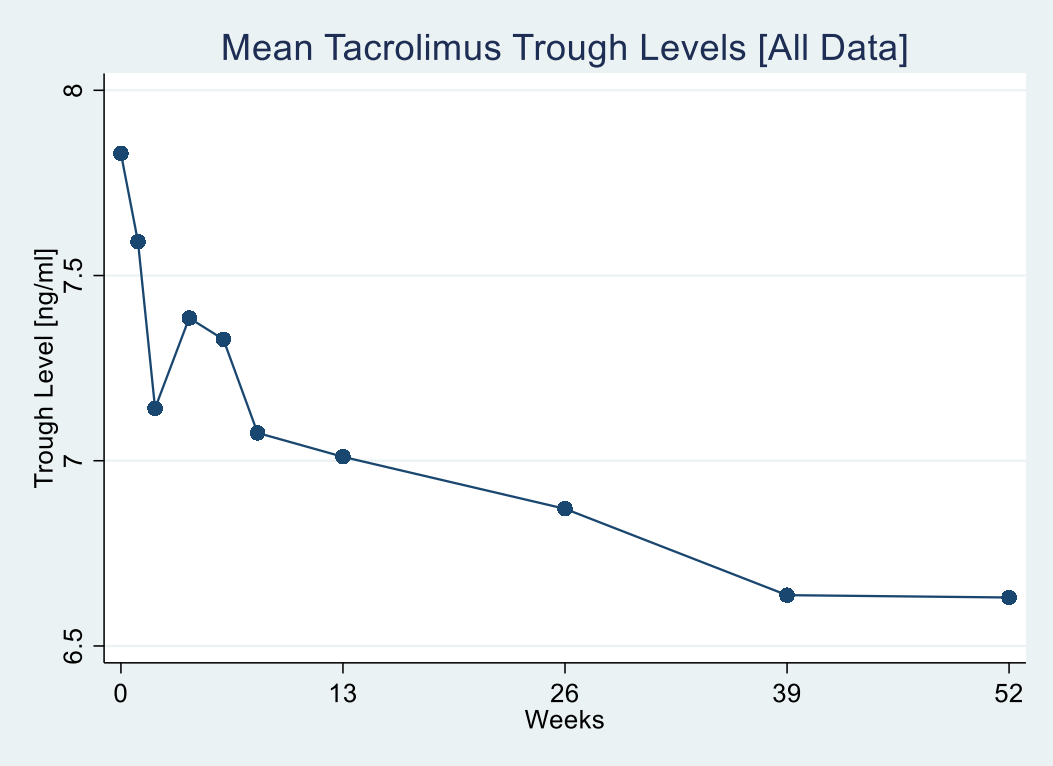


Figure 49: Mean cyclosporine trough levels over the study follow-up (All data)



Note: 'All data' refers to trough levels measured at the study protocol visits plus extra trough levels recorded over the study follow-up. Some patients have more than one trough level record per-protocol visit window; in this case, the average value was used based on the SAP document for SYMPHONY.

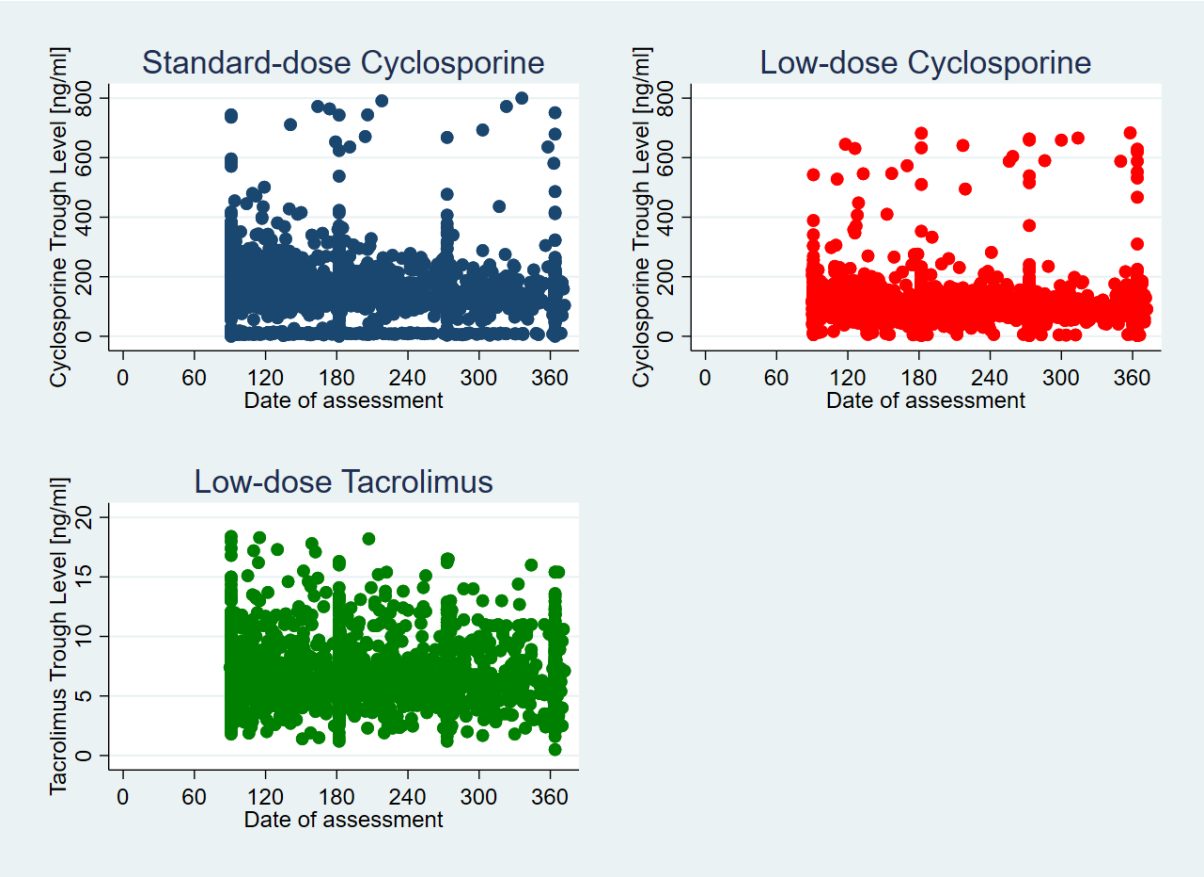
Figure 50: Mean tacrolimus trough levels over the study follow-up (All data)



Note: All data means trough levels measured at protocol visits + extra trough levels recorded over the study follow-up.

Figure 51 shows the distribution of trough levels data from protocol visits and extra visits by the date of assessment and treatment arm. These data were combined such that all the trough level data recorded between 3 and 12 months were used in the analysis for calculating adherence in the SYMPHONY trial.

Figure 51: Intra-patient variability for immunosuppressants by treatment group



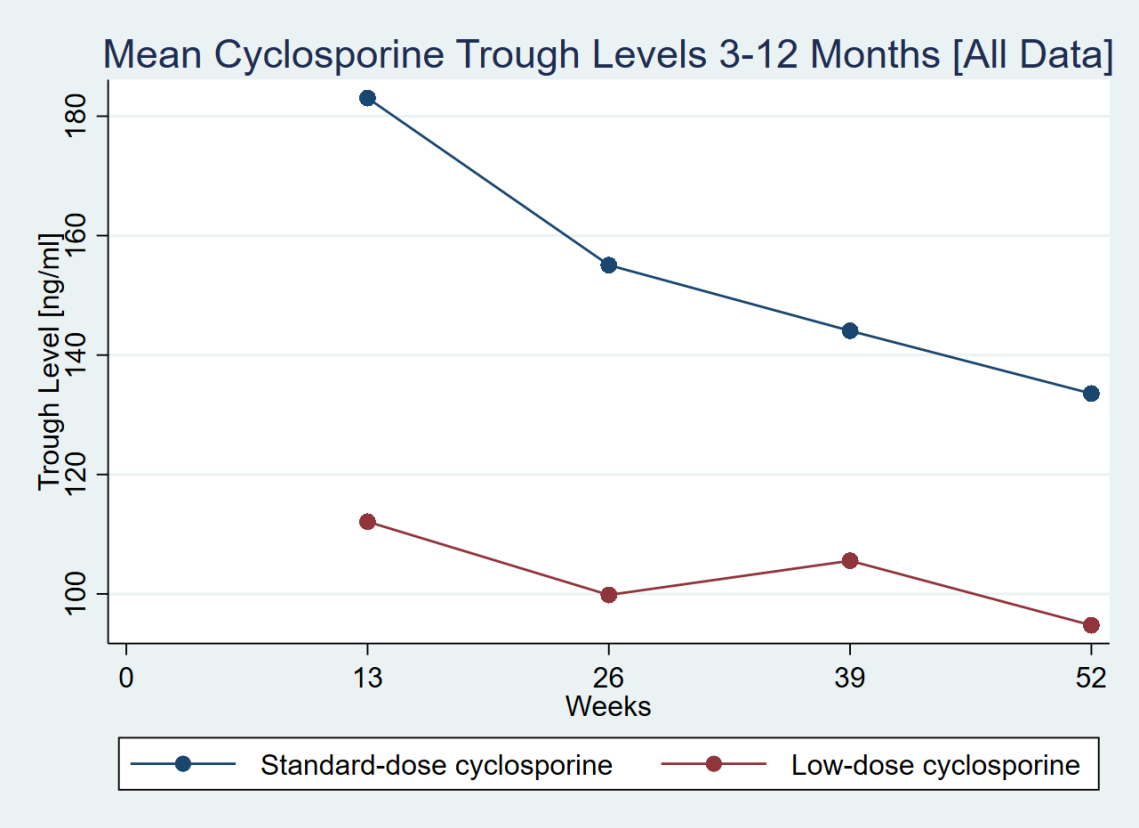
Data shown are the trough level records used in the analysis (including records from protocol visits and extra visits)

Of the 18,873 available trough level records, a total of 11,049 records measured between baseline and 3 months were excluded from the analysis. The exclusion was based on recommendations from published guides and discussions with clinicians.[137, 138] The justification for excluding trough levels recorded between 0-3 months post-transplantation include several factors including fluctuations due to infections, and intravenous steroid treatments see (Figures 47-50).[139] Therefore, a total of 7,824 drug concentration records measured between 3 and 12 months post-transplantation (as a more stable period) were included in the final analysis. Discarding trough levels data recorded between 0-3 months post-transplantation is consistent with the studies that analysed the real-world adherence

data. Among others, the key reasons for discarding data include fluctuations due to infections and intravenous steroid treatments.[139]

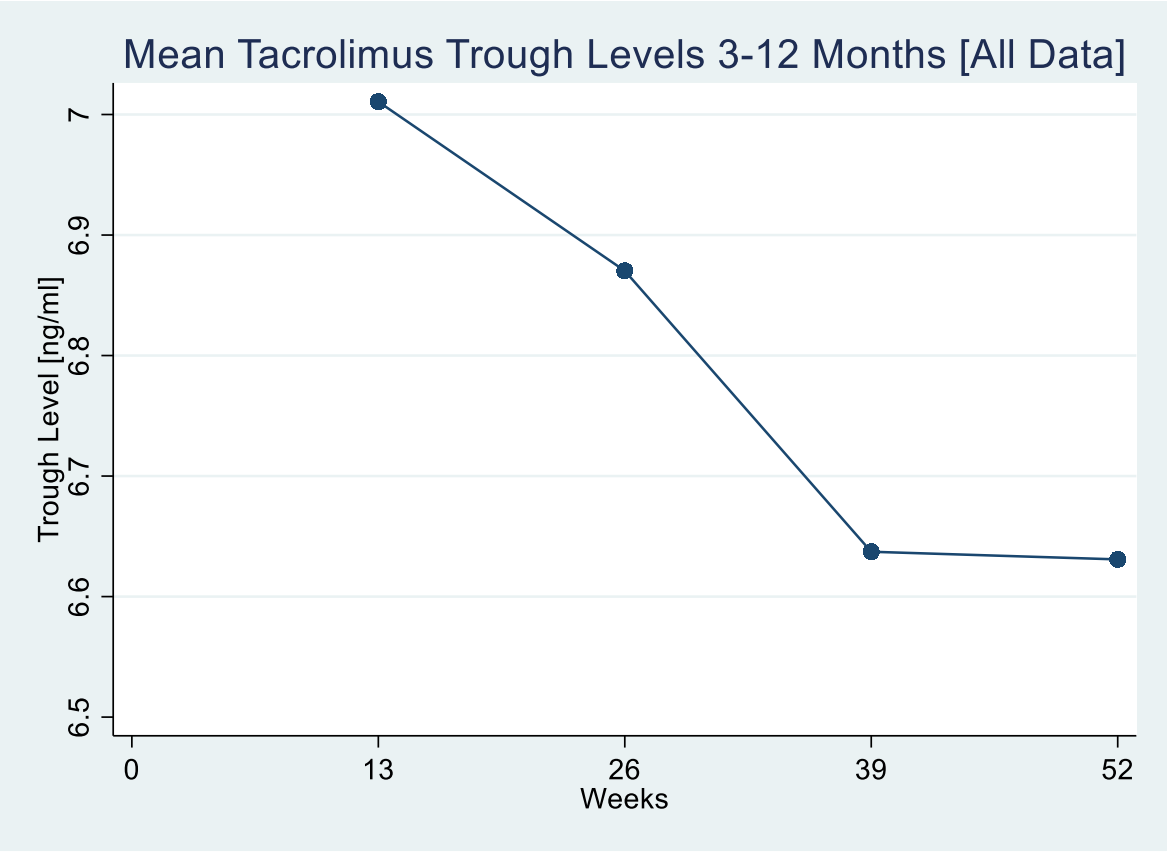
Figures 52 and 53 show the mean trough levels for data included in the analysis by the treatment arm. These represent the adherence observed in the SYMPHONY trial. These are data used for calculating the CV% and prediction of real-world non-adherence levels as described in Section 6.5.

Figure 52: Mean cyclosporine trough levels (3-12 months)



All data (protocol visits + extra records) measured between 3 to 12 months was used for calculating the coefficient of variation (CV%)

Figure 53: Mean tacrolimus trough levels (3-12 months)



The total and mean numbers of trough level records and the sampling frequency for each treatment group are provided in Table 17. This represents the adherence data used in the analysis.

Table 17: Mean number of trough level records excluding patients with less than three records (3-12 months)

Treatment group	**Number of patients with less than 3 records of trough levels: n (%)**	**Total number of trough level records**	**Mean number of records per patient**
Standard-dose cyclosporine (n=390)	83 (21.3)	2325	7.6
Low-dose cyclosporine (n=399)	68 (17.0)	2409	7.3
Low-dose tacrolimus (n=401)	48 (12.0)	2701	7.7

6.4.3.2 Calculating the coefficient of variation from trough levels data

Pharmacological treatments usually work if their concentration levels in the blood are maintained within a minimum and maximum window known as the 'therapeutic index' (the ratio of the highest dose that is acceptably safe to the lowest which is sufficient for the drug to be effective). The dose

that the patient needs to take regularly is usually calculated to maintain the drug concentration within that therapeutic window until the end of prescribing. However, due to patient non-adherence, the drug concentration level will fluctuate as a result of erratic use of the prescribed dosing regimen. This may lead to a situation where the drug concentration falls below that recommended level and ultimately results in a therapeutic failure. This fluctuation is usually measured using intra-patient variability (IPV) as part of therapeutic drug monitoring (TDM) as a standard in clinical practice for particular medications. Higher IPV indicates non-adherence to the prescribed dosing regimen depending on the cut-off point which differs by drug or classes of drugs.

IPV is defined as fluctuation in drug trough concentration levels for an individual patient over a specified follow-up period in which the dose was not changed. IPV is commonly quantified using the coefficient of variation (CV) for the pre-dose blood concentration of the immunosuppressant known as trough concentration (C_0). Trough levels are routinely used in practice for TDM of medications with a low therapeutic index such as cyclosporine and tacrolimus. The percentage of CV (CV%) for C_0 can be calculated using the following formula:[139]

$$CV\% = \left\{\frac{[(X_{mean} - X_1) + (X_{mean} - X_2) \ldots\ldots + (X_{mean} - X_n)]}{n}\right\} / X_{mean} \times 100 \qquad [36]$$

where X_{mean} is the average of all available trough concentrations (C_0) records measured over the analysis time, n is the number of records, X_1 is the individual's first trough level record, X_2 is the second record, etc… There are other formulas used in the literature to calculate CV%, but no clinically relevant differences have been demonstrated between the formula.

Once the CV% for each patient is calculated, a cut-off point is required to determine patient adherence to the prescribed dosing regimen. Patients with a CV% higher than the cut-off point are considered non-adherent in terms of implementation of the prescribed dosing regimen. Different approaches are commonly used in practice to determine the cut-off point including median CV and simultaneous measurement of adherence such as electronic monitoring. The key points are that standard cut-offs are widely accepted and that these vary by drug due to their different PKPD characteristics and forgiveness profiles.

<u>Calculating CV% in the trial</u>

In the base-case analysis, the CV% was calculated by dividing the analysis time into three separate time intervals. These were baseline to 3 months, 3-6 months and 6-12 months post-transplantation. Dividing the last time interval into two equal intervals of 3 months each was considered; however, this

was not applied. The main reason for this was the smaller frequency of trough level measurements during that period compared to the first 6 months resulting in higher proportions of patients with missing CV% which is not desirable. Nevertheless, some patients had missing data in terms of the calculated CV% in the primary analysis. In this case, patients who had fewer than three trough level records during specific follow-up time intervals did not provide sufficient data to calculate their CV% for that particular time interval. In these cases, the CV% could not be calculated as per the recommendations of published guides for the reasons mentioned in Section 6.4.3. No adjustment was applied for those patients effectively assuming they adhered to the prescribed dosing regimens which might not be the case.

Three sensitivity analyses relating to the approach for addressing discarded adherence data and real-word non-adherence estimates were conducted. The first sensitivity analysis used differential adherence levels based on a small study that used real-world data.[140] In that study, implementation adherence was higher for tacrolimus compare to cyclosporine with non-adherence levels of 26.7% and 36.4%, respectively. To perform this sensitivity analysis, a different cut-off point of CV% was used such that the predicted non-adherence proportions for each treatment arm are similar to those observed in the real-world study. The second sensitivity analysis assumed perfect adherence during the first 3 months post-transplantation as the trough levels data recorded during this period were discarded due to concerns of reliability as discussed above. The third sensitivity analysis adjusted for non-adherence for the period 6-12 months only (effectively discarding adherence data measured between baseline and 6 months). The latter approach is consistent with the real-world estimates of non-adherence and the level of missing adherence data for the last 6 months was lower than the interval 3-6 months due to longer follow-up. The last 6 months period also represents a more stable immunosuppression follow-up based on the CV% for trough levels.

A sensitivity analysis assuming that patients with no or less than three trough level records for each interval were considered but not performed. The justification was that a higher proportion of patients falls into that category and performing a sensitivity analysis with such a strong assumption will risk producing misleading cost-effectiveness results.

6.4.3.3 Primary and secondary outcome data

The primary outcome used in SYMPHONY was the estimated glomerular filtration rate (eGFR) 12 months after transplantation as defined in the study protocol. Secondary outcomes included time to biopsy-proven acute rejection (BPAR), graft survival, allograft dysfunction, treatment failure and death. In this case study, graft survival was the outcome of interest used in the analysis. The selection of graft

survival was made to match the outcome used in the adapted economic model for estimating the long-term adherence-adjusted cost-effectiveness of immunosuppressants in kidney transplantation. The primary and secondary outcomes data were used as additional checks for data consistency (alongside baseline characteristics).

The SYMPHONY graft survival data show that the total number of patients who lost their graft during the first year post-transplantation, including death, was 91 (standard-dose cyclosporine= 41, low-dose cyclosporine=27, tacrolimus=23). Data on graft survival censored for death with a functioning graft (DWFG) shows that the total number of graft loss was 67 (standard-dose cyclosporine = 31, low-dose cyclosporine= 22, tacrolimus= 14). All these numbers are consistent with the SYMPHONY CSR report.

6.4.3.4 Missing data

Missing data were checked and handled in the analysis performed in this case study. Missing data were particularly important if they are considered to affect the "no unmeasured confounding assumption" relied upon by adjustment methods applied in this case study. The SYMPHONY dataset was checked for missing data for the key variables used in the analysis. These include baseline covariates (age and gender) and time-dependent confounders (BMI and acute rejection).

Out of the 1,190 patients included in the analysis, 42 patients (Group A= 11, Group B= 11, Group B= 20) had missing age values. For BMI values at baseline, a total of 157 patients (Group A= 55, Group B= 43, Group B= 59) had missing data mainly due to missing weight values. Missing data for age and BMI at baseline were imputed using the mean values. For baseline covariates, using means instead of multiple imputation methods was considered more appropriate to maintain the prognostic balance, between treatment groups, generated by the randomisation procedure. There were no missing values for gender and acute graft rejection at baseline; and therefore, no imputation was required.

The imputation of missing values for time-dependent confounders at 3 and 6 months follow-up post-transplantation was handled using the last observation carried forward (LOCF) method. The LOCF imputation used all the available data from the study visits rather than data from the three analysis time points used in this case study. This meant that the nearest available observation was used to impute missing values. Alternative approaches for handling missing data for time-dependent confounding were considered (including mean values and multiple imputations) and the LOCF was selected as the most appropriate option for this particular case study. The decision was justified because BMI does not usually change much over a short period making the last observation value more likely to hold. In addition, missing data on time-dependent acute rejection was minimal.

6.5 Non-adherence to medications in the real world

A review was conducted to identify real-world estimates of IPV using the CV for tacrolimus and cyclosporine in kidney transplantation. The objective was to identify relevant papers that reported real-world estimates of CV% for trough concentrations as a proxy measure of non-adherence to tacrolimus and cyclosporine A (CsA) in kidney transplantation.

6.5.1 Review to identify real-world adherence levels

A citation search and reference list checking was undertaken on five key papers identified based on discussions with two clinical experts (WM and JF).[137-139, 141-143] In addition, a targeted author search on "Kahan, B. OR Johnston A." was performed based on advice from a clinical expert (WM). The citation search and reference list checking was undertaken using the Web of Science electronic database. Papers were included if they reported CV estimates for tacrolimus and/or cyclosporine regimens for adult kidney transplant recipients with a minimum follow-up of 12 months. Papers that reported CV for Sandimune® (the old cyclosporine formulation) were excluded because they are not similar to the formulation assessed in SYMPHONY and are no longer used in current clinical practice.

6.5.2 Real-world studies included

The search generated 703 records after the removal of duplicates. Following title screening, 647 records were excluded. A further 43 papers were excluded at the abstract screening stage and the remaining 13 papers were reviewed in full text. A final list of two relevant papers was included in the review.[139, 144] A summary of estimates obtained from the included papers is provided in Table 18.

.

6.5.2.1 Tacrolimus real-world non-adherence level

The Whalen et al.[139] study assessed variability in a tacrolimus regimen that is consistent with SYMPHONY, using data from the West of Scotland Electronic Renal Patient Record. Tacrolimus trough levels were measured at clinic visits. The analysis used tacrolimus trough levels (C_0) measured 6 to 12 months after kidney transplantation. The median CV of 15% was used as a cut-off point assigning 190 patients (50.5%) to the high variability group, suggesting implementation non-adherence. The key estimates are provided in Table 18. These estimates were used for predicting the real-world

implementation non-adherence for each individual patient allocated to tacrolimus in the SYMPHONY dataset.

The real-world adherence data is comparable to the SYMPHONY data in terms of frequency of sampling for trough level records. Sampling frequency refers to the number of available drug concentration records over a particular follow-up period. The sampling frequency ranged between 6.93-10.02 records (on average) in the real-world data compared to 7.3-7.7 records on average in the SYMPHONY dataset. Moreover, both the SYMPHONY trial and the real-world study included incident transplant recipients only and discarded trough levels data recorded at the first 3-6 months (depending on the analysis) providing more assurance that the two estimates of adherence levels are comparable.

6.5.2.2 Cyclosporine real-world non-adherence level

Jorga et al.[144] assessed the IPV of low-dose cyclosporine trough levels (C_0) among 102 stable kidney transplant recipients (minimum of 6 months post-transplantation) recruited from an outpatient clinic in the UK. The mean age of the study population was 50 years and the mean weight was 75 kg. The study found that 50% of patients were above the cut-off point for the CV of 26% for C_0. Estimates from this study were used as real-world non-adherence to cyclosporine in the adjusted analysis. No study was identified for the standard dose cyclosporine, therefore, it was assumed that they are similar to the low-dose formulation which is a conservative assumption. Non-adherence to the standard dose cyclosporine in the real world is likely to be higher than low-dose formulation due to the higher level of toxicity. However, this is a minor issue for the case study as the key difference in adherence levels are between the tacrolimus and cyclosporine regimens.

The real-world adherence data on cyclosporine is also comparable to the SYMPHONY trial data as both included incident transplant recipients and discarded the trough concentration levels recorded during the first 3-6 months. However, Jorga et al. does not provide the sampling frequency and therefore it was not possible to compare it with SYMPHONY dataset in this aspect.

Table 18: Estimates of real-world non-adherence to tacrolimus and cyclosporine

Author (Year of publication)	Tacrolimus/ Cyclosporine	Study population	Study context	Sample size (n)	IPV assessment period	CV% cut-off[a]	CV% estimates	Impact on clinical outcomes
Whalen et al. (2017)	Tacrolimus	Adult kidney transplant between January 2007 and December 2011	West of Scotland, UK	376	6-12 months	15%	– Percentage of patients in the low variability (LV) group (defined as CV% C_0 < 15%)= 49.5% – Percentage of patients in the high variability (HV) group (CV% C_0 ≥ 15)= 50.5%	Reduced risk of rejection-free survival among HV patients compared with LV patients at 12 months follow-up (HR=1.953; 95% CI: 1.234-3.093; p-value=0.0054).
Jorga et al. (2004)	Cyclosporine A (CsA, Neoral®)	Adult kidney transplant between January 2003 and January 2004	UK	102	12 months (minimum 6 months post-transplant)	26%	– Percentage of patients in the high variability (HV) group (defined as CV% C_0 > 26%) = 50%	Impact on clinical outcomes not assessed in this phase of the study

Note: C_0 = trough concentration (i.e. the drug concentration in a whole blood sample measured immediately before the next dose); C_{av} = average concentration (i.e. dosing interval corrected value).

a: This cut-off point was based on the median IPV among all patients.

6.6 Re-estimation of treatment effectiveness

6.6.1 Predicting real-world adherence within SYMPHONY dataset

The real-world implementation non-adherence for each individual patient for each time interval was predicted using the CV from observed trough levels measured in the SYMPHONY trial. The prediction was based on the pharmacokinetics data (trough levels) using a new cut-off point for CV% with SYMPNHY to achieve the adherence levels in the real world. In other words, I identified a new cut-off point of CV% for trough levels recorded in SYMPHONY above which every patient with higher CV% falls into the non-adherence group in the predicted adherence indicator. The predication method

satisfies two conditions: (a) the percentage of non-adherent patients matches the real-world non-adherence level; and (b) patients with poor prognosis (aged<24 years, male, high BMI and/or experienced acute rejection) have a higher probability of being non-adherent in the predicted adherence data. The predicted adherence indicator (a binary adherence variable) is then used in the application of g-methods to produce the adjusted effectiveness estimates for the economic model (see Section 6.6.2). The main advantage of this prediction approach is that is based on an objective measure of adherence using pharmacokinetics data measured within the trial. This prediction approach was informed by discussions with two clinicians (WM and JF).

Table 19 provides the real-world predicted non-adherence as a percentage of patients by treatment arm for the base-case analysis. The CV% cut-off point required to achieve the real-world estimates is also provided in the Table. The percentage of non-adherent between 3-6 months is low because a large proportion of patients do not have the minimum three data points required to calculate the CV% based on trough levels recorded in the trial. These patients were assumed adherent and will result in an underestimate of real-world predicted non-adherence within the SYMPHONY dataset. For the interval between 6 and 12 months, sufficient data on trough levels were available and therefore, non-adherence predicted for that interval could be considered as the best estimate of real-world adherence. This matches the analysis from the identified papers of real-world adherence levels as both studies estimated non-adherence for 6 months only (6-12 months post-transplantation) as discussed in Section 6.5.2.

Table 19: Real-world predicted non-adherence by treatment arm

Treatment arm	CV% cut-off for predicting real-world adherence	Percentage patient predicted as non-adherent
Non-adherence between 3-6 months		
Standard-dose cyclosporine	32.51	13.60
Low-dose cyclosporine	28.94	12.96
Tacrolimus	24.25	15.65
Non-adherence between 6-12 months		
Standard-dose cyclosporine	26.02	41.64
Low-dose cyclosporine	25.69	39.78
Tacrolimus	24.39	43.94

6.6.2 Adjusted and unadjusted analysis

The SYMPHONY dataset was first analysed using the standard ITT analysis to produce unadjusted effectiveness estimates. The adjusted analysis was performed using IPCW in the base case analysis and SNFTM and secondary analysis. The IPCW adjustment method was chosen to re-estimate effectiveness in the base-case analysis for this case study because its performance on coverage was superior to SNFTM across the relevant implementation non-adherence scenarios. In those scenarios, the IPCW produced about 10% higher coverage (about 85% coverage on average) compared to SNFTM which has produced around 75% coverage on average. The IPCW adjusted analysis used CV% estimates to predict the real-world non-adherence for each individual patient in each time interval in the dataset.

Clinical effectiveness estimates from these sensitivity analyses were obtained using the IPCW adjustment method. The effectiveness was estimated in a form of graft survivor functions including the SEs and 95% confidence intervals around the estimates. The choice of the form was influenced by how treatment effectiveness was incorporated into the original economic model[145]. The sensitivity analyses in terms of estimating adherence-adjusted treatment effectiveness were performed using three alternative approaches for addressing missing adherence data. The sensitivity analyses were as follows:

- Using differential non-adherence levels (as non-adherence percentages) based on a small single-centre study used real world data from Serbia (tacrolimus= 26.7%, cyclosporine= 36.4%).[140]
- Assuming perfect adherence between baseline and 3 months as the trough concentration levels data was discarded for this interval as discussed in Section 6.4.3.
- Adjusting for non-adherence between 6 and 12 months only as a more stable time interval for trough level records.

As expected, the real-world non-adherence levels identified from the review (Section 6.5) were higher than adherence levels observed in the SYMPHONY trial, although these are not very different between the tacrolimus and cyclosporine regimens. In other words, adherence to the real world is worse than adherence observed in the trial. Therefore, the differential adherence sensitivity analysis was performed to see the impact on the adjusted cost-effectiveness results compared to the base-case analysis. Clinical effectiveness estimates from these three sensitivity analyses were obtained using the IPCW adjustment method.

The clinical effectiveness results from the analysis performed including unadjusted and adjusted analyses are presented in Section 6.8.1. The adjusted survivor functions obtained from the sensitivity analyses were used in the adapted economic model to obtain the adherence-adjusted cost-effectiveness estimates (see Section 6.7 for more detail about the economic analysis).

The real world predicted adherence levels within the SYMPHONY dataset (described in Section 6.6.1) were used for applying the IPCW using the four steps described in Chapter 4 (Section 4.6.3). In brief, the IPCW analytical steps applied in this case study were: First, I censored observations at the beginning of the interval when they are predicted to be non-adherent in the real world. Note this is the point that makes their graft survival outcome worse as a contribution to the weighted pseudo-population. Second, I modelled the probability of non-adherence using 18 logistic models comprising 6 models for each arm (two models for each of the three intervals). These models were applied using the real-world predicted non-adherence with baseline confounders (age, gender); and again with both baseline and time-dependent confounders (age, gender, BMI, acute rejection) for generating the stabilised weights. Third, I computed the inverse probability of remaining uncensored and the stabilised IPCW weights. Finally, I applied survival analysis on the pseudo-population to obtain the adjusted graft survivor functions including the SEs and 95% confidence intervals around these estimates. Graft survival functions were censored for DWFG for running the economic model. I also generated the KM graft survival curves and RMSTs by the treatment arm.

The key thing that makes the accounting for real-world adherence levels was the use of predicted non-adherence in generating the IPCW weights. Because the predicted non-adherence was influenced by the confounders as a result of being generated from the objective pharmacokinetics data recorded in the trial (i.e. trough levels with different CV% threshold) it ended up creating a pseudo-population with people predicted to be non-adherent having a negative contribution on the overall average graft survival. That is because the part after the censoring point was the part that had a good outcome in the absence of real-world non-adherence as opposed to having a worse outcome in the simulations. In other words in this analysis approach, I censored those who were predicted to become non-adherent in the real world and found patients who remain unescorted and up-weighted to account for themselves and those censored.

A secondary analysis using the SNFTM with g-estimation was applied for estimating adjusted estimates of treatment effectiveness. This secondary analysis was undertaken for two main reasons. (1) Although the IPCW coverage performance was better than SNFTM in the simulation study, there was uncertainty around the coverage performance data, and given it is similar performance in terms of bias and ModSE, SNFTM could be considered as an alternative option; and (2) the analysis provides

additional information on the application of SNFTM (which was found to be the best-performing method across many scenarios) using real data in this case study. The latter will aid the transferability to other disease areas where the SNFTM might be the best option for adjusting for real-world non-adherence in future studies.

In the SNFTM adjusted analysis, I followed the following steps: First, I used the predicted non-adherence as a binary lagged variable within the g-estimation incorporating all confounders (age, gender, BMI and acute rejection) in the *stgest3* model by treatment arm. This step resulted in an acceleration factor (AF) which I used to shrink graft survival for those predicted non-adherent. This is the step where the negative contribution to the overall average graft survival comes as opposed to the positive contribution in the simulation study. In other words, the truncated part is the part that had a better outcome in the absence of real-world non-adherence (i.e. because graft survival observed in the trial was still the one used in the analysis dataset). This step ended up generating an estimated average graft survival that is worse than the one observed in the trial as produced by the ITT. Then, I generated the adjusted graft survivor functions censored for DWFG for running the economic model and the KM graft survival curves and RMSTs by treatment arm. The key contributing factor is the objective prediction of real-world non-adherence in the trial dataset that led to the truncation of a good part of graft survival for those predicted to be non-adherent. This ended up with worse outcomes (on average) as a result of accounting for real-world non-adherence.

As another secondary analysis, the clinical effectiveness in terms of RMSTs was estimated. These were not used in the economic model but were generated to provide more information in terms of comparing adjusted and unadjusted effectiveness estimates using the SYMPHONY trial dataset.

6.7 The economic model and cost-effectiveness analysis

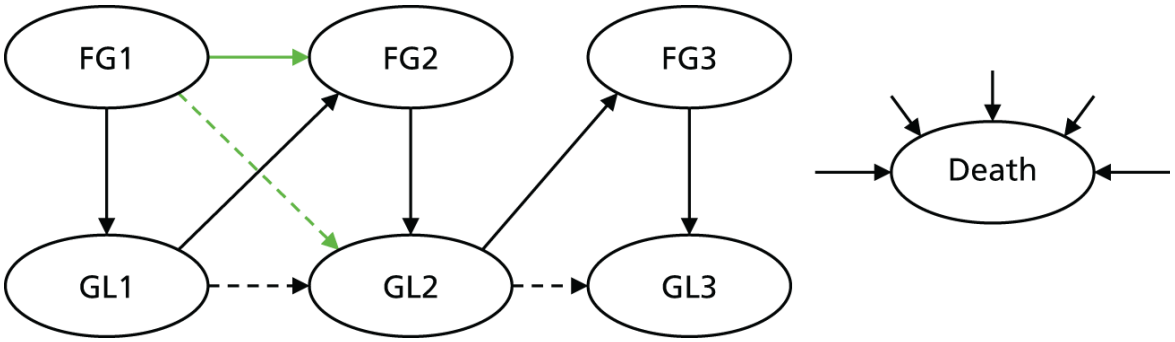
The economic analysis was performed using the adapted model for estimating adherence-adjusted cost-effectiveness of immunosuppressive therapy for adult kidney transplant recipients. The economic model including the adaptation, cost-effectiveness analysis and consideration of the impact on non-adherence on direct drug costs are described in the following subsections

6.7.1 Economic model

The economic model was originally built by the PenTAG group for assessing a range of 16 immunosuppressive regimens to update of NICE Technology Appraisal guidance (TA481).[120, 145] The overall structure of the case study involved an adaptation of this economic model which was focused

on the re-estimated treatment effectiveness in terms of graft loss in the first 12 months post-transplantation (see Section 6.7.2 for more detail on model adaptation). The structure of the economic model (Figure 54) was a discrete-time state transition model with three principal health states (functioning graft, graft loss, death) with up to two re-transplantations allowed.[145] The cycle length was 3 months and a 50-year time horizon was used in the economic analysis (up to a maximum cohort age of 100 years). The main outcome used in the economic evaluation was QALYs with graft survival and patient survival as key other outcomes used in the cost-utility analysis. The economic analysis uses the NHS and Personal Social Service perspective with both costs and QALYs discounted at 3.5% per year. Costs associated with each treatment option for each health state within the original economic model were estimated using various sources and these estimates were not updated in this case study.

Figure 54: Model structure



FG= functioning graft, GL= graft loss, dashed arrow represent primary non-functioning graft, green arrows represent pre-emptive re-transplantation.

Source: Reproduced from Jones-Hughes et al.[120] This is an open access article distributed under the terms of the CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium. This includes minor additions and formatting changes to the original.

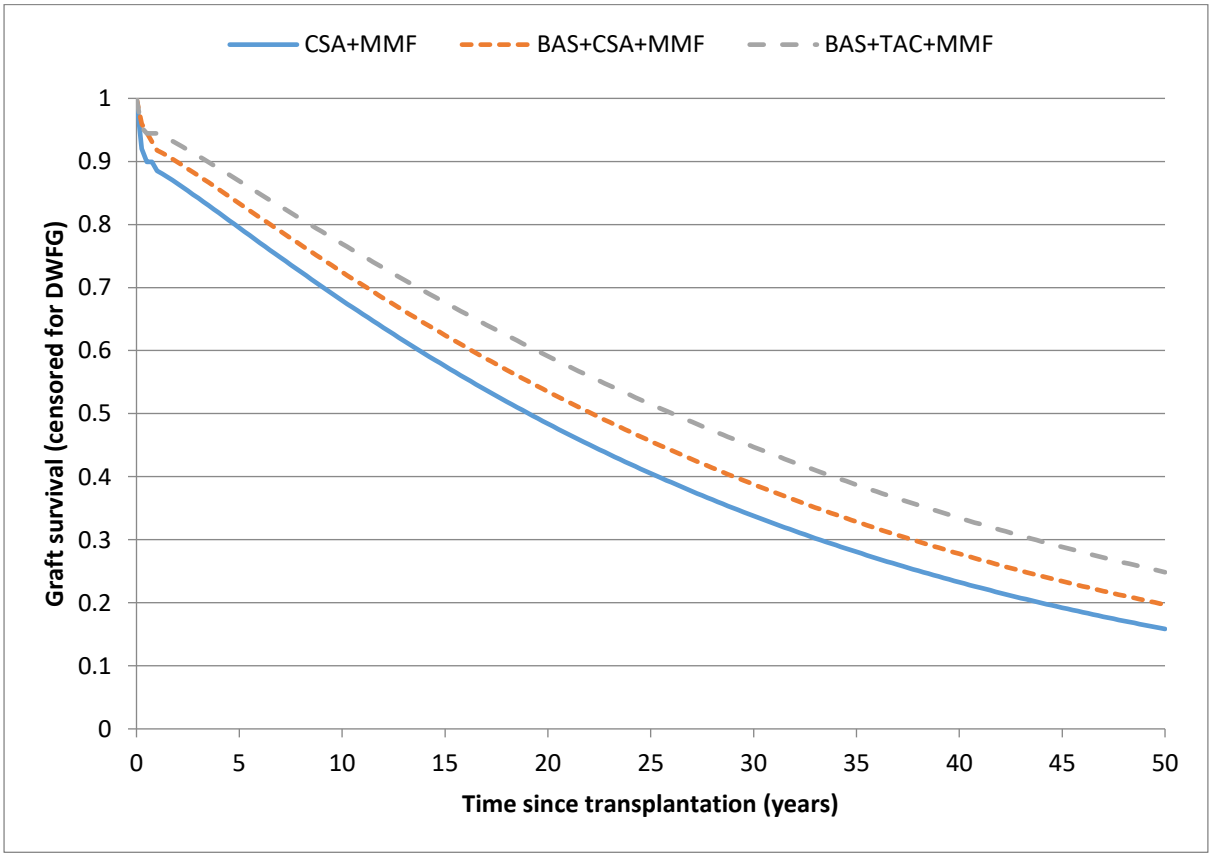
6.7.2 Adaptation of the economic model and extrapolation

The long-term treatment effect was estimated by extrapolating the adherence-adjusted treatment effect based on the analysis of SYMPHONY data undertaken in this study. The adherence-adjusted treatment effect was applied to the initial period of 12 months (the trial follow-up). The extrapolation beyond 12 months was based on data from the UK Transplant Registry dataset as used in the original model. The economic model adaptation involved using estimates from SYMPHONY data analysis in terms of graft survivor functions (censored for DWFG) at baseline, 3, 6, 9 and 12 months. These estimates of graft survivor functions replaced the values in the original economic model up to 12 months. Beyond that, I did not apply the SYMPHONY estimates in the original to the post-12month survival models because the risk comes from somewhere else. The risk for graft survival beyond 12

months come from the UKTR data that already reflects real world adherence levels. All other parameters, including drug costs, were not updated in the adapted model. By leaving everything else the same in the economic model, I was able to assess the sole impact of adjusting for real-world non-adherence levels versus not adjusting.

The IPCW adjusted graft survival curves in the first 12 months from SYMPHONY and the extrapolation is shown in Figure 55. The post-12 months' curves reflect graft survival from the UKTR data. These extrapolated estimates, as well as estimates of graft survival obtained from SYMPHONY, were used to populate the adapted economic model for estimating the relative cost-effectiveness of immunosuppressants. The model was run using 10,000 probabilistic sensitivity analysis (PSA) iterations to produce the cost-effectiveness estimates. The results in terms of NHBs were calculated at the cost-effectiveness thresholds of £20,000 per QALYs as described in the NICE Methods Guide.[60] The cost-effectiveness estimates are presented in terms of total discounted costs and QALYs, incremental costs and QALYs, and net health benefits (NHBs). The IPCW analysis is compared with the ITT analysis of the SYMPHONY data.

Figure 55: Graft survival extrapolation from the IPCW adjusted analysis



DWFG= death with functioning graft, CSA= Cyclosporine, MMF= mycophenolate, BAS= basiliximab, TAC= tacrolimus

6.7.3 Impact of non-adherence on direct drug costs

The impact of non-adherence on drug costs was considered for the adapted economic model. This issue has been investigated in other disease areas but no evidence was identified on immunosuppressive therapy after kidney transplantation. Consultations with two clinicians (WM and JF) and a Clinical Pharmacist (DG, Papworth Hospital NHS Foundation Trust, Papworth Everard, Cambridge) were conducted to investigate if drug costs for non-adhered patients are likely to differ from those adhered to in real practice. Based on the discussions, it seems that the NHS is likely to incur the full drug cost since patients are likely to get their prescriptions dispensed but not fully use their medications as prescribed as a result of non-adherence. Consequently, it was decided to use the full drug cost in the adapted economic model. However, it should be noted that this assumption might not be generalisable to other disease areas or other health care systems.

6.8 Results of the case study

6.8.1 Re-estimation of treatment effectiveness

The results in terms of unadjusted and adjusted clinical effectiveness estimates are presented in the following subsections.

6.8.1.1 Unadjusted analysis using ITT

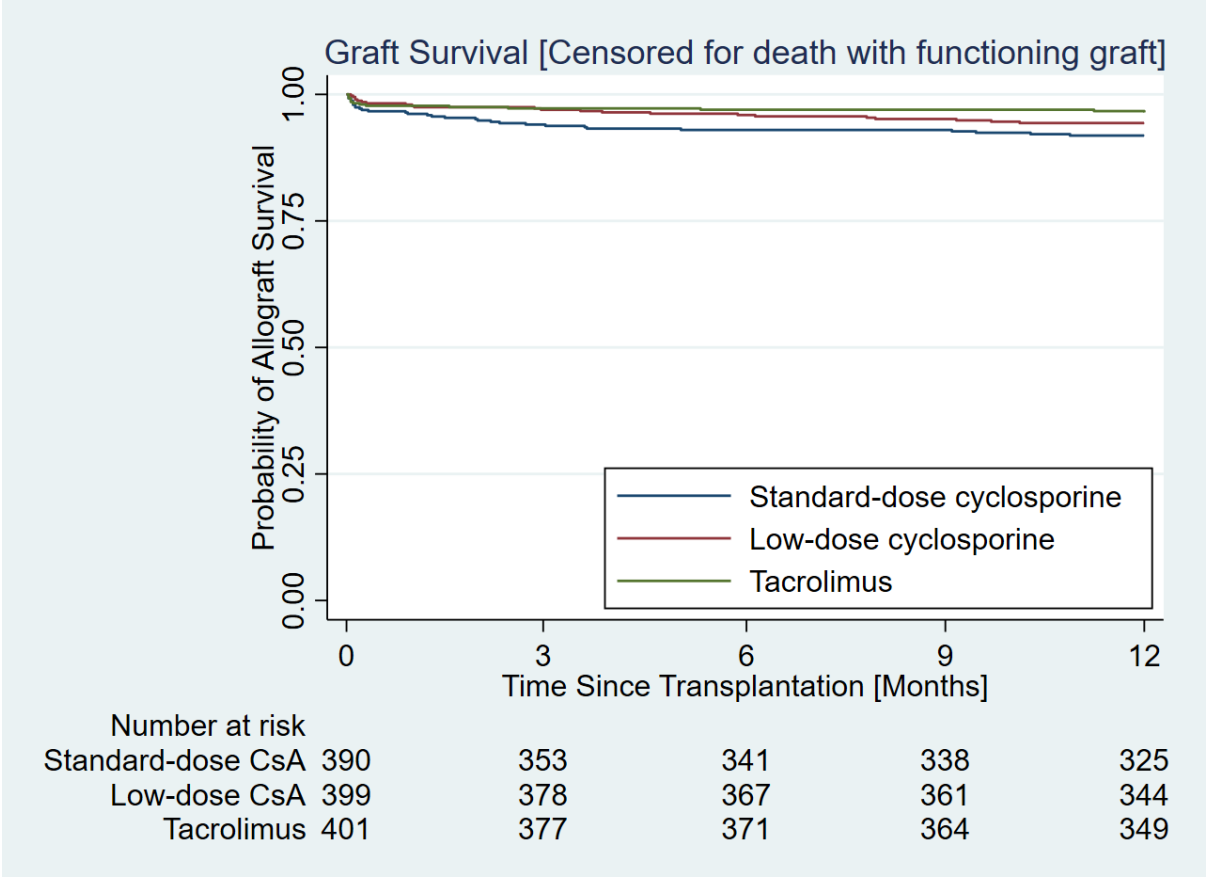
The graft survivor functions estimated for the ITT analysis of SYMPHONY data at baseline, 3, 6, 9 and 12 months are presented in Table 20. The survivor function estimates show that the probability of graft survival among patients in the two interventions groups (low-dose cyclosporine and tacrolimus) were higher than the standard-dose cyclosporine group. The probability of graft survival in the tacrolimus group was better than those of the low-dose cyclosporine group. These estimates from the unadjusted analysis are used to replace the values of survivor functions in the adapted economic model.

Table 20: ITT graft survivor functions censored for DWFG (unadjusted analysis)

	Time	Beg. Total	Fail	Survivor Function	SE	95% Confidence Interval	
Standard-dose cyclosporine							
	Baseline	390	0	1	-	-	
	Month 3	354	23	0.9404	0.012	0.9117	0.9600
	Month 6	342	4	0.9297	0.0131	0.8991	0.9512
	Month 9	339	0	0.9297	0.0131	0.8991	0.9512
	Month 12	325	4	0.9186	0.014	0.8863	0.9421
Low-dose cyclosporine							
	Baseline	399	0	1	-	-	
	Month 3	379	12	0.9696	0.0086	0.9471	0.9826
	Month 6	368	4	0.9592	0.01	0.9343	0.9748
	Month 9	362	3	0.9514	0.0109	0.9248	0.9687
	Month 12	344	3	0.9434	0.0117	0.9153	0.9624
Tacrolimus							
	Baseline	401	0	1	-	-	
	Month 3	379	11	0.9724	0.0082	0.9506	0.9846
	Month 6	372	1	0.9697	0.0086	0.9473	0.9827
	Month 9	365	0	0.9697	0.0086	0.9473	0.9827
	Month 12	349	2	0.9642	0.0094	0.9403	0.9787

The Kaplan-Meier (KM) curves on graft survival (censored for DWFG) obtained from the unadjusted ITT analyses of the SYMPHONY data is presented in Figure 56. The survival curves show that the tacrolimus regimen was superior to both low-dose and standard-dose cyclosporine regimens.

Figure 56: Graft survival censored for DWFG – ITT unadjusted analysis



Note: CsA= cyclosporine

6.8.1.2 Adjusted analysis using MSM with IPCW (base-case adjusted analysis)

The IPCW estimates of graft survivor functions (censored for DWFG) including the SEs and 95% confidence intervals are presented in

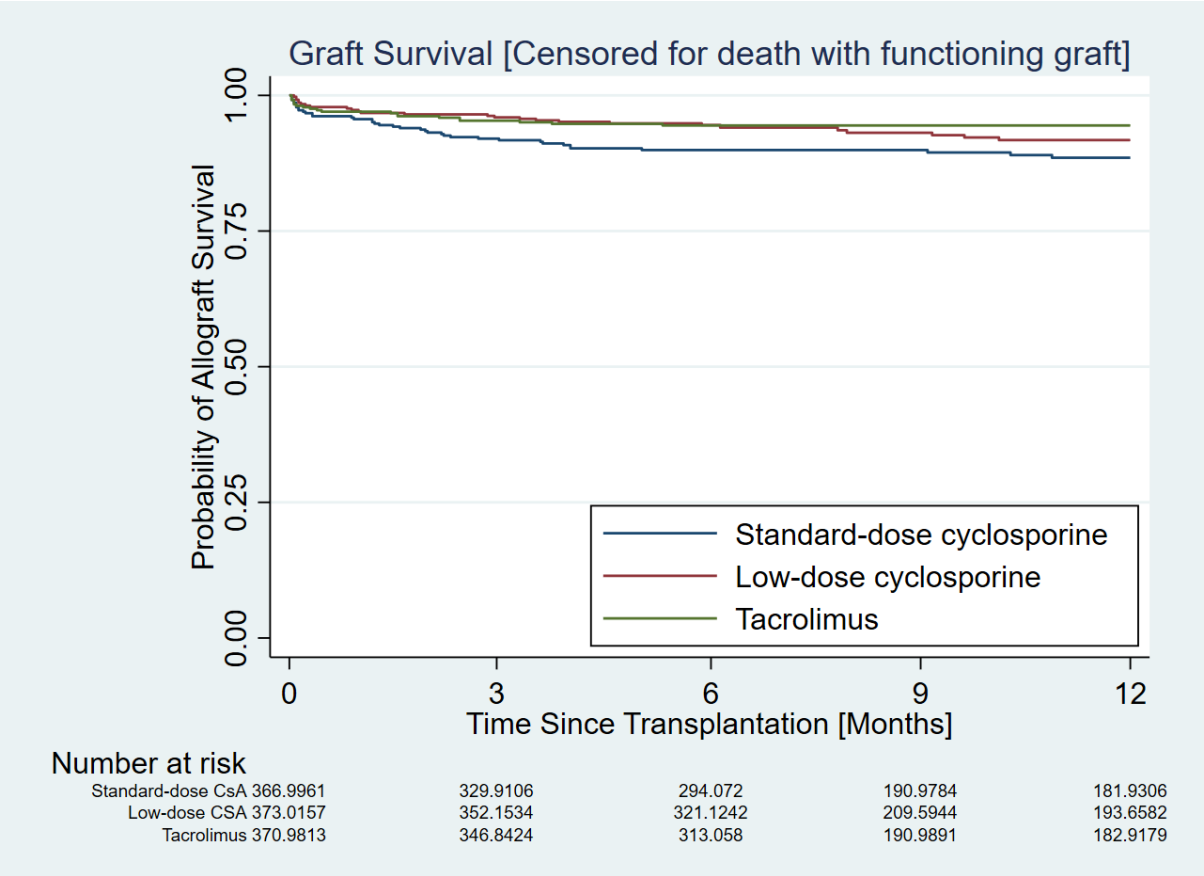
Table 21. These adjusted estimates were used in the adapted economic model to estimate the real-world adherence-adjusted cost-effectiveness of the alternative regimens.

Table 21: IPCW graft survivor functions censored for DWFG (base-case adjusted analysis)

	Time	Beg. Total	Fail	Survivor Function	SE	95% Confidence Interval
Standard-dose cyclosporine						
	Baseline	390	0	1	-	-
	Month 3	329.9	29.06	0.9203	0.0142	0.8874 0.9439
	Month 6	294.1	6.96	0.8993	0.0159	0.8631 0.9263
	Month 9	192.0	0	0.8993	0.0159	0.8631 0.9263
	Month 12	181.9	2.99	0.8851	0.0177	0.8452 0.9152
Low-dose cyclosporine						
	Baseline	399	0	1	-	-
	Month 3	352.2	15.00	0.9595	0.0102	0.9337 0.9754
	Month 6	321.1	4.82	0.9454	0.0119	0.9165 0.9645
	Month 9	210.6	3.24	0.9311	0.0141	0.8974 0.9541
	Month 12	193.7	2.99	0.9177	0.0159	0.8801 0.9438
Tacrolimus						
	Baseline	401	0	1	-	-
	Month 3	346.8	17.13	0.9534	0.011	0.9262 0.9707
	Month 6	313.1	2.96	0.9446	0.012	0.9154 0.9639
	Month 9	193.0	0	0.9446	0.012	0.9154 0.9639
	Month 12	182.9	0	0.9446	0.012	0.9154 0.9639

Figure 57 shows the KM curves on graft survival censored for DWFG from the adjusted IPCW analysis. The graphs are based on the survival analysis applied to the pseudo-population generated from the IPCW weights. The KM graphs show that the tacrolimus regimen still produces better graft survival compared to the alternative regimens, although the scale of health benefits is slightly lower when real-world implementation non-adherence was taken into account.

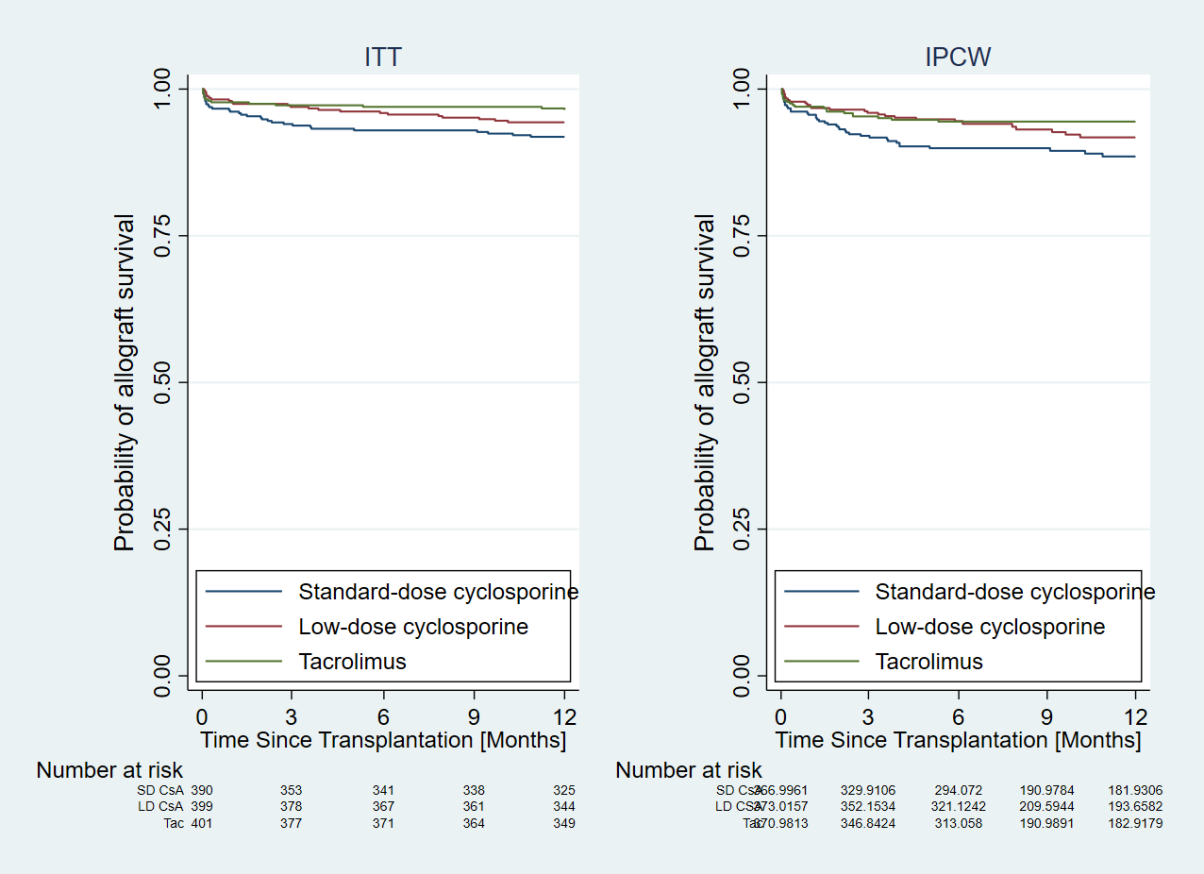
Figure 57: Graft survival censored for DWFG – IPCW adjusted base-case analysis



Note: The risk table shows numbers in decimals as a pseudo-population was used in the IPCW adjusted survival analysis

Figure 58 illustrates the graft survival from the standard ITT unadjusted analysis compared with the IPCW adjusted analysis.

Figure 58: Graft survival from the standard ITT unadjusted analysis and IPCW adjusted analysis



SD CsA= standard-dose cyclosporine, LD CsA= low-dose cyclosporine

Note: For the IPCW graft survival curve, the number at risk include decimals because the adjusted analysis is based on a pseudo-population of adherent patients up weighted to represent the entire study population

Table 22 provides the graft survivor functions comparing the ITT unadjusted estimates with the IPCW adjusted estimates by treatment arm.

Table 22: Unadjusted ITT estimates versus IPCW adjusted estimates of graft survivor functions

	Time	ITT	IPCW	% difference between IPCW and ITT
Standard-dose cyclosporine				
	Baseline	1	1	0
	Month 3	0.9404	0.9203	-2.14
	Month 6	0.9297	0.8993	-3.27
	Month 9	0.9297	0.8993	-3.27
	Month 12	0.9186	0.8851	-3.65
Low-dose cyclosporine				
	Baseline	1	1	0
	Month 3	0.9696	0.9595	-1.04
	Month 6	0.9592	0.9454	-1.44
	Month 9	0.9514	0.9311	-2.13
	Month 12	0.9434	0.9177	-2.72
Tacrolimus				
	Baseline	1	1	0
	Month 3	0.9724	0.9534	-1.95
	Month 6	0.9697	0.9446	-2.59
	Month 9	0.9697	0.9446	-2.59
	Month 12	0.9642	0.9446	-2.03

The coefficients generated from the 18 non-adherence models (logistic models to generate the IPCW weights) are provided in Appendix I, Table 43. Table 23 provides the IPCW weights including un-stabilised and stabilised weights generated from weighting models. In this table, the number of observations represents the uncensored observations for the three-time intervals, which are weighted to create the pseudo-population to which the survival analysis was applied to produce the adjusted clinical effectiveness estimates.

Table 23: Un-stabilised and stabilised IPCW weights by treatment arm

Treatment arm	n	Mean	SD	Min	Max
Weights (un-stabilised)					
Standard-dose cyclosporine	866	1.2570	0.2997	1.0188	2.2339
Low-dose cyclosporine	916	1.2571	0.2874	1.0066	2.2934
Tacrolimus	885	1.2982	0.3334	1.0411	2.3500
Stabilised weights					
Standard-dose cyclosporine	866	0.9998	0.0196	0.8436	1.0887
Low-dose cyclosporine	916	0.9997	0.0297	0.9135	1.2285
Tacrolimus	885	1.0001	0.0156	0.8664	1.0842

6.8.1.3 Sensitivity analysis

The IPCW estimates of graft survivor functions from sensitivity analyses are presented in Table 24. The graft survivor functions from the IPCW adjusted base-case analysis are also included for comparisons. The detailed results of graft survivor functions from all sensitivity analyses including SEs and the 95% confidence intervals around the estimates are provided in Appendix I, Table 44-46

Table 24: Graft survivor functions from sensitivity analyses (IPCW adjusted analysis)

	Time	**Base-case**	**Differential MNA**	**Assuming perfect adherence between 0-3 months**	**Adjusting for MNA between 6-12 months only**
Standard-dose cyclosporine					
	Baseline	1	1	1	1
	Month 3	0.9203	0.9203	0.9205	0.9205
	Month 6	0.8993	0.9	0.8995	0.8993
	Month 9	0.8993	0.9	0.8995	0.8993
	Month 12	0.8851	0.8882	0.8853	0.8855
Low-dose cyclosporine					
	Baseline	1	1	1	1
	Month 3	0.9595	0.9595	0.9595	0.9595
	Month 6	0.9454	0.9456	0.9454	0.9449
	Month 9	0.9311	0.9335	0.9311	0.9311
	Month 12	0.9177	0.9221	0.9177	0.9181
Tacrolimus					
	Baseline	1	1	1	1
	Month 3	0.9534	0.9534	0.9538	0.9538
	Month 6	0.9446	0.9502	0.9449	0.9448
	Month 9	0.9446	0.9502	0.9449	0.9448
	Month 12	0.9446	0.9502	0.9449	0.9448

Graft survival censored for DWFG from sensitivity analysis using differential real-world non-adherence is presented in Figure 59.

Figure 59: Graft survival censored for DWFG from sensitivity analysis - Differential real-world non-adherence levels

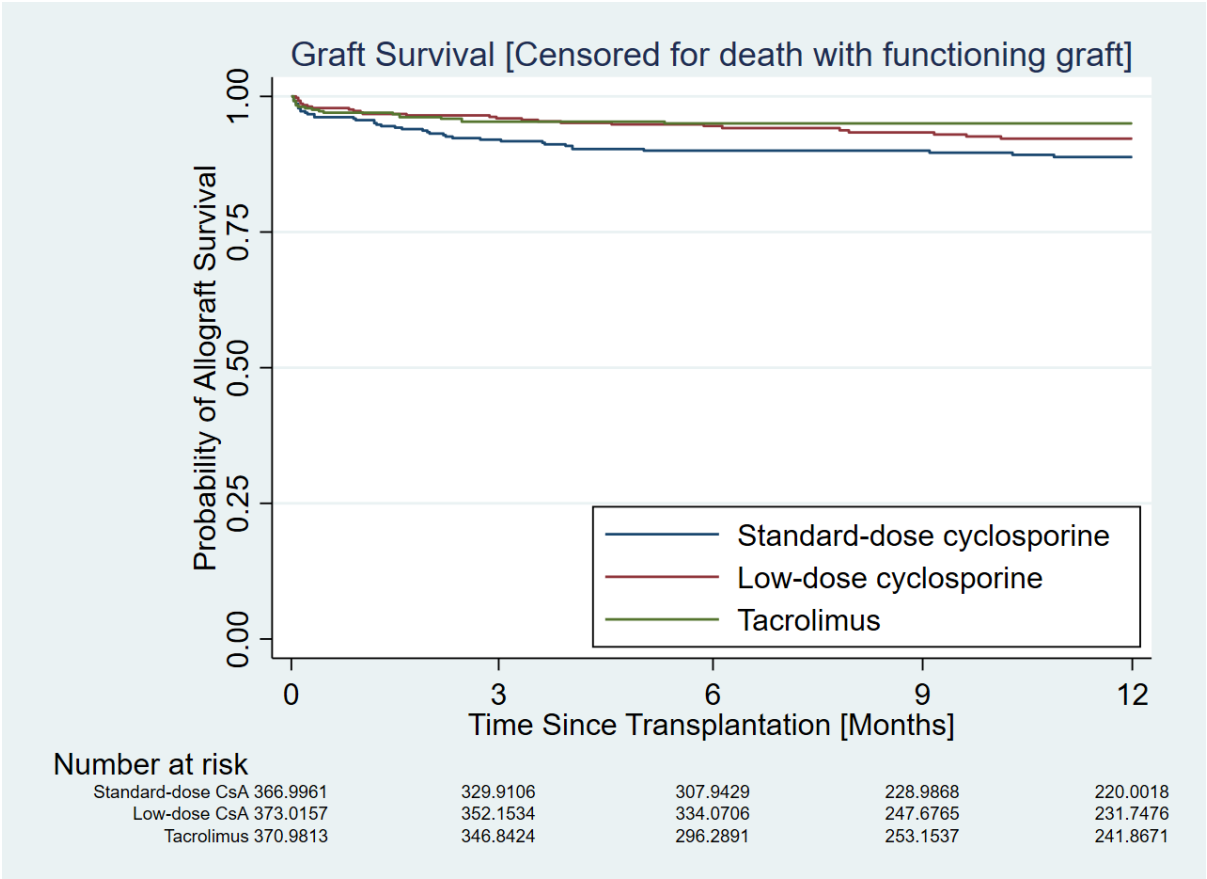


Figure 60 shows graft survival (censored for DWFG) from the IPCW sensitivity analysis assuming perfect adherence between baseline and 3 months post-transplantation.

Figure 60: Graft survival censored for DWFG - Sensitivity analysis assuming perfect adherence between 0- 3 months

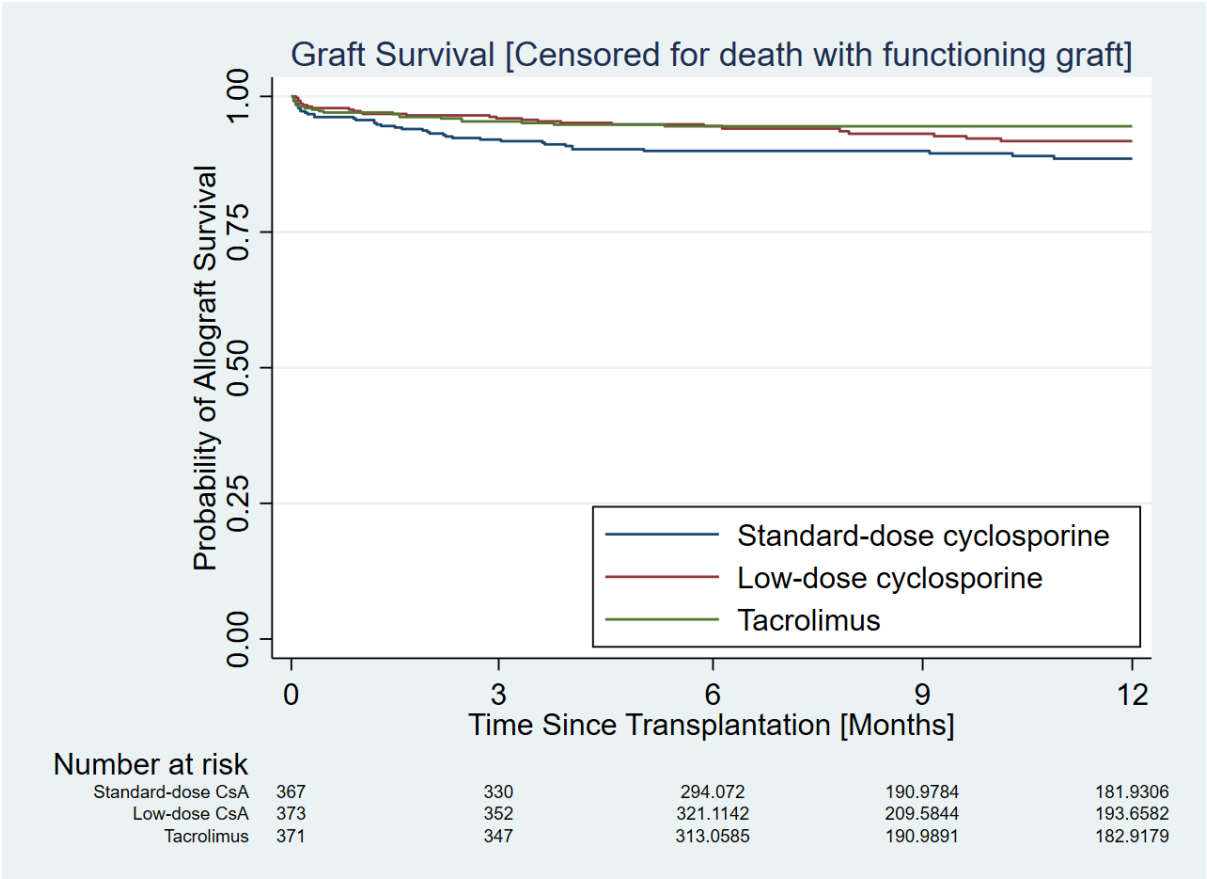
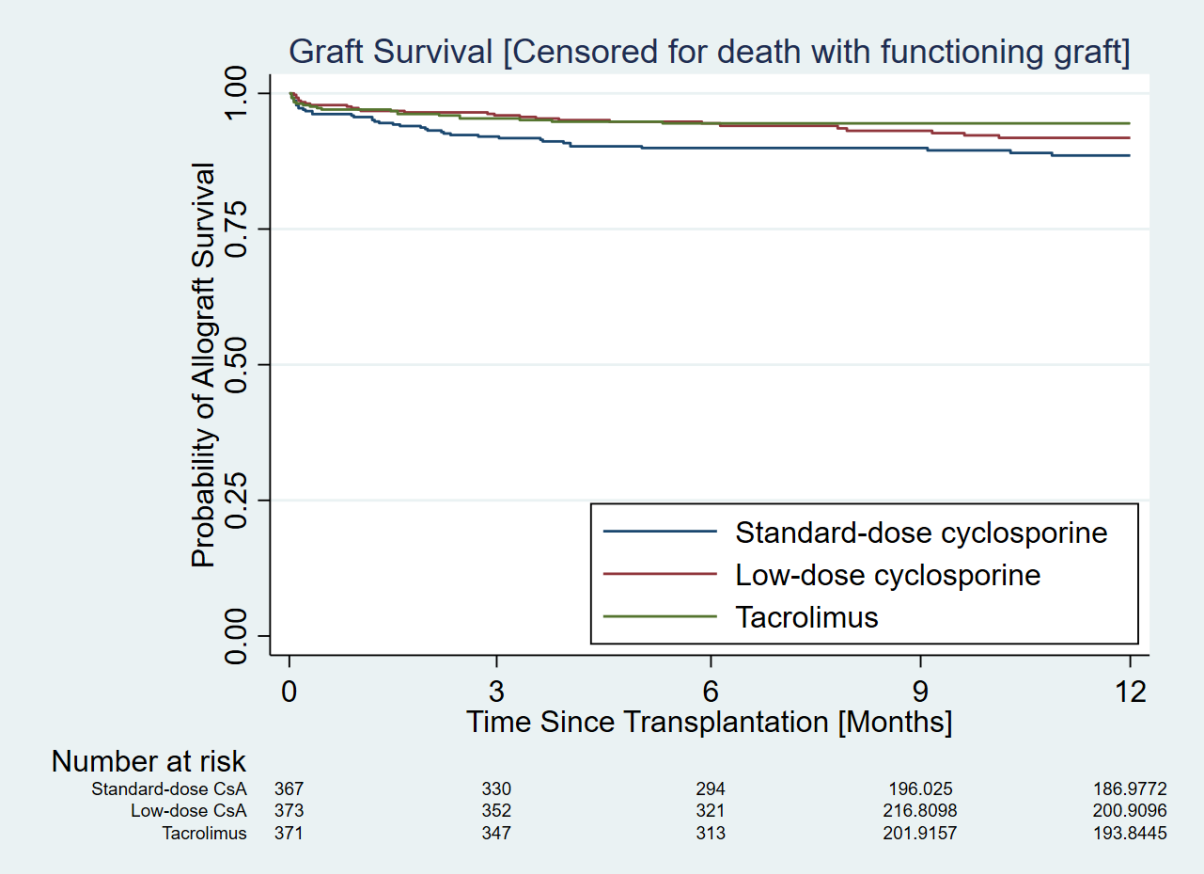


Figure 61 shows graft survival (censored for DWFG) from the IPCW sensitivity analysis adjusting for non-adherence between 6 and 12 months only.

Figure 61: Graft survival censored for DWFG - Sensitivity analysis adjusting for non-adherence between 6-12 months only



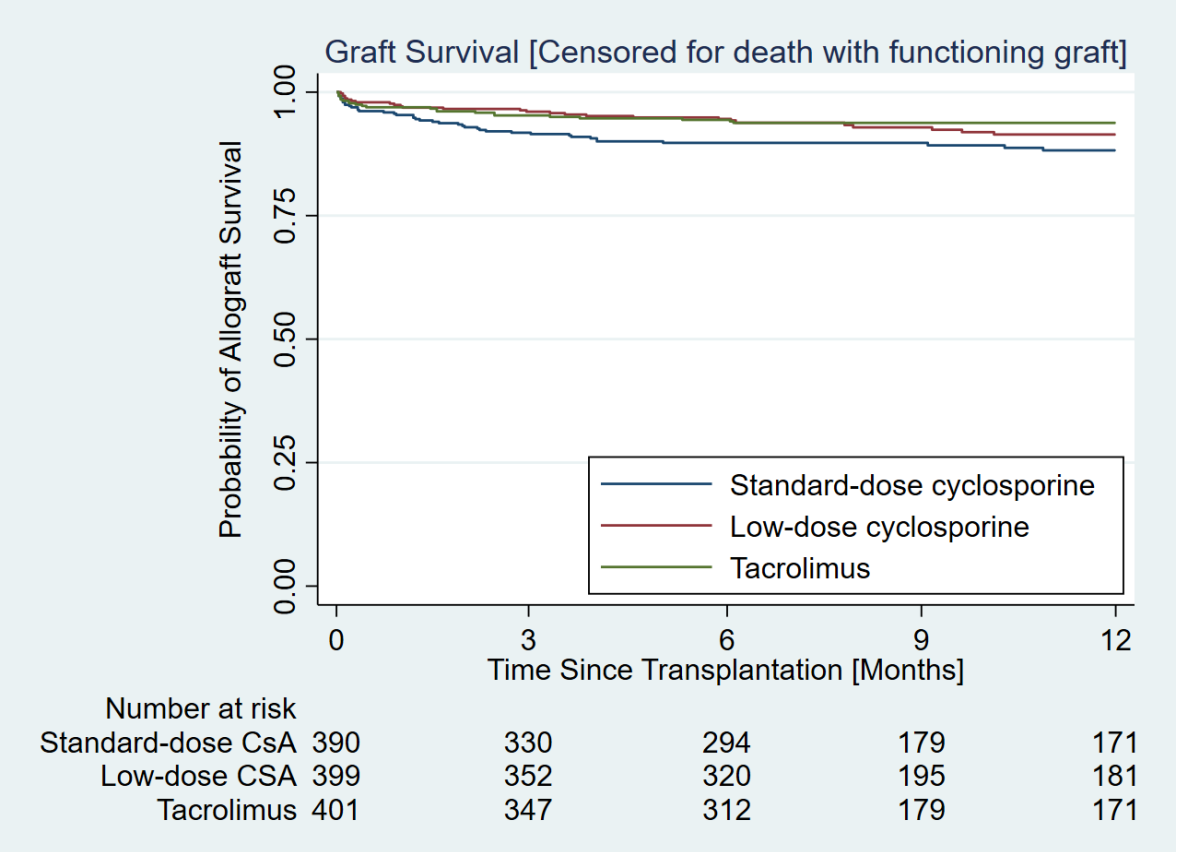
6.8.1.4 Adjusted Secondary analysis using SNFTM with G-estimation

The treatment effectiveness results in terms of graft survivor functions obtained from the secondary adjusted analysis using the SNFTM with g-estimation is presented in Table 25. The graft survival from the secondary SNFTM analysis is provided in Figure 62. The AF (value of psi) and the causal survival ratio (the value by which graft survival shrunk based on the impact of non-adherence) are provided in Appendix I, Table 47.

Table 25: SNFTM graft survivor functions censored for DWFG (secondary adjusted analysis)

	Time	Beg. Total	Fail	Survivor Function	SE	95% Confidence Interval
Standard-dose cyclosporine						
	Baseline	0	0	1	-	-
	Month 3	331	31	0.9178	0.0142	0.8852 0.9415
	Month 6	295	7	0.8973	0.0158	0.8615 0.9242
	Month 9	180	0	0.8973	0.0158	0.8615 0.9242
	Month 12	171	3	0.8821	0.0178	0.8420 0.9126
Low-dose cyclosporine						
	Baseline	0	0	1	-	-
	Month 3	353	15	0.9609	0.0099	0.9359 0.9762
	Month 6	321	5	0.9464	0.0117	0.9180 0.9651
	Month 9	196	4	0.931	0.0139	0.8978 0.9536
	Month 12	181	3	0.9165	0.016	0.8788 0.9428
Tacrolimus						
	Baseline	0	0	1	-	-
	Month 3	349	18	0.9529	0.0109	0.9262 0.9701
	Month 6	313	3	0.9439	0.0119	0.9151 0.9631
	Month 9	180	2	0.9377	0.0126	0.9077 0.9583
	Month 12	171	0	0.9377	0.0126	0.9077 0.9583

Figure 62: Graft survival censored for DWFG - secondary analysis using SNFTM with g-estimation



6.8.1.5 *RMST estimates from IPCW, SNFTM and ITT analyses*

The effectiveness results in terms of RMSTs based on IPCW compared to ITT estimates are provided in Table 26. RMST estimates from SNFTM compared to ITT estimates are provided in Table 27. These estimates were not used in the economic model, but they provide more information in terms of comparing adjusted and unadjusted effectiveness estimates. The results demonstrated that accounting for real-world adherence levels reduces the RMSTs by 1.65 to 2.53% compared to ITT analysis.

Table 26: Graft restricted mean survival times: unadjusted ITT estimates versus IPCW adjusted estimates

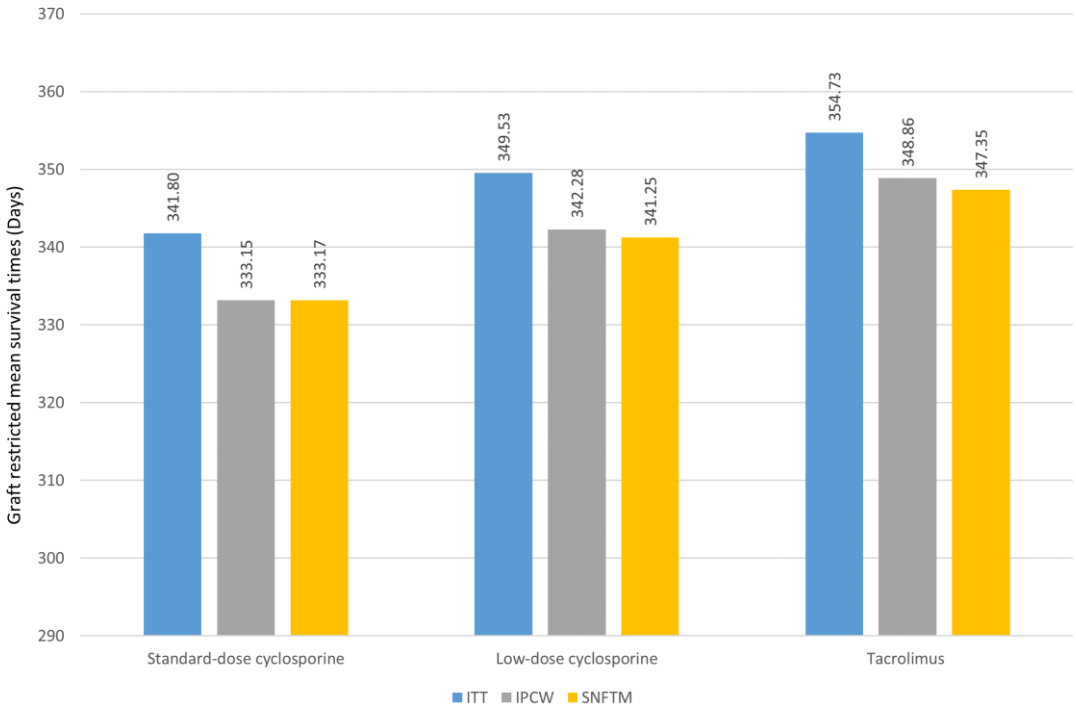
Treatment arm	RMST (SE) – ITT (Days)	RMST (SE) – IPCW (Days)	Difference in RMSTs between IPCW and ITT (Days)	Difference in RMSTs between IPCW and ITT (%)
Standard-dose cyclosporine	341.80 (3.888)	333.15 (4.649)	-8.65	-2.53
Low-dose cyclosporine	349.53 (2.027)	342.28 (2.518)	-7.26	-2.08
Tacrolimus	354.73 (2.363)	348.86 (3.238)	-5.87	-1.65

Table 27: Graft restricted mean survival times: unadjusted ITT estimates versus SNFTM adjusted estimates

Treatment arm	RMST (SE) – ITT (Days)	RMST (SE) – SNFTM (Days)	Difference in RMSTs between SNFTM and ITT (Days)	Difference in RMSTs between SNFTM and ITT (%)
Standard-dose cyclosporine	341.80 (3.888)	333.17 (4.607)	-8.62	-2.52
Low-dose cyclosporine	349.53 (2.027)	341.25 (2.498)	-8.28	-2.37
Tacrolimus	354.73 (2.363)	347.35 (3.346)	-7.38	-2.08

Figure 63 provides the RMSTs generated from IPCW, SNGTM and ITT by treatment group for comparison. The results show that ITT analysis overestimates treatment effectiveness when accounting for real-world adherence levels. The results also show that IPCW and SNFTM produce similar results using real data and this is consistent with the findings from the simulation study.

Figure 63: Graft restricted mean survival times from IPCW, SNFTM and ITT by treatment group



6.8.2 Cost-effectiveness results

The results in terms of treatment of cost-effectiveness estimates (unadjusted and adherence-adjusted estimates) are presented in the following subsections.

6.8.2.1 Unadjusted cost-effectiveness results

Table 28 presents the total lifetime discounted costs and QALYs, incremental costs and QALYs from a fully incremental analysis and NHBs for the three alternative immunosuppressive regimens from the unadjusted ITT analysis. The results from the unadjusted analysis show that the tacrolimus regimen is the most cost-effective option producing an NHB of 6.64 QALYs over the patient lifetime at the willingness to pay (WTP) threshold of £20,000 per QALY. This is followed by low-dose cyclosporine and standard-dose cyclosporine regimens producing NHBs of 6.10 and 5.75 QALYs per patient, respectively.

The results show tacrolimus regimen dominated low-dose and standard-dose cyclosporine (both cyclosporine regimens were more costly and less effective compared with tacrolimus).

Table 28: Net health benefits and incremental costs and QALYs from the ITT unadjusted analysis

Regimen	Total discounted cost	Total discounted QALYs	NHB (WTP £20K	Incremental costs	Incremental QALYs
Bas+Tac+MMF+ST	£90,830.29	11.0015	6.4600	-	-
Bas+CsA+MMF+ST	£97,713.35	10.9871	6.1015	£6,883.06	-0.0143
CsA+MMF+ST	£101,636.32	10.8342	5.7524	£3,922.97	-0.1529

Bas= Basiliximab, Tac= tacrolimus, MMF= mycophenolate, ST= corticosteroids, NHB= net health benefit, WTP= willingness to pay, QALYs= quality-adjusted life years.

6.8.2.2 Adherence-adjusted cost-effectiveness results from IPCW

Table 29 presents the adherence-adjusted cost-effectiveness results generated from the IPCW estimates. The results show that the NHBs estimated from the real-world adherence-adjusted analysis are smaller than those obtained from the ITT unadjusted analysis. For the tacrolimus regimen, the NHBs estimated from the adjusted analysis is reduced by 0.0596 QALYs per patient. For the low-dose cyclosporine regime, the reduction on NHBs when adjusting for real-world non-adherence is 0.0721 QALYs per patient.

The total discounted cost per patient is higher for the tacrolimus regimen (the most cost-effective option). For the tacrolimus regimen, the total discounted cost increased by £2,429 per patient when the real-world non-adherence levels are taken into account. The results show that the total discounted costs are also higher for other alternative regimens and the NHBs are smaller when adjusting for real-world non-adherence levels (see Table 29). The percentage change in NHBs compared to ITT is provided in Section 6.8.2.3 alongside all sensitivity analysis for comparison. There was no change in the cost-effectiveness conclusions predicted by the original model.

Table 29: Net health benefits and incremental costs and QALYs from IPCW adjusted analysis

Regimen	Total discounted cost	Total discounted QALYs	NHB (WTP £20K	Incremental costs	Incremental QALYs
Bas+Tac+MMF+ST	£93,259.20	10.9419	6.2790	-	-
Bas+CsA+MMF+ST	£100,697.51	10.9150	5.8801	£7,438.31	-0.0270
CsA+MMF+ST	£105,407.90	10.7455	5.4751	£4,710.39	-0.1695

Bas= Basiliximab, Tac= tacrolimus, MMF= mycophenolate, ST= corticosteroids, NHB= net health benefit, WTP= willingness to pay, QALYs= quality-adjusted life years.

When compared to the ITT unadjusted analysis, the IPCW adjusted cost-effectiveness results show a noticeable impact of real-world non-adherence on the incremental results. For low-dose cyclosporine

versus standard-dose cyclosporine regimens, an increase of 8.1% in incremental costs and a reduction of 88.2% in incremental QALYs were found compared to the ITT unadjusted analysis. For tacrolimus versus low-dose cyclosporine regimens, an increase of 20.1% in incremental cost and a reduction of 10.8% in incremental QALYs were found from the real-world adjusted estimates compared to the ITT unadjusted analysis.

6.8.2.3 Adherence-adjusted cost-effectiveness results from sensitivity analyses

Table 30 present the IPCW adjusted cost-effectiveness results from the sensitivity analyses with the results from the adjusted base-case analysis provided at the top of the table from comparison. The table also shows the percentage change in NHBs in the adjusted analysis compared to the ITT analysis. The case study findings demonstrated that accounting for real-world adherence levels lead to a reduction in the expected health benefits when adherence levels in the real world are worse than those observed in the trial.

Table 30: Net health benefits and incremental costs and QALYs from sensitivity analyses

Regimen	Total discounted cost	Total discounted QALYs	NHB (WTP £20K)	Incremental costs	Incremental QALYs	% Change in NHBs
Base-case analysis (IPCW adjusted)						
Bas+Tac+MMF+ST	£93,259.20	10.9419	6.2790	-	-	-2.80
Bas+CsA+MMF+ST	£100,697.51	10.9150	5.8801	£7,438.31	-0.0270	-3.63
CsA+MMF+ST	£105,407.90	10.7455	5.4751	£4,710.39	-0.1695	-4.82
Sensitivity analysis: Using differential real-world non-adherence levels						
Bas+Tac+MMF+ST	£92,512.17	10.9522	6.3266	-	-	-2.06
Bas+CsA+MMF+ST	£100,095.06	10.9218	5.9170	£7,582.89	-0.0304	-3.02
CsA+MMF+ST	£104,925.53	10.7489	5.5026	£4,830.48	-0.1729	-4.34
Sensitivity analysis: Assuming perfect adherence between 0-3 months						
Bas+Tac+MMF+ST	£93,149.31	10.9586	6.3012	-	-	-2.46
Bas+CsA+MMF+ST	£100,615.43	10.9301	5.8993	£7,466.13	-0.0286	-3.31
CsA+MMF+ST	£105,285.48	10.7604	5.4962	£4,670.04	-0.4031	-4.45
Sensitivity analysis: Adjusting for non-adherence between 6-12 months only						
Bas+Tac+MMF+ST	£93,172.47	10.9473	6.2886	-	-	-2.65
Bas+CsA+MMF+ST	£100,551.69	10.9213	5.8937	£7,379.22	-0.0260	-3.41
CsA+MMF+ST	£105,233.75	10.7513	5.4896	£4,682.06	-0.4040	-4.57

NHB= net health benefit, WTP= willingness to pay, Bas= basiliximab, TAC= tacrolimus, CsA= cyclosporine, MMF= mycophenolate, ST= corticosteroids, QALYs= quality-adjusted life years. IPCW= inverse probability of censoring weighting.

6.8.2.4 Adjusted cost-effectiveness results from SNFTM with g-estimation

Table 31 present the adjusted cost-effectiveness results from the SNFTM secondary analysis. The table shows the percentage change in NHBs compared to the results from the ITT analysis. The results are very close to the IPCW estimates (see Section 6.8.2.2).

Table 31: Net health benefits and incremental costs and QALYs from SNFTM adjusted analysis

Regimen	Total discounted cost	Total discounted QALYs	NHB (WTP £20K	Incremental costs	Incremental QALYs	% change in NHBs
Bas+Tac+MMF+ST	£93,979.35	10.9443	6.2453			-3.32
Bas+CsA+MMF+ST	£100,720.43	10.9299	5.8939	£6,741.08	-0.0143	-3.40
CsA+MMF+ST	£105,617.05	10.7555	5.4747	£4,896.62	-0.1744	-4.83

Bas= Basiliximab, Tac= tacrolimus, MMF= mycophenolate, ST= corticosteroids, NHB= net health benefit, WTP= willingness to pay, QALYs= quality-adjusted life years.

6.9 Discussion and conclusions

6.9.1 Summary of findings from the case study

The case study applied the g-methods (IPCW and SNFTM) for estimating the adherence-adjusted cost-effectiveness of immunosuppressive therapy among adult kidney transplant recipients in the UK. The analysis used individual-patient level data from SYMPHONY (a large multicentre RCT with data from 1,190 patients) combined with real-world implementation non-adherence levels for assessing three immunosuppressive regimens commonly used in the UK. The regimens assessed consisted of standard-dose cyclosporine, low-dose cyclosporine and tacrolimus as the maintenance immunosuppressants for which the non-adherence adjustment was applied in this analysis.

In this case study, graft survival functions were re-estimated by applying the g-methods on the SYMPHONY dataset. The IPCW adjusted estimates were used in the base-case economic analysis with SNFTM used in a secondary analysis to show how this adjustment method could be applied in real data. Further sensitivity analyses were performed using three different ways to deal with non-adherence data recorded during the first 6 months post-transplantation. The analyses used the CV% which was calculated from the trough levels data recorded in the SYMPHONY trial and combined these with real-world adherence levels to predict non-adherence for each individual patient in the trial dataset. Real-world adherence was predicted by using a new cut-off point for CV% above which patients were classified as non-adhered such that the overall probability of non-adherence in the predicted dataset matches the levels observed in the real world. Causal analysis using g-methods was then applied to produce the adherence-adjusted effectiveness estimates. Subsequently, the adjusted

effectiveness estimates were used in an adapted economic model to produce the adherence-adjusted cost-effectiveness of the alternative immunosuppressive regimens.

The cost-effectiveness results generated from the adjusted analysis were compared with estimates from the unadjusted ITT analysis of SYMPHONY trial data. The key findings show that unadjusted ITT analysis overestimates the NHBs and underestimates the total cost for all regimens. As an alternative approach, the g-methods provides a practical framework for correcting the cost-effectiveness estimates by taking into account the real-world adherence levels for each regimen. The case study demonstrated that it is possible to incorporate different levels of real-word adherence using the g-methods. This provides better cost-effectiveness evidence for resource allocation decision making relating to maintenance immunosuppressive therapy by taking into account their real-world adherence levels. This analysis framework is potentially transferable to other disease areas where non-adherence to the prescribed dosing regimens is identified as an issue.

The key findings of the case study show reduced net health benefits and increased costs per patient for the alternative regimens when real-world implementation non-adherence levels are taken into account in the cost-effectiveness analysis. The findings demonstrated that using estimates of treatment effectiveness adjusted for real-world non-adherence alters the results in terms of total lifetime discounted costs and QALYs and incremental costs and QALYs, although the cost-effectiveness conclusions have not changed in this particular case study. The tacrolimus regimen remains the most cost-effective immunosuppressive therapy in kidney transplantation, although the NHB estimated by the IPCW adjusted analysis is predicted to be smaller by 0.0596 QALYs with £2,429 higher cost per patient compared to the standard ITT analysis when the real-world adherence levels are taken into account.

The analysis showed that the adjustment methods make graft survival worse by accounting for the potentially higher levels of the predicted real-world non-adherence. This is consistent with the whole hypothesis that adherence levels in the real world are worse than trials and by accounting for it; we expect the treatment effectiveness to be worse than the ITT estimates. This is demonstrated by the findings of the case study presented in this chapter. The findings demonstrate that accounting for real-world non-adherence in the SYMPHONY trial dataset resulted in reductions in the expected graft survival times by about 6-9 days (depending on the treatment group) during the first 12 months post-transplantation. The impact on cost-effectiveness was reduced net health benefits with a higher cost per patient when the real-world adherence levels are taken into account. For low-dose cyclosporine versus standard-dose cyclosporine, the impact amounted to an 8.1% increase in incremental costs and 88.2% reduction in incremental QALYs compared to the unadjusted ITT analysis. For tacrolimus versus

low-dose cyclosporine regimens, the incremental cost per patient was higher by 20.1% with a 10.8% reduction in incremental QALYs compared to the ITT estimates.

In the economic model, the knock-on effects of non-adherence on costs are more likely to come from the worse outcome, which may result in higher total costs. This happens automatically in the economic model because people move to a different health state that has higher costs (e.g. graft loss health state) and a lower utility in terms of quality of life as a result of moving to kidney dialysis intensive treatment on a regular basis.

6.9.2 Strengths and limitations of the case study

There are strengths and limitations of this case study. The main strengths include the application of the adjustment methods (selected based on the new evidence of performance from the simulation study reported in Chapter 5) using data from a large well-conducted RCT using data from 1,190 patients with 12 months follow-up. The SYMPHONY trial dataset included the variables of interest required for the application of g-methods including measurement of the important baseline covariates (age and gender), time-dependent confounders (BMI and acute rejection) and a time-to-event outcome (i.e. time to graft loss). Based on the DAG (described in Section 6.3.4), it seems the important time-dependent confounders were collected in SYMPHONY; and therefore, the assumption of no unmeasured confounding is likely to be met. This is an important point for consideration in the design of future studies intending to adjust for non-adherence. Data on patient survival (time to death) was also available in the SYMPHONY dataset and this allowed for estimating the probability of graft survival censored for DWFG to match the parameters used in the adapted economic model.

The availability of individual patient-level records on trough levels for the three immunosuppressants measured at 10 study protocol visits as well as extra visits to clinics allowed for the real-world non-adherence to be predicted in a way that takes into account the relationships between adherence and important prognostic characteristics observed in the trial. These included time updated BMI and acute rejections. Recent estimates of real-world adherence levels using CV% from two UK studies that assessed SYMPHONY-style tacrolimus and low-dose cyclosporine regimens provided the data required for predicting real-world adherence within the SYMPHONY dataset. This allowed for the planned causal analysis to be performed on a randomised dataset for producing valid adherence-adjusted effectiveness estimates (i.e. the intended estimates) to populate the economic model. Another important strength of the case study was using an adaptation of a validated economic model which underpinned the recent update of NICE Technology Appraisal guidance for immunosuppressive therapy for adults in the UK (TA481).[120]

There are some limitations of the cases study. The CV% was used as a measure of adherence in the case study, although it is not the perfect or most recommended measure for implementation adherence and this may be considered as a limitation. The main reason for using CV% in this case study was that trough levels was the only adherence data measured in the SYMPHONY trial. Indeed, this also influenced the choice of CV% as a measure of real-world adherence so that the prediction within the SYMPHONY dataset could be performed. Moreover, CV% is commonly used and well accepted as a proxy measure of adherence in many disease areas including kidney transplantation. The real-world estimates of standard-dose cyclosporine were not identified from the review and this was assumed similar to the low-dose cyclosporine which is a conservative assumption.

When considering the generalisability of the findings from this case study to adjust for non-adherence beyond kidney transplantation, a range of adherence measures are available for consideration depending on the disease area, type of medications assessed and study population. Based on preliminary findings from an ongoing ESPACOMP (International Society for Medication Adherence) Delphi panel, the following adherence measures are recommended for each type of non-adherence:[146]

- Initiation adherence: prescription refill data or questionnaires
- Implementation adherence: Electronic monitoring devices (e.g. Medication Event Monitoring System [MEMS])
- Persistence adherence: questionnaires or Medication Possession Ratio (MPR)

Initiation and persistence types of non-adherence are generally expressed as binary variables derived from the abovementioned measures, and these could be implemented for adjusting effectiveness and cost-effectiveness in a similar way applied in the simulation study (Chapters 4-5). While implementation non-adherence data such as trough levels could be analysed to derive a binary variable as applied in this case study, alternative approaches might be needed to transform data collected by other types of adherence measures, such as MEMS, to a binary variable to allow for adjustment methods to be applied.

Another limitation associated with using CV% is the ability to use a different measure of adherence to predict real-world adherence levels within the randomised dataset. The approach used was discussed with clinical advisors and they considered that it would provide a realistic prediction of real-world adherence within the trial dataset. However, this approach implicitly assumes that the pharmacokinetics data collected in the trial is a good predictor of real-world adherence (patients with higher CV% in the trial are more likely to become non-adherent in the real world) which may not hold. It is also very difficult to assess the accuracy of the prediction and this is likely to be an issue for other prediction methods.

As an alternative approach, the marginal standardisation method might be considered in scenarios where trough levels data (such as the data used in this case study) is not available.[147] The marginal standardisation method is considered a special case of the g-computation method.[147] While this method has not been tested in this case study due to the type of adherence data available in SYMPHONY, the method relies on predicting the marginal probabilities of non-adherence for each individual patient over the distribution of observed confounders and a counterfactual adherence level. This could be applied by setting the probability of adherence to a counterfactual value (real-world adherence) for all patients in the RCT dataset. Then, confounder-adjusted logistic regression could be fitted using the observed data to compute the predicted probabilities of non-adherence for all patients at the observed values of confounders and the newly assigned (counterfactual) adherence level. This will create real-world predicted probabilities of adherence in the RCT dataset and then estimates of effectiveness could be obtained by applying the same approach used in this case study.

Another limitation relates to medication adherence data recorded in the trial during the first 3 months post-transplantation. To calculate the CV% from trough levels in SYMPHONY, at least three data points within each time interval (e.g. 0-3, 3-6 and 6-12 months post-transplantation) is required to generate a valid adherence estimate. This means for each individual patient in trial, a minimum of nine equally distributed trough level records is required. In SYMPHONY, some patients do not meet this data requirement and as a result, missing CV% was generated for those patients. Moreover, for all patients, the trough levels data recorded between baseline and 3 months post-transplantation were discarded based on recommendations from published guides and discussions with two clinicians. The implications were that all patients in the SYMPHONY dataset had missing CV% generated at the first interval (baseline to 3 months). The decision for discarding that data was due to several factors including fluctuations due to infections and intravenous steroid treatments.[139] The implication of these two related issues is that non-adherence adjustment was not applied for those patients which effectively means assuming they adhered to implement the prescribed treatment and this should be considered as a limitation. However, three sensitivity analyses were performed to address the issue. The key learning point from this issue highlights the importance of collecting medication adherence data in the trial to allow for the non-adherence adjusted analysis to be performed properly and avoid making such assumptions.

The issue of multiple medication adherence (MMA) was not addressed in this case study which could be considered as a limitation given that multiple treatments were used within each immunosuppressive regimen assessed in the case study. Addressing MMA in the analyses undertaken in this case study was not possible because adherence data for the other medications (BAS, MMF, ST) were not collected in the SYMPHONY trial. This issue of dealing with MMA has been identified as an

area with no consensus exists in the health economics literature as characterised by a recent ISPOR special interest group report.[135] This can be considered as an area for future research when it comes to accounting for non-adherence to multiple medications on the cost-effectiveness of prescribed chronic medications.

Dichotomised adherence variable was incorporated in the adjusted analyses using the g-methods in which a patient was either adhered on not adhered. However, there are alternative approaches used in the health economics literature for incorporating implementation non-adherence, although these were not related to the g-methods. These include stratifying patients by different levels of adherence (e.g. >80%, 60-80%, or <60% as characterised earlier) and then assign reduced treatment effects for patients with lower levels of adherence. This approach might not be appropriate given the way the dosing schedules are designed to maintain the drug concentration with the therapeutic window to achieve the intended therapeutic effect. In other wards for particular medications, adherence below 80% is more likely to result in a therapeutic failure rather than a reduced treatment effect. This approach was heavily criticised in the medication adherence literature. MEMS is the recommended measure of implementation adherence (based on the ESPACOMP Delphi panel) which is different from the CV% used in this case study.[146] The analysis of medication adherence data collected using MEMS tools is generally presented in a form of percentages of tablets used over a specific time interval. Then a cut-off point will be required to assign each patient above the threshed to a binary non-adherence variable in the same way applied in this case study using the CV%. The key point is the requirement for objective determination of the cut-off point to avoid introducing bias in the analysis. However, this approach has not been empirically assessed in this case study, and therefore, incorporating implementation non-adherence as a non-binary variable could be an interesting area for future research. Another related limitation is that the censoring mechanism used in the adjusted analysis has not previously been tested. This was based on the predicted non-adherence as opposed to censoring patients who actually did not adhere to the treatment in the trial. Further research should also consider assessing the censoring mechanism alongside the real-world non-adherence prediction methods, ideally in a well-conducted simulation study.

Although the discarded medication adherence data recorded between 0-3 months post-transplantation could be considered as a limitation of this case study, three approaches were used in terms of sensitivity analyses to address the issue. These analyses suggested that the adherence-adjusted cost-effectiveness results generated from the base-case analysis are robust. This indicates that the data on CV% from SYMPHONY based on trough levels measured between 3 and 12 months were adequate for performing the planned adjusted analyses in this case study. Furthermore, and based on discussion with clinicians, it could be argued that most graft losses that occur at the first

three months post-transplantation (the interval where we do not have usable trough levels data) are more likely to be due to technical rejections rather than non-adherence to the prescribed immunosuppressive therapy.

Missing data on the baseline and time-dependent confounders could be considered as a limitation, although this was a minor issue in the SYMPHONY dataset. With age missing data of only 3.5% (42 out of 1,190 patients) and BMI at baseline missing data of 13.2% (157 out 1,190 patients) which are evenly distributed by treatment arm. Missing data also applies to time-dependent confounders where BMI was the main variable that had missing data. Nevertheless, missing data for baseline and time-dependent confounders were imputed using acceptable approaches; and therefore, the risk of introducing bias as a result of missing data is minimal in this particular case study. The data imputation allowed for the full dataset of all patients randomised in the three relevant arms of the SYMPHONY trial to be included in both adjusted and unadjusted analyses.

In this case study, the LOCF method was used to impute missing data relating to time-dependent confounders and this was considered the most appropriate option for this particular case study. To aid transferability of this analysis for other disease areas where the LOCF method might not be appropriate, the alternative approaches are briefly discussed here with references provided from the methodological literature further information. Missing data is a common problem in clinical trials and real-world data and the methodological literature on dealing with this issue is well established.[148-151] Guidance and tools to address the issue which could be considered including multiple imputations and complete case analysis.[150]

Finally, subgroup analyses are commonly performed in the analysis of clinical trials; however, this was not planned for this case study. The main issue with subgroup analyses is that trials are generally not powered to detect a difference in treatment effect based on a smaller sample size specifying a subgroup of participants. Therefore, evidence from subgroup analysis is unlikely to be used to inform resource e allocation decision making in health care. However, these subgroup analyses are commonly used to inform future research where the technology might show potential for working better among a specific group of the study population. In the simulation study, the performance of the g-methods was negatively affected by the small sample size leading to higher bias. G-methods may also run into convergence problems when a small sample size is used as observed in the simulation study. This is likely to be the case for subgroup analyses, although this has not been assessed in this case study (see Chapter 5).

6.9.3 Conclusions

Based on the analysis undertaken in this case study, a number of conclusions can be drawn about the application of g-methods for estimating adherence-adjusted clinical effectiveness and cost-effectiveness of prescribed chronic medications. First, adjusting for real-world implementation non-adherence for immunosuppressive regimens commonly used in the UK did result in reduced estimates of health benefits and increased costs per patient compared to using the standard ITT estimates in cost-effectiveness analysis. The impact of real-world non-adherence on incremental results was noticeable in this case study. For low-dose cyclosporine versus standard-dose cyclosporine, the impact amounted to an 8.1% increase in incremental costs and 88.2% reduction in incremental QALYs compared to the unadjusted ITT analysis. For tacrolimus versus low-dose cyclosporine regimens, the impact was higher in terms of increased incremental cost by 20.1% with a 10.8% reduction in incremental QALYs compared to the unadjusted ITT analysis.

Second, g-methods provide a practical framework for incorporating real-world adherence levels to predict the cost-effectiveness of alternative regimes using individual patient-level data from an RCT dataset. Availability of the required data including measurements of adherence to the prescribed medications, baseline and time-dependent confounding within the RCT dataset as well as data on real-world adherence levels is crucial to implementing the adherence-adjusted analysis. Third, the type of non-adherence measure (e.g. CV% based on PK data) represents a key element in predicting real-world adherence in this particular case study and there is some uncertainty in the prediction method. Finally, missing data, including data on adherence to medication as well as baseline and time-dependent confounding represent a risk as g-methods rely on data from these variables to produce valid estimates of adherence-adjusted effectiveness and cost-effectiveness results. Predicted non-adherence between 6 and 12 months post-transplantation was close to real-world estimates; however, potential differences between an adherer in the adjusted trial dataset, compared to an adherer in the real world, is possible given that I had to change the CV% cut-off in the trial to get the desired non-adherence.

Chapter 7: A methodological framework to account for the impact of non-adherence on the cost-effectiveness of prescribed chronic medications

7.1 Introduction

Chapter 6 presented the application of g-methods in a case study for estimating adherence-adjusted cost-effectiveness of immunosuppressive therapy in kidney transplantation using data from the SYMPHONY trial and real-world non-adherence within an adapted economic model. Evidence generated from the simulation study (Chapter 5) was used for selecting the appropriate non-adherence adjustment methods for application in the case study. The MSM with IPCW and SNFTM with g-estimation (g-methods) were applied in the case study. The chapter described the analytical steps and considerations for applying these g-methods using real data.

This chapter presents the work undertaken in Stage 4 of this doctoral research project aimed to develop a methodological framework to account for the impact of non-adherence on the cost-effectiveness of prescribed chronic medications. In this context, a "methodological framework" is defined as a tool to guide researchers through a sequence of steps and considerations for estimating adherence-adjusted cost-effectiveness to inform resource allocation decision making in health care.[152] The methodological framework provides a set of recommendations based on new evidence generated and presented in this thesis complemented with existing evidence from the literature. The purpose of the adjustment is to allow the estimation of treatment effects associated with adherence levels from outside the trial (i.e. real-world adherence levels), and this has a direct impact on the cost-effectiveness estimates. The chapter aims to answer the following research question: *"How should economic evaluations incorporate the impact of non-adherence using evidence from both RCTs and the real world?"*

This chapter puts forward a seven-stage methodological framework to account for the impact of real-world non-adherence on the cost-effectiveness of prescribed chronic medications. The framework is relevant to the context of survival analysis (studies with time-to-event outcomes) and HTA; and therefore, does not address studies with continuous or categorical outcomes.

The methodological framework is built on the work undertaken in Chapters 2-6. It has also been informed by other literature in the medical and health economics research fields. The latter was used

to inform important considerations as part of the framework (e.g. study population and the impact of non-adherence on direct drug costs).

This chapter is structured into the following sections. Section 7.2 describes the methods used for developing the methodological framework. Section 7.3 presents the methodological framework including the essential recommendations and considerations that could be followed to adjust for patient non-adherence in future economic evaluations. Section 7.4 provides the implications of the framework for health economics analysis plans (HEAPs). Section 7.5 discusses the methodological framework including the strengths and limitations and suggests key areas for future research.

7.2 Methods for developing the methodological framework

McMeeken et al.[152] published a recent paper entitled *"How methodological frameworks are being developed: evidence from a scoping review"*. The authors reviewed a set of 30 papers published in the last 10 years that reported a range of approaches for developing methodological frameworks. The authors concluded that no formal guidance exists, but three phases were suggested to inform the development of future methodological frameworks based on an overall consensus on approaches found from their scoping review. The development of the methodological framework presented in this chapter was informed by that review including the three-phase development process.

The methods used for developing the methodological framework followed the following three phases: 1) identifying existing frameworks and evidence to inform the development of the methodological framework; 2) developing the methodological framework; 3) informing the methodological framework using a case study on kidney transplantation. Phase 3 also discusses the transferability of the methodological framework to other disease areas. The following subsections describe the methods used for developing the methodological framework in greater detail.

7.2.1 Identifying evidence to inform the methodological framework

The process started with identifying evidence to inform the development of the methodological framework. This phase has two parts comprising: (a) a review to identify any existing methodological frameworks; and (b) a review of methodological findings from Chapters 2-6 of this thesis (i.e. the systematic review, the simulation study and the case study). These two parts are described in more detail in the following subsections.

7.2.1.1 Review to identify existing frameworks

To identify any existing methodological frameworks, a scoping review was conducted at an early stage when the plan of investigation for this doctoral research project was developed in 2017. The review was based on a citation search which was undertaken using Web of Science and Scopus electronic databases for three key papers (identified as relevant and highly cited papers in the literature) which address the issue of non-adherence.[26, 32, 69] The review identified five methodological approaches used for incorporating non-adherence in cost-effectiveness analysis as characterised by Hughes and colleagues.[32] In this context, a framework is different from an approach. A framework implies some set of steps that need to be followed to achieve a particular outcome. In contrast, the approaches identified by that review were all separate ways of attempting to account for non-adherence in pharmacoeconomic evaluations. These methodological approaches are partly based on structural model designs (e.g. Markov models and Decision trees) for implementation in pharmacoeconomic evaluations rather than on adjusting for non-adherence in treatment effectiveness estimates (see Chapter 1, Section 1.6.2).

Given the limitations of these approaches discussed in Chapter 1 (Sections 1.4 and 1.6), the early scoping review concluded that no methodological framework exists. Nevertheless, the findings from the scoping review helped in shaping the design of the systematic review of methodological papers undertaken and reported in this thesis (Chapter 2).

To update the findings from the early scoping review, a new literature review was carried out in 2021 to identify any existing methodological frameworks. A search strategy was developed and used to identify existing frameworks for addressing patient non-adherence to medications in the context of economic evaluation and HTA. The search was conducted using OVID MEDLINE(R) and Web of Science databases from inception to June 2021. The search used three sets of relevant terms which were combined with "AND": (1) terms for the framework (framework, methodology, guidelines, recommendations, or methods); (2) a comprehensive list of medication adherence terms similar to those used in the methodological systematic review reported in Chapter 2 of this thesis, and (3) terms for HTA (cost-effectiveness or economic evaluation or health technology assessment). All of these terms were limited to the title in the search to help identify relevant articles with a focus on methodological frameworks to account for non-adherence in the HTA context. The search terms used were informed by those used in the early scoping review, complemented with additional terms based on the McMeeken et al.[152] paper. The detailed search terms used in the updated review and the results

in terms of the number of records generated from each step of the searches are provided in Appendix A.

The database searches generated a total of 41 records of potentially relevant papers (MEDLINE(R) = 12, Web of Science = 29). Following de-duplication, a total of 31 potentially relevant papers was identified. Twenty-five records were excluded at the title screening stage, and then the abstracts for the remaining six records were reviewed. One abstract was identified as relevant.[153] The abstract is entitled "Framework for real-world economic evaluation by incorporating implementation parameters" and was published in 2016. This was a conference abstract and no full paper related to the abstract was identified. This abstract was missed by the early scoping review but has been picked up by the updated review because additional terms that were informed by the recently published guide were used as part of this iterative process.[152]

The abstract authors proposed a framework to allow step-wise considerations of net benefits assuming three possible scenarios (perfect world, real-world and improved world) in terms of implementation adherence. The proposed framework specifically aimed to aid early-stage economic evaluations using three hypothetical scenarios. The first scenario (perfect world) was designed to predict if the treatment could be cost-effective. The second scenario aims to predict cost-effectiveness in real-world adherence levels. The third scenario aims to predict a situation between perfect adherence and real-world adherence levels. Interestingly, the "clinical trial world", which is probably somewhere between perfect and real worlds was missing from this framework, because in reality that is the starting point with trial outcomes that the analyst would want to adjust up (to perfect) or won (to real) from. Furthermore, the authors reported using a Markov model to test their proposed framework in evaluating direct hearing aid provision versus provision by referral, incorporating patient adherence and professional uptakes as parameters in the economic model. The abstract reported cost savings based on the predicted improvement in adherence and concluded that the framework could help in terms of using real-world economic evaluations to inform policy decisions.

The proposed framework falls into the category of structural approaches to incorporate adherence in economic models. This is similar to the approaches identified earlier when the proposal for this doctoral research project was developed (as discussed above). The limitations of these approaches were discussed in greater detail in Chapter 1 (Section 1.4 and 1.6), and these limitations informed the motivation for undertaking this doctoral research. In brief, these approaches make simplistic assumptions about the causal relationships between non-adherence and treatment effect which may produce misleading cost-effectiveness estimates. This is because the impact of non-adherence on the

treatment effect is very complex and more likely to be determined by the interplay between several factors including the type and levels of non-adherence, the nature of the disease, the prognostic characteristics and characteristics of the medication (e.g. PKPD). Also, it is not consistent over time – for example, a patient may start with 95% adherence in the first year, but be lower than this in the next year. Therefore, more complex methods are needed to deal with different types of adherence and multiple time-invariant and time-variant confounding factors.

To investigate further, an author search was conducted to establish whether the first author of the relevant abstract had published any methodological framework at any point in time that addresses the issue of non-adherence. The author search was undertaken via Web of Science on "Grutters JP" [153] without any restrictions in terms of dates or other search terms. A total of 14 papers were found by the author search and after title screening, no relevant paper reporting a methodological framework was identified.

In summary, no existing relevant methodological framework was identified from the reviews. Therefore, the gap in the methodological literature in terms of accounting for non-adherence in economic evaluation remains.

7.2.1.2 Recommended non-adherence adjustment methods to inform the methodological framework

The recommendation of non-adherence adjustment methods in the framework is based on findings from three linked studies undertaken in this research as reported in Chapters 2-6 and these are briefly summarised in this section.

First, a systematic review of the methodological literature identified 12 statistical methods for adjusting estimates of treatment effectiveness for patient non-adherence. The review concluded that four methods (g-methods [IPCW, SNFTM, RPSFTM] and PKPD) appear to be more appropriate to estimate treatment effects in the presence of real-world non-adherence.[61]

Second, a simulation study assessed a subset of two g-methods (IPCW and SNFTM) and two simple methods and excluded the RPSFTM and PKPD methods. The PKPD method was excluded, as it requires the specification of the surrogate-final endpoint relationship which adds an additional layer of complexity and uncertainty. This needs a different set of DGMs rather than those applied in the simulation study, therefore, it was not possible to directly compare it with IPCW and SNFTM and this has been identified as a key area for further research. The RPSFTM only works to adjust for non-adherence for one-arm studies or studies with a placebo control arm. However, the simulation study

was designed to assess the performance of adjustment methods in RCTs with two active treatments and hence the RPSFTM was also excluded from the assessment in the simulation study. In the end, two g-methods (IPCW and SNFTM) and two simple methods (ITT and PP) were assessed in the simulation study. The key findings show both g-methods are the best performing methods for adjusting estimates of treatment effectiveness for non-adherence.

Third, a case study that applied the two best-performing g-methods (IPCW and SNFTM) and estimated the adherence-adjusted cost-effectiveness of immunosuppressants in kidney transplantation (see Section 7.2.3)

Therefore, the methodological framework was informed by evidence from the findings of these three linked studies. The recommendation for using the non-adherence adjustment methods as part of the methodological framework is provided in Section 7.3.4. The framework is meant to be "live" and updated as and when further research is done (see Section 7.5.3) as I have not been able to cover everything within this doctoral research project.

7.2.2 Developing the methodological framework

Seven stages were identified as key elements to outline the methodological framework. The stages were formulated from the practical experience of applying the adjustment methods in this doctoral research. This was further informed by the design of existing methodological frameworks developed to address other issues to improve the quality of economic models in the HTA context.[154, 155] The initial outline of the methodological framework was discussed with supervisors and based on comments, a second version of the framework was shared for further comments. The second version of the framework was further discussed with a clinician (WM) and a leading expert in medication adherence research (DH). All comments received were used to update the methodological framework and the final version is presented in this chapter.

The methodological framework put forward in this thesis involves seven stages: (1) understanding the clinical context; (2) identifying and measuring the relevant types of non-adherence; (3) specifying other data requirements and making assumptions explicit; (4) selecting the appropriate non-adherence adjustment method (5); adjusting treatment effectiveness for real-world non-adherence levels; (6) considering the impact of non-adherence on direct treatment costs; and (7) estimating and reporting adherence-adjusted cost-effectiveness of treatments. A description of the seven stages and the recommendations within each stage are presented in Section 7.3.

7.2.3 Informing the methodological framework using a case study

A case study on kidney transplantation (Chapter 6) was used to jointly formulate and apply the framework using real data. The case study was used to show how the methods could be applied in practice to produce cost-effectiveness estimates adjusted for real-world non-adherence. The study used individual patient-level data (IPD) from a large multicentre RCT (The SYMPHONY trial) with real-world adherence levels estimates obtained from the literature within an adapted economic model. The case study was used to inform the development of the methodological framework presented in this chapter.

A key feature of the methodological framework is its potential transferability to other disease areas. Although the framework was developed and applied to a single case study in kidney transplantation, it is transferable to other chronic diseases where non-adherence to prescribed regimens is identified as an important issue and provided that the endpoint is time-to-event. However, I argue that most aspects of the framework would still be appropriate for use in studies with non-time-to-event outcomes, although the analyst will need different methods to make the non-adherence adjustment to the treatment effectiveness estimates (e.g. different variants of g-methods).[156]

The methodological framework will apply to all types of non-adherence as characterised by the ABC medication adherence taxonomy (initiation, implementation, persistence), and therefore, it is fully aligned with that widely used taxonomy.[1] The recommendations and considerations in the methodological framework are generic; therefore, will apply to virtually any chronic disease with long-term use of prescribed pharmacological interventions where non-adherence is identified as an issue. However, it should be noted that the methodological framework has not yet been applied to other disease areas. The application of the methodological framework to other disease areas is encouraged and this thesis and the publications that will come out of it will form useful resources to aid the transferability. This in turn will help in evaluating the methodological framework in practice; however, this evaluation falls outside the scope of this thesis and represents an area for further research.

7.3 A seven-stage methodological framework to account for the impact of non-adherence on the cost-effectiveness of chronic medications

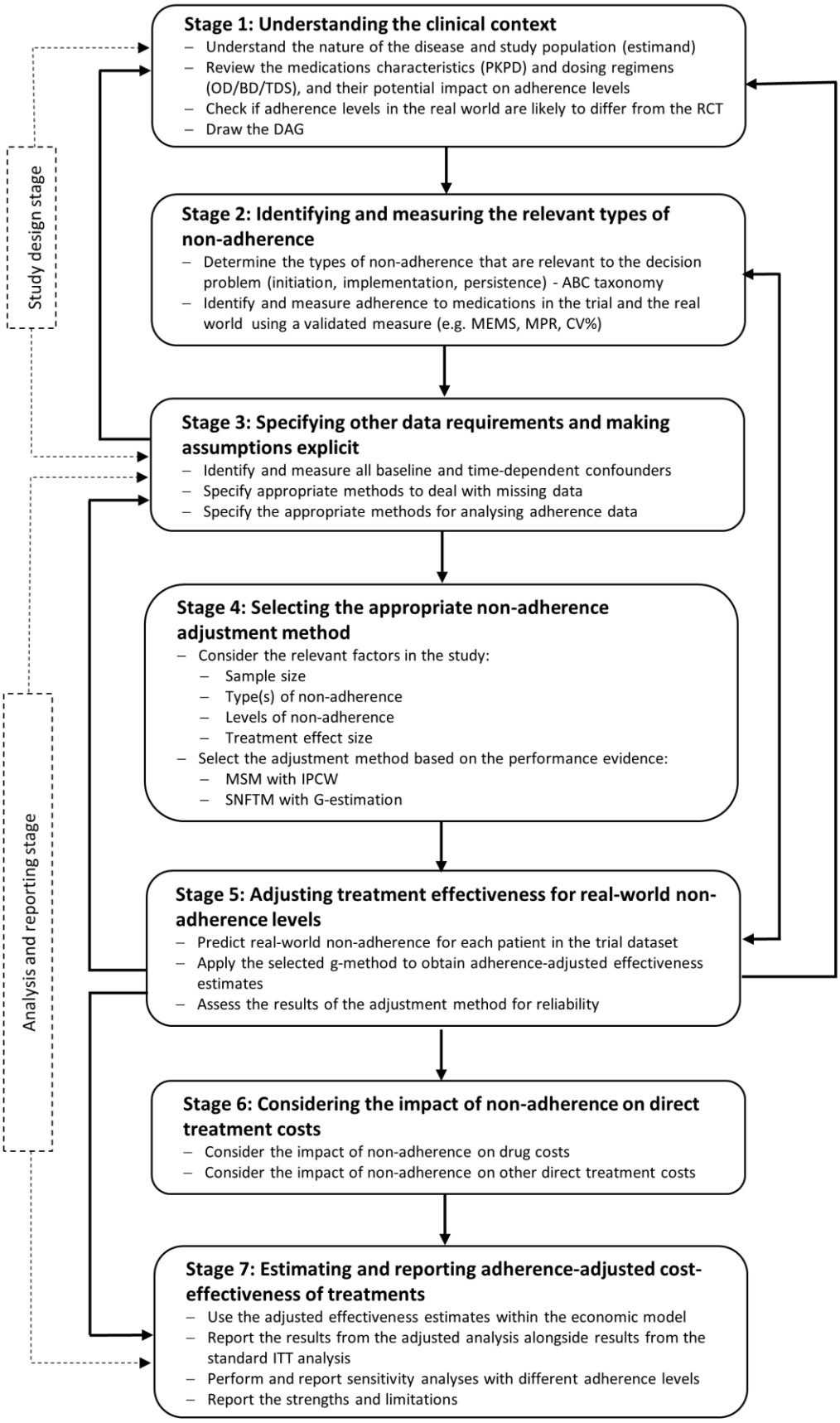
The methodological framework is presented in Figure 64. The stages in the methodological framework flow diagram are linked with arrows illustrating the process that could be followed to adjust for real-world non-adherence levels in economic evaluations. Although the stages are generally sequential in

terms of the steps that need to be followed, iterations may be necessary as illustrated by the arrows shown on both sides of the framework diagram (see Figure 64). The stages of the methodological framework are described in further detail in Sections 7.3.1 to 7.3.7 below.

The recommendations indicate specific measures that need to be taken based on findings from this research; whereas considerations identify important issues that may be considered with further measures taken based on what the investigator/analyst find in their particular study. For example, as a recommendation, the DAG should be drawn in for each study to apply this framework. An example of a consideration relates to the impact of non-adherence on drug costs where this will be drug/disease-specific; so the economist should consider it and include it if applicable. In the subsequent explanations, recommendations are *italicised* to make them more prominent.

The considerations and recommendations cover the whole process of designing and analysing RCTs for adjusting for non-adherence in economic evaluations. These are split into two parts: (a) Stages 1-3 cover nine recommendations for the RCT design (e.g. deciding what adherence measure is relevant and then collecting data on it, deciding what variables are relevant and collecting data on them); and (b) Stages 4-7 cover eight recommendations for analysing RCTs after data collection and these also apply to analysing existing RCT datasets (e.g., selecting and applying the appropriate adjustment method to produce adherence-adjusted effectiveness and cost-effectiveness estimates). Therefore, considerations and recommendations covered in Stages 1-3 should be undertaken at an early stage, ideally when the study is designed to ensure that the data required to perform the adherence-adjusted analysis is collected correctly. Some recommendations in Stage 3 are relevant to both study design and analysis and therefore the arrows are pointing to both parts in the framework diagram.

Figure 64: A seven-stage methodological framework to account for the impact of non-adherence on the cost-effectiveness of chronic medications in the context of time-to-event outcomes



7.3.1 Stage 1: Understanding the clinical context

The first stage in the methodological framework involves understanding the clinical context to help inform the study design in terms of adjusting for non-adherence in the analysis. This stage involves four recommendations as shown in the methodological framework flow diagram (Figure 64). These are described in the following subsections.

a) Understand the nature of the disease and the characteristics of the study population

These may include prognosis, comorbidities and patients age groups with a focus on understanding how these prognostic characteristics may influence patient adherence to the prescribed dosing regimens of the medications used to treat these diseases or conditions. The age distribution of the study population is particularly important as different age groups are likely to have different patterns of adherence for certain types of treatments such as maintenance immunosuppressive therapy after kidney transplantation. For example, in the case study, it was noted that patients aged 19-24 years have significantly higher probabilities of non-adherence to immunosuppressants compared to older patients (25-44 years).[157] Age-related non-adherence differences are also evident in other disease areas such as chronic heart failure.[158] Furthermore, polypharmacy for patients with comorbidities is an important factor for consideration as it has a significant impact on medication adherence behaviours. The key point is that polypharmacy can lead to higher levels of non-adherence because patients have greater numbers of tablets to take which increases the chance of missing doses.[159] Moreover, it is important to bear in mind that medication adherence is more likely to be determined by individual beliefs and social influences.[2]

Another important consideration is specifying the estimand of interest (defined in Chapter 1, Section 1.3.3) that the study investigators intend to use to make inferences about the study population.[14, 15] Specification of the estimand of interest is particularly important as it forms a key element in terms of the characteristics of the appropriate adjustment methods. Specifying the estimand of interest involves four attributes: 1) the population, 2) the outcome variable or endpoint, 3) the specification of how to deal with intercurrent events, and 4) the population-level summary of the outcome variable.[61]

This recommendation should be informed by reviewing the existing evidence from the relevant literature and discussions with clinicians, patient groups and other stakeholders involved in specifying the decision problem (e.g. NICE).

b) Review the medications characteristics and dosing regimens and their potential impact on adherence levels

Medication characteristics include the PKPD information for the prescribed regimens and the dosing schedule that the patient needs to follow to achieve the intended treatment effect. The medication PKPD characteristics have a direct impact on the relationship between non-adherence and treatment effect.

Another important consideration relates to the dosing schedule of the prescribed treatments. Different treatments have varying dosing schedules depending on the formulation (e.g. once daily [OD], twice a day [BD] or three times a day [TDS]) and some evidence suggests that the dosing schedule have a direct impact on patient adherence levels with multiples doses a day leading to potentially higher levels of non-adherence. For example, findings from a recent clinical trial of 219 patients found that implementation adherence to once-daily modified-release tacrolimus was 88.2% compared to 78.8% for twice-daily immediate-release tacrolimus formulation among kidney transplant recipients.[130] However, this is complicated and some have argued that missing 11.8% of doses of a once-daily drug is worse than missing 21.2% of doses of a twice-daily drug, in terms of drug exposure.

The relationship between non-adherence and the treatment effect is very complex; therefore, understanding the medications characteristics and dosing regimens will help to inform the selection of the appropriate data collection tools at the study design stage. Particularly, it will inform the selection of the appropriate tools for measuring medication adherence in the trial and the real world required to undertake the adjusted analysis. For example, if the drugs have a very low therapeutic index with therapeutic drug monitoring used as a normal in clinical practice then trough levels data might be available and in that case, CV% could be used as a measure of adherence. Although, this might not be the most accurate measure for implementation adherence, using electronic monitoring devices (e.g. MEMS) is rarely applicable in routine clinical practice. If the medications under consideration are regularly dispensed using prescription refills and the data from electronic patient records available for access by the analyst for both the trial and the real-world settings, then the MPR might be used.[160] This choice of the appropriate measure of adherence could be achieved by reviewing the relevant literature and/or discussions with clinicians and clinical pharmacists at the study design stage. It should also be noted that CV% might not be validated as a proxy measure of adherence in many medications outside the domain of CNIs.

Another important characteristic of medications is drug forgiveness profiles and their potential impact on adherence levels. The concept of drug forgiveness (defined in Chapter 1, Section 1.3.8) is characterised by a threshold of adherence level above which the intended therapeutic effect of the treatment is maintained.[21] However, the arbitrary choice of threshold (e.g. 80%) has been heavily criticised in the literature, and is likely to be inappropriate for most medicines. The adherence threshold could be determined objectively based on pharmacokinetic data such as CV% of trough concentration levels as applied in the case study.

There is a relationship between drug forgiveness and the impact of non-adherence on clinical outcomes.[21, 22] In other words, the outcome in the context of sub-optimal implementation is dependent on forgiveness. However, the drug forgiveness profile differs between medication classes and that is why it is important to understand it for all the medications under consideration for non-adherence adjustment. For example, patient non-adherence to drugs with a low therapeutic index (the ratio of the highest dose that is acceptably safe to the lowest which is sufficient for the drug to be effective) such as tacrolimus and cyclosporine (assessed in the case study) and oral contraceptive pills have far more negative consequences on treatment effect compared to drugs with a high therapeutic index such as aspirin. Understanding the relationship between adherence levels and drug forgiveness provides additional information about the strength of the relationship between non-adherence and treatment effect. This in turn will inform the selection of the appropriate method based on the performance evidence reported in chapter 5. This is based on is some evidence from the simulation study that shows a stronger relationship leads to higher bias especially when combined with a smaller sample size although there was no significant difference between SNFTM and IPCW (See Stage 5 of the framework for more information).

c) Check if adherence levels in the real world are likely to differ from the clinical trial

Evidence suggests that adherence levels in the real world are likely to differ from clinical trials and that is why HTA needs to account for real-world adherence levels in cost-effectiveness analysis as this methodological framework aims to achieve.[34] If adherence levels in the real world are similar to those observed in the trial, then predicting real-world adherence may not be necessary. There might also be some situations where adherence levels in the real world are better than those observed in trials, for instance, when an adherence improvement intervention is applied after the clinical effectiveness evidence from trials was published. In this situation, the prediction of real-world adherence levels in the RCT dataset may also be applied using the methods proposed in this methodological framework. Therefore, it is important to understand the difference in adherence levels between trials and the real world for the study population and medications under consideration.

There is also the issue of how the analyst knows what adherence levels to a new drug would be if it is not yet available in the real world. In this case, the analyst could apply this recommendation based on expert opinion or adherence to similar drugs in terms of formulations and dosing regimens that might be used. If adherence levels in the real world and the trial are not different, or if they cannot be estimated due to lack of data, there is no need to go any further for that particular case.

d) Draw the DAG

It is recommended to draw the DAG (Directed-Acyclic Graph) at the study design stage to conceptualise the causal links between adherence, baseline and time-dependent confounders, and the outcome of interest (i.e. time-to-event).[13] The DAG illustrates the assumptions about the relationships between these variables in the dataset. This should be the basis for identifying the data requirements in terms of the important variables and follow-up time points to allow for the adherence-adjusted analysis to be undertaken. The task for drawing the DAG will require evidence from the literature about the important prognostic factors and their relationships with non-adherence and the outcome of interest. The evidence from the literature could be complemented by discussions with clinicians to understand these causal relationships. More importantly, there is a clear role for the patient and public involvement (PPI) to play here given that adherence is a behaviour issue. The task of drawing the DAG will be based on an iterative process between this stage and the preceding two stages, which are all relating to the study design stage. This is clearly illustrated by the arrow to the left-hand side of the framework diagram (Figure 64).

7.3.2 Stage 2: Identifying and measuring the relevant types of non-adherence

This second stage involves two recommendations as described below.

a) Determine the types of non-adherence that are relevant to the decision problem

The methodological framework is aligned with the ABC medication adherence taxonomy and the associated operational definition.[1, 161] Based on an understanding of the clinical context, the important types of non-adherence (initiation, implementation and/or persistence) that may be present should be identified. The framework applies to the three types of non-adherence. However, some types of non-adherence may be more important than others for a particular study population, and therefore, may not warrant the incorporation in the adjustment analysis. For instance, based on the practical application of the adjustment methods in the case study, initiation non-adherence to immunosuppressants was considered less important in kidney transplantation as most patients

initiate the prescribed treatment and dosing is supervised during the inpatient stay. For this study population, in particular, implementation was identified as the most important type of non-adherence and that is why it was the only type of non-adherence included in the case study.

For other diseases or conditions, different types of non-adherence or a combination of more than one type of non-adherence might be important for incorporation into the adjusted effectiveness and cost-effectiveness analyses. This information could be identified from the relevant literature and/or discussion with clinicians and other stakeholders. For example, both persistence and implementation non-adherence to treatments were identified as important for the management of type 2 diabetes (T2D). In this example, Guerci et al revealed that persistence and implementation of non-adherence to T2D medications can have profound negative consequences to patients such as the increased risk of complications in the long term and mortality.[162] In addition, economic consequences are also likely which may include higher healthcare resource utilisation and increased costs. In this case, it might be important to adjust for both persistence and implementation non-adherence in cost-effectiveness analysis.

b) Identify and measure adherence to medications in the trial and the real world

To apply the methodological framework for adjusting for non-adherence, IPD data on adherence to medications in the trial is required alongside data on confounders and the time-to-event outcome. The recommended adherence measure for each type of adherence included in the study should be used. Depending on the type of non-adherence being investigated, the study design should consider the appropriate follow-up time points and time intervals such that the adherence data is aligned with time-dependent confounders' data. These are important points for consideration at the early stage of the study design.

Data on adherence levels in normal healthcare practice will be required so that real-world adherence levels could be predicted for each patient in the trial dataset. Real-world adherence data could come from observational studies or Registry databases. While IPD data on real-world adherence will be useful, this could come from published estimates if these are aligned with the adherence data collected in the trial in terms of type of non-adherence, adherence measure, length of follow-up and follow-up time points. The latter approach was applied in the kidney transplantation case study to which this methodological framework was applied.

A range of adherence measures is available including subjective and objective measures. These include pill count, electronic medication packaging (EMP) devices, self-reporting and assessment by

clinicians.[7] Using an invalidated measure of adherence could introduce a risk of bias in the analysis, which may result in generating misleading adherence-adjusted cost-effectiveness estimates. Therefore, it is important to choose a validated measure to capture data on adherence to medications for each type of non-adherence included in the study. This recommendation applies to the measurement of adherence in clinical trials and the real world, which are both essential for the adjusted analysis. It should be noted that real-world data on MEMS is exceedingly hard to find; and the more invasive the measure.

The following adherence measures are recommended for each type of non-adherence based on preliminary findings from an ongoing ESPACOMP Delphi panel:[146]

- Initiation: prescription refill data or questionnaires
- Implementation: Electronic monitoring devices (e.g. MEMS)
- Persistence: prescription-based MPR

7.3.3 Stage 3: Specifying other data requirements and making assumptions explicit

The third stage covers three recommendations as described below. The framework has implications for HEAPs that have been identified as an important development and these are provided in Section 7.4.

a) Identify and measure all baseline and time-dependent confounders

Identification of all potential baseline and time-dependent confounders is crucial to generate valid adherence-adjusted effectiveness and cost-effectiveness estimates. This step will ensure that the collected data satisfies the assumption of "no unmeasured confounding" relied upon up by the adjustment methods recommended in this methodological framework (i.e. g-methods). The task of identifying confounders will be based on the DAG drawn in the previous step. The potential confounders could then be assessed when the data becomes available at the analysis stage as a confirmatory step. Assessment of the potential confounders could also be done using datasets from previous studies conducted on the same study population if the analyst has access to such IPD data.

To my knowledge, there is no way for testing for "unmeasured confounding assumption" and the analyst can only work with the data that have been collected – hence that is why it is important to consider this at the RCT design stage as recommended by this framework. There are several approaches to test for confounding using IPD data. These may include fitting simple regressions

between these potential confounders, non-adherence variables and the outcome of interest and checking the regression output tables. However, these simple regressions could be biased themselves, so the analyst should rely on expert opinion and DAGs with these regressions used as confirmatory checks for selecting the final list of confounders to include the adjusted analysis. For this approach, the inclusion of the final confounding variables must not be restricted to prognostic factors with relationships that generate statistically significant associations, but instead, some degree of association would be sufficient to include the confounder in the adjusted analysis.

The key data items required to apply the framework include randomisation variable, baseline characteristics, follow-up time points, medication adherence data (initiation, implementation and/or persistence), baseline and time-dependent data for any potentially prognostic variables, loss to follow-up, patient survival, and time-to-event outcomes. In addition, data on medication adherence levels in the real world (either IPD or summary estimates) are also required to perform the adjusted analysis.

b) Specify appropriate methods to deal with missing data

Missing data is a common problem in clinical trials, observational data and real-world data. However, the methodological literature on dealing with missing data is well established with published guidance and tools to address the issue that could be considered and ultimately followed.[148-151] The selection of method for handling missing data will depend on the assumption around the missing data mechanism. Generally, these are classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).[151] These may include multiple imputations, last observation carried forward (LOCF), imputation using mean values from the available data, or complete case analysis.[150] The important variables to investigate for missing data include medication adherence data, baseline covariates and time-dependent confounders including their values at baseline.

The key point is that the analyst will need information on "confounders", that is prognostic characteristics that inform non-adherence and the outcome. In some instances in other contexts (e.g. adjusting for treatment switches instigated by clinicians), if a lab test for a particular prognostic variable was not carried out, it can simply be analysed as missing and/or use LOCF to impute the value of that test as if the information on that prognostic variable was not available to a clinician it cannot be a confounder. However, that will be different in the context of non-adherence, where a prognostic variable that was not measured by a hospital test could still influence a patient's decision not to adhere to the prescribed treatment. This will need careful consideration to select the appropriate method to deal with missing data for performing the adjusted analysis.

c) Specify the appropriate methods for analysing adherence data

Depending on the type of measure used to capture medication adherence data, methods for analysing adherence data should be specified based on the ABC taxonomy and the associated operational definitions.[1, 161] For instance, if adherence data is collected using drug trough level records (as used in the case study), the analysis may involve calculating the IPV for each patient at each time interval and then the CV% for that interval need to be calculated. Then, a cut-off point will be required to assign a binary non-adherence indicator to each patient with a CV% above that threshold so that it could be used as a proxy measure of adherence in the adjusted analysis. Dichotomising a continuous variable measuring implementation adherence is not the best strategy but the adjustment methods recommended within this framework require a binary variable which could be considered as a limitation. Other types of adherence data such as prescription refill data, pill count and data captured by MEMS will all require data analysis to produce the adherence variable required for adjusted analysis. The ESPACOMP Medication Adherence Data Analysis Working Group has recommended some tools and recommendations for standardising and analysing adherence data and this could be considered among other sources.

7.3.4 Stage 4: Selecting the appropriate non-adherence adjustment method

The fourth stage involves two steps as described below.

a) Consider the relevant factors in the study

This step recommends considering the relevant factors listed below as they apply to the study. This will help in selecting the appropriate non-adherence adjustment method as described in the next recommendation.

- Sample size
- Type(s) of non-adherence
- Level of non-adherence
- Treatment effect size

Although there is some evidence of impact from the above-mentioned factors, alongside other factors such as the relationship between non-adherence and treatment effect and existence of time-dependent treatment effect, on the performance of g-methods, a significant impact was evident for

the sample size, the type and level of non-adherence and treatment effect size (See Chapter 5 for more detail).

b) Select the recommended non-adherence adjustment method based on the performance evidence

For implementation and persistence non-adherence, both SNFTM and IPCW are recommended for use although the bias associated with SNFTM was slightly higher in the simulation study reported in Chapter 5. In studies where a large treatment effect size is expected/observed, the SNFTM with g-estimation is likely to be the best option, based on the results of the simulation study. A smaller sample size is expected to lead to more bias compared to data from large studies, but there is no evidence to prefer one g-method over the other based on the simulation evidence reported in this thesis. In addition to the findings of the simulation study, there is some evidence (and theoretical rationale) for why IPCW is more sensitive than SNFTM with small sample sizes (because it involves weighing, and when you get high weights the method becomes prone to higher error and convergence problems).[128] Some other minor differences might influence the selection between IPCW and SNFTM when considering the combination of factors listed in (a) above for each particular study. However, it should be noted that both g-methods perform well and neither is superior for all situations based on the simulation evidence reported in Chapter 5.

For initiation non-adherence, SNFTM with g-estimation is recommended followed by IPCW, although both methods could be applied as the performance differences between them are minor. In studies with larger sample sizes, SNFTM would be the best option especially when the anticipated treatment effect size is large. However, it should be noted that the performance evidence generated from the simulation study showed that IPCW produces better coverage than SNFTM, although the performance results were very close in terms of unbiasedness and ModSE. There has been some uncertainty around this particular finding and there is an argument that SNFTM coverage could have been better if bootstrapping was applied in the simulation. While the simulation demonstrated that both IPCW and SNFTM have generally performed well, often it was generally not clear exactly which method will be best, and so it will be necessary to apply both and compare results. Therefore, it is recommended that both adjustment methods should be applied, if possible, and then further assess the adjusted results as described in Stage 5 of this framework.

7.3.5 Stage 5: Adjusting treatment effectiveness for real-world non-adherence levels

The fifth stage of the framework involves adjusting treatment effectiveness for real-world non-adherence levels using the following three steps.

a) Predict real-world non-adherence for each patient in the trial dataset

This analytical step will use real-world adherence levels estimated in Stage 2. The prediction of real-world non-adherence will be based on the adherence data observed in the trial taking into account the prognostic relationship between adherence level and the confounding variables in the dataset. The prediction should be performed such that the overall proportion of non-adhered patients matches the real-world estimates. As a sense-check, the analyst should look at the predicted non-adherence to make sure the prediction has face validity. For example, the predicted data should show that patients with a poor prognosis (in terms of the prognostic confounding variables) are more likely to fall into the non-adhered group. There are different ways for doing this step depending on the measure of adherence used in the dataset and the available data. For instance, when trough levels data is available (as in used the kidney transplantation case study) a new cut-off point for the CV% could be used to predict real-world non-adherence in the RCT dataset.

An alternative approach for predicting real-world non-adherence in the trial dataset is to apply the marginal standardisation method.[147] This method is considered as a special case of g-computation (described in more detail in Chapter 6, Section 6.9.2). In brief, the method predicts the marginal probabilities of non-adherence for each patient over the distribution of the confounders and the counterfactual adherence level (e.g. real-world adherence level). Then, confounder-adjusted logistic regression could be used to compute the predicted probabilities of non-adherence in the trial dataset. This method has not been applied in the case study, as a more appropriate prediction method for that particular case study was applied utilising drug concentration levels that were available within the SYMPHONY trial dataset. Therefore, more research will be valuable here to assess the appropriateness of this prediction method when applying the framework in other case studies in the future.

b) Apply the selected g-method to obtain adherence-adjusted effectiveness estimates

This step involves applying the g-method (IPCW or SNFTM) to the trial dataset (including the real-world predicted non-adherence) to generate valid adherence-adjusted clinical effectiveness estimates. The detailed analytical steps for applying each of the g-methods (including the Stata code) are provided in Chapters 4-6.

c) Assess the results of the adjustment method for reliability

Following the application of the selected adjustment method in the previous step, the analyst should assess the adjusted effectiveness results generated for reliability. For example, the analyst might apply IPCW and get large weights, making it very unreliable. On the other hand, when applying the SNFTM, the g-estimation might not work. Furthermore, the results could be very sensitive to different model specifications. The latter is particularly relevant to the untestable assumption of no unmeasured confounding relied upon by the adjustment method. Discussion with clinical experts is key here as explained as part of the process for drawing the DAG (see Section 7.3.1).

Sullivan et al. have recently published a paper providing recommendations for reporting adjusted analysis in the context of treatment switching.[163] Some of their recommendation relating to the application of IPCW are relevant to the context of adjusting for non-adherence and these have been adapted for consideration to assess the adherence-adjusted results using this framework. These include: (a) checking whether stabilised or un-stabilised weights were used; (b) checking the specification of the statistical models used to calculate the weights (e.g. pooled logistic models); (c) checking how missing data on the baseline and time-dependent confounders were dealt with in the weighting models; (d) checking and documenting the coefficients and the associated SEs and 95% confidence intervals generated from the weighting models; and (e) summarising the distribution of weights used to create the pseudo-population for the applying the adjusted survival analysis.

For the SNFTM with g-estimation, assessing the results may include checking how multiple observations per individual patient are dealt with in the model (e.g. including all observations which may risk overstating statistical significance or using first observation to get statistical significance right at the expense of precision). Other checks may include checking whether grid search (or interval bisection search) was used to run the g-estimation and checking the results using the Z graph to ensure that the g-estimation process has worked properly.

7.3.6 Stage 6: Considering the impact of non-adherence on direct treatment costs

Stage six of the framework involves two considerations as described below.

a) Consider the impact of non-adherence on drug costs

There are arguments in the literature as to whether direct treatment costs could be lower as a result of patient non-adherence to their prescribed medications.[164] The argument is that patients do not

take enough quantities of the prescribed treatments by not getting their prescriptions dispensed. This means there could be cost savings because fewer packs are dispensed. The counter-argument is that most patients get their regular prescriptions dispensed, but ended up not using these treatments as a result of non-adherence which leads to wastage. In other words, patients may not take the drugs according to the recommended schedule, but the full costs of used and unused medicines are incurred by the NHS (or the payer in other healthcare systems). This assumption is unlikely to apply as much to non-persistence when prescriptions are most likely to stop being dispensed; and certainly not to non-initiation. The latter argument is particularly relevant to immunosuppressants in the kidney transplantation case study reported in Chapter 6 in this thesis. This assumption was based on discussions with two clinicians (WM and JF) and a clinical pharmacist (DG). In the UK NHS, this means the healthcare system will not save costs as a result of patient non-adherence to their prescribed treatments in particular disease areas such as immunosuppressants used after kidney transplantation or antihypertensive drugs. Hence, the full drug costs were used in the economic analysis undertaken in the case study.

However, it should be noted that the impact of non-adherence on drug costs will depend on many factors including the disease area, the type of non-adherence and the class of medications and how the pharmacological treatments are financed in a particular health care system. Therefore, this should be considered when adjusting for non-adherence in cost-effectiveness analysis and any assumptions fully justified.

b) Consider the impact of non-adherence on other direct treatment costs

This point relates to considering the impact of patient non-adherence on other direct treatment costs (other than drugs). This may include administration costs for medications administered by a healthcare professional (e.g. in visits to the clinics). Patients who do not attend their appointments to get their prescribed medications administered may incur other direct treatment costs (e.g. cost of unattended visits). Again this will depend on how the health care system is organised and financed. All these cost-consequences resulting from patient non-adherence should be considered to make informed decisions on whether or not to include them in the economic analysis. This is a very real cost, probably hard to measure. For example in the context of kidney transplantation, non-adherent patients are more likely to miss the clinic appointments, but they also fail to attend for blood tests, forget to omit their tacrolimus before the drug level which then needs to be repeated and all these incur costs to the health system. This consideration could be achieved by consultations with pharmacists, patient groups, clinicians and other stakeholders.

The impact of non-adherence on other non-direct costs such as fewer adverse events, but maybe worse control of the disease and, therefore, more general practitioner (GP)/hospital appointments is very important. However, the assumption is that this will come out of the economic model through different transition probabilities to different health states after adjusting the treatment effectiveness for non-adherence. Also very relevant is the healthy adherer effect – people who are most adherent to medications are likely to be adherent to a good diet and healthy living and vice versa, meaning greater ill-health (and associated costs) among non-adherent patients for reasons other than not taking their medicines.

7.3.7 Stage 7: Estimating and reporting adherence-adjusted cost-effectiveness of treatments

Stage seven of the methodological framework involves estimating and reporting the results from the adherence-adjusted cost-effectiveness analysis. This covers the following four recommendations.

a) Use the adjusted effectiveness estimates within the economic model

This will depend on the estimand of interest as estimated by the adjustment method. The form of the treatment effect estimates generated from the analysis will depend on how the economic model is conceptualised and structured.

The adjusted effectiveness estimates may be obtained in a form of adjusted hazard ratios (HRs), restricted mean survival times (RMSTs) or survivor functions depending on how the economic model is conceptualised. For example, in the kidney transplantation case study to which was carried out to inform the development of this methodological framework, treatment effectiveness was incorporated in a form of adjusted graft survivor functions for each of the alternative treatment options estimated at baseline, 3, 6, 9 and 12 months.

As another example in a different disease area, HRs including the 95% confidence intervals are used to incorporate the treatment effectiveness in a 6-health-state Markov model to assess the cost-effectiveness of docetaxel and paclitaxel-containing chemotherapy regimens (taxanes) compared with standard (non-taxane) treatment for adjuvant treatment of early breast cancer.[165, 166] While the treatment effect was not adjusted for real-world adherence levels in that economic model (as in most economic evaluations), the example show a different form of treatment effectiveness incorporated in economic models. In other words, HRs were used in this example compared with graft survivor functions used in the case study economic model.

b) Report the results from adjusted analysis alongside results from the standard ITT analysis

It is recommended that the cost-effectiveness estimates from the adherence-adjusted analysis are reported alongside results from the standard ITT analysis. If the adherence-adjusted cost-effectiveness estimates were considered the most important for the resource allocation decision-making, then reporting the adjusted analysis as primary analysis with ITT reported as secondary analysis might be preferred. It is also good practice to report the type of medication adherence incorporated in the adjusted analysis including the operational definition for each type, method of measurement and results of non-adherence levels in the trial and the real world as recommended by other existing guidelines.[126, 161]

c) Perform and report sensitivity analyses with different adherence levels

It is also recommended that sensitivity analyses with different adherence levels, given that the analyst may not be able to predict real-world adherence levels accurately (especially for new drugs). Furthermore, the analyst could also consider a threshold analysis (i.e. what percentage adherence would be needed for the drug to be cost-effective). The latter will provide useful additional information for decision-makers in terms of investing in interventions (e.g. patient education) to improve medication adherence in the real world.[167]

d) Report the strengths and limitations

It is important to report the strengths and limitations of the adjusted analysis and these will vary between studies. These may include the assumptions used to perform the adjusted analysis and the sensitivity of the results to deviations from those assumptions. Among others, these may include the assumption of no unmeasured confounding, the quality of medication adherence data, adherence measurement errors, and missing data. This will improve transparency in the application of the methodological framework.

Prerequisites to apply the framework

To apply the methodological framework, three prerequisites need to be met, otherwise, the framework cannot be used. These are: (a) identification and measurement of patient non-adherence based on the temporal phases of initiation, implementation and persistence as defined by the ABC taxonomy; (b) identification and measurement of baseline and time-dependent confounders, and (c) access to IPD data from an RCT with adherence metrics and prognostic characteristics plus real-world adherence data.

7.4 Implications for health economics analysis plans

The methodological framework has implications for HEAPs that have been identified as an important development. The key items within the framework that may be specified within a HEAP are listed below:

- Detail the DAG illustrating the causal relationships between adherence, baseline and time-dependent confounders and the outcome with the diagram included in the HEAP document.
- Specify the estimand, including the population-level summary form for estimating the adherence-adjusted treatment effectiveness for incorporation into the economic model.
- Specify the types of non-adherence included.
- Justify and describe the tool for measuring adherence in the trial and the real world (if applicable).
- Specify the methods for dealing with missing data on adherence metrics and confounders.
- Outline how the impact of non-adherence on direct treatment costs will be estimated (if applicable).
- Describe any planned sensitivity analyse (e.g. analyses using different levels of adherence)
- Describe how the adherence-adjusted effectiveness and cost-effectiveness results will be reported.

7.5 Discussion, areas for future research and conclusion

7.5.1 Summary of the methodological framework

This chapter presented a seven-stage methodological framework to account for the impact of non-adherence on the cost-effectiveness of prescribed chronic medications. The framework provides guidance on the steps that could be followed at the study design stage, and the analysis and reporting stage for adjusting the cost-effectiveness treatments for real-world adherence levels.

In summary:

a) The methodological framework was developed in a systematic way using a recently published "how-to" guide.
b) The framework involves seven stages with Stages 1-3 covering recommendations for RCT designs and Stages 4-7 covering recommendations for analysing RCTs to produce adherence-adjusted effectiveness and cost-effectiveness estimates.

c) The stages cover a series of considerations and 17 recommendations with the difference between them being recommendations indicate specific measures that need to be taken based on findings from this research; whereas considerations, specify important issues that may be considered with further actions taken based on what the investigators/analyst find in their particular study.
d) The framework has implications for HEAPs that have also been identified; particularly, to specify the plan for adherence-adjusted analysis in these documents.
e) The framework is intended to be "live" in nature as it needs to be informed by further research in terms of application in different case studies for evaluation and extension to incorporate additional adjustment methods and endpoints.

The methodological framework put forward in this thesis is expected to help in improving the overall quality of economic models used for estimating the cost-effectiveness of chronic medications in the context of time-to-event outcomes and HTA. The main contribution of the methodological framework is that it builds on four linked studies reported in this thesis to provide a systematic approach for incorporating real-world non-adherence in economic evaluations.

The methodological framework is aligned with the widely used ABC medication adherence taxonomy characterised by three types (initiation, implementation, persistence) to aid the application of the framework in future economic evaluations.

The framework recommends g-methods (IPCW and SNFTM) for obtaining valid clinical effectiveness estimates adjusted for patient adherence for incorporation into the economic model to estimate adherence-adjusted cost-effectiveness of treatments. The recommendations for using g-methods was informed by performance evidence generated from the simulation study reported in this thesis. It should be noted that the other two methods (RPSFTM and PKPD) appear to be appropriate for HTA but these methods have not been assessed in the simulation study and this should be considered as a limitation.

The development of the methodological framework was informed by a recently published guide on "how to develop methodological frameworks".[152]

7.5.2 Strengths and limitations

The methodological framework has several strengths. First, the development of the framework was informed by a recently published guide on how to develop a methodological framework which itself was based on a review of 30 papers involving some form of a methodological framework that has been

developed in the last 10 years. Second, the framework was developed based on evidence from four linked studies reported in this thesis. Third, the framework was developed from the points of view of an investigator and an analyst in providing practical recommendations and considerations that could be followed from the study design stage to reporting the findings for any study with a time-to-event outcome considering to account for real-world non-adherence in the cost-effectiveness of chronic medications. Finally, the framework aimed at providing a set of recommendations and considerations that could be followed systematically to adjust for non-adherence in economic evaluations, although some of these recommendations might not be applicable or feasible to implement for each particular study.

There are some limitations associated with the methodological framework. First, the framework is focused on studies time-to-event outcomes, and therefore, could not be used in studies with other types of outcomes or endpoints (e.g. continuous or categorical outcomes) if those are the primary outcomes of interest. However, most of the considerations and recommendations in the framework would remain the same non-time-to-event outcomes. The analytical methods would differ, and therefore, some other factors that determine how the methods are applied may differ. Second, ideally, new methodological frameworks are evaluated using multiple case studies and/or Delphi panels and this has not been undertaken in this study. A Delphi panel exercise was considered at the stage of developing the NIHR Doctoral Research Fellowship application that has led to funding this PhD, but it was excluded to keep the scope of the project manageable within the maximum three years allowed by this funding stream. However, the framework was applied to a case study in kidney transplantation which in itself informed the formulation of the framework. In addition, an earlier version of the methodological framework was discussed with supervisors, a clinician (WM, a Consultant Nephrologist) and a leading expert in medication adherence research (DH). Finally, two methods identified as appropriate for the HTA context (PKPD and RPSFTM) are not included in the recommended list of adjustment methods as these were not directed assessed in the simulation study.

The PKPD method was excluded as it requires a different set of DGMs from those applied in the simulation study, therefore, it was not possible to directly compare it with IPCW and SNFTM. The RPSFTM only works to adjust for non-adherence in one-arm studies or studies with placebo-control arms; however, the simulation study was designed to assess the performance of adjustment methods in RCTs with two active treatments, and hence, the method was excluded from the simulation study and the methodological framework.

The PKPD method has been assessed to adjust the cost-effectiveness of treatments for varying levels of patient non-adherence in a recent doctoral thesis by Hill-McManus et al,[91] and therefore, it is

available as an option for consideration. However, it should be noted that the PKPD performance versus g-methods has not been assessed in this doctoral research which could be considered as a limitation. Future research is recommended to directly assess the performance of the PKPD method versus g-methods for potential addition as a recommended adjustment method in a future extension of the framework.

An important limitation is the fact that many chronic disease treatments have no direct or indirect assessment of adherence, even in RCTs. Application of the framework when medication adherence data both in the RCT and the real world is not available will not be possible. This limitation is relevant to applying the framework in post hoc analysis to the great majority of clinical scenarios where adherence measures are not used or available. The absence of this data is common and will be a limitation of how widely this framework can be applied. The framework addressed this in the flow diagram prospective study design to ensure that the data required to account for non-adherence are considered at the study design stage and collected in the trial.

7.5.3 Areas for future research

Although the methodological framework has been applied in a single case study on kidney transplantation as part of this doctoral research project, it has not yet been applied in other case studies. The kidney transplantation case study involved a range of analyses including obtaining estimates of real-world adherence levels, analysis of adherence data from a large multicentre RCT, predicting real-world adherence for each patient in the trial dataset, estimating adherence-adjusted effectiveness estimates and incorporating these estimates in an economic model for producing adherence-adjusted cost-effectiveness. Given the practical experience of applying the framework in the kidney transplantation case study, the framework is potentially applicable to other case studies and disease areas where non-adherence to prescribed medications is identified as an issue. Therefore, future research is recommended in terms of other cases studies to apply this methodological framework. The intention of encouraging future applications in other studies is to evaluate and refine the framework as it continues to improve based on application in practice. I will look for opportunities to apply this methodological framework in future projects and other researchers are also encouraged to apply the framework in their studies.

Another recommended area for key future research involves assessing non-adherence adjustment methods for studies with other types of outcomes or endpoints (i.e. continuous and categorical outcomes). This will contribute to the improvement of the methodological framework over time and

ultimately address challenges and shortcomings for much wider types of studies in the context of economic evaluation and HTA.

Further research is recommended to assess the performance of PKPD and RPSFTMs in a direct comparison with IPCW and SNFTM using a well-conducted simulation study. This will provide performance evidence for the potential addition of these methods as alternative adjustment methods in the future extension of the framework.

Alternative methods for predicting non-adherence such as the marginal standardisation method is recommended.[147] This will provide further information in terms of the accuracy of the predicted real-world adherence levels within the RCT dataset.

Assessing the impact of using alternative adherence measures (e.g. MEMS, MPR) and the value of collecting better adherence data in the trial and the real world is also recommended as an important area for future research. There is an ongoing Delphi panel supported by the ESPACOMP and the final report from that panel might provide specific recommendations in terms of what measure should be used for each type of non-adherence. Therefore, the analyst may consider recommendations from that panel report when it becomes available.

7.5.4 Conclusion

To conclude, the systematic approach put forward through this methodological framework may help in capturing the important elements that need to be considered at an early stage of the study design. The framework provides guidance for academic researchers (health economists and economic modellers) and the pharmaceutical industry at the analysis and reporting stage to account for real-world non-adherence levels in economic evaluations for HTA. This will lead to better cost-effectiveness evidence to inform resource allocation decision making, and ultimately leads to improvements in population health.

Chapter 8: Discussions, recommendations and conclusions

8.1 Introduction

This chapter provides a recap of the full thesis, highlights the contributions in the context of other research, provides the recommendations for accounting for non-adherence in cost-effectiveness models and draws the overall conclusions. Section 8.2 provides a review of the principal aim of this doctoral research and addresses the research questions outlined in the introductory chapter. The section summarises the key findings from all previous chapters (Chapter 1-7) to make coherent recommendations to account for the impact of non-adherence on the cost-effectiveness of prescribed chronic medications. Section 8.3 provides a summary of the contributions made by this thesis in the context of the health economics literature. Section 8.4 discusses the strengths and limitations of this research. Section 8.5 provides the recommendations in a form of a methodological framework and the plan for dissemination for use in other studies. Section 8.6 outlines the directions for future research and possible extension to the current methodological framework. Section 8.7 provides the conclusions on the whole thesis.

8.2 Thesis overview

The overall aim of this doctoral research study was to develop a methodological framework to account for patient non-adherence to prescribed medications for chronic conditions when undertaking economic evaluations for HTA. The focus is on methods for adjusting estimates of treatment effectiveness for real-world non-adherence levels in studies with time-to-event outcomes and further incorporation in cost-effectiveness analysis to produce better evidence for resource allocation decision-making in health care.

This research involved four linked stages: (1) a systematic review of methodological papers that identified 12 non-adherence adjustment methods and assessed their suitability for use in the context of HTA; (2) a simulation study that assessed the relative performance of a subset of four adjustment methods across a range of 90 scenarios using simulated RCT datasets; (3) a case study that applied the g-methods and estimated adherence-adjusted estimates of the cost-effectiveness of maintenance immunosuppressive therapy for adult kidney transplant recipients in the UK; and (4) the development of the methodological framework put forward in this thesis to account for the impact of real-world non-adherence on the cost-effectiveness of prescribed medications.

This thesis is comprised of eight chapters (including this final chapter) and addressed three research questions as summarised in the following paragraphs with a particular focus on the methodological framework presented in Chapter 7.

Chapter 1

Chapter 1 set out the content and structure of the thesis. The chapter defined non-adherence to medications (based on the ABC taxonomy) and the key concepts used throughout the thesis (including estimands, confounding, and DAGs). Then, the problem of patient non-adherence to medications and its importance in the context of HTA was explained. This was followed by identifying the gap in the literature in terms of methodological frameworks for correcting the clinical effectiveness and cost-effectiveness of treatments for patient non-adherence. The research questions addressed in this thesis and the objectives of the doctoral research were outlined. The chapter then outlined the potential solutions by summarising the existing evidence in terms of accounting for non-adherence in estimating the clinical effectiveness and cost-effectiveness of treatments. The chapter concluded by putting the research questions into context and outlined the expected contribution of this research.

Chapter 2

Chapters 2 and 3 presented the work undertaken in Stage 1 of this research (the systematic review). Chapter 2 addressed research question 1: "*What are the key methodological approaches used to account for the impact of non-adherence on the effectiveness and cost-effectiveness of health technologies used in chronic conditions with time-to-event outcomes?"*.

The chapter presented 12 methods and 8 extensions to those methods identified for adjusting estimates of treatment effectiveness for patient non-adherence in the context of time-to-event outcomes and HTA.[61] These methods were identified from a systematic review of methodological papers using the 'pearl growing' technique and 2-stage iterative search approach across seven databases. Only three out of 20 included papers looked at methods for adjusting cost-effectiveness estimates for non-adherence with the remaining 17 papers focused on methods for adjusting estimates of treatment effect. This provided further evidence about the gap in the health economics literature in terms of methods for accounting for non-adherence in cost-effectiveness models.

The chapter also put forward a new taxonomy of methods to increase understanding of the concept behind each identified method and its relation to other methods in terms of estimands and estimators. In the proposed taxonomy, adjustment methods are broadly classed into four groups: (1) simple methods that do not appropriately adjust for patient non-adherence, (2) principal stratification

methods for estimating the CACE estimand, (3) g-methods which are based on the counterfactual outcome framework, and (4) pharmacometrics-based methods using PKPD analysis. The chapter described the identified non-adherence adjustment methods (including their extensions) based on a narrative synthesis undertaken in this research. The description of methods covered the origin of the method, theoretical characteristics and application in a simulation study and/or a case study. The chapter highlighted that not all methods produce the same estimand and that each method makes specific assumptions (e.g. the no unmeasured confounding assumption and the exclusion restriction assumption) with associated limitations.

Chapter 3

Chapter 3 compared the alternative adjustment methods based on existing evidence identified by the systematic review. This included methods compared empirically, in simulation studies and/or cases studies based on evidence from the literature and the narrative synthesis undertaken. The chapter assessed the appropriateness of adjustment methods for the HTA context based on three criteria developed in this doctoral research. The criteria were: (i) the suitability of the estimand for HTA, (ii) the types of non-adherence the method is capable of dealing with, and (iii) whether it is possible to use the method to account for real-world non-adherence. The chapter concluded that g-methods (MSM with IPCW, SNFTM with g-estimation, and RPSFTM with g-estimation) and PKPD method are more appropriate than the alternative methods for adjusting estimates of treatment effectiveness for real-world non-adherence. The chapter also provided justification for the selection of a subset of four methods for further assessment in a simulation study (Chapters 4-5). These included two g-methods (IPCW and SNFTM) and two simple methods (ITT and PP). I recognise that the choice of criteria and the assessment of methods against them involve an element of subjectivity, but I argue the list of selected methods is the product of my interpretation of evidence relating to the needs of resource allocation decision-makers and the relative merits of statistical methods identified.

Chapter 4

Chapters 4 and 5 presented the work undertaken in Stage 2 of this research (the simulation study). Chapter 4 described the design and implementation of the simulation study using the ADEMP structural approach for planning simulation studies. The simulation study was designed to assess the performance of a subset of non-adherence adjustment methods identified in Stage 1 across a range of 90 scenarios with 1900 simulations each. The chapter provided the specification of scenarios assessed in the simulations based on the DGMs that covered alternative representations of factors covering the type of non-adherence (initiation, implementation and persistence), level of non-

adherence, sample size, the pattern of hazards, treatment effect size, relationship between treatment effect and adherence level, and the existence of any time-dependent treatment effect. The chapter outlined the detailed steps undertaken to run the simulations and provided the Stata code for implementing the simulation program.

Chapter 5

Chapter 5 addressed research question 2: *"What is the relative performance of the alternative methods in estimating the impact of non-adherence on treatment effectiveness?"*

The chapter provided new simulation evidence on the performance of non-adherence adjustment methods across a range of 90 scenarios across the three types of non-adherence. Methods performance was assessed according to the bias, MSE, EmpSE, ModSE, and overage with the difference in RMSTs used as a primary estimand. The simulation study demonstrated that g-methods (SNFTM and IPCW) are the best-performing methods in terms of unbiasedness and ModSE for adjusting estimates of treatment effect in the presence of implementation and persistence non-adherence. For initiation non-adherence, SNFTM was found to be the best-performing method in terms of unbiasedness and ModSE. The findings also showed that the PP method performed well in many scenarios, although its estimand is different from the ITT and g-methods estimands as it is not marginalised to the entire study population. However, it should be noted that the performance of IPCW, SNFTM and PP are very close and often there are only minor differences, depending on the type of non-adherence and performance measure. The findings from the simulations demonstrated that ITT was generally the worst method when estimating treatment effect in the presence of non-adherence because the purpose of the ITT is not to adjust for non-adherence based on the estimand. The latter finding is generally consistent with existing simulation evidence in the methodological literature and my prior expectations.

Chapter 6

Chapter 6 presented the work undertaken in Stage 3 of this research (the case study). The chapter focused on the application of g-methods using a real dataset from a large multicentre RCT (SYMPHONY trial) with data from 1,190 patients and 12 months follow-up. The chapter described the methods and analytical steps used for re-estimating the treatment effectiveness adjusted for real-world non-adherence levels and presented the estimated effectiveness results. The chapter then described the incorporation of adjusted effectiveness estimates into an adapted economic model to account for the impact of real-world non-adherence to immunosuppressants among adult kidney transplant

recipients in the UK. The case study estimated that the low-dose tacrolimus regimen (the current standard treatment in the NHS) is predicted to be less effective with a higher cost per patient when the real-world adherence levels are taken into account. The incremental cost per patient was higher by 20.1% with a 10.8% reduction in incremental QALYs compared to the ITT analysis. However, in this case study, adjusting for real-world adherence did not change the original conclusions of the economic analysis, as low-dose tacrolimus remains highly cost-effective, irrespective of whether the adjustment is applied.

Chapter 7

Chapter 7 addressed research question 3: *"How should economic evaluations incorporate the impact of non-adherence using evidence from both RCTs and real-world data?*

The chapter presented the work undertaken in Stage 4 of this research (development of the methodological framework). The chapter put forward a new methodological framework to account for the impact of real-world non-adherence on the cost-effectiveness of prescribed chronic medications in studies with time-to-event outcomes. The chapter described the three phases used to develop the framework comprising: (1) a review to identify existing frameworks, (2) development of the methodological framework, and (3) transferability of the framework to other disease areas beyond the kidney transplantation case study. The chapter then described the seven stages of the methodological framework covering 17 recommendations and considerations from the study design stage to the application of the g-methods to adjust for real-world non-adherence levels in economic evaluations for HTA and reporting. The chapter concluded with the recommendations for the next steps, including the plan for disseminating the methodological framework and key areas for future research. The latter specified the directions for future research in terms of adding other adjustment methods to the framework that have not been assessed in this research and addressing other types of outcomes/endpoints (i.e. continuous and categorical outcomes).

Further observations

The current approaches used in health economic modelling tend to rely on simplistic approaches to attempt to model the impact of patient non-adherence in economic evaluations. These approaches were discussed in greater detail in Chapters 1 and 7, but generally, they share the same characteristics and limitations in terms of making strong assumptions about the causal relationship between patient non-adherence and treatment effect. For example, some approaches assigned reduced treatment effects proportional to the level of non-adherence (e.g. 20% reduction in treatment effect for 20%

lower levels of adherence). This is a very risky approach as it does not take into account the complexity of the relationship between medication adherence and treatment effect. The latter is more likely to be affected by many factors such as the PKPD characteristics of the medications and drug forgiveness profiles that differ between drugs. Consequently, these simplistic approaches are more likely to produce misleading cost-effectiveness results, which may lead to suboptimal allocation of scarce health resources for the NHS and other healthcare systems around the world.

As an alternative approach, the methodological framework presented in this thesis provides a systematic approach and a practical step-by-step guide to model the impact of non-adherence on the cost-effectiveness of prescribed medications. Rather than relying on the abovementioned simplistic approaches used in the health economics literature, the development of the present methodological framework was underpinned by the work undertaken in the four stages of this research as reported in Chapters 1-7. The framework was built on the key principle that non-adherence to medications is primarily a clinical issue and accounting for it in economic evaluations should first be based on adjusting the clinical effectiveness estimates. Then, the impact of patient non-adherence on direct treatment costs (including drug costs) should also be considered and accounted for, if applicable, as recommended by the framework.

To apply the methodological framework, three prerequisites need to be met, otherwise, the framework cannot be used. These prerequisites are: (a) identification and measurement of patient non-adherence should be based on the temporal phases of initiation, implementation and persistence as defined by the ABC taxonomy; (b) identification and measurement of baseline and time-dependent confounders is crucial for adjusting the clinical and cost-effectiveness of treatments for patient non-adherence, and (c) access to individual patient-level data from an RCT with adherence metrics and prognostic characteristics is essential for undertaking the adjusted analysis to account for real-world adherence levels in economic evaluations.

The application of the methodological framework will require a change in practice in terms of the way non-adherence to medications is incorporated into health economic models. It may also require earlier consideration of data requirements at the study design phase of trials. This means moving away from approaches that rely on making strong assumptions about the causal relationship between non-adherence and treatment effects that risk producing misleading cost-effectiveness evidence towards a systematic approach that relies on considering all the essential elements provided by the framework proposed in this thesis. The systematic approach provided by the framework covers the essential recommendations and considerations from the study design stage (including the collection of data

required to adjust for patient non-adherence) to the analysis stage by applying the g-methods to model the relationship between non-adherence and treatment effectiveness to the final stage reporting of reporting the adherence-adjusted cost-effectiveness results. This framework will help to generate more robust clinical effectiveness estimates for further incorporation into the economic model to produce cost-effectiveness estimates adjusted for real-world non-adherence levels.

The methodological framework is intended to provide key stakeholders with the logical steps and tools required to design their studies in a way that make them amenable to adjustments for changes in adherence levels observed in the trial (e.g. if lower adherence levels in the real world are likely). In this context, the framework targets a specific group of stakeholders including clinical trialists, health economists and economic modellers involved in the design of clinical trials and associated economic evaluations. The framework also targets analysts involved in designing, building and populating health economic models for HTA (i.e. researchers from academia including the technology assessment groups (TAGs), Health Economics and Outcomes Research (HEOR) consultancy firms and the pharmaceutical industry) to help improve the overall quality of economic models.

The methodological framework is characterised by two main aspects: (a) recommendations and considerations related to the design of data collection before the trial starts, where the economist believes that adherence in the real world is likely to differ from what will happen in that trial; and (b) recommendations and considerations for applying non-adherence adjustment methods after the trial has finished for producing the intended cost-effectiveness estimates. The detailed recommendations and considerations within the seven stages of the framework are outlined based on these two aspects as illustrated in the framework flow diagram and the associated descriptions reported in Chapter 7. The dissemination plan (reported in Section 8.6) will make this framework more accessible to the abovementioned groups to ensure that the correct steps could be followed to account for non-adherence in future economic evaluations.

8.3 Contribution to knowledge

Previous research

When this project started, the gap in the health economics literature in terms of incorporating the impact of non-adherence on the cost-effectiveness of treatments was evident. This gap was identified based on an early scoping review undertaken in 2017 as reported in Chapter 1. A need for further research to resolve this issue was identified at that stage and that ultimately motivated me to develop my doctoral research fellowship application that led to undertaking this research.[3, 33] Based on a recent

review undertaken in 2021 as part of Stage 4 of this research, I concluded that no methodological framework exists to address this important issue. Therefore, it seems the gap in the methodological literature in terms of accounting for non-adherence in economic evaluation remains.

The current status of research in this area included six approaches relied upon by health economists and economic modellers for adjusting the cost-effectiveness of treatments for patient non-adherence. Five of these approaches were identified by the first scoping review as described in Chapter 1. A further sixth approach was identified by the 2021 review conducted to identify any existing frameworks to inform the development of the present methodological framework as reported in Chapter 7.[153] It should be noted that these six approaches are trying to model the impact of non-adherence using structural models (Decision trees or Markov models) based on simplifications and strong assumptions about the causal links between adherence levels and treatment effect rather than adjusting the treatment effectiveness for patient non-adherence using complex methods (e.g. g-methods) to obtain valid estimates.

In this context, a methodological framework is different from an approach. A framework would usually imply some set of steps that need to be followed to achieve a particular outcome. In contrast, the approaches identified by that review were all separate ways of trying to account for non-adherence in cost-effectiveness models. Thus, the existing approaches do not provide a systematic process that a methodological framework provides to address non-adherence in economic evaluations. The limitations of these approaches are discussed in greater detail in Chapters 1 and 7.

To put this research into context by summarising the status of existing evidence, I summarise the progress made in the health economics literature over the past two decades in this paragraph. The methodological challenges in terms of the need for improved modelling of the impact of patient non-adherence in economic evaluations were first raised by Hughes and colleagues in a paper published in *PharmacoEconomics* in 2001.[27] Seven years later and despite the increased attention to modelling the impact of non-adherence, Muszbek and colleagues published a paper in the *International Journal of Clinical Practice* in 2008 that updated the status of research and suggested that the evidence was inconclusive and that further research is warranted to resolve the issue.[3] In 2016, Hughes and colleagues published another paper in *PharmacoEconomics* that provided a recent update in terms of methodological challenges at that point and concluded that further research is still needed for improvement.[33] I met with the first author of that paper (DH, who is also an advisor for this doctoral research) and he provided useful insights and comments on the design of this research. DH also provided further advice on the systematic review work (Chapters 2-3) during a one-week study visit

to Bangor University which was arranged in 2018 as part of the training element of my NIHR fellowship project. DH also commented on an early version of the methodological framework and provided insights that led to further improvements.

Furthermore, regarding the existing research around the relevant adjustment methods, Latimer and colleagues assessed the g-methods in several simulation studies for adjusting treatment effectiveness in the presence of treatment switching.[101, 128, 129] The simulations assessed the performance of IPCW, SNFTM, PP and ITT (among other methods) and demonstrated that IPCW and SNFTM are superior to ITT and PP as reported in Chapter 5. Although the context of treatment switching is different from medication adherence, the adjustment methods that can be used are similar in both settings. The key point from those papers is that g-methods work to adjust for other important problems such as treatment switching, and this is generally consistent with the findings from research reported in this thesis. In relation to this point, if the analyst is adjusting for non-adherence, then censoring needs to reflect non-adherence as opposed to other causal mechanisms, such as switching. However, if the trial includes both non-adherence and switching, the analysis suggests that both could be modelled simultaneously using g-methods. This idea has not been tested in this research and could be an interesting area for future research.

In conclusion, there is a gap in the health economics literature in terms of assessment methods to model the impact of model non-adherence in economic evaluations. In addition, the current status of research in this area suggests that no framework exists to account for non-adherence in cost-effectiveness analysis.

Contribution of this work

Motivated by the evident gap in the health economics literature in this area, this thesis:

a) Put forward a seven-stage methodological framework to account for the impact of patient non-adherence on the cost-effectiveness of prescribed chronic medications in the context of time-to-event outcomes and HTA. The methodological framework is expected to improve the overall quality of health economic models by providing guidance to clinical trialists, academic researchers (including health economists and economic modellers) and the pharmaceutical industry to account for real-world non-adherence levels in cost-effectiveness analyses. This will provide better evidence for healthcare decision-makers which will lead to improvements in decision making, and ultimately lead to improvements in population health through better resource allocation in health care.

b) Identified 12 methods and 8 extensions to those methods for adjusting estimates of treatment effectiveness for patient non-adherence, and assessed their suitability for use in the context of HTA.[61] The review was published in Medical Decision Making and concluded that g-methods and PKPD are the most appropriate methods to account for real-world adherence levels in HTA.[61]

c) Proposed a new taxonomy of methods for adjusting estimates of treatment effectiveness for patient non-adherence in the context of studies with time-to-event outcomes.[61] The taxonomy is expected to increase understanding of the concepts behind each non-adherence adjustment method in terms of estimands and estimators.

d) Provided new simulation evidence on the performance of four alternative non-adherence adjustment methods (two g-methods and two simple methods) across a range of 90 scenarios across all types of non-adherence (initiation, implementation and persistence). The simulation is superior to previous comparable studies; it was based on international guidelines for planning simulation studies,[113, 115] aligned with the influential ABC taxonomy for medication adherence,[1] included the simple censoring PP method, used five performance measures and four estimands and a nested loop plot to assess performance patterns. Overall, the findings demonstrated that g-methods performed consistently better than ITT in terms of unbiasedness and ModSE across all types of non-adherence.

e) Applied the g-methods in a case study and provided new evidence on the adherence-adjusted cost-effectiveness of maintenance immunosuppressive therapy for adult kidney transplant recipients in the UK taking into account their adherence levels in the real world. The case study demonstrated that the impact of lower levels of adherence to these treatments in the real world leads to a reduction in the NHBs and an increase in the average cost per patient when the results are compared with the estimates from the standard ITT analysis.

8.4 Strengths and limitations

There are a number of strengths and limitations of this research. In terms of the strengths, the development of the methodological framework put forward in this thesis was informed by a recent guide on how to develop a methodological framework.[152] Based on that guide, a new literature review was undertaken in 2021 to identify any existing methodological framework as discussed in Chapter 7. The review used more search terms and strategies (compared with the initial scoping review) based on learning from previous stages of this research and the recent guidance by McMeeken et al.[152] paper. In the new review, two major databases (MEDLINE[R] and Web of Science) were searched

systematically and the review concluded that no methodological framework exists. Thus, the methodological framework put forward in this thesis is likely to be the first one that could be followed to account for real-world non-adherence levels in economic evaluations using the g-methods.

Another key strength is that the development of the methodological framework was informed by a systematic approach that utilised four linked studies undertaken in this doctoral research as summarised in Section 8.2. The key strengths of each stage of this research are highlighted in the following paragraphs.

In Stage 1, the systematic review of methodological papers reported in Chapter 2 used novel iterative search techniques and followed international guidelines, with methods that were pre-specified in a published protocol.[62] The findings from the systematic review, and the associated work in terms of assessing the suitability adjustment method for the HTA context (Chapter 3), have informed the design of the simulation study.

In Stage 2, the simulation study (Chapters 4-5) followed a pre-specified protocol and the study design was based on the best available international guidelines.[1, 113, 115] The simulation study design was aligned with the ABC medication adherence taxonomy to facilitate transferability.[1] The simulations provided evidence on nuances on relationships between prognostic variables, patient mix, and adherence to medication over time, and the ability of g-methods to model these relationships. The recommendations from the simulations were applied in the case study.

In Stage 3, the case study in kidney transplantation (Chapter 6) provided a practical example of applying the g-methods for generating adherence-adjusted estimates of clinical effectiveness and cost-effectiveness of treatment. The study used individual patient-level data from the SYMPHONY study (a large, multicentre RCT with data from 1,190 patients and 12 months follow-up) which is considered as one of the strengths of this research.[116] The impact of real-world non-adherence on the cost-effectiveness of treatments was assessed by incorporating the adjusted clinical effectiveness estimates in a validated economic model that underpinned the recent update of NICE Technology Appraisal guidance for immunosuppressive therapy for adults in the UK (TA85).[120]

In Stage 4, the development of the methodological framework (Chapter 7) was informed by the work undertaken in Stages 1-3. This was further complemented by a discussion with two clinicians and a leading expert in medication adherence research. The comments received from these experts on an early version of the methodological framework were used to amend the framework to the version presented within this thesis.

The research reported in this thesis has some limitations. First, only one case study was used to inform the development of the methodological framework. Application in multiple case studies was considered at an early stage when the application for the doctoral fellowship that led to funding this research was developed. However, the final plan of investigation implemented in this research included a single case study as reported in Chapter 6 to ensure that this doctoral research project is completed within the funded period. Nevertheless, the development of the methodological framework was informed by the practical experience of applying the adjustment methods in the kidney transplantation case study. This was complemented by the consultations with clinicians and an expert in medication adherence research as discussed above. The McMeeken et al.[152] guide for developing a methodological framework suggested undertaking an evaluation of the framework by application in studies and/or evaluation based on a Delphi panel.

Another limitation is that the adjustment methods recommended in the methodological framework are only applicable to studies with time-to-event outcomes. This is because the methods identified by the systematic review (Chapter 2) and further assessed in the simulation study (Chapters 4-5) are focused on time-event outcomes and therefore the framework does not apply to continuous or categorical outcomes. However, I argue that most aspects of the current methodological framework put forward in this thesis would still be appropriate for non-time-to-event outcomes, although the analyst will need to use different methods to make the non-adherence adjustment to the treatment effectiveness estimates. In most cases, these will be different variants of the same methods recommended in the present framework. These may include structural nested mean models [SNMM], structural nested distribution models [SNDM] and structural nested logistic models [SNLM]).[156] However, these methods have not been assessed in this research and therefore this should be considered as a limitation. Apart from the recommended adjustment methods, most of the recommendations and considerations outlined in the framework to adjust for non-adherence in studies with time-to-event outcomes are likely to also be applicable to studies with non-time-to-event outcomes.

A further limitation relates to the exclusion of the PKPD method from the assessment in the simulation study. The PKPD method was identified as appropriate to adjust for real-world non-adherence in the HTA context;[61] however, it was excluded from the simulations because it requires a different study design in terms of DGMs in order to directly compare it with the g-methods. The PKPD method has been assessed to adjust the cost-effectiveness of treatments for varying levels of patient non-adherence in a recent doctoral thesis by Hill-McManus et al.[91] The method was applied in a case study that assessed the impact of non-adherence on the cost-effectiveness of 4 options of dual urate-

lowering therapy (ULT) using three patterns of medication adherence (100%, 80% and 50% adherence levels) using a linked PKPD model. That research has demonstrated that the method could be used for adjusting the cost-effectiveness estimates for real-world adherence levels. Therefore, the PKPD method is available as an option for consideration as an alternative adjustment method based on evidence from the Hill-McManus work. However, it should be noted that the PKPD performance versus g-methods has not been assessed in this doctoral research which could be considered as a limitation. Future research is recommended to directly assess the performance of the PKPD method versus g-methods for potential addition to the methodological framework.

Other limitations include using only one method for predicting real-world non-adherence, dealing with intermittent non-adherence, the performance of g-methods in the presence of unmeasured confounding, non-convergence issues and availability of medication adherence data in RCTs. These issues are briefly discussed in this paragraph. First, in the case study, only one method for predicting non-adherence was tested (i.e. adjusting the CV% cut-off point). This limitation is discussed further in Section 8.6 as further research is needed on it. Second, implementation non-adherence using a non-binary variable was not examined and this will apply to a significant proportion of patients. The key point is that using a binary variable to capture implementation non-adherence risks losing information if granular data on intermittent non-adherence (e.g. from MEMS) is available for the analyst. Third, the performance of g-methods in the presence of unmeasured confounders seems important but this has not been assessed in the simulation study which should be considered as a limitation. Fourth, the issue of non-convergence with small sample sizes was not fully investigated beyond the small sample size of 120 patients in an RCT. Therefore, using g-methods for analysing trial data with a smaller sample size might be problematic and this issue needs further investigation in future simulation studies. Fifth, in many chronic disease treatments, primary licensing RCTs have not routinely included assessments of adherence and data on real-world adherence data is limited. In these situations, it will be challenging to apply the framework in the absence of adherence data. This should be considered as a limitation on how widely this framework could be used in practice. The last issue is particularly relevant to economic evaluations using existing RCT datasets.

Finally, multiple medication adherence (MMA) has been identified as an important issue when considering adjusting for non-adherence for regimens that include a combination of multiple treatments.[135] This issue relates to "polypharmacy" which is a major issue in many disease states and has a contribution to medication adherence behaviours which is too complicated to address. For example, in the SYMPHONY trial used in the case study, non-adherence to MMF used as part of the immunosuppressive regimen evaluated in the trial may have been better or worse than with the CNI

(i.e. tacrolimus or cyclosporine). This issue was identified as a limitation in the case study since adherence data on the other medications within the maintenance immunosuppressive regimens were not collected in the SYMPHONY trial. This is likely to be the case in many studies, and therefore, the issue warrants further research to specify methods for measuring MMA. Further research to assess how these could be incorporated into the present methodological framework should also be considered as an important area for future research.

8.5 Recommendations

The main recommendations from this work are provided in a form of a methodological framework that I put forward in Chapter 7, which I will not repeat here. Wider recommendations to apply the framework are provided in this section. These are of relevance to clinical trialists, the pharmaceutical industry, health economists and economic modellers involved in the design of clinical trials that are intended to investigate the clinical effectiveness and cost-effectiveness of pharmacological interventions for the treatment of chronic diseases or conditions. The recommendations are outlined below.

First, although the development of the methodological framework was informed by a single case study in kidney transplantation, the framework is potentially transferable to other disease areas. Therefore, future applications across a range of case studies in different disease areas will provide more information about the applicability of the framework in those contexts. An evaluation of the framework will be needed to test its validity and transferability to other disease areas. Therefore, I recommend and encourage the abovementioned stakeholders to apply the framework in their own studies. This recommendation applies to studies with time-to-event outcomes intending to account for real-world non-adherence levels in economic evaluations.

Second, although the framework is focused on chronic medications, I argue that many of its recommendations will be applicable to other studies assessing the cost-effectiveness of medication used in the treatment of non-chronic diseases as well as non-pharmacological interventions where an appropriate measurement of adherence, prognostic characteristics and a time-to-event outcome are collected in the trial. Therefore, I recommend using the framework to consider the relevant elements as they apply to each particular study at the design, analysis and/or reporting stages. This in turn will provide additional information relating to the application of the framework in practice which in turn will help in evaluating and refining the framework for future improvements.

Finally, I recommend that health economists and economic modellers should document the process of their plan to adjust for non-adherence at the study designs stage. To achieve this, I recommend

using the health economics analysis plan (HEAP) as a vehicle to document these processes to aid the discussion with other stakeholders involved including clinicians, statisticians and resource allocation decision-makers.

8.6 Areas for further development and future research

Areas for future research that emerged from each of the four studies undertaken in this research (Stages 1-4) have been reported within each chapter. In this section, I focus on those relating to the further development of the work presented in this doctoral thesis.

Dissemination and evaluation of the methodological framework

The development of the methodological framework was informed by a recent guide for developing a methodological framework, literature reviews, a simulation study and a case study. However, further evaluation is recommended for improvements in the future once the current framework is applied in practice in multiple case studies. The current framework will be disseminated widely to make it more accessible to clinical trialists, academic researchers and the pharmaceutical industry in order to be considered and used in their future studies. The wider dissemination will be achieved via peer-reviewed publications from the work that informed the development of the framework (i.e. simulation study, case study and the development of the methodological framework). The dissemination plan also includes publishing a tutorial-style paper showing how the g-methods could be applied in practice with detailed analytical steps and analysis code which will be based on the work undertaken in the simulation study and case study as reported in Chapters 4-6 of this thesis. The systematic review work was published in Medical Decision Making journal,[61] and the following further manuscripts are in preparation based on this work, including the target journals:

Alshreef A, Latimer N, Tappenden P, Dixon S. Assessing methods for adjusting estimates of treatment effectiveness for patient non-adherence in the context of time-to-event outcomes and health technology assessment: a simulation study. Medical Decision Making. (In preparation).

Alshreef A, Tappenden P, Latimer N, McKane W, Fotheringham J, Dixon S. A case study assessing the cost-effectiveness of maintenance immunosuppressive therapy for kidney transplantation in adults: accounting for non-adherence using data from a randomised trial and the real world within a decision-analytic model. Value in Health. (In preparation).

Alshreef A, Tappenden P, Latimer N, McKane W, Hughes D, Dixon S. A methodological framework to account for the impact of non-adherence on the cost-effectiveness of prescribed chronic medications. Value in Health. (In preparation).

Furthermore, I am planning to look for opportunities to present the framework at various workshops, meetings and national and international conferences. These will include seminars to health economics groups at various universities. Presentations at other universities and conferences would be a good opportunity for peer review and would also help identify possible participants for an expert workshop or Delphi exercise to revise the Framework and prioritise areas for future work. I will also deposit my PhD thesis as an open-access document in the White Rose Online Research repository.

Although the case study was focused is on kidney transplantation, I envisage that the methodological framework will apply to any chronic disease area in studies with a time-to-event outcome. Therefore, I will look for opportunities to apply the framework in future case studies in other disease areas and other researchers are also encouraged to apply the methodological framework in their studies. The application of the framework across a range of case studies will provide additional information to evaluate it with the intention of refining and further improvement. The evaluation could be achieved by conducting in-depth interviews with investigators and researchers who apply the framework in their own studies in the future.

Extending the framework to incorporate other adjustment methods

Two adjustments methods (PKPD and RPSFTM with g-estimation) were identified as appropriate for use in HTA but not fully recommended in the methodological framework. The two methods represent viable options for inclusion in the framework, but they were excluded from the simulation study as discussed earlier, and therefore, further research is required regarding the assessment of their performance. The PKPD has been assessed in a recent doctoral thesis as discussed earlier; however, it was not directly assessed for performance against the g-methods. Similarly, the RPSFTM was identified as appropriate for adjusting for real-world adherence levels, but its performance in this context was not assessed in the simulation study reported in Chapters 4-5. Therefore, both methods are currently in the framework for further consideration. The implications will be minimal as two g-methods (IPCW and SNFTM) are recommended in the current framework based on the simulation evidence. Overall, the potential addition of PKPD and RPSFTM falls into the category of key areas for future research and possible extensions to the current framework.

Extending the framework to cover other types of outcomes

The current framework is focused on studies with time-to-event outcomes. The intention is to extend the framework in the future by adding other adjustment methods to account for non-adherence in the context of other outcomes including continuous and categorical outcomes. The advantage is that

many of the elements in the current framework will apply to studies other outcomes. The argument is that the current framework could still be applied in those contexts but the limitation is that other variants of g-methods (e.g. SNMM and SNDM) and other methods may be needed to do the adjustments for non-adherence.[156] Therefore, future research is warranted to assess the performance of these methods in a well-conducted simulation study. It is my intention that the current methodological framework will evolve to incorporate other appropriate adjustment methods and outcomes. This will ensure that the framework will cater for most types of studies used to evaluate a wide range of medications across different disease areas and conditions.

Further research to assess methods for predicting real-word non-adherence

Only one method for predicting real-world non-adherence levels was used based on CV% of trough levels measured within the trial. There is a limitation associated with using CV% as the method cannot be used to predict real-world adherence levels within the randomised dataset if trough levels data are not available. As an alternative approach, the marginal standardisation method might be considered in scenarios where trough levels data are not available.[147] The marginal standardisation method is considered a special case of the g-computation method and has the potential to add value to the methodological framework. Therefore, further research is needed to assess these methods and potentially alternative prediction methods could also be identified and assessed. The censoring mechanism based on the predicted non-adherence should also be considered for assessment in future research alongside the prediction method, ideally in a well-conducted simulation study.

Further research to assess adjustment for multiple medication adherence

Assessing the impact of MMA on the clinical effectiveness and cost-effectiveness of prescribed chronic medications was identified as a key area for future research. The issue is that there is no consensus on how MMA could be measured as characterised by a recent report of the ISPOR Medication Adherence and Persistence Special Interest Group.[135] Once this issue is resolved, future research should consider how to deal with MMA in the methodological framework which will lead to further improvements.

8.7 Conclusions

The overall aim of this research was to develop a methodological framework to account for patient non-adherence to prescribed chronic medications with time-to-event outcomes when undertaking economic evaluations for HTA. The methodological framework has been developed and reported in

this thesis to provide a systematic approach for incorporating real-world non-adherence in economic evaluations.

When this research started, there was no methodological framework to account for the impact of non-adherence on the cost-effectiveness of medications. The research reported in this thesis has attempted to fill that gap and provides a significant contribution to the health economic literature. The framework is based on using more complex methods (g-methods) for adjusting estimates of treatment effectiveness in the presence of patient non-adherence and further account for it in cost-effectiveness models to produce better evidence for resource allocation decision making in health care.

The systematic approach put forward through this framework should help in capturing the important elements that need to be considered at an early stage of the study design to account for real-world adherence levels in economic evaluations when these differ from adherence levels observed in the trial. The framework should also help in avoiding inconsistencies between economic studies incorporating patient non-adherence at the analysis and reporting stage. This will lead to standardised and comparable adherence-adjusted cost-effectiveness estimates in the medium- to long term.

The methodological framework provides practical recommendations for investigators, analysts and the pharmaceutical industry interested in accounting for patient non-adherence in their studies. The key benefits of the framework will be realised at the study design stage and when the trial is completed and the analysis begins and final results are reported. At the study design stage, the benefits involve the identification of the key data requirements include medication adherence data and information about baseline and time-dependent confounding, which are essential for undertaking the adjusted analysis. At the analysis stage, the framework provides a systematic approach including the selection of the appropriate adjustment methods and analytical steps that should be followed to account for real-world non-adherence levels. The framework is aligned with the temporal phases of medication adherence (initiation, implementation and persistence) as defined by the ABC taxonomy to aid transferability and consistency with existing international guidelines.[1]

The main contribution of this original research is that the framework put forward in this thesis provides guidance for academic researchers (health economists and economic modellers) and the pharmaceutical industry to account for real-world non-adherence levels in economic evaluations for HTA. The framework will contribute to improvement in the overall quality of economic models used for estimating cost-effectiveness. This will lead to better cost-effectiveness evidence to inform

resource allocation decision making, improvements in patients' quality of life, cost savings to healthcare systems and ultimately leads to improvements in population health.

References

1. Vrijens B, Geest SD, Hughes DA, Przemyslaw K, Demonceau J, Ruppar T, et al. A new taxonomy for describing and defining adherence to medications. Br J Clin Pharmacol. 2012;73(5):691–705.

2. Andrzejczyk A, Clyne W, Geest SD, Demonceau J, Dobbels F, Fargher E, et al. Ascertaining barriers for compliance: policies for safe, effective and cost-effective use of medicines in Europe. 2012.

3. Muszbek N, Brixner D, Benedict A, Keskinaslan A, Khan ZM. The economic consequences of noncompliance in cardiovascular disease and related conditions: a literature review. Int J Clin Pract. 2008;62(2):338–51.

4. Elliott RA, Shinogle JA, Peele P, Bhosle M, Hughes DA. Understanding Medication Compliance and Persistence from an Economics Perspective. Value Health. 2008;11(4):600–10.

5. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? A systematic review of definitions. BMC Geriatrics. 2017;17(230).

6. Gellad WF, Thorpe CT, Steiner JF, Voils CI. The myths of medication adherence. Pharmacoepidemiol Drug Saf. 2017;26(12):1437-41.

7. Lam WY, Fresco P. Medication Adherence Measures: An Overview. Biomed Res Int. 2015:1-12.

8. Sperber CM, Samarasinghe SR, Lomax GP. An upper and lower bound of the Medication Possession Ratio. Patient Prefer Adherence. 2017;11:1469–78.

9. Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 Statist Sci. 1990;5(4):465-72.

10. Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol. 1974;66:688–701.

11. Robins JM, Greenland S, Hu F-C. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. J Am Stat Assoc. 1999;94(447):687-700.

12. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effect. Math Model. 1986;7(9):1393-512.

13. Hernán MA, Robins JM. Causal Inference: What If. Boston: Boca Raton: Chapman & Hall/CRC; 2020.

14. Akachaa M, Bretza F, Ohlssen D, Rosenkran G, Schmidli H. Estimands and Their Role in Clinical Trials. Stat Biopharm Res. 2017;9(3):268-71.

15. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials London: European Medicines Agency; 2017 [Available from: https://www.ema.europa.eu/documents/scientific-guideline/draft-ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical_en.pdf.

16. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. Stata J. 2002;2(2):164–82.

17. Aalen O, Røysland K, Gran J, Kouyos R, Lange T. Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. Stat Methods Med Res. 2016;25(5):2294–314.

18. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417–20.

19. Willems S, Schat A, Noorden Mv, Fiocco M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Stat Methods Med Res. 2018;27(2):323–35.

20. Lin D, Robins J, Wei L. Comparing two failure time distributions in the presence of dependent censoring. Biometrika. 1996;83(2):381-93.

21. Morrison A, Stauffer ME, Kaufma AS. Relationship Between Adherence Rate Threshold and Drug 'Forgiveness'. Clin Pharmacokinet. 2017;56:1435–40.

22. Osterberg LG, Urquhart J, Blaschke TF. Understanding Forgiveness: Minding and Mining the Gaps Between Pharmacokinetics and Therapeutics. Clin Pharmacol Ther. 2010;88(4):457–9.

23. Dartois V. Drug Forgiveness and Interpatient Pharmacokinetic Variability in Tuberculosis. J Infect Dis. 2011;204(12):1827–9.

24. Cleemput I, Kesteloot K, DeGeest S. A review of the literature on the economics of noncompliance. Room for methodological improvement. Health Policy. 2002;59:65–94.

25. Barber N. Should we consider non-compliance a medical error? Qual Saf Health Care. 2002;11:81–4.

26. Vermeire E, Hearnshaw H, Royen PV, Denekens J. Patient adherence to treatment: three decades of research. A comprehensive review. J Clin Pharm Ther. 2001;26(5):331–42.

27. Hughes DA, Bagust A, Haycox A, Walley T. Accounting for Noncompliance in Pharmacoeconomic Evaluations. Pharmacoeconomics. 2001;19(12):1185-97.

28. Negus SS, Banks ML. Pharmacokinetic—Pharmacodynamic (PKPD) Analysis with Drug Discrimination. Curr Top Behav Neurosci. 2018;39:245–59.

29. Hughes D, Hughes DA, Sridhar FG, Elliott R. The assessment, determinants & economics of medication compliance & persistence. 10th Annual European Congress of the International Society for Pharmacoeconomics and Outcomes Research; Dublin, Ireland: ISPOR; 2007.

30. Orr A, Orr D, Willis S, Holmes M, Britton P. Patient perceptions of factors influencing adherence to medication following kidney transplant. Psychol Health Med. 2007;12(4):509-17.

31. Trueman P, Lowson K, Blighe A, Meszaros A, Wright D, Glanville J, et al. Evaluation of the Scale, Causes and Costs of Waste Medicines: The Economic Impact of Poor Compliance.: York Health Economics Consortium and School of Pharmacy, University of London; 2010.

32. Hughes D, Cowell W, Koncz T, Cramer J. Methods for Integrating Medication Compliance and Persistence in Pharmacoeconomic Evaluations. Value Health. 2007;10(6):498–509.

33. Hughes D, Charles J, Dawoud D, Edwards RT, Holmes E, Jones C, et al. Conducting Economic Evaluations Alongside Randomised Trials: Current Methodological Issues and Novel Approaches. Pharmacoeconomics. 2016.

34. Shroufi A, Powles J. Adherence and chemoprevention in major cardiovascular disease: a simulation study of the benefits of additional use of statins. J Epidemiol Community Health. 2010;64(2):109-13.

35. Kadambi A, Leipold RJ, Kansal AR, Sorensen S, Getsios D. Inclusion of Compliance and Persistence in Economic Models: Past, Present and Future. Appl Health Econ Health Policy. 2012;10(6):365-79.

36. Robins JM. Correction for non-compliance in equivalence trials. Stat Med. 1998;17(3):269-302.

37. Robins JM, Tsiatis AA. Correcting for noncompliance in randomized trials using rank preserving structural failure time models. Commun Stat-Theory Methods. 1991;20(8):2609-31.

38. Kubo Y, Sterling LR, Parfrey PS, Gill K, Mahaffey KW, Gioni I, et al. Assessing the treatment effect in a randomized controlled trial with extensive non-adherence: the EVOLVE trial. Pharm Stat. 2015;14(3):242-51.

39. Korhonen PA, Laird NM, Palmgren J. Correcting for non-compliance in randomized trials: An application to the ATBC Study. Stat Med. 1999;18(21):2879-97.

40. Zhang M, Tsiatis AA, Davidian M, Pieper KS, Mahaffey KW. Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation: the SYNERGY trial. Biostatistics. 2011;12(2):258-69.

41. Latimer N. Systematic review of statistical methods for adjusting survival estimates in the presence of treatment crossover. Sheffield University of Sheffield; 2012.

42. Hughes D, Dubois D. Cost-effectiveness analysis of extended-release formulations of oxybutynin and tolterodine for the management of urge incontinence. Pharmacoeconomics. 2004;22(16):1047-59.

43. Suárez PG, Floyd K, Portocarrero J, Alarcón E, Rapiti E, Ramos G, et al. Feasibility and cost-effectiveness of standardised second-line drug treatment for chronic tuberculosis patients: a national cohort study in Peru. Lancet. 2002;359(9322):1980-9.

44. Fischer K, Goetghebeur E, Vrijens B, White IR. A structural mean model to allow for noncompliance in a randomized trial comparing 2 active treatments. Biostatistics. 2011;12(2):247–57.

45. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost‐effectiveness analysis. Health Econ. 2002;11(5):415-30.

46. Hemels MEH, Kasper S, Walter E, Einarson TR. Cost-effectiveness of escitalopram versus citalopram in the treatment of severe depression. Ann Pharmacother. 2004;38(6):954-60.

47. Edwards N, Rupnow M, Pashos C, Botteman M, Diamond R. Cost-effectiveness model of long-acting risperidone in schizophrenia in the US. Pharmacoeconomics. 2005;23(3):299-314.

48. Haby MM, Tonge B, Littlefield L, Carter R, Vos T. Cost-effectiveness of cognitive behavioural therapy and selective serotonin reuptake inhibitors for major depression in children and adolescents. Aust N Z J Psychiatry. 2004;38(8):579–91.

49. Donnelly M, Haby MM, Carter R, Andrews G, Vos T. Cost-effectiveness of dexamphetamine and methylphenidate for the treatment of childhood attention deficit hyperactivity disorder. Aust N Z J Psychiatry. 2004;38(8):592–601.

50. O'Brien BJ, Goeree R, Bernard L, Rosner A, Williamson T. Cost-effectiveness to tolterodine for patients with urge incontinence who discontinue initial therapy with oxybutynin: A Canadian perspective. Clin Ther. 2001;23(12):2038-49.

51. Chisholm D. Cost-effectiveness of First-line Antiepileptic Drug Treatments in the Developing World: A Population-level Analysis. Epilepsia. 2005;46(5):751-9.

52. Jasmer R, Snyder D, Chin D, Hopewell P, Cuthbert S, AntonioPaz E, et al. Twelve months of isoniazid compared with four months of isoniazid and rifampin for persons with radiographic evidence of previous tuberculosis: an outcome and cost-effectiveness analysis. Am J Respir Crit Care Med. 2005;162(5):1648-52.

53. Chue PS, Heeg B, Buskens E, van-Hout BA. Modelling the impact of compliance on the costs and effects of long-acting risperidone in Canada. Pharmacoeconomics. 2005;23 Suppl 1:62-74.

54. Caro J. Pharmacoeconomic analyses using discrete event simulation. Pharmacoeconomics. 2005;23(4):1179-2027.

55. Maclean JR, Pfister M, Zhou Z, Roy A, Tuomari VA, Heifets M. Quantifying the impact of nonadherence patterns on exposure to oral immunosuppressants. Ther Clin Risk Manag. 2011;7:149–56.

56. Pfister M, Labbé L, Hammer SM, Mellors J, Bennett KK, Rosenkranz S, et al. Population Pharmacokinetics and Pharmacodynamics of Efavirenz, Nelfinavir, and Indinavir: Adult AIDS Clinical Trial Group Study 398. Antimicrob Agents Chemother. 2003;47(1):130–7.

57. Labbé L, Verotta D. A non-linear mixed effect dynamic model incorporating prior exposure and adherence to treatment to describe long-term therapy outcome in HIV-patients. J Pharmacokinet Pharmacodyn. 2006;33(4):519–42.

58. Kenna LA, Labb L, Barrett JS, Pfister M. Modeling and Simulation of Adherence: Approaches and Applications in Therapeutics. AAPS J. 2005;7(2):390-407.

59. Hughes DA, Walley T. Predicting “real world” effectiveness by integrating adherence with pharmacodynamic modelling. Clin Pharmacol Ther. 2003;74(1):1–8.

60. NICE. Guide to the methods of technology appraisal: Processes and methods guides London: National Institute for Health and Care Excellence; 2013 [Available from: https://www.nice.org.uk/process/pmg9/chapter/foreword.

61. Alshreef A, Latimer N, Tappenden P, Wong R, Hughes D, Fotheringham J, 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. Med Decis Making. 2019;39(8):910–25.

62. 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;41(5):567–82.

63. Centre for Reviews and Dissemination. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: Centre for Reviews and Dissemination, University of York; 2008.

64. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444-55.

65. Alshreef A, Dixon S, Latimer N, Tappenden P, McKane W. Systematic review of methods to account for patient non-adherence on estimating treatment effect and cost-effectiveness of chronic medications: Review protocol York: PROSPERO; 2018 [Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018095544.

66. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56(3):779-88.

67. Wang W, Husan F, Chow SC. The impact of patient compliance on drug concentration profile in multiple doses. Stat Med. 1996;15(6):659-69.

68. Vrijens B, Goetghebeur E, de Klerk E, Rode R, Mayer S, Urquhart J. Modelling the association between adherence and viral load in HIV-infected patients. Stat Med. 2005;24(17):2719-31.

69. Butler JA, Roderick P, Mullee M, Mason JC, Peveler RC. Frequency and impact of nonadherence to immunosuppressants after renal transplantation: a systematic review. Transplantation. 2004;77(5):769–89.

70. Yale University. Yale MeSH Analyzer New Haven: Yale University; 2018 [Available from: http://mesh.med.yale.edu/.

71. Online-Utility. Online Utility Text Analyzer 2018 [Available from: https://www.online-utility.org/text/analyzer.jsp.

72. Moher D, Liberati A, Tetzlaff J, Altman D. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264-9.

73. Frangakis CE, Rubin DB. Principal Stratification in Causal Inference. Biometrics. 2002;58(1):21-9.

74. Yu W, Chen K, Sobel ME, Ying ZL. Semiparametric transformation models for causal inference in time-to-event studies with all-or-nothing compliance. J R Stat Soc Series B Stat Methodol. 2015;77(2):397-415.

75. Wu Y, Zhao L, Hou Y, Li K, Zhou X. Correcting for non-compliance in randomized non-inferiority trials with active and placebo control using structural models. Stat Med. 2015;34(6):950-65.

76. Cuzick J, Sasieni P, Myles J, Tyrer J. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. J R Stat Soc Series B Stat Methodol. 2007;69:565-88.

77. Lin JY, Ten Have TR, Bogner HR, Elliott MR. Baseline patient characteristics and mortality associated with longitudinal intervention compliance. Stat Med. 2007;26(28):5100-15.

78. Li S, Gray RJ. Estimating treatment effect in a proportional hazards model in randomized clinical trials with all-or-nothing compliance. Biometrics. 2016;72(3):742-50.

79. Loeys T, Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. Biometrics. 2003;59(1):100-5.

80. Baker SG. Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. J Am Stat Assoc. 1998;93(443):929-34.

81. Nie H, Cheng J, Small DS. Inference for the effect of treatment on survival probability in randomized trials with noncompliance and administrative censoring. Biometrics. 2011;67(4):1397-405.

82. Gao X, Zheng M. Estimating the causal effects in randomized trials for survival data with a cure fraction and non compliance. Commun Stat-Theory Methods. 2017;46(8):4065-87.

83. Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of NonrandomizedTreatments. J Am Stat Assoc. 2001;96(454):440-8.

84. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis-carinii pneumonia on the survival of aids patients. Epidemiology. 1992;3(4):319-36.

85. Loeys T, Vansteelandt S, Goetghebeur E. Accounting for correlation and compliance in cluster randomized trials. Stat Med. 2001;20(24):3753-67.

86. Korhonen P, Palmgren J. Effect modification in a randomized trial under non-ignorable non-compliance: an application to the alpha-tocopherol beta-carotene study. J R Stat Soc Ser C Appl Stat. 2002;51:115-33.

87. Loeys T, Goetghebeur E. Baseline information in structural failure time estimators for the effect of observed treatment compliance. Stat Med. 2002;21(9):1173-88.

88. Matsui S. Analysis of times to repeated events in two-arm randomized trials with noncompliance and dependent censoring. Biometrics. 2004;60(4):965-76.

89. White IR, Goetghebeur EJT. Clinical trials comparing two treatment policies: Which aspects of the treatment policies make a difference? Stat Med. 1998;17(3):319-39.

90. Pink J, Pirmohamed M, Lane S, Hughes DA. Cost-Effectiveness of Pharmacogenetics-Guided Warfarin Therapy vs. Alternative Anticoagulation in Atrial Fibrillation. Clin Pharmacol Ther. 2014;95(2):199-207.

91. Hill-McManus D, Marshall S, Soto E, Lane S, Hughes D. Impact of non-adherence and flare resolution on the cost effectiveness of treatments for gout: Application of a linked pharmacometric/pharmacoeconomic model. Value Health 2018;21(12):1373-81.

92. Cox DR. Regression Models and Life-Tables. J R Stat Soc Series B Stat Methodol. 1972;34(2):187-220.

93. Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomised clinical trials. Stat Med. 1997;16(9):1017-29.

94. Lin JY, Ten Have TR, Elliott MR. Nested Markov compliance class model in the presence of time-varying noncompliance. Biometrics. 2009;65(2):505-13.

95. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. Ann Stat. 1997;25(1):305-27.

96. Rubin DB. More powerful randomization-based p-values in double-blind trials with non-compliance. Stat Med. 1998;17(3):371-85.

97. Imbens GW, Rubin DB. Estimating outcome distributions for compliers in instrumental variables models. Rev Econ Stud. 1997;64(4):555-74.

98. Robins J, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. Biometrika. 2004;91(4):763-83.

99. Lok JJ. Causal Inference: Structural Nested Models. Summer Short Course "An Introduction to Causal Inference"2018.

100. Latimer NR, Henshall C, Siebert U, Bell H. Treatment Switching: Statistical and Decision-Making Challenges and Approaches. Int J Technol Assess Health Care. 2016;32(3):160–6.

101. 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. Stat Methods Med Res. 2020;29(10):2900–18.

102. Hamberg AK, Wadelius M, Lindh JD, Dahl ML, Padrini R, Deloukas P, et al. A pharmacometric model describing the relationship between warfarin dose and INR response with respect to variations in CYP2C9, VKORC1, and age. Clin Pharmacol Ther. 2010;87(6):727-34.

103. Odondi Lo, McNamee R. Performance of statistical methods for analysing survival data in the presence of non-random compliance. Stat Med. 2010;29(29):2994-3003.

104. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received - is it really an option. Stat Med. 1991;10(10):1595-605.

105. Mark SD, Robins JM. Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. Stat Med. 1993;12(17):1605-28.

106. Robins JM, Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high-dose versus low-dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc. 1994;89(427):737-49.

107. Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: An application in a clinical trial of unresectable non-small-cell lung cancer. Stat Med. 2004;23(13):2005-22.

108. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2009;28(12):1725-38.

109. Hernán MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. Epidemiology. 2000;11(5):561-70.

110. Fewell Z, Hern´an MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. Stata J. 2004;4(4):402–20.

111. Schwab J, Lendle S, Petersen M, Laan Mvd, Gruber S. Longitudinal Targeted Maximum Likelihood Estimation Vienna: The R Foundation; 2018 [Available from: https://cran.r-project.org/web/packages/ltmle/ltmle.pdf.

112. Wojciechowski J, Hopkins A, Upton R. Interactive Pharmacometric Applications Using R and the Shiny Package. CPT Pharmacometrics Syst Pharmacol. 2015;4(3):146–59.

113. Morris TP, White I, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-102.

114. Russell CL, Conn VS, Ashbaugh C, Madsen R, Hayes K, Ross G. Medication Adherence Patterns in Adult Renal Transplant Recipients. Res Nurs Health. 2006;29:521–32.

115. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25(24):4279–92.

116. Ekberg H, Tedesco-Silva H, Demirbas A, Vítko Š, Nashan B, Gürkan A, et al. Reduced Exposure to Calcineurin Inhibitors in Renal Transplantation. N Engl J Med. 2007;357(25):2562-75.

117. Kennedy C, Mann CB. "RANDOMIZE: Stata module to create random assignments for experimental trials, including blocking, balance checking, and automated rerandomization. Boston College Department of Economics; 2015.

118. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Stat Med. 2013;32(23):4118–34.

119. Crowther MJ, Lambert PC. Simulating complex survival data. Stata J. 2012;12(4):674-87.

120. Jones-Hughes T, Snowsill T, Haasova M, Coelho H, Crathorne L, Cooper C, et al. Immunosuppressive therapy for kidney transplantation in adults: a systematic review and economic model. Health Technol Assess. 2016;20(62):1-594.

121. Pizzo HP, Ettenger RB, Gjertson DW, Reed EF, Zhang J, Gritsch HA, et al. Sirolimus and tacrolimus coefficient of variation is associated with rejection, donor-specific antibodies, and nonadherence. Pediatr Nephrol. 2016;31:2345–52.

122. Rodrigo E, Segundo DS, Fernández-Fresnedo G, López-Hoyos M, Benito A, Ruiz JC, et al. Within-Patient Variability in Tacrolimus Blood Levels Predicts Kidney Graft Loss and Donor-Specific Antibody Development. Transplantation. 2016;100(11):2479-85.

123. White IR. simsum: Analyses of simulation studies including Monte Carlo error. Stata J. 2010;10(3):369-85.

124. Rücker G, Schwarzer G. Presenting simulation results in a nested loop plot. BMC Med Res Methodol. 2014;14(129):1-8.

125. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013;13(152):1-15.

126. Geest SD, Zullig LL, Dunbar-Jacob J, Helmy R, Hughes DA, Wilson IB, et al. ESPACOMP Medication Adherence Reporting Guideline (EMERGE). Ann Intern Med. 2018;169(1):30-5.

127. Latimer NR, Abrams KR, Sieber U. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. BMC Med Res Methodol. 2019;19(69):1-19.

128. Latimer NR, Abrams KR, Lambert PC, Morden JP, Crowther MJ. Assessing methods for dealing with treatment switching in clinical trials: A follow-up simulation study. Stat Methods Med Res. 2018;27(3):765–84.

129. Latimer NR, Abrams K, Lambert P, Crowther M, Wailoo A, Morden J, et al. Adjusting for treatment switching in randomised controlled trials – A simulation study and a simplified two-stage method. Stat Methods Med Res. 2017;26(2):724–51.

130. Kuypers DRJ, Peeters PC, Sennesael JJ, Kianda MN, Vrijens B, Kristanto P, et al. Improved Adherence to Tacrolimus Once-Daily Formulation in Renal Recipients: A Randomized Controlled Trial Using Electronic Monitoring. Transplantation. 2013;95(2):333-40.

131. Albano L, Banas B, Klempnauer JL, Glyda M, Viklicky O, Kamar N. OSAKA Trial: A Randomized, Controlled Trial Comparing Tacrolimus QD and BD in Kidney Transplantation. Transplantation. 2013;96(10):897-903.

132. Haynes R, Harden P, Judge P, Blackwell L, Emberson J, Landray MJ, et al. Alemtuzumab-based induction treatment versus basiliximab-based induction treatment in kidney transplantation (the 3C Study): a randomised trial. Lancet. 2014;384(9955):1684-90.

133. Akobeng AK. Understanding randomised controlled trials. Arch Dis Child.90(8):840–4.

134. Sun Z-J, Du X, Su L-L, Zhang X-D, Wang W. Efficacy and Safety of Basiliximab Versus Daclizumab in Kidney Transplantation: A Meta-Analysis. Transplant Proc. 2015;47(8):2439-45.

135. Pednekar PP, Agh T, Malmenas M, Raval AD, Bennett BM, Borah BJ, et al. Methods for Measuring Multiple Medication Adherence: A Systematic RevieweReport of the ISPOR Medication Adherence and Persistence Special Interest Group. Value Health. 2019;22(2):139-56.

136. Shuker N, Gelder T, Hesselink DA. Intra-patient variability in tacrolimus exposure: Causes, consequences for clinical management. Transplant Rev. 2015;29:78-84.

137. Brunet M, van Gelder T, Asberg A, Haufroid V, Hesselink DA, Langman L, et al. Therapeutic Drug Monitoring of Tacrolimus-Personalized Therapy: Second Consensus Report. Ther Drug Monit. 2019;41(3):261-307.

138. Kuypers DRJ. Intrapatient Variability of Tacrolimus Exposure in Solid Organ Transplantation: A Novel Marker for Clinical Outcome. Clin Pharmacol Ther. 2020;107(2):347-58.

139. Whalen HR, Glen JA, Harkins V, Stevens KK, Jardine AG, Geddes CC, et al. High Intrapatient Tacrolimus Variability Is Associated With Worse Outcomes in Renal Transplantation Using a Low-Dose Tacrolimus Immunosuppressive Regime. Transplantation. 2017;101(2):430-6.

140. Lalić J, Veličković-Radovanović R, Mitić B, Paunović G, Cvetković T. Immunosuppressive Medication Adherence in Kidney Transplant Patients. Med Princ Pract. 2014;23(4):351-6.

141. Goodall DL, Willicombe M, McLean AG, Taube D. High Intrapatient Variability of Tacrolimus Levels and Outpatient Clinic Nonattendance Are Associated With Inferior Outcomes in Renal Transplant Patients. Transplant Direct. 2017;3(8):e192.

142. Gonzales HM, McGillicuddy JW, Rohan V, Chandler JL, Nadig SN, Dubay DA, et al. A comprehensive review of the impact of tacrolimus intrapatient variability on clinical outcomes in kidney transplantation. Am J Transplant. 2020;20(8):1969-83.

143. He X, Johnston A. Variable cyclosporine exposure: A risk factor for chronic allograft nephropathy and graft loss? Transplant Proc. 2004;36(5):1321-6.

144. Jorga A, Holt DW, Yaqoob M, Whittaker C, Johnston A. A survey to determine the blood concentration of cyclosporine 2 hours postdose in stable renal transplant patients. Transplant Proc. 2004;36(10):3239-41.

145. Snowsill TM, Moore J, Mota REM, Peters JL, Jones-Hughes TL, Huxley NJ, et al. Immunosuppressive agents in adult kidney transplantation in the National Health Service: a model-based economic evaluation. Nephrol Dial Transplant. 2017;32:1251–9.

146. Kronish I. Measuring the multiple dimensions of medication non-adherence: Findings from a Delphi survey of adherence experts. 22th ESPACOMP conference; Dublin, Ireland: International Society for Medication Adherence; 2018.

147. Muller CJ, MacLehose RF. Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. Int J Epidemiol. 2014;43(3):962-70.

148. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzib P. Strategies for Dealing with Missing Data in Clinical Trials: From Design to Analysis. Yale J Biol Med. 2013;86(3):343–58.

149. Ibrahim JG, Chu H, Chen M-H. Missing Data in Clinical Studies: Issues and Methods. J Clin Oncol. 2012;30(26):3297–303.

150. Estruch R, Ros E, Salas-Salvadó J, Covas M-I, Corella D, Arós F, et al. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. N Engl J Med. 2018;378(25):e34.

151. Faria R, Gomes M, Epstein D, White IR. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. Pharmacoeconomics. 2014;32(12).

152. McMeekin N, Wu O, Germeni E, Briggs A. How methodological frameworks are being developed: evidence from a scoping review. BMC Med Res Methodol. 2020;20(1):173.

153. Grutters J, Joore M. A Framework for Real-World Economic Evaluation by Incorporating Implementation Parameters. 2008.

154. Tappenden P, Chilcott J, Brennan A, Squires H, Stevenson M. Whole Disease Modeling to Inform Resource Allocation Decisions in Cancer: A Methodological Framework. Value Health. 2012;15(8):1127–36.

155. Squires H, Chilcott J, Akehurst R, Burr J, Kelly MP. A Framework for Developing the Structure of Public Health Economic Models. Value Health. 2016;19(5):588-601.

156. Robins JM. Correcting for noncompliance in randomized trials using structural nested mean models. Commun Stat-Theory Methods. 1994;23(8):2379-412.

157. Takemoto SK, Pinsky BW, Schnitzler MA, Lentine KL, Willoughby LM, Burroughs TE, et al. A retrospective analysis of immunosuppression compliance, dose reduction and discontinuation in kidney transplant recipients. Am J Transplant. 2007;7(12):2704-11.

158. Krueger K, Botermann L, Schorr SG, Griese-Mammen N, Laufs U, Schulz M. Age-related medication adherence in patients with chronic heart failure: A systematic literature review. Int J Cardiol. 2015;1(184):728-35.

159. Marcum ZA, Gellad WF. Medication Adherence to Multi-Drug Regimens. Clin Geriatr Med. 2012;28(2):287–300.

160. Anghel LA, Farcas AM, Oprean RN. An overview of the common methods used to measure treatment adherence. Med Pharm Rep. 2019;92(2):117–22.

161. Dima AL, Allemann SS, Dunbar-Jacob J, Hughes DA, Vrijens B, Wilson IB. TEOS: A framework for constructing operational definitions of medication adherence based on Timelines-Events-Objectives-Sources. Br J Clin Pharmacol. 2021;87(6):2521-33.

162. Guerci B, Chanan N, Kaur S, Jasso-Mosqueda JG, Lew E. Lack of Treatment Persistence and Treatment Nonadherence as Barriers to Glycaemic Control in Patients with Type 2 Diabetes. Diabetes Ther. 2019;10(2):437–49.

163. Sullivan TR, Latimer NR, Gray J, Sorich MJ, Salter AB, Karnon J. Adjusting for Treatment Switching in Oncology Trials: A Systematic Review and Recommendations for Reporting. Value Health. 2020;23(3):388–96.

164. Pennington M, McCrone P. Does Non-Adherence Increase Treatment Costs in Schizophrenia? Pharmacoeconomics. 2018;36(8):941–55.

165. Alshreef A, MacQuilkan K, Dawkins B, Riddin J, Ward S, Meads D, et al. Cost-Effectiveness of Docetaxel and Paclitaxel for Adjuvant Treatment of Early Breast Cancer: Adaptation of a Model-Based Economic Evaluation From the United Kingdom to South Africa. Value Health Reg Issues. 2019;19:65-74.

166. Ward S, Simpson E, Davis S, Hind D, Rees A, Wilkinson A. Taxanes for the adjuvant treatment of early breast cancer: systematic review and economic evaluation. Health Technol Assess. 2007;11(40):1-144.

167. Kini V, Ho PM. Interventions to Improve Medication Adherence: A Review. JAMA. 2018;320(23):2461-73.

Appendices

Appendix A: Search strategies used in the scoping and systematic reviews

A1: Search strategy used for identifying existing evidence

Citation search using Web of Science and Scopus electronic databases for three key papers (identified as relevant and highly cited papers in the literature) which address the issue of non-adherence.[26, 32, 69] This search generated 1156 records which were screened by title to identify the relevant papers. Articles were included if they had one of the following terms in the title: 'economic evaluation', 'cost-effectiveness', 'chronic disease' or 'chronic condition'. Following title screening, 1066 records were excluded. A further 67 records were excluded based on abstract screening and the remaining 23 articles were reviewed in full text. An additional 20 papers were identified through hand searching reference lists of included studies and snowballing (tracking down references) and experts' advice; and therefore, 43 papers in total were reviewed.

A2: First search iteration used in the systematic review

Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily, Ovid MEDLINE and Versions(R) 1946 to February 9 2018

#	Terms	Results
1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).ti.	120596
2	*Models, Structural/	2122
3	*models, statistical/	28208
4	*models, economic/ or *models, econometric/	4387
5	1 and (2 or 3 or 4)	245

Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily, Ovid MEDLINE and Versions(R) 1946 to March 28 2018 – Adding "Models, Biological" MeSH heading.

3rd April 2018

1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).ti.	121991
2	*Models, Structural/	2128
3	*Models, Statistical/	28482
4	*models, economic/ or *models, econometric/	4416
5	*Models, Biological/	97476
6	1 and (2 or 3 or 4)	249
7	1 and 5	625
8	limit 7 to humans	324
9	8 not 6	316

Embase 1974 to 2018 March 30
3rd April 2018

1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).ti.	159985
2	*structural model/	153
3	*statistical model/	21073
4	*economic model/	491
5	*biological model/	56706
6	1 and (2 or 3 or 4)	143
7	1 and 5	381
8	limit 7 to human	156
9	6 or 8	298

Cochrane Library
3rd April 2018

#1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal):ti (Word variations have been searched)	12301
#2	MeSH descriptor: [Models, Structural] this term only	25
#3	MeSH descriptor: [Models, Statistical] this term only	1577
#4	MeSH descriptor: [Models, Economic] this term only	1578
#5	MeSH descriptor: [Models, Econometric] this term only	470
#6	MeSH descriptor: [Models, Biological] this term only	2370
#7	#1 and (#2 or #3 or #4 or #5)	52
#8	#1 and #6	16
#9	#7 or #8	67

Econlit 1886 to May 3, 2018
8th May 2018

1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).ti.	8632
2	model*.ti.	90932
3	(structural or statistical or economic or econometric or biological).ti.	114247
4	1 and 2 and 3	31

Web of Science
8th May 2018

# 1	TITLE: ((compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal))	256,870
# 2	TITLE: (model*)	1,892,481
# 3	TITLE: ((structural or statistical or economic or econometric or biological))	706,793
# 4	#3 AND #2 AND #1	240

Scopus
8th May 2018

#1	(TITLE ((compliance OR adherence OR pharmacoadherence OR persistence OR persistency OR concordance OR initiation OR implementation OR noncompliance)) OR TITLE ((nonadherence OR nonpersistence OR discontinuation OR pharmionics OR therapeutic AND alliance OR patient AND irregularity OR treatment AND refusal)))	321,353
#2	TITLE (model*)	2,338,498

#3	TITLE ((structural OR statistical OR economic OR econometric OR biol ogical))	905,167
#4	((TITLE ((compliance OR adherence OR pharmacoadherence OR persistence OR persistency OR concordance OR initiation OR implementation OR noncompliance)) OR TITLE ((nonadherence OR nonpersistence OR discontinuation OR pharmionics OR therapeutic AND alliance OR patient AND irregularity OR treatment AND refusal)))) AND (TITLE (model*)) AND (TITLE ((structural OR statistical OR economic OR econometric OR biological)))	323

A3: Second search iteration used in the systematic review

Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily, Ovid MEDLINE and Versions(R) 1946 to May 16, 2018

22nd May 2018

1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).ti.	123742
2	*Models, Structural/	2129
3	*Models, Statistical/	28851
4	*models, economic/ or *models, econometric/	4476
5	*Models, Biological/	98419
6	1 and (2 or 3 or 4)	250
7	1 and 5	634
8	limit 7 to humans	329
9	8 not 6	321
10	*Survival Analysis/	2666
11	*Proportional Hazards Models/	1811
12	*Linear Models/	2498
13	*Logistic Models/	1675
14	Biometry/mt [Methods]	4244
15	Randomized Controlled Trials as Topic/sn [Statistics & Numerical Data]	4789
16	Cost-Benefit Analysis/sn [Statistics & Numerical Data]	981
17	Economics, Pharmaceutical/sn [Statistics & Numerical Data]	144
18	or/10-17	17625
19	pharmacometric*.tw.	388
20	causal inference.tw.	1455
21	proportional hazards.ti.	411
22	structural model*.ti.	1624
23	proportional hazards model*.ab.	20752
24	structural nested model*.ab.	27
25	marginal structural model*.ab.	518
26	structural proportional hazards.ab.	3
27	structural accelerated failure.ab.	7
28	compliance class model*.ab.	2
29	preserving structural failure.ab.	31
30	rank preserving structural.ab.	31
31	accelerated failure time.ab.	479
32	or/19-31	25187
33	18 or 32	41976
34	1 and 33	616
35	limit 34 to humans	523
36	35 not 9	520

Embase 1974 to 2018 May 21
22nd May 2018

1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).ti.	162336
2	*structural model/	160
3	*statistical model/	21307
4	*economic model/	506
5	*biological model/	57036
6	1 and (2 or 3 or 4)	146
7	1 and 5	385
8	limit 7 to human	157
9	6 or 8	302
10	*survival analysis/	645
11	*proportional hazards model/	1385
12	10 or 11	2017
13	pharmacometric*.tw.	568
14	causal inference.tw.	1544
15	proportional hazards.ti.	418
16	structural model*.ti.	1691
17	proportional hazards model*.ab.	32565
18	structural nested model*.ab.	26
19	marginal structural model*.ab.	681
20	structural proportional hazards.ab.	3
21	structural accelerated failure.ab.	9
22	compliance class model*.ab.	2
23	preserving structural failure.ab.	98
24	rank preserving structural.ab.	98
25	accelerated failure time.ab.	552
26	or/13-25	37594
27	1 and (12 or 26)	702
28	27 not 9	691

Web of Science
22nd May 2018

# 1	TITLE: ((compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal))	257,485
# 2	TITLE: (model*)	1,896,217
# 3	TITLE: ((structural or statistical or economic or econometric or biological))	708,343
# 4	#3 AND #2 AND #1	240
# 5	TI=("survival analysis")	3,080
# 6	TI=("proportional hazards model*")	513
# 7	TI=("linear model*")	6,683
# 8	TI=("logistic model*")	789

# 9	TOPIC: (pharmacometric*)	464
# 10	TS=("causal inference")	3,246
# 11	TI=("proportional hazards")	823
# 12	TI=("structural model*")	4,117
# 13	TI=("proportional hazards model*")	513
# 14	TOPIC: ("proportional hazards model*")	20,669
# 15	TOPIC: ("structural nested model*")	43
# 16	TOPIC: (“marginal structural model*”)	1,192
# 17	TOPIC: (“structural proportional hazards”)	3
# 18	TOPIC: ("structural accelerated failure")	9
# 19	TOPIC: ("compliance class model*")	4
# 20	TOPIC: ("preserving structural failure")	37
# 21	TOPIC: ("rank preserving structural")	38
# 22	TOPIC: ("accelerated failure time")	849
# 23	#22 OR #21 OR #20 OR #19 OR #18 OR #17 OR #16 OR #15 OR #14 OR #13 OR #12 OR #11 OR #10 OR #9 OR #8 OR #7 OR #6 OR #5	40,195
# 24	#23 AND #1	575
# 25	#24 not #4	519

MathSciNet
23rd May 2018
27 records

1.	"Title=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation)"	7806	Compliance in title
2.	"Title=(noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics)"	91	
3.	'Title=("therapeutic alliance" or "patient irregularity" or "treatment refusal")'	0	
4.	"Title=(model*)"	21005	Model in title
5.	"Title=(structural or statistical or economic or econometric or biological)"	36577	
6.	“Review Text=(survival analysis or proportional hazards model* or linear model* or logistic model*)"	507	Second iteration model terms
7.	"Review Text=(pharmacometric* or “causal inference“)"	439	
8.	"Review Text=(proportional hazards or structural model* or proportional hazards model* or structural nested model*)"	22	
9.	'Review Text=(marginal structural model* or structural proportional hazards or structural accelerated failure or compliance class model*)'	0	
10.	"Review Text=(preserving structural failure or rank preserving structural or accelerated failure time)"	5	
11.	"Title=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation) AND Title=(model*) AND Title=(structural or statistical or economic or econometric or biological)"	16	1st search
12.	"Title=(noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics) AND Title=(model*) AND	3	

	Title=(structural or statistical or economic or econometric or biological)"		
13.	Title=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation) AND Review Text=(survival analysis or proportional hazards model* or linear model* or logistic model*)" '	1	2nd search (a)
14.	'Title=(noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics) AND Review Text=(survival analysis or proportional hazards model* or linear model* or logistic model*)'	0	
15.	"Title=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation) AND Review Text=(pharmacometric* or “causal inference“)"	3	2nd search (b)
16.	'Title=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation) AND Review Text=(proportional hazards or structural model* or proportional hazards model* or structural nested model*)'	0	
17.	'Title=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation) AND Review Text=(preserving structural failure or rank preserving structural or accelerated failure time)'	0	
18.	"Title=(noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics) AND Review Text=(pharmacometric* or “causal inference“)"	3	
19.	'Title=(noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics) AND Review Text=(proportional hazards or structural model* or proportional hazards model* or structural nested model*)'	0	
20.	"Title=(noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics) AND Review Text=(preserving structural failure or rank preserving structural or accelerated failure time)"	1	
21.	or/11-20	27	

Appendix B: Details of included and excluded papers

Table 32: Details of included papers: application of methods in a simulation study or a case study, disease area and interventions assessed

ID	Reference	Method	Type of study	Disease area	Interventions compared	Outcome
01	Robins et al., 1992[84]	SNFTM	Case study	AIDS	Low- versus high dose AZT	Time to death
02	Baker, 1998[80]	IV with LE	Case study	Breast cancer	Screening vs usual care	Cost-effectiveness
03	White and Goetghebeur, 1998[89]	RPSFTM	Case study	Hypertension	Diuretics, beta-blockers or placebo	Time to CV death
04	Korhonen et al., 1999[39]	AT	Case study	Lung cancer	Alpha-tocopherol vs beta-carotene supplementation	Time to incidence of lung cancer
05	Robins and Finkelstein, 2000[66]	MSW (IPCW)	Case study	AIDS	Bactrim vs Aerosolised Pentamidine	Time to death
06	Hernan et al., 2001[83]	MSM (IPTW)	Case study	AIDS	AZT and PCP Prophylaxis vs no treatment initiation	Time to death
07	Loeys et al., 2001[85]	RPSFTM	Simulation study, Case study	Vitamin A deficiency	Vitamin A supplement vs placebo	Time to death
08	Korhonen and Palmgren, 2002[86]	RPSFTM	Simulation study, Case study	Lung cancer	Beta-carotene supplement vs placebo	Treatment-free survival
09	Loeys and Goetghebeur, 2002[87]	RPSFTM	Simulation study, Case study	Leukaemia	Bone marrow transplantation vs conventional chemotherapy	Time to death
10	Loeys and Goetghebeur, 2003[79]	C-PROPHET	Simulation study, Case study	Colorectal cancer	Surgical resection followed by chemotherapy vs surgical resection alone	Time to death
11	Matsui, 2004[88]	RPSFTM	Simulation study, Case study	Acute myeloid leukaemia	Macrophage colony-stimulating factor vs placebo	Time to blood count recovery
12	Cuzick et al., 2007[76]	CPH with PLE	Simulation study	hypothetical	Hypothetical	Survival time
13	Lin et al., 2007[77]	MCC	Case study	Depression	Meeting with health specialist vs usual care	Time to death
14	Nie et al., 2011[81]	IV with PNEMLE	Simulation study, Case study	Breast cancer	Screening with three annual follow up visits vs usual care	Time to death
15	Pink et al., 2014[90]	PKPD	Simulation study	Atrial fibrillation	Warfarin vs Apixaban vs Rivaroxaban vs Dabigatran	INR, cost-effectiveness
16	Yu et al., 2015[74]	ITT	Simulation study, Case study	Breast cancer	Four yearly screenings vs Usual care	Time to death
17	Wu et al., 2015[75]	PP	Simulation study, Case study	Depression	Active treatment vs placebo	Time to first remission
18	Li and Gray, 2016[78]	Wtd PP	Simulation study, Case study	Breast cancer	CMFP (cyclophosphamide/methotrexate/ fluorouracil/prednisone) vs observation	Disease-free survival
19	Gao and Zheng, 2017[82]	IV with MLE	Simulation study, Case study	Breast cancer	Screening vs usual care	Time to death
20	Hill-Mcmanus et al. 2018[91]	PKPD	Simulation study	Gout	Dual urate-lowering therapy (ULT) with allopurinol vs febuxostat	SuA level, Cost-effectiveness

- Details of excluded papers

A total of 130 papers were excluded at the full-text eligibility assessment stage. The numbers of excluded papers by reason of exclusion are presented in Table 33. The leading reasons for exclusion are non-time-to-event outcomes (continuous outcomes=36, binary outcomes=31). Other common reasons for exclusion are: application of methods already known without extension (n=27), and methods adjusting for the impact of non-adherence on PKPD only without assessing patient outcomes (n=13) as they do not meet the second inclusion criterion.

The methods discussed in each excluded paper were identified and reported in Table 33. The majority of excluded papers discussed IV methods (n=21), AT analysis (n=18), and SNMs (n=17). Bayesian inference methods, the traditional PP analysis and C-PROPHET were represented in the least number of excluded papers (5, 4 and 2, respectively). Papers reported other methods (n=16) included less common methods such as the Generalised Endogenous Treatment (GET) model, the Grizzle Model (GM) and the Generalised Grizzle Model (GGM). See Table 34 for the number of papers excluded per each method identified.

Table 33: Details of excluded papers: number of papers by reason for exclusion

Reason for exclusion	**Number of excluded papers**
Continuous outcomes	36
Binary outcomes	31
Categorical outcomes	8
No patient outcome	16
PKPD only without assessment of patient outcomes	13
Economic evaluations	12
Application of Known method without extension	27
Non-methodological papers	6
Not peer-reviewed abstracts	3
Not met the definition of non-adherence	5
Assessed associations, not causation	2
Theoretical papers with no application in a simulation or a case study	2
Comparisons of known methods	9

Note: Papers may be excluded for more than one reason

Table 34: Methods discussed in the excluded papers

Method	**Number of excluded papers**
Structural Nested Models (SNMs)	17
Marginal Structural Models (MSMs)	7
Inverse Probability of Censoring Weighting (IPCW)	8
Accelerated Failure Time Models (AFTMs) and Rank-Preserving Structural Failure Time Models (RPSFTMs)	7
G-estimation	7
Instrumental Variable (IV) methods	21
Compliers Average Causal Effect (CACE)	12
Compliers PROPortional Hazards of Treatment (C-PROPHET) Model	2
Cox Proportional Hazards (CPH) Models	9
Bayesian inference methods	5
Pharmacokinetics and Pharmacodynamics (PKPD) Methods	12
Intention-To-Treat (ITT) Analysis	8
Per-Protocol (PP) analysis	4
As Treated (AT) analysis	18
Economic models	12
Other methods	16
Non-causal model	11

Note: papers may have discussed more than one method. Other methods include a range of heterogeneous methods which are not common among methods identified in this review

Appendix C: Specification of scenarios evaluated in the simulation study

Table 35: Specification of scenarios assessed in the simulation across implementation, persistence and initiation non-adherence

Adjustment Scenario No.	Truth Scenario No.	Sample size	Survival time Data-generating Model (DGM)	Type and level of non-adherence	Relationship between treatment effect and non-adherence	Time-dependent treatment effect	Treatment effect size	Non-adherence adjustment methods assessed
1	1	450	Standard PSM - Weibull	Low implementation	Strong	No	Small	ITT, PP, IPCW, SNFTM
2	2	450	Two-component Weibull Mixture	Low implementation	Strong	No	Small	ITT, PP, IPCW, SNFTM
3	1	450	Standard PSM - Weibull	High implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
4	2	450	Two-component Weibull Mixture	High implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
5	1	450	Standard PSM - Weibull	Low implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
6	2	450	Two-component Weibull Mixture	Low implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
7	3	450	Standard PSM - Weibull	High implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
8	4	450	Two-component Weibull Mixture	High implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
9	3	450	Standard PSM - Weibull	Low implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
10	4	450	Two-component Weibull Mixture	Low implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
11	5	450	Standard PSM - Weibull	High implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
12	6	450	Two-component Weibull Mixture	High implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
13	5	450	Standard PSM - Weibull	Low implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
14	6	450	Two-component Weibull Mixture	Low implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
15	7	450	Standard PSM - Weibull	High implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
16	8	450	Two-component Weibull Mixture	High implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
17	7	450	Standard PSM - Weibull	Low implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
18	8	450	Two-component Weibull Mixture	Low implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
19	9	120	Standard PSM - Weibull	High implementation	Strong	No	Small	ITT, PP, IPCW, SNFTM
20	10	120	Two-component Weibull Mixture	High implementation	Strong	No	Small	ITT, PP, IPCW, SNFTM
21	9	120	Standard PSM - Weibull	Low implementation	Strong	No	Small	ITT, PP, IPCW, SNFTM
22	10	120	Two-component Weibull Mixture	Low implementation	Strong	No	Small	ITT, PP, IPCW, SNFTM
23	9	120	Standard PSM - Weibull	High implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
24	10	120	Two-component Weibull Mixture	High implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
25	9	120	Standard PSM - Weibull	Low implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM
26	10	120	Two-component Weibull Mixture	Low implementation	Weak	No	Small	ITT, PP, IPCW, SNFTM

27	11	120	Standard PSM - Weibull	High implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
28	12	120	Two-component Weibull Mixture	High implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
29	11	120	Standard PSM - Weibull	Low implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
30	12	120	Two-component Weibull Mixture	Low implementation	Strong	No	Large	ITT, PP, IPCW, SNFTM
31	13	120	Standard PSM - Weibull	High implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
32	14	120	Two-component Weibull Mixture	High implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
33	13	120	Standard PSM - Weibull	Low implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
34	14	120	Two-component Weibull Mixture	Low implementation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
35	15	120	Standard PSM - Weibull	High implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
36	16	120	Two-component Weibull Mixture	High implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
37	15	120	Standard PSM - Weibull	Low implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
38	16	120	Two-component Weibull Mixture	Low implementation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
39	1	450	Standard PSM - Weibull	Low persistence	Strong	No	Small	ITT, PP, IPCW, SNFTM
40	2	450	Two-component Weibull Mixture	Low persistence	Strong	No	Small	ITT, PP, IPCW, SNFTM
41	1	450	Standard PSM - Weibull	High persistence	Weak	No	Small	ITT, PP, IPCW, SNFTM
42	2	450	Two-component Weibull Mixture	High persistence	Weak	No	Small	ITT, PP, IPCW, SNFTM
43	4	450	Two-component Weibull Mixture	High persistence	Strong	No	Large	ITT, PP, IPCW, SNFTM
44	3	450	Standard PSM - Weibull	Low persistence	Strong	No	Large	ITT, PP, IPCW, SNFTM
45	4	450	Two-component Weibull Mixture	Low persistence	Strong	No	Large	ITT, PP, IPCW, SNFTM
46	5	450	Standard PSM - Weibull	High persistence	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
47	6	450	Two-component Weibull Mixture	High persistence	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
48	5	450	Standard PSM - Weibull	Low persistence	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
49	6	450	Two-component Weibull Mixture	Low persistence	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
50	10	120	Two-component Weibull Mixture	High persistence	Strong	No	Small	ITT, PP, IPCW, SNFTM
51	12	120	Two-component Weibull Mixture	High persistence	Strong	No	Large	ITT, PP, IPCW, SNFTM
52	12	120	Two-component Weibull Mixture	Low persistence	Strong	No	Large	ITT, PP, IPCW, SNFTM
53	16	120	Two-component Weibull Mixture	High persistence	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
54	15	120	Standard PSM - Weibull	Low persistence	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
55	16	120	Two-component Weibull Mixture	Low persistence	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
56	14	120	Two-component Weibull Mixture	High persistence	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
57	1	450	Standard PSM - Weibull	Low initiation	Strong	No	Small	ITT, PP, IPCW, SNFTM
58	2	450	Two-component Weibull Mixture	Low initiation	Strong	No	Small	ITT, PP, IPCW, SNFTM
59	1	450	Standard PSM - Weibull	High initiation	Weak	No	Small	ITT, PP, IPCW, SNFTM
60	2	450	Two-component Weibull Mixture	High initiation	Weak	No	Small	ITT, PP, IPCW, SNFTM
61	1	450	Standard PSM - Weibull	Low initiation	Weak	No	Small	ITT, PP, IPCW, SNFTM
62	2	450	Two-component Weibull Mixture	Low initiation	Weak	No	Small	ITT, PP, IPCW, SNFTM
63	3	450	Standard PSM - Weibull	High initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM

64	4	450	Two-component Weibull Mixture	High initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
65	3	450	Standard PSM - Weibull	Low initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
66	4	450	Two-component Weibull Mixture	Low initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
67	5	450	Standard PSM - Weibull	High initiation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
68	6	450	Two-component Weibull Mixture	High initiation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
69	5	450	Standard PSM - Weibull	Low initiation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
70	6	450	Two-component Weibull Mixture	Low initiation	Strong	Yes	Large	ITT, PP, IPCW, SNFTM
71	7	450	Standard PSM - Weibull	High initiation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
72	8	450	Two-component Weibull Mixture	High initiation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
73	7	450	Standard PSM - Weibull	Low initiation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
74	8	450	Two-component Weibull Mixture	Low initiation	Weak	Yes	Small	ITT, PP, IPCW, SNFTM
75	9	120	Standard PSM - Weibull	High initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
76	10	120	Two-component Weibull Mixture	High initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
77	9	120	Standard PSM - Weibull	Low initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
78	10	120	Two-component Weibull Mixture	Low initiation	Strong	No	Large	ITT, PP, IPCW, SNFTM
79	9	120	Standard PSM - Weibull	High initiation	Weak	No	Large	ITT, PP, IPCW, SNFTM
80	10	120	Two-component Weibull Mixture	High initiation	Weak	No	Large	ITT, PP, IPCW, SNFTM
81	9	120	Standard PSM - Weibull	Low initiation	Weak	No	Large	ITT, PP, IPCW, SNFTM
82	10	120	Two-component Weibull Mixture	Low initiation	Weak	No	Large	ITT, PP, IPCW, SNFTM
83	13	120	Standard PSM - Weibull	High initiation	Strong	Yes	Small	ITT, PP, IPCW, SNFTM
84	14	120	Two-component Weibull Mixture	High initiation	Strong	Yes	Small	ITT, PP, IPCW, SNFTM
85	13	120	Standard PSM - Weibull	Low initiation	Strong	Yes	Small	ITT, PP, IPCW, SNFTM
86	14	120	Two-component Weibull Mixture	Low initiation	Strong	Yes	Small	ITT, PP, IPCW, SNFTM
87	15	120	Standard PSM - Weibull	High initiation	Weak	Yes	Large	ITT, PP, IPCW, SNFTM
88	16	120	Two-component Weibull Mixture	High initiation	Weak	Yes	Large	ITT, PP, IPCW, SNFTM
89	15	120	Standard PSM - Weibull	Low initiation	Weak	Yes	Large	ITT, PP, IPCW, SNFTM
90	16	120	Two-component Weibull Mixture	Low initiation	Weak	Yes	Large	ITT, PP, IPCW, SNFTM

Note: For each set of scenarios numbered in the first column, there will be one large dataset (truth scenario numbered in the second column) which be simulated using 1 million

Appendix D: Simulation study supplementary tables

Table 36: Sample size for clinical trials assessed maintenance immunosuppression after kidney transplantation

Author (Year)	**n**
Grinyo 2009	1529
Vincenti 2010	686
Waller 2002	102
Laskow 1996	120
Baboolal 2002	51
Campos 2002	166
Margreiter 2002	560
Sadek 2002	477
Mayer 1997	448
Tricontinental MMF renal study 1996	497
Yang 1999	60
Schaefer 2006	80
Hardinger 2005	200
Rowshani 2006	126
Weimer 2006	81
Tedesco-Silva 2010	783
Barsoum 2007	113
Anil Kumar 2005	150
Mendez 2005	361
Sampaio 2008	100
Martinez-Mier 2006	41
Nafar 2012	100
Anil Kumar 2008	200
Vincenti 2005	218
Ferguson 2011	89
Flechner 2002	61
Vacher-Coponat 2012	289
Büchler 2007	145
Charpentier 2003	83
Merville 2004	71
Durrbach 2010	578
Lorber 2005	583
Bertoni 2011	106
Gallon 2006	83
Guba 2010	140
Lebranchu 2009	192
Vítko 2005	588
Larson 2006	162
Chadban 2013	126
Flechner 2011	450

Appendix E: Performance of methods across scenarios

Table 37: Performance of methods across implementation non-adherence scenarios

No.	Successful nsim	Truth	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Bias (%)	MSE (%)	Model SE (%)	Empirical SE (%)	Coverage (%)	Successful estimation (%)
						Lower	Upper							
1	1872	0.108	ITT	0.135	0.032	0.072	0.197	0.026	24.26	1.54	29.37	28.86	86.49	100.00
			PP	0.131	0.037	0.059	0.202	0.022	20.61	1.68	33.74	33.53	90.33	100.00
			IPCW	0.128	0.037	0.056	0.201	0.020	18.24	1.62	34.05	34.14	91.08	100.00
			SNFTM	0.129	0.031	0.067	0.190	0.020	18.52	1.84	28.81	36.85	83.49	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.66	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.07	0.86	-
2	1874	0.063	ITT		0.030	0.046	0.164	0.042	66.06	4.18	47.48	47.26	71.08	100.00
			PP	0.083	0.033	0.019	0.146	0.019	30.43	2.29	51.47	51.82	90.72	100.00
			IPCW	0.084	0.033	0.018	0.149	0.020	31.94	2.43	52.57	53.09	90.72	100.00
			SNFTM	0.082	0.028	0.028	0.136	0.019	30.07	2.40	43.46	53.71	82.55	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.67	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.05	-
3	1863	0.108	ITT	0.136	0.031	0.075	0.196	0.027	25.18	1.56	28.49	28.39	85.72	100.00
			PP	0.124	0.034	0.056	0.192	0.016	14.34	1.27	31.80	31.13	93.34	100.00
			IPCW	0.123	0.035	0.054	0.192	0.015	13.66	1.34	32.50	32.38	92.22	100.00
			SNFTM	0.121	0.031	0.061	0.181	0.013	11.60	1.41	28.22	34.20	88.35	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.66	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.07	0.86	-
4	1855	0.063	ITT	0.091	0.027	0.038	0.145	0.028	44.12	2.34	42.90	41.74	83.29	100.00
			PP	0.074	0.029	0.016	0.132	0.010	16.54	1.51	46.53	45.93	93.96	100.00
			IPCW	0.074	0.030	0.015	0.134	0.011	17.49	1.62	48.06	47.48	94.07	100.00
			SNFTM	0.072	0.026	0.021	0.122	0.008	13.34	1.49	40.82	46.58	90.84	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.04	0.55	-
			max MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.87	-
5	1859	0.108	ITT	0.131	0.031	0.070	0.192	0.022	20.51	1.33	28.70	28.34	88.81	100.00
			PP	0.123	0.036	0.052	0.194	0.015	13.57	1.38	33.52	33.05	92.79	100.00
			IPCW	0.121	0.037	0.049	0.193	0.012	11.42	1.36	33.92	33.57	93.01	100.00

			SNFTM	0.120	0.031	0.060	0.180	0.012	10.74	1.65	28.33	37.52	84.51	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.59	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.07	0.84	-
6	1845	0.063	ITT	0.091	0.028	0.036	0.146	0.028	43.56	2.41	44.09	43.63	83.31	100.00
			PP	0.073	0.031	0.012	0.135	0.010	15.79	1.72	49.49	49.66	93.82	100.00
			IPCW	0.075	0.032	0.011	0.138	0.011	17.74	1.86	51.05	51.20	93.88	100.00
			SNFTM	0.072	0.026	0.021	0.124	0.009	14.41	1.80	41.26	51.38	87.59	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.05	0.56	-
			max MCSE	-	-	-	-	-	0.08	0.00	0.00	0.05	0.87	-
7	1840	0.169	ITT	0.209	0.031	0.148	0.270	0.040	23.94	1.52	18.39	18.08	75.22	100.00
			PP	0.198	0.034	0.131	0.264	0.029	17.29	1.18	20.13	19.96	86.63	100.00
			IPCW	0.196	0.035	0.128	0.264	0.027	16.07	1.16	20.57	20.75	87.61	100.00
			SNFTM	0.196	0.030	0.136	0.255	0.027	16.02	1.17	17.93	20.94	82.17	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.77	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.01	-
8	1841	0.095	ITT	0.148	0.028	0.094	0.203	0.054	56.71	3.88	29.49	29.63	51.33	100.00
			PP	0.120	0.029	0.062	0.177	0.025	26.28	1.54	31.01	30.57	87.34	100.00
			IPCW	0.120	0.030	0.061	0.179	0.025	26.77	1.62	31.79	31.53	87.51	100.00
			SNFTM	0.119	0.026	0.068	0.170	0.024	25.64	1.55	27.40	31.31	81.31	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.77	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.16	-
9	1845	0.169	ITT	0.202	0.031	0.140	0.263	0.033	19.65	1.25	18.65	18.82	81.52	100.00
			PP	0.198	0.036	0.128	0.268	0.029	17.34	1.25	21.27	20.94	87.37	100.00
			IPCW	0.194	0.036	0.123	0.265	0.026	15.26	1.17	21.46	21.46	89.05	100.00
			SNFTM	0.195	0.030	0.136	0.255	0.027	15.79	1.34	18.07	23.41	80.05	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.73	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	0.93	-
10	1822	0.095	ITT	0.150	0.029	0.094	0.207	0.056	58.88	4.14	30.59	30.15	51.54	100.00
			PP	0.120	0.031	0.059	0.181	0.025	26.69	1.70	33.09	32.97	86.66	100.00
			IPCW	0.121	0.032	0.058	0.184	0.026	27.61	1.82	33.80	34.08	85.84	100.00
			SNFTM	0.120	0.026	0.068	0.171	0.025	26.25	1.90	27.85	36.32	76.89	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.80	-

			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.17	-
11	1826	0.155	ITT	0.195	0.031	0.135	0.256	0.040	26.04	1.63	20.02	19.27	74.92	100.00
			PP	0.185	0.034	0.119	0.252	0.030	19.46	1.28	21.93	21.08	86.04	100.00
			IPCW	0.183	0.035	0.115	0.251	0.028	18.08	1.28	22.39	22.38	87.51	100.00
			SNFTM	0.182	0.030	0.122	0.241	0.027	17.33	1.40	19.56	24.62	79.30	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.77	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	1.01	-
12	1829	0.087	ITT	0.141	0.028	0.086	0.196	0.054	61.47	4.21	32.11	32.20	51.67	100.00
			PP	0.113	0.029	0.055	0.171	0.026	29.38	1.79	33.75	34.42	84.96	100.00
			IPCW	0.114	0.030	0.054	0.173	0.026	29.87	1.89	34.62	35.63	84.64	100.00
			SNFTM	0.112	0.026	0.060	0.163	0.024	27.57	1.82	29.83	36.30	79.33	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.84	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.17	-
13	1855	0.155	ITT	0.188	0.031	0.126	0.250	0.033	21.30	1.31	20.26	19.80	82.32	100.00
			PP	0.185	0.036	0.115	0.256	0.030	19.64	1.38	23.17	22.43	86.85	100.00
			IPCW	0.182	0.036	0.111	0.253	0.027	17.48	1.31	23.36	23.27	88.73	100.00
			SNFTM	0.182	0.031	0.122	0.242	0.027	17.49	1.47	19.69	25.39	78.81	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.73	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	0.95	-
14	1825	0.087	ITT	0.143	0.029	0.086	0.200	0.055	63.10	4.47	33.29	33.62	51.40	100.00
			PP	0.114	0.031	0.053	0.176	0.027	30.80	2.04	36.01	37.17	84.60	100.00
			IPCW	0.115	0.032	0.052	0.178	0.028	31.90	2.17	36.83	38.29	85.04	100.00
			SNFTM	0.113	0.026	0.061	0.165	0.026	29.59	2.22	30.28	40.84	74.30	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.83	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.17	-
15	1801	0.093	ITT	0.121	0.031	0.061	0.182	0.028	30.44	1.94	33.18	34.06	83.73	100.00
			PP	0.110	0.034	0.043	0.178	0.017	18.71	1.70	37.07	38.42	90.62	100.00
			IPCW	0.110	0.035	0.041	0.180	0.017	18.72	1.78	37.90	39.51	90.34	100.00
			SNFTM	0.107	0.031	0.047	0.167	0.014	15.05	2.02	32.93	44.14	83.95	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.69	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	0.87	-
16	1811	0.054	ITT	0.081	0.027	0.028	0.135	0.027	49.10	2.77	50.23	51.79	82.11	100.00

			PP	0.065	0.030	0.007	0.123	0.011	19.81	1.90	54.50	55.64	92.88	100.00
			IPCW	0.065	0.031	0.006	0.125	0.011	20.33	2.03	56.17	57.60	93.21	100.00
			SNFTM	0.064	0.026	0.013	0.114	0.009	16.77	2.03	47.74	58.76	87.24	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.59	-
			max MCSE	-	-	-	-	-	0.08	0.00	0.00	0.05	0.90	-
17	1854	0.093	ITT	0.116	0.031	0.055	0.177	0.023	24.78	1.69	33.42	34.68	86.79	100.00
			PP	0.110	0.036	0.039	0.181	0.017	17.94	1.83	39.05	40.50	91.48	100.00
			IPCW	0.108	0.037	0.036	0.180	0.015	15.87	1.83	39.54	41.36	91.91	100.00
			SNFTM	0.105	0.031	0.045	0.166	0.012	13.07	2.33	33.08	48.36	81.45	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.63	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	0.90	-
18	1803	0.054	ITT	0.081	0.028	0.026	0.136	0.027	49.22	2.80	51.57	52.30	84.03	100.00
			PP	0.065	0.031	0.003	0.127	0.011	19.71	2.17	57.88	59.98	92.62	100.00
			IPCW	0.066	0.032	0.003	0.130	0.012	21.80	2.34	59.58	61.94	92.68	100.00
			SNFTM	0.064	0.026	0.013	0.116	0.010	18.37	2.58	48.15	66.38	83.47	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.61	-
			max MCSE	-	-	-	-	-	0.09	0.00	0.00	0.06	0.87	-
19	1625	0.108	ITT	0.146	0.059	0.031	0.261	0.038	34.69	4.39	54.26	53.37	90.65	100.00
			PP	0.134	0.065	0.007	0.261	0.026	23.72	4.38	59.71	58.99	92.98	100.00
			IPCW	0.133	0.081	-0.025	0.291	0.025	22.91	7.01	75.13	77.12	92.43	100.00
			SNFTM	0.133	0.058	0.020	0.246	0.024	22.43	4.74	53.33	62.24	88.55	100.00
			min MCSE	-	-	-	-	-	0.14	0.02	0.01	0.10	0.63	-
			max MCSE	-	-	-	-	-	0.21	0.03	0.03	0.15	0.79	-
20	1773	0.064	ITT	0.111	0.057	0.000	0.222	0.047	74.61	8.41	89.19	87.62	87.65	100.00
			PP	0.086	0.060	-0.031	0.203	0.022	35.25	6.19	93.96	92.24	93.74	100.00
			IPCW	0.087	0.073	-0.056	0.230	0.024	37.26	9.42	116.33	115.96	93.17	100.00
			SNFTM	0.088	0.052	-0.015	0.191	0.024	38.13	6.70	82.77	95.38	87.70	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.58	-
			max MCSE	-	-	-	-	-	0.18	0.02	0.03	0.12	0.78	-
21	1671	0.108	ITT	0.134	0.062	0.013	0.255	0.026	23.57	3.95	56.97	55.63	92.22	100.00
			PP	0.130	0.071	-0.009	0.269	0.022	20.29	4.79	65.54	63.33	94.61	100.00
			IPCW	0.126	0.084	-0.038	0.290	0.018	16.25	6.91	77.87	78.23	94.14	100.00

			SNFTM	0.128	0.060	0.010	0.246	0.020	18.22	5.33	55.81	67.79	87.49	100.00
			min MCSE	-	-	-	-	-	0.15	0.01	0.01	0.10	0.55	-
			max MCSE	-	-	-	-	-	0.21	0.02	0.02	0.15	0.81	-
22	1776	0.064	ITT	0.107	0.059	-0.008	0.222	0.044	68.52	8.31	92.41	91.58	88.40	100.00
			PP	0.085	0.064	-0.040	0.209	0.021	33.54	6.88	100.40	98.53	92.91	100.00
			IPCW	0.086	0.074	-0.059	0.232	0.023	36.06	9.60	118.64	117.56	93.35	100.00
			SNFTM	0.085	0.053	-0.019	0.190	0.022	34.36	7.46	84.15	102.80	85.92	100.00
			min MCSE	-	-	-	-	-	0.14	0.01	0.01	0.10	0.59	-
			max MCSE	-	-	-	-	-	0.18	0.02	0.04	0.13	0.83	-
23	1697	0.108	ITT	0.132	0.060	0.015	0.249	0.024	22.26	3.62	55.15	53.41	93.93	100.00
			PP	0.120	0.067	-0.011	0.251	0.012	11.10	4.09	61.80	60.48	95.05	100.00
			IPCW	0.117	0.083	-0.045	0.279	0.009	8.11	6.91	76.83	79.47	92.52	100.00
			SNFTM	0.117	0.059	0.001	0.233	0.009	7.94	4.32	54.57	62.65	90.93	100.00
			min MCSE	-	-	-	-	-	0.14	0.01	0.01	0.10	0.53	-
			max MCSE	-	-	-	-	-	0.21	0.03	0.03	0.15	0.70	-
24	1785	0.064	ITT	0.092	0.053	-0.012	0.196	0.029	45.37	5.75	83.84	83.66	91.32	100.00
			PP	0.076	0.058	-0.037	0.188	0.012	18.92	5.50	90.89	91.14	94.17	100.00
			IPCW	0.076	0.071	-0.063	0.214	0.012	19.24	8.48	112.64	113.96	93.61	100.00
			SNFTM	0.076	0.050	-0.023	0.174	0.012	19.15	5.79	79.29	93.52	89.41	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.55	-
			max MCSE	-	-	-	-	-	0.17	0.02	0.03	0.12	0.73	-
25	1739	0.108	ITT	0.130	0.060	0.012	0.248	0.022	20.22	3.64	55.69	54.32	93.27	100.00
			PP	0.121	0.070	-0.017	0.259	0.013	11.86	4.60	65.17	64.10	94.59	100.00
			IPCW	0.116	0.082	-0.045	0.277	0.008	7.11	6.62	76.44	77.87	94.08	100.00
			SNFTM	0.120	0.059	0.004	0.236	0.012	10.82	5.43	54.88	69.99	87.18	100.00
			min MCSE	-	-	-	-	-	0.14	0.01	0.01	0.10	0.54	-
			max MCSE	-	-	-	-	-	0.20	0.02	0.02	0.14	0.80	-
26	1806	0.064	ITT	0.092	0.055	-0.015	0.199	0.029	45.19	5.98	85.99	85.84	91.36	100.00
			PP	0.075	0.061	-0.045	0.195	0.011	17.80	5.96	96.52	95.20	94.13	100.00
			IPCW	0.077	0.071	-0.063	0.217	0.013	21.00	8.30	113.50	112.36	94.46	100.00
			SNFTM	0.076	0.051	-0.024	0.175	0.012	19.16	6.43	80.08	98.81	87.38	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.54	-

			max MCSE	-	-	-	-	-	0.17	0.02	0.03	0.12	0.78	-
27	1622	0.169	ITT	0.210	0.060	0.093	0.328	0.042	24.81	3.18	35.56	35.64	88.10	100.00
			PP	0.199	0.066	0.070	0.327	0.030	17.77	3.08	39.04	38.92	91.68	100.00
			IPCW	0.194	0.082	0.033	0.355	0.025	15.10	4.70	49.14	50.61	91.55	100.00
			SNFTM	0.196	0.058	0.082	0.310	0.028	16.37	3.16	34.61	40.07	87.85	100.00
			min MCSE	-	-	-	-	-	0.15	0.02	0.01	0.11	0.69	-
			max MCSE	-	-	-	-	-	0.21	0.03	0.03	0.15	0.81	-
28	1801	0.095	ITT	0.149	0.054	0.043	0.256	0.054	56.94	6.04	57.35	55.72	83.73	100.00
			PP	0.119	0.057	0.008	0.231	0.024	25.40	3.93	60.19	59.07	92.73	100.00
			IPCW	0.120	0.070	-0.017	0.257	0.025	26.05	5.66	74.42	72.58	94.16	100.00
			SNFTM	0.120	0.050	0.022	0.219	0.025	26.53	4.19	53.03	60.86	88.01	100.00
			min MCSE	-	-	-	-	-	0.12	0.01	0.01	0.09	0.55	-
			max MCSE	-	-	-	-	-	0.16	0.02	0.03	0.12	0.87	-
29	1727	0.169	ITT	0.203	0.061	0.084	0.322	0.034	20.39	2.96	36.05	36.62	90.21	100.00
			PP	0.198	0.069	0.063	0.334	0.030	17.71	3.38	41.19	41.11	91.60	100.00
			IPCW	0.191	0.082	0.030	0.351	0.022	13.02	4.57	48.76	50.42	91.84	100.00
			SNFTM	0.195	0.059	0.080	0.310	0.027	15.78	3.54	34.85	43.01	84.83	100.00
			min MCSE	-	-	-	-	-	0.15	0.02	0.01	0.11	0.66	-
			max MCSE	-	-	-	-	-	0.20	0.03	0.02	0.14	0.86	-
30	1819	0.095	ITT	0.151	0.056	0.041	0.262	0.056	58.73	6.63	59.34	59.32	83.34	100.00
			PP	0.120	0.061	0.001	0.240	0.025	26.55	4.60	64.22	64.30	92.58	100.00
			IPCW	0.122	0.071	-0.017	0.261	0.027	28.42	6.35	75.06	76.58	92.35	100.00
			SNFTM	0.122	0.051	0.022	0.222	0.027	28.18	4.95	53.90	66.43	86.64	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.61	-
			max MCSE	-	-	-	-	-	0.17	0.02	0.02	0.12	0.87	-
31	1608	0.155	ITT	0.193	0.060	0.075	0.311	0.038	24.30	3.32	38.79	39.38	89.49	100.00
			PP	0.183	0.066	0.054	0.313	0.028	18.25	3.41	42.55	43.21	91.79	100.00
			IPCW	0.178	0.083	0.015	0.340	0.023	14.65	5.44	53.93	57.39	90.72	100.00
			SNFTM	0.179	0.059	0.065	0.294	0.024	15.71	3.57	37.86	45.33	87.56	100.00
			min MCSE	-	-	-	-	-	0.15	0.02	0.01	0.11	0.68	-
			max MCSE	-	-	-	-	-	0.22	0.03	0.03	0.16	0.82	-
32	1802	0.088	ITT	0.139	0.055	0.031	0.246	0.051	57.78	6.22	62.40	61.23	84.74	100.00

			PP	0.110	0.057	-0.003	0.222	0.022	24.90	4.21	65.53	64.65	93.17	100.00
			IPCW	0.112	0.070	-0.024	0.249	0.025	27.90	6.47	80.46	81.18	92.44	100.00
			SNFTM	0.111	0.051	0.012	0.210	0.024	26.77	4.49	57.77	66.34	88.57	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.59	-
			max MCSE	-	-	-	-	-	0.17	0.02	0.03	0.12	0.85	-
33	1665	0.155	ITT	0.186	0.061	0.067	0.306	0.031	20.23	3.15	39.25	40.31	90.81	100.00
			PP	0.184	0.070	0.047	0.320	0.029	18.50	3.77	44.94	45.71	91.95	100.00
			IPCW	0.179	0.082	0.018	0.339	0.024	15.25	5.25	53.18	56.16	91.77	100.00
			SNFTM	0.182	0.059	0.066	0.297	0.027	17.22	3.96	38.03	47.52	85.59	100.00
			min MCSE	-	-	-	-	-	0.15	0.02	0.01	0.11	0.67	-
			max MCSE	-	-	-	-	-	0.21	0.03	0.03	0.15	0.86	-
34	1811	0.088	ITT	0.140	0.057	0.029	0.251	0.052	59.19	6.72	64.69	64.38	85.26	100.00
			PP	0.111	0.061	-0.009	0.231	0.023	26.26	4.84	69.99	69.44	93.43	100.00
			IPCW	0.114	0.071	-0.025	0.254	0.026	30.13	6.84	81.68	82.94	93.04	100.00
			SNFTM	0.111	0.051	0.011	0.212	0.023	26.74	5.23	58.58	72.37	86.20	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.58	-
			max MCSE	-	-	-	-	-	0.17	0.02	0.02	0.12	0.83	-
35	1633	0.093	ITT	0.117	0.060	0.000	0.234	0.024	25.76	4.61	64.08	65.45	91.92	100.00
			PP	0.105	0.067	-0.026	0.235	0.012	12.61	5.26	71.66	74.12	93.63	100.00
			IPCW	0.102	0.083	-0.061	0.265	0.009	9.40	8.89	90.23	97.30	92.84	100.00
			SNFTM	0.101	0.059	-0.015	0.217	0.008	8.13	5.83	63.55	78.75	87.94	100.00
			min MCSE	-	-	-	-	-	0.15	0.01	0.01	0.11	0.60	-
			max MCSE	-	-	-	-	-	0.22	0.03	0.03	0.16	0.81	-
36	1785	0.055	ITT	0.084	0.053	-0.020	0.189	0.030	54.11	6.78	97.86	97.44	91.71	100.00
			PP	0.068	0.058	-0.045	0.181	0.013	24.05	6.22	105.96	104.03	94.57	100.00
			IPCW	0.067	0.071	-0.071	0.206	0.013	23.50	9.18	131.19	127.61	95.46	100.00
			SNFTM	0.067	0.050	-0.031	0.166	0.013	23.23	6.43	92.57	106.01	90.76	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.49	-
			max MCSE	-	-	-	-	-	0.16	0.02	0.03	0.12	0.69	-
37	1754	0.093	ITT	0.113	0.060	-0.005	0.230	0.019	20.80	4.37	64.59	65.32	93.16	100.00
			PP	0.105	0.070	-0.032	0.243	0.012	13.25	5.58	75.58	76.28	94.18	100.00
			IPCW	0.102	0.082	-0.060	0.263	0.009	9.42	8.09	89.01	92.73	93.22	100.00

			SNFTM	0.102	0.059	-0.015	0.218	0.008	9.04	5.96	63.86	79.48	86.89	100.00
			min MCSE	-	-	-	-	-	0.15	0.01	0.01	0.10	0.56	-
			max MCSE	-	-	-	-	-	0.21	0.03	0.02	0.15	0.81	-
38	1822	0.055	ITT	0.084	0.055	-0.024	0.191	0.029	53.09	7.18	100.39	101.67	91.44	100.00
			PP	0.069	0.061	-0.051	0.189	0.015	26.77	7.34	112.64	112.86	93.91	100.00
			IPCW	0.072	0.071	-0.068	0.211	0.017	31.10	10.13	132.08	132.67	93.58	100.00
			SNFTM	0.068	0.051	-0.031	0.168	0.014	24.77	7.70	93.42	116.21	87.54	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.56	-
			max MCSE	-	-	-	-	-	0.17	0.02	0.03	0.12	0.77	-

Table 38: Performance of methods across persistence non-adherence scenarios

No.	Successful nsim	Truth	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Percent bias	MSE	Model SE	Empirical SE	Coverage (%)	Successful estimation (%)
						Lower	Upper							
39	1817	0.108	ITT	0.131	0.032	0.068	0.195	0.023	21.18	29.29	29.71	1.42	88.77	100.00
			PP	0.130	0.035	0.060	0.199	0.021	19.52	32.04	32.61	1.53	90.64	100.00
			IPCW	0.127	0.036	0.057	0.198	0.019	17.54	33.20	33.31	1.53	91.19	100.00
			SNFTM	0.128	0.031	0.067	0.190	0.020	18.34	37.35	28.94	1.88	83.32	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.66	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	0.87	-
40	1812	0.063	ITT	0.104	0.030	0.044	0.163	0.040	63.44	47.61	48.02	3.98	73.01	100.00
			PP	0.083	0.031	0.022	0.145	0.020	31.76	49.19	49.72	2.17	90.89	100.00
			IPCW	0.084	0.032	0.021	0.147	0.021	33.00	50.09	50.69	2.28	90.89	100.00
			SNFTM	0.083	0.028	0.029	0.138	0.020	31.64	53.45	43.74	2.44	82.73	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.68	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.04	-
41	1824	0.108	ITT	0.138	0.031	0.077	0.200	0.030	27.46	28.52	28.91	1.70	83.94	100.00
			PP	0.127	0.034	0.061	0.193	0.019	17.22	30.50	31.06	1.33	91.94	100.00
			IPCW	0.129	0.035	0.059	0.198	0.020	18.73	33.08	32.62	1.57	90.52	100.00
			SNFTM	0.125	0.031	0.064	0.185	0.016	15.21	32.62	28.46	1.40	88.43	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.64	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	0.86	-

42	1811	0.063	ITT	0.099	0.028	0.044	0.154	0.035	55.87	43.08	44.37	3.15	76.59	100.00
			PP	0.078	0.029	0.021	0.135	0.015	23.85	45.39	45.95	1.66	92.55	100.00
			IPCW	0.078	0.030	0.019	0.138	0.015	23.69	47.62	48.07	1.79	92.44	100.00
			SNFTM	0.078	0.026	0.026	0.129	0.014	22.57	45.95	41.72	1.66	88.85	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.05	0.62	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.00	-
43	1830	0.095	ITT	0.147	0.028	0.092	0.203	0.053	55.57	30.78	29.76	3.82	52.51	100.00
			PP	0.120	0.029	0.064	0.176	0.025	26.89	30.76	30.21	1.58	85.57	100.00
			IPCW	0.120	0.030	0.062	0.178	0.025	26.69	32.20	31.40	1.66	86.07	100.00
			SNFTM	0.120	0.026	0.069	0.171	0.025	26.68	31.81	27.53	1.63	80.11	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.81	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.17	-
44	1809	0.169	ITT	0.195	0.032	0.133	0.257	0.027	15.78	18.72	18.84	1.01	86.07	100.00
			PP	0.196	0.035	0.128	0.263	0.027	16.04	20.32	20.54	1.13	88.23	100.00
			IPCW	0.192	0.035	0.123	0.262	0.024	14.18	21.16	20.93	1.09	89.17	100.00
			SNFTM	0.194	0.031	0.134	0.254	0.026	15.17	22.57	18.13	1.25	80.71	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.73	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	0.93	-
45	1813	0.095	ITT	0.148	0.029	0.090	0.206	0.053	56.25	31.27	31.00	3.92	55.16	100.00
			PP	0.121	0.030	0.062	0.181	0.027	28.17	31.70	32.03	1.70	86.38	100.00
			IPCW	0.122	0.031	0.062	0.183	0.027	28.89	32.60	32.63	1.80	85.05	100.00
			SNFTM	0.122	0.027	0.069	0.174	0.027	28.36	34.27	28.09	1.87	78.49	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.81	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.05	1.17	-
46	1804	0.155	ITT	0.192	0.031	0.130	0.253	0.037	23.60	19.26	20.17	1.44	78.88	100.00
			PP	0.183	0.033	0.119	0.248	0.028	18.34	20.44	21.38	1.17	86.92	100.00
			IPCW	0.184	0.035	0.116	0.252	0.029	18.79	22.21	22.46	1.31	86.31	100.00
			SNFTM	0.181	0.030	0.121	0.241	0.026	16.77	22.08	19.61	1.19	82.59	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.79	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	0.96	-
47	1827	0.087	ITT	0.140	0.028	0.084	0.195	0.052	59.65	32.72	32.39	4.05	54.30	100.00
			PP	0.114	0.029	0.057	0.170	0.026	29.81	33.03	32.89	1.73	85.77	100.00
			IPCW	0.113	0.030	0.055	0.172	0.026	29.61	34.39	34.25	1.80	86.75	100.00

			SNFTM	0.113	0.026	0.061	0.164	0.025	28.76	33.81	30.01	1.72	81.17	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.79	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.17	-
48	1818	0.155	ITT	0.183	0.032	0.120	0.245	0.028	17.85	20.14	20.50	1.12	85.31	100.00
			PP	0.183	0.035	0.115	0.251	0.028	18.05	21.62	22.39	1.23	87.02	100.00
			IPCW	0.180	0.035	0.111	0.249	0.025	16.10	22.64	22.83	1.20	89.05	100.00
			SNFTM	0.181	0.031	0.121	0.241	0.026	16.82	25.03	19.79	1.41	80.42	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.73	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	0.93	-
49	1828	0.087	ITT	0.140	0.030	0.083	0.198	0.053	60.51	34.42	33.76	4.24	55.91	100.00
			PP	0.115	0.030	0.056	0.175	0.028	31.81	35.85	34.87	2.01	84.08	100.00
			IPCW	0.116	0.031	0.055	0.177	0.028	32.46	36.97	35.55	2.12	84.14	100.00
			SNFTM	0.115	0.027	0.062	0.167	0.027	31.15	38.46	30.59	2.14	75.93	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.85	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.16	-
50	1768	0.064	ITT	0.109	0.057	-0.003	0.221	0.046	71.96	88.23	89.93	8.23	87.16	100.00
			PP	0.086	0.058	-0.028	0.200	0.023	35.59	90.05	91.69	5.95	93.50	100.00
			IPCW	0.086	0.076	-0.062	0.234	0.023	35.58	122.69	120.91	10.36	92.39	100.00
			SNFTM	0.089	0.053	-0.015	0.192	0.025	39.38	92.71	83.32	6.44	88.69	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.59	-
			max MCSE	-	-	-	-	-	0.19	0.02	0.04	0.13	0.80	-
51	1774	0.095	ITT	0.148	0.055	0.040	0.256	0.053	55.21	56.87	57.94	5.98	84.50	100.00
			PP	0.120	0.056	0.011	0.229	0.025	25.87	58.60	58.76	3.90	92.84	100.00
			IPCW	0.121	0.072	-0.021	0.263	0.026	26.88	78.90	77.66	6.61	93.20	100.00
			SNFTM	0.122	0.051	0.023	0.222	0.027	28.52	60.33	53.38	4.24	87.37	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.60	-
			max MCSE	-	-	-	-	-	0.18	0.02	0.04	0.13	0.86	-
52	1823	0.095	ITT	0.149	0.057	0.037	0.261	0.054	56.51	59.34	60.27	6.39	83.87	100.00
			PP	0.122	0.059	0.006	0.238	0.027	28.33	61.59	62.29	4.37	92.38	100.00
			IPCW	0.126	0.071	-0.013	0.265	0.031	32.15	75.95	75.47	6.47	92.20	100.00
			SNFTM	0.125	0.052	0.023	0.226	0.030	31.09	62.82	54.51	4.67	86.56	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.62	-

			max MCSE	-	-	-	-	-	0.17	0.02	0.03	0.12	0.86	-
53	1806	0.055	ITT	0.092	0.055	-0.016	0.200	0.037	68.58	102.42	101.13	8.29	88.93	100.00
			PP	0.074	0.057	-0.038	0.186	0.019	35.50	105.81	104.69	6.79	93.24	100.00
			IPCW	0.073	0.074	-0.072	0.219	0.019	34.46	139.05	138.41	11.20	93.51	100.00
			SNFTM	0.074	0.052	-0.027	0.175	0.019	35.37	108.38	94.93	7.09	89.76	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.58	-
			max MCSE	-	-	-	-	-	0.18	0.02	0.04	0.13	0.74	-
54	1615	0.093	ITT	0.112	0.061	-0.008	0.232	0.019	20.33	65.25	65.76	4.35	93.25	100.00
			PP	0.107	0.068	-0.026	0.240	0.014	14.84	73.61	73.20	5.25	94.37	100.00
			IPCW	0.101	0.085	-0.065	0.267	0.008	8.57	97.63	91.76	8.94	92.57	100.00
			SNFTM	0.106	0.060	-0.012	0.223	0.013	13.72	75.62	64.44	5.50	88.73	100.00
			min MCSE	-	-	-	-	-	0.15	0.01	0.01	0.11	0.57	-
			max MCSE	-	-	-	-	-	0.23	0.03	0.03	0.16	0.79	-
55	1831	0.055	ITT	0.089	0.057	-0.023	0.201	0.034	63.03	104.51	104.53	8.13	89.84	100.00
			PP	0.072	0.060	-0.046	0.190	0.017	32.04	110.67	110.55	7.24	93.72	100.00
			IPCW	0.074	0.073	-0.068	0.216	0.019	35.18	136.64	134.17	10.86	93.77	100.00
			SNFTM	0.073	0.052	-0.030	0.176	0.018	33.62	112.27	96.30	7.49	88.91	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.56	-
			max MCSE	-	-	-	-	-	0.17	0.02	0.03	0.12	0.73	-
56	1799	0.088	ITT	0.139	0.055	0.031	0.247	0.051	58.19	61.85	62.99	6.33	85.49	100.00
			PP	0.112	0.056	0.002	0.222	0.024	27.33	63.11	64.01	4.15	93.00	100.00
			IPCW	0.114	0.073	-0.029	0.256	0.026	29.44	85.19	84.16	7.13	93.36	100.00
			SNFTM	0.114	0.051	0.014	0.214	0.026	29.67	64.45	58.22	4.42	89.61	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.59	-
			max MCSE	-	-	-	-	-	0.18	0.02	0.04	0.13	0.83	-

Table 39: Performance of methods across initiation non-adherence scenarios

No.	Successful nsim	Truth	Method	Mean estimate	SE of mean	95% Confidence interval		Bias	Percent bias	MSE	Model SE	Empirical SE	Coverage (%)	Successful estimation (%)
						Lower	Upper							
57	1815	0.108	ITT	0.162	0.032	0.100	0.225	0.054	49.83	3.61	29.27	29.14	59.23	100.00
			PP	0.132	0.035	0.063	0.201	0.023	21.64	1.62	32.29	32.09	89.86	100.00
			IPCW	0.131	0.037	0.058	0.204	0.023	21.08	1.79	34.41	34.78	90.58	100.00
			SNFTM	0.129	0.030	0.069	0.188	0.020	18.54	1.97	28.10	38.38	80.77	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.69	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	1.15	-
58	1830	0.063	ITT	0.126	0.029	0.068	0.184	0.063	99.11	7.54	46.57	45.66	44.04	100.00
			PP	0.079	0.030	0.020	0.138	0.016	24.88	1.77	47.71	46.73	91.97	100.00
			IPCW	0.080	0.032	0.018	0.142	0.017	26.30	1.97	50.01	49.18	91.53	100.00
			SNFTM	0.081	0.026	0.029	0.133	0.017	27.61	2.17	41.67	51.63	82.79	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.64	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.05	1.16	-
59	1805	0.108	ITT	0.136	0.031	0.076	0.196	0.027	25.11	1.56	28.19	28.38	85.32	100.00
			PP	0.118	0.033	0.052	0.183	0.009	8.69	1.11	30.77	30.82	94.35	100.00
			IPCW	0.119	0.035	0.050	0.188	0.011	9.78	1.32	32.66	33.46	92.96	100.00
			SNFTM	0.113	0.030	0.054	0.172	0.004	4.07	1.33	27.72	34.86	87.76	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.54	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	0.83	-
60	1837	0.063	ITT	0.088	0.026	0.036	0.139	0.024	38.14	2.00	41.57	41.27	85.57	100.00
			PP	0.068	0.028	0.013	0.122	0.004	6.57	1.23	44.00	43.54	94.83	100.00
			IPCW	0.069	0.029	0.011	0.126	0.005	8.21	1.38	46.15	45.89	95.05	100.00
			SNFTM	0.066	0.025	0.017	0.115	0.003	4.55	1.30	39.48	45.12	91.29	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.04	0.51	-
			max MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.82	-
61	1805	0.108	ITT	0.148	0.031	0.088	0.209	0.040	36.91	2.36	28.50	28.51	73.74	100.00
			PP	0.123	0.035	0.055	0.191	0.014	13.33	1.31	32.09	32.13	93.19	100.00
			IPCW	0.122	0.037	0.050	0.193	0.013	12.37	1.42	33.72	34.01	93.52	100.00
			SNFTM	0.116	0.030	0.057	0.175	0.007	6.83	1.68	27.82	38.72	85.04	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.58	-

			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	1.04	-
62	1814	0.063	ITT	0.102	0.027	0.048	0.155	0.038	60.72	3.46	43.26	42.18	72.66	100.00
			PP	0.071	0.029	0.013	0.128	0.007	11.34	1.43	46.35	46.15	93.88	100.00
			IPCW	0.071	0.031	0.011	0.132	0.008	12.60	1.59	48.67	48.54	94.38	100.00
			SNFTM	0.069	0.025	0.020	0.119	0.006	9.59	1.57	40.04	48.93	88.81	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.04	0.54	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.05	-
63	1817	0.169	ITT	0.213	0.031	0.152	0.273	0.044	26.21	1.70	18.27	18.00	70.83	100.00
			PP	0.188	0.033	0.123	0.252	0.019	11.32	0.83	19.43	19.12	91.25	100.00
			IPCW	0.188	0.035	0.119	0.257	0.020	11.69	1.00	20.96	21.40	90.53	100.00
			SNFTM	0.185	0.029	0.127	0.242	0.016	9.50	0.93	17.51	21.47	85.58	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.66	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.07	-
64	1811	0.095	ITT	0.149	0.027	0.095	0.203	0.054	57.27	3.88	28.96	28.65	49.31	100.00
			PP	0.109	0.028	0.055	0.163	0.014	15.13	0.99	29.11	28.54	91.99	100.00
			IPCW	0.110	0.029	0.053	0.166	0.015	16.07	1.10	30.49	30.12	91.44	100.00
			SNFTM	0.112	0.025	0.063	0.160	0.017	17.79	1.19	26.36	30.58	84.54	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.04	0.64	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.17	-
65	1803	0.169	ITT	0.230	0.032	0.168	0.293	0.062	36.57	2.84	18.87	18.66	51.36	100.00
			PP	0.201	0.034	0.134	0.269	0.033	19.37	1.30	20.41	19.86	84.36	100.00
			IPCW	0.200	0.037	0.128	0.272	0.031	18.51	1.41	21.87	22.19	85.14	100.00
			SNFTM	0.198	0.030	0.139	0.256	0.029	17.42	1.55	17.71	24.78	74.43	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.84	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	1.18	-
66	1796	0.095	ITT	0.181	0.030	0.123	0.240	0.087	91.45	8.88	31.55	31.83	18.21	100.00
			PP	0.120	0.030	0.062	0.179	0.026	26.97	1.62	31.44	31.37	87.92	100.00
			IPCW	0.121	0.031	0.061	0.182	0.027	28.15	1.77	32.79	32.78	87.31	100.00
			SNFTM	0.124	0.026	0.073	0.175	0.030	31.35	2.23	27.54	37.00	73.16	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.77	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.05	-
67	1801	0.155	ITT	0.200	0.031	0.139	0.260	0.045	28.74	1.85	19.91	19.12	69.91	100.00

			PP	0.175	0.033	0.110	0.239	0.020	12.61	0.89	21.21	20.38	91.39	100.00
			IPCW	0.176	0.035	0.106	0.245	0.021	13.35	1.10	22.83	23.08	91.12	100.00
			SNFTM	0.172	0.030	0.114	0.230	0.017	11.09	1.01	19.11	23.00	85.56	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.66	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.06	1.08	-
68	1812	0.087	ITT	0.142	0.028	0.088	0.196	0.054	62.21	4.27	31.55	31.76	48.45	100.00
			PP	0.102	0.028	0.048	0.157	0.015	16.95	1.19	31.71	32.77	91.11	100.00
			IPCW	0.103	0.029	0.046	0.160	0.016	18.17	1.31	33.27	34.20	90.78	100.00
			SNFTM	0.105	0.025	0.056	0.155	0.018	20.51	1.41	28.72	34.51	83.28	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.67	-
			max MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	1.17	-
69	1777	0.155	ITT	0.217	0.032	0.155	0.280	0.062	40.25	3.15	20.53	20.26	49.75	100.00
			PP	0.188	0.034	0.120	0.256	0.033	21.28	1.42	22.24	21.46	84.86	100.00
			IPCW	0.188	0.037	0.116	0.260	0.033	21.21	1.58	23.81	23.87	85.31	100.00
			SNFTM	0.185	0.030	0.126	0.244	0.030	19.36	1.67	19.31	26.48	74.51	100.00
			min MCSE	-	-	-	-	-	0.07	0.01	0.00	0.05	0.84	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.01	0.07	1.19	-
70	1776	0.087	ITT	0.174	0.030	0.115	0.233	0.087	99.21	9.66	34.36	34.73	18.30	100.00
			PP	0.115	0.030	0.056	0.174	0.027	31.43	1.88	34.25	34.18	84.91	100.00
			IPCW	0.116	0.031	0.055	0.177	0.029	32.82	2.07	35.72	35.90	84.63	100.00
			SNFTM	0.119	0.026	0.067	0.170	0.031	35.68	2.59	30.00	41.12	70.55	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.85	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.06	1.08	-
71	1791	0.093	ITT	0.120	0.031	0.061	0.180	0.027	29.45	1.89	32.88	34.19	83.70	100.00
			PP	0.103	0.033	0.037	0.168	0.010	10.45	1.40	35.90	37.36	92.74	100.00
			IPCW	0.104	0.036	0.035	0.174	0.011	12.05	1.68	38.21	40.71	92.35	100.00
			SNFTM	0.098	0.030	0.039	0.157	0.005	5.04	1.76	32.34	43.18	85.37	100.00
			min MCSE	-	-	-	-	-	0.08	0.00	0.00	0.05	0.61	-
			max MCSE	-	-	-	-	-	0.09	0.01	0.00	0.07	0.87	-
72	1794	0.054	ITT	0.078	0.026	0.026	0.130	0.023	43.13	2.36	48.71	49.87	85.56	100.00
			PP	0.059	0.028	0.004	0.114	0.004	7.95	1.56	51.50	52.95	94.15	100.00
			IPCW	0.060	0.029	0.002	0.117	0.005	9.87	1.76	53.97	56.03	93.65	100.00

			SNFTM	0.058	0.025	0.009	0.107	0.003	6.19	1.70	46.23	55.54	89.97	100.00
			min MCSE	-	-	-	-	-	0.06	0.00	0.00	0.05	0.55	-
			max MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.83	-
73	1796	0.093	ITT	0.134	0.031	0.073	0.194	0.041	43.54	2.89	33.20	34.74	73.22	100.00
			PP	0.108	0.035	0.040	0.177	0.015	16.46	1.67	37.40	39.02	91.59	100.00
			IPCW	0.108	0.037	0.037	0.180	0.015	16.37	1.83	39.28	41.18	91.76	100.00
			SNFTM	0.101	0.030	0.041	0.160	0.007	8.01	2.10	32.49	46.84	83.02	100.00
			min MCSE	-	-	-	-	-	0.08	0.00	0.00	0.05	0.65	-
			max MCSE	-	-	-	-	-	0.10	0.01	0.00	0.07	1.04	-
74	1804	0.054	ITT	0.092	0.028	0.039	0.146	0.038	69.97	4.12	50.61	51.84	71.56	100.00
			PP	0.062	0.029	0.004	0.120	0.007	13.67	1.84	54.24	56.57	93.18	100.00
			IPCW	0.063	0.031	0.002	0.123	0.009	15.67	2.01	56.80	58.78	93.35	100.00
			SNFTM	0.061	0.025	0.011	0.110	0.006	11.40	2.02	46.84	59.93	87.20	100.00
			min MCSE	-	-	-	-	-	0.07	0.00	0.00	0.05	0.59	-
			max MCSE	-	-	-	-	-	0.08	0.01	0.00	0.05	1.06	-
75	1421	0.108	ITT	0.210	0.060	0.093	0.328	0.102	94.27	12.99	55.20	55.71	58.69	100.00
			PP	0.185	0.064	0.060	0.309	0.076	70.48	9.22	58.86	59.53	76.07	100.00
			IPCW	0.181	0.090	0.004	0.358	0.073	67.10	13.57	84.52	89.62	82.30	100.00
			SNFTM	0.180	0.057	0.068	0.292	0.072	66.31	9.02	52.72	62.74	71.01	100.00
			min MCSE	-	-	-	-	-	0.16	0.03	0.01	0.11	1.01	-
			max MCSE	-	-	-	-	-	0.26	0.05	0.05	0.18	1.31	-
76	1638	0.064	ITT	0.150	0.054	0.045	0.256	0.087	136.73	16.33	84.76	83.75	62.39	100.00
			PP	0.110	0.054	0.004	0.216	0.046	73.08	7.88	85.19	84.10	86.87	100.00
			IPCW	0.113	0.073	-0.030	0.256	0.050	77.95	12.96	117.25	119.68	88.95	100.00
			SNFTM	0.115	0.049	0.020	0.210	0.052	81.47	8.83	76.70	85.22	78.75	100.00
			min MCSE	-	-	-	-	-	0.13	0.02	0.01	0.09	0.78	-
			max MCSE	-	-	-	-	-	0.19	0.03	0.05	0.13	1.20	-
77	1344	0.108	ITT	0.230	0.062	0.109	0.350	0.121	112.02	16.92	56.89	55.47	50.67	100.00
			PP	0.201	0.067	0.070	0.332	0.093	85.41	11.80	61.67	60.00	70.54	100.00
			IPCW	0.202	0.090	0.026	0.378	0.093	86.19	16.60	83.67	88.88	78.79	100.00
			SNFTM	0.199	0.058	0.086	0.311	0.090	83.26	12.37	53.20	67.02	61.38	100.00
			min MCSE	-	-	-	-	-	0.16	0.04	0.01	0.12	1.11	-

			max MCSE	-	-	-	-	-	0.26	0.06	0.04	0.19	1.36	-
78	1714	0.064	ITT	0.183	0.058	0.069	0.297	0.119	187.89	27.78	91.97	91.81	46.27	100.00
			PP	0.123	0.058	0.009	0.237	0.060	93.79	10.86	91.96	91.13	82.73	100.00
			IPCW	0.128	0.075	-0.019	0.275	0.064	101.39	15.98	119.64	121.95	84.89	100.00
			SNFTM	0.129	0.050	0.030	0.228	0.065	102.40	12.73	79.75	97.72	70.89	100.00
			min MCSE	-	-	-	-	-	0.14	0.02	0.01	0.10	0.87	-
			max MCSE	-	-	-	-	-	0.19	0.04	0.03	0.13	1.20	-
79	1398	0.108	ITT	0.207	0.059	0.092	0.322	0.099	91.07	12.23	54.39	54.72	60.09	100.00
			PP	0.182	0.063	0.058	0.307	0.074	68.42	8.83	58.62	58.89	77.11	100.00
			IPCW	0.180	0.088	0.007	0.353	0.071	65.85	13.24	82.62	88.82	81.60	100.00
			SNFTM	0.178	0.057	0.067	0.289	0.070	64.26	8.63	52.44	61.94	71.39	100.00
			min MCSE	-	-	-	-	-	0.16	0.03	0.01	0.11	1.04	-
			max MCSE	-	-	-	-	-	0.26	0.05	0.05	0.18	1.31	-
80	1701	0.064	ITT	0.140	0.052	0.038	0.242	0.077	120.46	13.47	82.09	81.81	68.96	100.00
			PP	0.107	0.053	0.002	0.211	0.043	67.67	7.28	84.24	82.95	88.07	100.00
			IPCW	0.109	0.072	-0.032	0.250	0.045	71.53	12.19	115.61	118.61	89.18	100.00
			SNFTM	0.110	0.048	0.017	0.203	0.046	73.12	7.89	75.42	84.15	81.25	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.76	-
			max MCSE	-	-	-	-	-	0.18	0.03	0.05	0.13	1.12	-
81	1471	0.108	ITT	0.227	0.058	0.113	0.340	0.118	109.24	15.96	53.52	52.94	47.31	100.00
			PP	0.195	0.064	0.070	0.320	0.087	80.07	10.52	58.89	57.44	71.65	100.00
			IPCW	0.193	0.083	0.029	0.356	0.084	77.95	13.50	77.73	79.97	79.95	100.00
			SNFTM	0.191	0.055	0.083	0.298	0.082	76.02	10.32	50.67	61.21	64.17	100.00
			min MCSE	-	-	-	-	-	0.15	0.03	0.01	0.11	1.04	-
			max MCSE	-	-	-	-	-	0.23	0.05	0.03	0.16	1.30	-
82	1715	0.064	ITT	0.168	0.056	0.059	0.277	0.105	165.09	22.19	87.68	87.61	52.24	100.00
			PP	0.116	0.057	0.005	0.228	0.053	83.14	9.36	90.08	88.46	85.66	100.00
			IPCW	0.120	0.073	-0.024	0.263	0.056	88.39	13.85	116.82	118.28	86.60	100.00
			SNFTM	0.120	0.049	0.024	0.217	0.057	89.57	10.34	77.74	90.85	75.80	100.00
			min MCSE	-	-	-	-	-	0.13	0.02	0.01	0.10	0.82	-
			max MCSE	-	-	-	-	-	0.18	0.03	0.03	0.13	1.21	-
83	1313	0.155	ITT	0.127	0.060	0.009	0.245	-0.028	-17.85	3.07	38.85	40.79	90.94	100.00

			PP	0.105	0.065	-0.022	0.232	-0.050	-32.28	4.51	42.00	43.19	86.75	100.00
			IPCW	0.110	0.092	-0.071	0.291	-0.045	-28.96	7.49	60.33	63.20	90.47	100.00
			SNFTM	0.103	0.059	-0.012	0.218	-0.052	-33.55	4.85	37.80	44.77	80.50	100.00
			min MCSE	-	-	-	-	-	0.17	0.02	0.01	0.12	0.79	-
			max MCSE	-	-	-	-	-	0.27	0.05	0.05	0.19	1.09	-
84	1680	0.088	ITT	0.091	0.054	-0.015	0.197	0.003	3.94	3.17	61.70	59.96	95.00	100.00
			PP	0.061	0.055	-0.047	0.170	-0.026	-30.04	4.13	63.19	61.62	91.96	100.00
			IPCW	0.066	0.076	-0.082	0.214	-0.022	-25.05	7.51	90.32	89.02	92.83	100.00
			SNFTM	0.063	0.050	-0.034	0.161	-0.024	-27.71	4.12	56.84	62.65	89.23	100.00
			min MCSE	-	-	-	-	-	0.13	0.01	0.01	0.09	0.53	-
			max MCSE	-	-	-	-	-	0.19	0.03	0.28	0.14	0.76	-
85	1416	0.155	ITT	0.150	0.059	0.035	0.265	-0.005	-3.43	2.32	37.95	38.51	94.99	100.00
			PP	0.118	0.065	-0.009	0.246	-0.037	-23.60	3.71	42.01	42.82	92.02	100.00
			IPCW	0.127	0.087	-0.043	0.297	-0.028	-17.91	5.66	56.38	57.70	93.01	100.00
			SNFTM	0.117	0.057	0.006	0.227	-0.039	-24.83	4.26	36.53	46.18	83.12	100.00
			min MCSE	-	-	-	-	-	0.16	0.01	0.01	0.11	0.58	-
			max MCSE	-	-	-	-	-	0.24	0.03	0.03	0.17	1.00	-
86	1715	0.088	ITT	0.117	0.058	0.004	0.230	0.029	32.90	4.77	65.76	65.94	91.08	100.00
			PP	0.069	0.059	-0.046	0.185	-0.019	-21.28	4.15	67.38	65.40	93.29	100.00
			IPCW	0.075	0.076	-0.075	0.225	-0.013	-14.41	7.11	88.28	88.87	93.46	100.00
			SNFTM	0.072	0.051	-0.028	0.172	-0.015	-17.62	4.28	58.33	67.54	88.80	100.00
			min MCSE	-	-	-	-	-	0.14	0.01	0.01	0.10	0.60	-
			max MCSE	-	-	-	-	-	0.19	0.02	0.03	0.13	0.76	-
87	1415	0.093	ITT	0.192	0.059	0.076	0.307	0.098	105.75	14.32	63.36	64.81	59.51	100.00
			PP	0.168	0.064	0.043	0.292	0.075	80.27	10.45	68.31	69.15	77.24	100.00
			IPCW	0.170	0.090	-0.006	0.346	0.077	82.73	17.15	97.71	107.61	81.30	100.00
			SNFTM	0.164	0.057	0.053	0.276	0.071	76.34	10.11	61.23	70.93	73.29	100.00
			min MCSE	-	-	-	-	-	0.16	0.03	0.01	0.11	1.04	-
			max MCSE	-	-	-	-	-	0.27	0.06	0.05	0.19	1.30	-
88	1696	0.055	ITT	0.134	0.052	0.031	0.236	0.079	144.81	16.26	95.94	93.99	67.57	100.00
			PP	0.100	0.053	-0.005	0.205	0.046	83.39	8.82	98.43	96.00	87.32	100.00
			IPCW	0.103	0.073	-0.039	0.245	0.048	88.60	14.93	136.37	139.73	88.72	100.00

			SNFTM	0.104	0.048	0.010	0.198	0.049	90.09	9.75	88.21	98.76	79.48	100.00
			min MCSE	-	-	-	-	-	0.12	0.02	0.01	0.09	0.77	-
			max MCSE	-	-	-	-	-	0.19	0.03	0.07	0.13	1.14	-
89	1461	0.093	ITT	0.210	0.060	0.092	0.328	0.117	125.65	18.83	64.83	66.60	50.65	100.00
			PP	0.179	0.066	0.049	0.309	0.086	92.41	12.98	71.35	73.49	73.03	100.00
			IPCW	0.181	0.089	0.007	0.356	0.088	94.85	17.72	96.69	100.20	80.08	100.00
			SNFTM	0.174	0.057	0.062	0.287	0.081	87.31	12.89	61.50	78.87	65.23	100.00
			min MCSE	-	-	-	-	-	0.16	0.04	0.01	0.11	1.04	-
			max MCSE	-	-	-	-	-	0.24	0.06	0.04	0.17	1.31	-
90	1778	0.055	ITT	0.164	0.056	0.054	0.273	0.109	199.64	27.52	102.59	102.85	49.72	100.00
			PP	0.110	0.057	-0.002	0.223	0.056	102.37	11.58	105.29	103.65	84.25	100.00
			IPCW	0.115	0.074	-0.030	0.259	0.060	110.13	16.95	137.57	137.61	87.01	100.00
			SNFTM	0.114	0.049	0.017	0.210	0.059	108.28	12.66	90.82	107.13	72.48	100.00
			min MCSE	-	-	-	-	-	0.13	0.02	0.01	0.09	0.80	-
			max MCSE	-	-	-	-	-	0.18	0.03	0.04	0.13	1.19	-

Appendix F: Results for secondary estimands

Table 40: Results for secondary estimands across implementation non-adherence scenarios

No.	Method	True RMST0	RMST0 estimate	SE of mean RMST0	True RMST0	RMST1 estimate	SE of mean RMST1	True RMST0	HR Estimate	SE of mean HR
1	ITT	0.608	0.471	0.023	0.716	0.606	0.023	0.649	0.630	0.071
	PP		0.537	0.027		0.668	0.025		0.624	0.085
	IPCW		0.545	0.027		0.673	0.026		0.627	0.087
	SNFTM		0.547	0.023		0.676	0.023		0.625	0.074
2	ITT	0.806	0.684	0.023	0.869	0.790	0.020	0.649	0.614	0.088
	PP		0.758	0.025		0.841	0.021		0.628	0.121
	IPCW		0.757	0.026		0.840	0.021		0.625	0.121
	SNFTM		0.759	0.022		0.841	0.022		0.627	0.103
3	ITT	0.608	0.534	0.023	0.716	0.670	0.022	0.649	0.608	0.071
	PP		0.573	0.026		0.697	0.024		0.624	0.085
	IPCW		0.580	0.026		0.703	0.024		0.622	0.088
	SNFTM		0.583	0.023		0.704	0.023		0.628	0.077
4	ITT	0.806	0.748	0.022	0.869	0.839	0.018	0.649	0.597	0.095
	PP		0.785	0.023		0.859	0.019		0.629	0.122
	IPCW		0.784	0.024		0.858	0.020		0.629	0.124
	SNFTM		0.788	0.021		0.860	0.021		0.635	0.109
5	ITT	0.608	0.516	0.022	0.716	0.647	0.022	0.649	0.626	0.072
	PP		0.566	0.027		0.689	0.025		0.631	0.089
	IPCW		0.572	0.027		0.693	0.025		0.634	0.091
	SNFTM		0.579	0.023		0.699	0.023		0.634	0.077
6	ITT	0.806	0.733	0.022	0.869	0.824	0.019	0.649	0.616	0.095
	PP		0.781	0.025		0.855	0.020		0.639	0.129
	IPCW		0.779	0.025		0.853	0.021		0.637	0.128
	SNFTM		0.785	0.021		0.857	0.021		0.639	0.109
7	ITT	0.608	0.496	0.023	0.777	0.705	0.022	0.484	0.462	0.056
	PP		0.549	0.026		0.747	0.023		0.459	0.065
	IPCW		0.556	0.026		0.752	0.023		0.458	0.067
	SNFTM		0.555	0.023		0.751	0.023		0.459	0.059
8	ITT	0.806	0.707	0.023	0.901	0.855	0.017	0.482	0.440	0.071
	PP		0.766	0.024		0.885	0.017		0.454	0.092
	IPCW		0.764	0.025		0.884	0.018		0.454	0.094
	SNFTM		0.766	0.022		0.885	0.022		0.453	0.082
9	ITT	0.608	0.471	0.023	0.777	0.672	0.022	0.484	0.487	0.057
	PP		0.536	0.027		0.734	0.024		0.467	0.068
	IPCW		0.543	0.027		0.738	0.025		0.470	0.070
	SNFTM		0.546	0.023		0.741	0.023		0.467	0.059
10	ITT	0.806	0.684	0.023	0.901	0.835	0.018	0.482	0.465	0.072
	PP		0.758	0.026		0.878	0.019		0.466	0.098
	IPCW		0.757	0.026		0.878	0.019		0.466	0.098

	SNFTM		0.760	0.022		0.880	0.022		0.467	0.083
11	ITT	0.609	0.497	0.023	0.764	0.693	0.022	0.518	0.488	0.058
	PP		0.550	0.026		0.735	0.023		0.485	0.067
	IPCW		0.557	0.026		0.740	0.023		0.485	0.070
	SNFTM		0.558	0.023		0.740	0.023		0.489	0.062
12	ITT	0.807	0.707	0.023	0.894	0.848	0.018	0.517	0.464	0.074
	PP		0.766	0.024		0.879	0.018		0.481	0.096
	IPCW		0.765	0.025		0.878	0.018		0.482	0.098
	SNFTM		0.767	0.022		0.879	0.022		0.486	0.086
13	ITT	0.609	0.472	0.023	0.764	0.660	0.022	0.518	0.513	0.060
	PP		0.537	0.027		0.722	0.024		0.493	0.071
	IPCW		0.544	0.027		0.726	0.025		0.496	0.072
	SNFTM		0.548	0.023		0.730	0.023		0.495	0.062
14	ITT	0.807	0.684	0.023	0.894	0.827	0.019	0.517	0.490	0.074
	PP		0.759	0.026		0.873	0.019		0.490	0.101
	IPCW		0.757	0.026		0.872	0.020		0.490	0.101
	SNFTM		0.761	0.022		0.874	0.022		0.492	0.086
15	ITT	0.610	0.534	0.023	0.703	0.655	0.022	0.694	0.643	0.074
	PP		0.573	0.026		0.683	0.024		0.661	0.088
	IPCW		0.579	0.026		0.689	0.024		0.658	0.091
	SNFTM		0.583	0.023		0.690	0.023		0.668	0.081
16	ITT	0.807	0.749	0.022	0.862	0.830	0.018	0.696	0.640	0.100
	PP		0.786	0.024		0.851	0.019		0.672	0.128
	IPCW		0.785	0.024		0.850	0.020		0.673	0.130
	SNFTM		0.790	0.021		0.853	0.021		0.677	0.115
17	ITT	0.610	0.517	0.023	0.703	0.633	0.022	0.694	0.662	0.075
	PP		0.566	0.027		0.675	0.025		0.667	0.093
	IPCW		0.571	0.027		0.679	0.025		0.670	0.095
	SNFTM		0.580	0.023		0.685	0.023		0.676	0.081
18	ITT	0.807	0.734	0.022	0.862	0.815	0.019	0.696	0.655	0.099
	PP		0.782	0.025		0.847	0.020		0.679	0.135
	IPCW		0.780	0.025		0.846	0.021		0.676	0.135
	SNFTM		0.787	0.021		0.851	0.021		0.680	0.115
19	ITT	0.608	0.496	0.042	0.716	0.642	0.042	0.662	0.605	0.130
	PP		0.549	0.048		0.683	0.044		0.620	0.154
	IPCW		0.565	0.058		0.699	0.057		0.626	0.203
	SNFTM		0.559	0.043		0.692	0.043		0.621	0.140
20	ITT	0.805	0.702	0.044	0.868	0.813	0.038	0.674	0.589	0.173
	PP		0.762	0.047		0.848	0.038		0.630	0.232
	IPCW		0.758	0.056		0.845	0.048		0.648	0.288
	SNFTM		0.763	0.042		0.851	0.042		0.623	0.205
21	ITT	0.608	0.471	0.044	0.716	0.605	0.045	0.662	0.640	0.140
	PP		0.537	0.052		0.667	0.049		0.637	0.170
	IPCW		0.554	0.061		0.680	0.058		0.653	0.212
	SNFTM		0.550	0.044		0.678	0.044		0.640	0.149

22	ITT	0.805	0.679	0.045	0.868	0.786	0.040	0.674	0.624	0.175
	PP		0.755	0.050		0.840	0.041		0.647	0.246
	IPCW		0.750	0.057		0.837	0.048		0.665	0.286
	SNFTM		0.757	0.042		0.843	0.042		0.647	0.210
23	ITT	0.608	0.538	0.044	0.716	0.671	0.043	0.662	0.622	0.142
	PP		0.577	0.050		0.697	0.046		0.645	0.172
	IPCW		0.594	0.060		0.711	0.058		0.660	0.226
	SNFTM		0.589	0.044		0.706	0.044		0.651	0.156
24	ITT	0.805	0.744	0.042	0.868	0.836	0.035	0.674	0.614	0.190
	PP		0.782	0.046		0.858	0.037		0.656	0.250
	IPCW		0.778	0.054		0.854	0.047		0.684	0.312
	SNFTM		0.786	0.040		0.862	0.040		0.653	0.222
25	ITT	0.608	0.515	0.043	0.716	0.646	0.043	0.662	0.636	0.142
	PP		0.566	0.052		0.687	0.048		0.650	0.179
	IPCW		0.582	0.060		0.698	0.056		0.670	0.220
	SNFTM		0.580	0.044		0.701	0.044		0.651	0.155
26	ITT	0.805	0.730	0.043	0.868	0.822	0.037	0.674	0.630	0.189
	PP		0.779	0.048		0.854	0.039		0.666	0.265
	IPCW		0.774	0.055		0.851	0.046		0.682	0.305
	SNFTM		0.783	0.041		0.859	0.041		0.662	0.225
27	ITT	0.608	0.496	0.044	0.776	0.707	0.042	0.493	0.465	0.110
	PP		0.549	0.050		0.748	0.044		0.466	0.128
	IPCW		0.568	0.061		0.762	0.055		0.474	0.173
	SNFTM		0.559	0.044		0.755	0.044		0.465	0.116
28	ITT	0.805	0.705	0.045	0.900	0.854	0.034	0.500	0.450	0.142
	PP		0.765	0.047		0.884	0.034		0.474	0.189
	IPCW		0.761	0.056		0.881	0.043		0.492	0.236
	SNFTM		0.766	0.042		0.886	0.042		0.469	0.167
29	ITT	0.608	0.471	0.044	0.776	0.674	0.043	0.493	0.491	0.112
	PP		0.536	0.052		0.735	0.047		0.475	0.135
	IPCW		0.554	0.061		0.744	0.054		0.488	0.168
	SNFTM		0.549	0.044		0.744	0.044		0.475	0.117
30	ITT	0.805	0.681	0.045	0.900	0.832	0.036	0.500	0.477	0.143
	PP		0.757	0.050		0.877	0.037		0.488	0.200
	IPCW		0.753	0.057		0.875	0.042		0.502	0.231
	SNFTM		0.758	0.043		0.880	0.043		0.481	0.168
31	ITT	0.609	0.499	0.044	0.764	0.691	0.042	0.528	0.500	0.116
	PP		0.550	0.050		0.734	0.044		0.499	0.135
	IPCW		0.568	0.061		0.746	0.056		0.512	0.184
	SNFTM		0.562	0.045		0.742	0.045		0.503	0.124
32	ITT	0.806	0.707	0.045	0.894	0.846	0.034	0.537	0.484	0.150
	PP		0.768	0.047		0.877	0.034		0.512	0.200
	IPCW		0.764	0.056		0.876	0.043		0.522	0.246
	SNFTM		0.768	0.042		0.880	0.042		0.505	0.177
33	ITT	0.609	0.473	0.044	0.764	0.659	0.043	0.528	0.524	0.118

	PP		0.538	0.052		0.722	0.047		0.508	0.142
	IPCW		0.555	0.061		0.734	0.055		0.514	0.174
	SNFTM		0.551	0.044		0.733	0.044		0.505	0.123
34	ITT	0.806	0.683	0.045	0.894	0.823	0.037	0.537	0.510	0.151
	PP		0.760	0.050		0.871	0.037		0.524	0.213
	IPCW		0.755	0.057		0.870	0.043		0.532	0.243
	SNFTM		0.763	0.042		0.874	0.042		0.521	0.180
35	ITT	0.609	0.538	0.044	0.702	0.655	0.042	0.706	0.663	0.149
	PP		0.578	0.050		0.683	0.045		0.690	0.181
	IPCW		0.594	0.060		0.696	0.059		0.710	0.242
	SNFTM		0.592	0.044		0.693	0.044		0.700	0.166
36	ITT	0.806	0.745	0.042	0.861	0.829	0.036	0.722	0.648	0.198
	PP		0.784	0.046		0.852	0.037		0.688	0.259
	IPCW		0.781	0.054		0.848	0.047		0.713	0.322
	SNFTM		0.789	0.040		0.856	0.040		0.689	0.231
37	ITT	0.609	0.519	0.044	0.702	0.632	0.043	0.706	0.680	0.151
	PP		0.569	0.052		0.674	0.048		0.693	0.189
	IPCW		0.582	0.060		0.684	0.056		0.711	0.231
	SNFTM		0.583	0.044		0.685	0.044		0.701	0.165
38	ITT	0.806	0.731	0.043	0.861	0.814	0.037	0.722	0.666	0.197
	PP		0.779	0.048		0.849	0.039		0.693	0.272
	IPCW		0.775	0.055		0.846	0.046		0.704	0.310
	SNFTM		0.785	0.041		0.854	0.041		0.696	0.234

Table 41: Results for secondary estimands across persistence non-adherence scenarios

No.	Method	True RMST0	RMST0 estimate	SE of mean RMST0	True RMST1	RMST1 estimate	SE of mean RMST1	True HR	HR Estimate	SE of mean HR
39	ITT	0.608	0.477	0.023	0.716	0.608	0.023	0.649	0.639	0.072
	PP		0.537	0.026		0.667	0.025		0.626	0.083
	IPCW		0.545	0.026		0.673	0.025		0.628	0.086
	SNFTM		0.547	0.023		0.675	0.023		0.627	0.075
40	ITT	0.806	0.683	0.023	0.869	0.787	0.020	0.649	0.621	0.090
	PP		0.756	0.025		0.839	0.020		0.625	0.116
	IPCW		0.755	0.025		0.839	0.021		0.624	0.117
	SNFTM		0.757	0.022		0.841	0.022		0.625	0.103
41	ITT	0.608	0.520	0.023	0.716	0.658	0.022	0.649	0.608	0.071
	PP		0.563	0.025		0.690	0.023		0.620	0.081
	IPCW		0.572	0.026		0.701	0.025		0.611	0.086
	SNFTM		0.571	0.023		0.696	0.023		0.623	0.075
42	ITT	0.806	0.730	0.022	0.869	0.828	0.019	0.649	0.590	0.092
	PP		0.776	0.023		0.854	0.019		0.620	0.115
	IPCW		0.774	0.024		0.853	0.020		0.624	0.120
	SNFTM		0.778	0.021		0.855	0.021		0.621	0.105
43	ITT	0.806	0.707	0.023	0.901	0.854	0.018	0.482	0.446	0.072
	PP		0.764	0.024		0.884	0.017		0.454	0.089
	IPCW		0.763	0.024		0.883	0.018		0.459	0.094
	SNFTM		0.764	0.022		0.884	0.022		0.454	0.082
44	ITT	0.608	0.477	0.023	0.777	0.672	0.023	0.484	0.500	0.059
	PP		0.537	0.026		0.733	0.024		0.472	0.066
	IPCW		0.544	0.026		0.737	0.024		0.474	0.069
	SNFTM		0.546	0.023		0.740	0.023		0.469	0.059
45	ITT	0.608	0.683	0.023	0.777	0.831	0.019	0.484	0.476	0.073
	PP		0.755	0.025		0.877	0.018		0.465	0.093
	IPCW		0.754	0.025		0.876	0.019		0.465	0.094
	SNFTM		0.756	0.022		0.878	0.022		0.464	0.082
46	ITT	0.609	0.503	0.023	0.764	0.694	0.022	0.518	0.495	0.059
	PP		0.550	0.025		0.733	0.022		0.489	0.066
	IPCW		0.558	0.026		0.742	0.024		0.482	0.071
	SNFTM		0.557	0.023		0.738	0.023		0.491	0.062
47	ITT	0.807	0.707	0.023	0.894	0.847	0.018	0.517	0.470	0.075
	PP		0.765	0.024		0.878	0.017		0.481	0.093
	IPCW		0.763	0.024		0.876	0.019		0.486	0.098
	SNFTM		0.765	0.022		0.878	0.022		0.484	0.086
48	ITT	0.609	0.478	0.023	0.764	0.660	0.023	0.518	0.524	0.061
	PP		0.538	0.026		0.721	0.024		0.498	0.069
	IPCW		0.545	0.026		0.725	0.024		0.501	0.072
	SNFTM		0.547	0.023		0.728	0.023		0.498	0.062
49	ITT	0.807	0.683	0.024	0.894	0.823	0.019	0.517	0.500	0.076

	PP		0.756	0.025		0.871	0.019		0.491	0.097
	IPCW		0.754	0.025		0.870	0.019		0.491	0.098
	SNFTM		0.757	0.022		0.872	0.022		0.492	0.086
50	ITT	0.805	0.702	0.045	0.868	0.812	0.038	0.674	0.595	0.176
	PP		0.761	0.046		0.847	0.037		0.629	0.224
	IPCW		0.757	0.057		0.843	0.052		0.670	0.318
	SNFTM		0.761	0.042		0.850	0.042		0.620	0.203
51	ITT	0.805	0.703	0.045	0.900	0.851	0.034	0.500	0.458	0.145
	PP		0.763	0.046		0.883	0.033		0.475	0.183
	IPCW		0.758	0.058		0.879	0.046		0.498	0.250
	SNFTM		0.763	0.042		0.885	0.042		0.467	0.166
52	ITT	0.805	0.679	0.046	0.900	0.828	0.037	0.500	0.486	0.146
	PP		0.753	0.048		0.876	0.036		0.483	0.190
	IPCW		0.748	0.058		0.874	0.043		0.493	0.228
	SNFTM		0.753	0.043		0.878	0.043		0.473	0.164
53	ITT	0.806	0.727	0.044	0.861	0.819	0.037	0.722	0.637	0.190
	PP		0.774	0.045		0.848	0.037		0.671	0.241
	IPCW		0.769	0.057		0.842	0.051		0.708	0.330
	SNFTM		0.775	0.041		0.849	0.041		0.673	0.221
54	ITT	0.609	0.503	0.044	0.702	0.615	0.044	0.706	0.687	0.151
	PP		0.558	0.050		0.665	0.047		0.691	0.179
	IPCW		0.576	0.061		0.677	0.059		0.722	0.243
	SNFTM		0.569	0.044		0.675	0.044		0.691	0.160
55	ITT	0.806	0.707	0.044	0.861	0.796	0.038	0.722	0.667	0.191
	PP		0.768	0.047		0.840	0.039		0.688	0.254
	IPCW		0.764	0.056		0.838	0.047		0.708	0.307
	SNFTM		0.769	0.042		0.842	0.042		0.685	0.222
56	ITT	0.806	0.705	0.045	0.894	0.844	0.035	0.537	0.486	0.151
	PP		0.764	0.046		0.876	0.034		0.507	0.192
	IPCW		0.759	0.058		0.873	0.046		0.529	0.261
	SNFTM		0.764	0.042		0.878	0.042		0.500	0.174

Table 42: Results for secondary estimands across initiation non-adherence scenarios

No.	Method	True RMST0	RMST0 estimate	SE of mean RMST0	True RMST0	RMST1 estimate	SE of mean RMST1	True RMST0	HR Estimate	SE of mean HR
57	ITT	0.608	0.482	0.023	0.716	0.645	0.023	0.649	0.565	0.065
	PP		0.561	0.027		0.693	0.024		0.609	0.083
	IPCW		0.570	0.028		0.701	0.025		0.607	0.090
	SNFTM		0.573	0.023		0.701	0.023		0.614	0.074
58	ITT	0.806	0.690	0.023	0.869	0.816	0.019	0.649	0.538	0.081
	PP		0.777	0.025		0.856	0.019		0.616	0.118
	IPCW		0.775	0.025		0.855	0.020		0.614	0.123
	SNFTM		0.775	0.021		0.856	0.021		0.613	0.104
59	ITT	0.608	0.555	0.023	0.716	0.691	0.022	0.649	0.599	0.071
	PP		0.591	0.025		0.709	0.023		0.631	0.085
	IPCW		0.598	0.026		0.717	0.025		0.625	0.091
	SNFTM		0.603	0.023		0.716	0.023		0.641	0.079
60	ITT	0.806	0.765	0.021	0.869	0.853	0.017	0.649	0.590	0.097
	PP		0.798	0.023		0.866	0.018		0.641	0.122
	IPCW		0.796	0.023		0.865	0.019		0.639	0.127
	SNFTM		0.801	0.020		0.867	0.020		0.644	0.113
61	ITT	0.608	0.525	0.023	0.716	0.673	0.022	0.649	0.580	0.068
	PP		0.581	0.027		0.704	0.023		0.624	0.086
	IPCW		0.588	0.027		0.710	0.025		0.623	0.092
	SNFTM		0.596	0.023		0.712	0.023		0.638	0.079
62	ITT	0.806	0.739	0.022	0.869	0.841	0.018	0.649	0.566	0.090
	PP		0.792	0.024		0.863	0.019		0.636	0.125
	IPCW		0.790	0.025		0.862	0.020		0.635	0.129
	SNFTM		0.795	0.020		0.865	0.020		0.639	0.111
63	ITT	0.608	0.515	0.023	0.777	0.727	0.021	0.484	0.445	0.055
	PP		0.574	0.026		0.761	0.022		0.464	0.065
	IPCW		0.580	0.027		0.768	0.023		0.459	0.072
	SNFTM		0.581	0.023		0.765	0.023		0.466	0.060
64	ITT	0.806	0.721	0.023	0.901	0.870	0.017	0.482	0.418	0.070
	PP		0.784	0.023		0.893	0.016		0.461	0.093
	IPCW		0.783	0.024		0.893	0.017		0.460	0.098
	SNFTM		0.781	0.021		0.893	0.021		0.456	0.084
65	ITT	0.608	0.460	0.023	0.777	0.690	0.022	0.484	0.440	0.053
	PP		0.548	0.027		0.749	0.023		0.452	0.064
	IPCW		0.556	0.028		0.756	0.024		0.449	0.070
	SNFTM		0.558	0.023		0.755	0.023		0.453	0.058
66	ITT	0.806	0.658	0.024	0.901	0.839	0.018	0.482	0.408	0.064
	PP		0.765	0.025		0.886	0.017		0.451	0.092
	IPCW		0.764	0.026		0.885	0.018		0.450	0.095
	SNFTM		0.760	0.022		0.885	0.022		0.447	0.080
67	ITT	0.609	0.516	0.023	0.764	0.715	0.021	0.518	0.471	0.058

	PP		0.575	0.026		0.749	0.022		0.493	0.068
	IPCW		0.581	0.027		0.757	0.024		0.486	0.075
	SNFTM		0.582	0.023		0.754	0.023		0.495	0.063
68	ITT	0.807	0.721	0.023	0.894	0.863	0.017	0.517	0.443	0.073
	PP		0.785	0.023		0.887	0.017		0.492	0.098
	IPCW		0.783	0.024		0.886	0.018		0.492	0.102
	SNFTM		0.782	0.021		0.887	0.021		0.485	0.088
69	ITT	0.609	0.461	0.023	0.764	0.678	0.022	0.518	0.463	0.055
	PP		0.549	0.027		0.737	0.023		0.479	0.067
	IPCW		0.556	0.028		0.744	0.024		0.475	0.073
	SNFTM		0.559	0.023		0.744	0.023		0.481	0.060
70	ITT	0.807	0.658	0.024	0.894	0.832	0.019	0.517	0.428	0.066
	PP		0.765	0.025		0.880	0.017		0.474	0.095
	IPCW		0.763	0.026		0.879	0.019		0.473	0.098
	SNFTM		0.760	0.022		0.879	0.022		0.471	0.083
71	ITT	0.610	0.556	0.023	0.703	0.676	0.022	0.694	0.638	0.075
	PP		0.592	0.026		0.695	0.023		0.674	0.089
	IPCW		0.598	0.026		0.702	0.025		0.668	0.096
	SNFTM		0.604	0.023		0.702	0.023		0.685	0.083
72	ITT	0.807	0.766	0.021	0.862	0.844	0.018	0.696	0.632	0.102
	PP		0.799	0.023		0.858	0.018		0.687	0.129
	IPCW		0.797	0.023		0.857	0.020		0.685	0.134
	SNFTM		0.802	0.020		0.859	0.020		0.690	0.118
73	ITT	0.610	0.525	0.023	0.703	0.659	0.022	0.694	0.616	0.071
	PP		0.580	0.027		0.689	0.023		0.664	0.090
	IPCW		0.587	0.027		0.695	0.025		0.660	0.096
	SNFTM		0.597	0.023		0.698	0.023		0.682	0.083
74	ITT	0.807	0.739	0.022	0.862	0.832	0.018	0.696	0.603	0.094
	PP		0.793	0.024		0.855	0.019		0.681	0.132
	IPCW		0.791	0.025		0.854	0.020		0.677	0.135
	SNFTM		0.796	0.020		0.857	0.020		0.685	0.117
75	ITT	0.608	0.517	0.045	0.716	0.727	0.041	0.662	0.456	0.110
	PP		0.577	0.050		0.761	0.042		0.478	0.132
	IPCW		0.593	0.066		0.774	0.061		0.495	0.213
	SNFTM		0.588	0.044		0.768	0.044		0.482	0.123
76	ITT	0.805	0.717	0.045	0.868	0.867	0.033	0.674	0.429	0.141
	PP		0.782	0.046		0.892	0.032		0.481	0.191
	IPCW		0.777	0.058		0.890	0.045		0.509	0.284
	SNFTM		0.779	0.041		0.895	0.041		0.462	0.168
77	ITT	0.608	0.460	0.045	0.716	0.690	0.043	0.662	0.446	0.104
	PP		0.549	0.052		0.750	0.044		0.460	0.128
	IPCW		0.563	0.067		0.765	0.060		0.461	0.189
	SNFTM		0.562	0.044		0.761	0.044		0.459	0.115
78	ITT	0.805	0.653	0.047	0.868	0.836	0.036	0.674	0.418	0.127
	PP		0.762	0.049		0.885	0.033		0.465	0.186

	IPCW		0.756	0.060		0.884	0.044		0.480	0.252
	SNFTM		0.758	0.043		0.886	0.043		0.454	0.161
79	ITT	0.608	0.531	0.044	0.716	0.737	0.040	0.662	0.454	0.110
	PP		0.583	0.049		0.765	0.042		0.479	0.133
	IPCW		0.598	0.065		0.778	0.059		0.492	0.208
	SNFTM		0.594	0.044		0.772	0.044		0.483	0.123
80	ITT	0.805	0.735	0.043	0.868	0.875	0.032	0.674	0.437	0.146
	PP		0.787	0.045		0.894	0.031		0.486	0.195
	IPCW		0.782	0.057		0.891	0.044		0.519	0.291
	SNFTM		0.788	0.040		0.898	0.040		0.470	0.173
81	ITT	0.608	0.482	0.043	0.716	0.709	0.040	0.662	0.439	0.100
	PP		0.561	0.050		0.756	0.042		0.464	0.125
	IPCW		0.576	0.062		0.769	0.055		0.471	0.182
	SNFTM		0.576	0.042		0.767	0.042		0.464	0.113
82	ITT	0.805	0.684	0.046	0.868	0.853	0.034	0.674	0.418	0.131
	PP		0.773	0.048		0.889	0.033		0.473	0.192
	IPCW		0.767	0.059		0.887	0.044		0.495	0.261
	SNFTM		0.772	0.042		0.892	0.042		0.458	0.166
83	ITT	0.608	0.532	0.044	0.776	0.660	0.042	0.493	0.640	0.145
	PP		0.584	0.050		0.689	0.044		0.686	0.176
	IPCW		0.598	0.065		0.708	0.066		0.687	0.274
	SNFTM		0.595	0.044		0.698	0.044		0.689	0.163
84	ITT	0.805	0.739	0.043	0.900	0.830	0.036	0.500	0.624	0.191
	PP		0.793	0.045		0.854	0.036		0.710	0.261
	IPCW		0.787	0.056		0.853	0.051		0.740	0.381
	SNFTM		0.793	0.040		0.857	0.040		0.700	0.235
85	ITT	0.608	0.483	0.043	0.776	0.633	0.042	0.493	0.600	0.128
	PP		0.561	0.050		0.680	0.044		0.657	0.165
	IPCW		0.572	0.062		0.699	0.061		0.635	0.227
	SNFTM		0.576	0.042		0.693	0.042		0.659	0.149
86	ITT	0.805	0.688	0.046	0.900	0.805	0.038	0.500	0.584	0.169
	PP		0.779	0.048		0.848	0.037		0.691	0.259
	IPCW		0.773	0.058		0.848	0.050		0.705	0.347
	SNFTM		0.778	0.041		0.851	0.041		0.679	0.225
87	ITT	0.609	0.532	0.044	0.764	0.723	0.041	0.528	0.486	0.116
	PP		0.583	0.049		0.751	0.042		0.514	0.140
	IPCW		0.594	0.065		0.764	0.061		-	0.220
	SNFTM		0.594	0.044		0.758	0.044		0.517	0.129
88	ITT	0.806	0.736	0.043	0.894	0.869	0.032	0.537	0.459	0.151
	PP		0.789	0.045		0.889	0.032		0.513	0.202
	IPCW		0.783	0.057		0.886	0.045		-	0.303
	SNFTM		0.789	0.040		0.892	0.040		0.499	0.181
89	ITT	0.609	0.485	0.045	0.764	0.695	0.042	0.528	0.471	0.110
	PP		0.564	0.052		0.744	0.043		0.501	0.139
	IPCW		0.576	0.066		0.758	0.060		0.494	0.199

	SNFTM		0.580	0.044		0.754	0.044		0.505	0.126
90	ITT	0.806	0.684	0.046	0.894	0.847	0.034	0.537	0.434	0.134
	PP		0.774	0.048		0.885	0.033		0.497	0.199
	IPCW		0.768	0.059		0.883	0.044		0.513	0.268
	SNFTM		0.774	0.042		0.887	0.042		0.487	0.174

Appendix G: Code used in the simulation program

The code used to run the simulation study in Stata MP (version 15.1) is presented below. The first set of code is for generating the truth for Scenario 1 with factor specifications as follows: sample size (n=450), standard PSM DGM, perfect implementation non-adherence, no time-dependent treatment effect and large treatment effect size. These factors were amended to run the simulation program for each scenario as specified in Appendix A. The second set of code incorporate non-adherence and applies the alternative adjustment methods with factors amended across 90 scenarios.

Simulation program to generate the truth

```
*** Install Stata packages***
ssc install randomize
ssc install survsim
ssc install moremata
ssc install rcsgen
ssc install stpm2
ssc install simsum
net from https://www.pclambert.net/downloads/standsurv
install stgest3  // manual installation
install nlplot     // manual installation

*** Truth 1***
clear
capture program drop truth1
program define truth1, rclass
version 15.1
scalar drop _all

*** Run the simulation code to produce RCT datasets with perfect adherence and estimates dataset
quietly {
 local nobs 450    // number of observations in each simulated data set
 set varabbrev off

              set coeftabresults off //  runs faster
  drop _all

  *** declare sample size
  set obs `nobs'

              * Generate baseline covariates
              gen id= _n
              gen random = uniform()
              gen age = cond(random < 0.55, 1, 0)

              * Generate time-varying covariate at baseline
              gen hBMI0 = rbinomial(1,0.60)

              * Randomise observation to two groups "1= experimental and 2= control"
              randomize, groups(2) balance(age hBMI0)
              recode _assignment (1 = 0) (2 = 1), gen(trt)

              ** Generate hBMI at 4 months influenced by MNA0 and hBMI0
              gen hBMI1=rbinomial(1,0.30) if hBMI0==1
              replace hBMI1=rbinomial(1,0.20) if hBMI0==0

              ** Generate hBMI at 8 months influenced by MNA1 and hBMI1
              gen hBMI2=rbinomial(1,0.30) if hBMI1==1
              replace hBMI2=rbinomial(1,0.20) if hBMI1==0

*** DGMs for generating survival times using survsim with a user-defined hazard function incorporating both baseline and time-dependent
covariates
```

```
capture: survsim stime event, loghazard(-1.2:*0.2:*#t:^(0.2:-1) :* (hBMI0:* 0.35) :* (0:<=#t:<0.3333333) :+ (hBMI1:* 0.35) :* ///
 (0.3333333:<=#t:<0.6666667) :+ (hBMI2:* 0.35) :* (0.6666667:<=#t:<=1)) cov(trt -0.45 age 0.25) maxt(1)

* Declare the data to be survival data
stset stime, failure(event=1) id(id)

        replace hBMI1=. if stime<= float(0.3333333)
        replace hBMI2=. if stime<= float(0.6666667)

        ***Estimate the truth
        capture stcox trt
        if (e(converged)>0) return scalar conv_hr=1
        else return scalar conv_hr= 0
        if (_rc>0) return scalar error_hr=1
        else return scalar error_hr=0
        if (e(converged)>0) {
              return scalar hr=exp(_b[trt])
                                    return scalar hr_se=exp(_b[trt])*_se[trt]
        }
        else {
              return scalar hr=.
                                    return scalar hr_se=.
        }

        capture stpm2 trt, scale(h) df(2) lininit nolog eform
        if (e(converged)>0) return scalar conv_rmst=1
        else return scalar conv_rmst= 0
        if (_rc>0) return scalar error_rmst=1
        else return scalar error_rmst=0
        if (e(converged)>0) {
                                    gen tt1 = 1 in 1
                                    standsurv, at1(trt 0) at2(trt 1) ci se timevar(tt1) contrast(difference) rmst atvars(rmst_trt0
rmst_trt1) contrastvar(rmst_diff)
                                    summ rmst_diff, meanonly
                                    return scalar rmstdiff= r(mean)
                                    summ rmst_diff_se, meanonly
                                    return scalar rmstdiff_se= r(mean)
                                    summ rmst_diff_lci, meanonly
                                    return scalar rmstdiff_lci= r(mean)
                                    summ rmst_diff_uci, meanonly
                                    return scalar rmstdiff_uci= r(mean)
                                    summ rmst_trt0, meanonly
                                    return scalar rmst0= r(mean)
                                    summ rmst_trt0_se, meanonly
                                    return scalar rmst_trt0_se=r(mean)
                                    summ rmst_trt0_lci, meanonly
                                    return scalar rmst0_lci= r(mean)
                                    summ rmst_trt0_uci, meanonly
                                    return scalar rmst0_uci= r(mean)
                                    summ rmst_trt1, meanonly
                                    return scalar rmst1= r(mean)
                                    summ rmst_trt1_se, meanonly
                                    return scalar rmst_trt1_se=r(mean)
                                    summ rmst_trt1_lci, meanonly
                                    return scalar rmst1_lci= r(mean)
                                    summ rmst_trt1_uci, meanonly
                                    return scalar rmst1_uci= r(mean)
        }
        else {
                                    return scalar rmstdiff=.
                                    return scalar rmstdiff_se=.
                                    return scalar rmstdiff_lci=.
                                    return scalar rmstdiff_uci=.
                                    return scalar rmst0=.
                                    return scalar rmst_trt0_se=.
                                    return scalar rmst0_lci=.
                                    return scalar rmst0_uci=.
                                    return scalar rmst1=.
                                    return scalar rmst_trt1_se=.
                                    return scalar rmst1_lci=.
                                    return scalar rmst1_uci=.
```

```
                        }
}
end

set rng mt64s   // set the  stream 64-bit Mersenne Twister
set rngstream 11       //set the stream of rng
simulate hr=r(hr)  ///
hr_se=r(hr_se) ///
conv_hr=r(conv_hr) ///
error_hr=r(error_hr) ///
rmstdiff= r(rmstdiff) ///
rmstdiff_se = r(rmstdiff_se) ///
rmstdiff_lci= r(rmstdiff_lci) ///
rmstdiff_uci= r(rmstdiff_uci) ///
rmst0= r(rmst0) ///
rmst_trt0_se= r(rmst_trt0_se) ///
rmst0_lci= r(rmst0_lci) ///
rmst0_uci= r(rmst0_uci) ///
rmst1= r(rmst1) ///
rmst_trt1_se= r(rmst_trt1_se) ///
rmst1_lci= r(rmst1_lci) ///
rmst1_uci= r(rmst1_uci) ///
conv_rmst=r(conv_rmst) ///
error_rmst=r(error_rmst), ///
reps(1000000) seed(13183) saving(estimates1, replace): truth1

use estimates1, clear
gen idrep= _n   // generate idrep number
order idrep, first
gen dgm= 1   // generate dgm number
order dgm, after(idrep)

***rename variable names to sensible names
rename rmst_trt0_se rmst0_se
rename rmst_trt1_se rmst1_se

*** Order variables in the estimates dataset
order hr, after(rmst1_uci)
order hr_se, after(hr)
order conv_hr, after(error_rmst)
order error_hr, after(conv_hr)

*** Label variables and values
label variable idrep "Rep num"
label variable dgm "Data-generating mechanism"
label variable rmstdiff "Difference in Restricted Mean Survival Times"
label variable rmstdiff_se "Standard Error of the Difference in Restricted Mean Survival Times"
label variable rmstdiff_lci "RMST 95% CI: Upper bound"
label variable rmstdiff_uci "RMST 95% CI: Lower bound"
label variable rmst0 "Restricted Mean Survival Time 'Control Group'"
label variable rmst0_se "Standard Error of RMST 'Control Group'"
label variable rmst0_lci "RMST 95% CI: Lower bound 'Control Group'"
label variable rmst0_uci "RMST 95% CI: Upper bound 'Control Group'"
label variable rmst1 "Restricted Mean Survival Time 'Exp Group'"
label variable rmst1_se "Standard Error of RMST 'Exp Group'"
label variable rmst1_lci "RMST 95% CI: Lower bound 'Exp Group'"
label variable rmst1_uci "RMST 95% CI: Upper bound 'Exp Group'"
label variable hr "Hazard Ratio"
label variable hr_se "Standard Error of Hazard Ratio"
label variable conv_hr "HR model converged"
label variable error_hr "Error - HR model"
label variable conv_rmst "RMST model converged"
label variable error_rmst "Error - RMST model"
label define nylab 0 "No" 1 "Yes"
label values conv_hr conv_rmst error_hr error_rmst nylab
label define dgmlab 1 "Standard PSM - Weibull" 2 "Two-component Weibull Mixture"
label values dgm dgmlab

*** Save the truth estimates dataset
save truth1, replace
```

Simulation program to apply non-adherence adjustment methods

```
*** Non-Adherence Adjustment: Scenario 1 ***
clear
capture program drop mysimc1
program define mysimc1, rclass
version 15.1
scalar drop _all

**Run the simulation to produce non-adherence adjusted estimates dataset
quietly {
 local nobs 450                    // number of observations in each simulated data set
 set varabbrev off

        set coeftabresults off //  runs faster
  drop _all

  * declare sample size
  set obs `nobs'

        * Generate baseline covariates
        gen id= _n
        gen random = uniform()
        gen age = cond(random < 0.55, 1, 0)

        * Generate time-varying covariate at baseline
        gen hBMI0 = rbinomial(1,0.60)

        * Randomise observation to two groups "1= experimental and 2= control"
        randomize, groups(2) balance(age hBMI0)
        recode _assignment (1 = 0) (2 = 1), gen(trt)

        * MNA0: Generate time-varying non-adherence (implementation) and covariates at follow-up time 1 (Month 4 for tdc, 0-4 Months for MNA)
        *** Exp Group
        *** This assumes people with high BMI and age <24 have 22.5% chance of MNA.
        *** This risk is reduced to 15% if they have high BMI but age> 24 and 10% risk if age<24 but normal BMI
        *** For people with normal BMI and age>24, the probability of MNA is 5%
        gen MNA0=rbinomial(1,0.225) if (age==1 & hBMI0==1 & trt==1)
        replace MNA0=rbinomial(1,0.15) if (age==0 & hBMI0==1 & trt==1)
        replace MNA0=rbinomial(1,0.10) if (age==1 & hBMI0==0 & trt==1)
        replace MNA0=rbinomial(1,0.05) if (age==0 & hBMI0==0 & trt==1)

        *** Control Group
        *** This assumes people with high BMI and age <24 have 30% chance of MNA (higher than in the control group).
        *** This risk is reduced to 20% if they have high BMI but age> 24 and 10% risk if age<24 but normal BMI
        *** For people with normal BMI and age>24, the probability of MNA is 5%
        replace MNA0=rbinomial(1,0.30) if (age==1 & hBMI0==1 & trt==0)
        replace MNA0=rbinomial(1,0.20) if (age==0 & hBMI0==1 & trt==0)
        replace MNA0=rbinomial(1,0.10) if (age==1 & hBMI0==0 & trt==0)
        replace MNA0=rbinomial(1,0.05) if (age==0 & hBMI0==0 & trt==0)

        ** Generate hBMI at 4 months influenced by MNA0 and hBMI0
        *** This assumes people high BMI at baseline and non-adhered between 0-4 months will have 90% chance to have high BMI at Month 4 and this risk is reduced to 30% among adhered
        *** People with normal BMI at baseline with MNA0=1 have 60% chance of moving to high BMI category and this risk is reduced to 20% among adhered people
        *** The strength of the relationship between previous hBMI/MNA and subsequent hBMI is assumed to be constant over time
        *** The strgenth of hBMI/MNA relationships is assumed to be similar between the two arms
        gen hBMI1=rbinomial(1,0.90) if (MNA0==1 & hBMI0==1)
        replace hBMI1=rbinomial(1,0.30) if (MNA0==0 & hBMI0==1)
        replace hBMI1=rbinomial(1,0.60) if (MNA0==1 & hBMI0==0)
        replace hBMI1=rbinomial(1,0.20) if (MNA0==0 & hBMI0==0)


        * MNA1: Generate time-varying non-adherence and covariates at follow-up time 2 (Month 8 for tdc, 4-8 Months for MNA)
        *** EXP GROUP
        *** This assumes people with high BMI and age <24 have 45% chance of MNA.
        *** This risk is reduced to 30% if they have high BMI but age> 24 and 15% risk if age<24 but normal BMI
```

```
*** For people with normal BMI and age>24, the probability of MNA is 7.5%
gen MNA1=rbinomial(1,0.45) if (age==1 & hBMI1==1 & trt==1)
replace MNA1=rbinomial(1,0.30) if (age==0 & hBMI1==1 & trt==1)
replace MNA1=rbinomial(1,0.15) if (age==1 & hBMI1==0 & trt==1)
replace MNA1=rbinomial(1,0.075) if (age==0 & hBMI1==0 & trt==1)

*** Control Group
*** This assumes people with high BMI and age <24 have 60% chance of MNA.
*** This risk is reduced to 40% if they have high BMI but age> 24 and 20% risk if age<24 but normal BMI
*** For people with normal BMI and age>24, the probability of MNA is 10%
replace MNA1=rbinomial(1,0.60) if (age==1 & hBMI1==1 & trt==0)
replace MNA1=rbinomial(1,0.40) if (age==0 & hBMI1==1 & trt==0)
replace MNA1=rbinomial(1,0.20) if (age==1 & hBMI1==0 & trt==0)
replace MNA1=rbinomial(1,0.10) if (age==0 & hBMI1==0 & trt==0)
replace MNA1=1 if MNA0==1

** Generate hBMI at 8 months influenced by MNA1 and hBMI1
** Similar probabilities as hBMI1 at 4 months
gen hBMI2=rbinomial(1,0.90) if (MNA1==1 & hBMI1==1)
replace hBMI2=rbinomial(1,0.30) if (MNA1==0 & hBMI1==1)
replace hBMI2=rbinomial(1,0.60) if (MNA1==1 & hBMI1==0)
replace hBMI2=rbinomial(1,0.20) if (MNA1==0 & hBMI1==0)

* MNA2: Generate time-varying non-adherence at follow-up time 3 (8-12 Months)
*** EXP GROUP
*** This assumes people with high BMI and age <24 have 45% chance of MNA.
*** This risk is reduced to 30% if they have high BMI but age> 24 and 15% risk if age<24 but normal BMI
*** For people with normal BMI and age>24, the probability of MNA is 7.5%
gen MNA2=rbinomial(1,0.45) if (age==1 & hBMI2==1 & trt==1)
replace MNA2=rbinomial(1,0.30) if (age==0 & hBMI2==1 & trt==1)
replace MNA2=rbinomial(1,0.15) if (age==1 & hBMI2==0 & trt==1)
replace MNA2=rbinomial(1,0.075) if (age==0 & hBMI2==0 & trt==1)

*** Control Group
*** This assumes people with high BMI and age <24 have 60% chance of MNA
*** This risk is reduced to 40% if they have high BMI but age> 24 and 20% risk if age<24 but normal BMI
*** For people with normal BMI and age>24, the probability of MNA is 10%
replace MNA2=rbinomial(1,0.60) if (age==1 & hBMI2==1 & trt==0)
replace MNA2=rbinomial(1,0.40) if (age==0 & hBMI2==1 & trt==0)
replace MNA2=rbinomial(1,0.20) if (age==1 & hBMI2==0 & trt==0)
replace MNA2=rbinomial(1,0.10) if (age==0 & hBMI2==0 & trt==0)
replace MNA2=1 if MNA1==1

*** Generate admin censoring time at 1 year (End of study follow-up)
gen admin = 1

* DGMs for generating survival times using survsim with a user-defined hazard function incorporating both baseline and time-
dependent covariates
* The model use coefficient values of 0.35 for BMI and 0.40 for non-adherence
        capture: survsim stime event, loghazard(-1.2:*0.2:*#t:^(0.2:-1) :* (hBMI0:* 0.35) :* (0:<=#t:<0.3333333) :+  ///
         (MNA0:* 0.40) :* (0:<=#t:<0.3333333) :+ (hBMI1:* 0.35) :* (0.3333333:<=#t:<0.6666667) :+ ///
        (MNA1:* 0.40) :* (0.3333333:<=#t:<0.6666667) :+ (hBMI2:* 0.35) :* (0.6666667:<=#t:<=1) :+ ///
        (MNA2:* 0.40) :* (0.6666667:<=#t:<=1)) cov(trt -0.45 age 0.25) maxt(1)

*** Declare the data to be survival data
stset stime, failure(event=1) id(id)

        replace hBMI1=. if stime<= float(0.3333333)
        replace hBMI2=. if stime<= float(0.6666667)

        replace MNA2 = . if stime<0.6666667
        replace MNA1 = . if stime<0.3333333

        ***************************************
        * Method 1: ITT
        ***************************************
        preserve
        capture stcox trt age
        if (e(converged)>0) return scalar conv_hr_method1=1
        else return scalar conv_hr_method1= 0
        if (_rc>0) return scalar error_hr_method1=1
```

```
else return scalar error_hr_method1=0

if (e(converged)>0) {
	return scalar hr_method1=exp(_b[trt])
		return scalar hr_se_method1=exp(_b[trt])*_se[trt]
}
else {
	return scalar hr_method1=.
		return scalar hr_se_method1=.
}

capture stpm2 trt age, scale(h) df(2) lininit nolog eform vce(robust)
if (e(converged)>0) return scalar conv_rmst_method1=1
else return scalar conv_rmst_method1= 0
if (_rc>0) return scalar error_rmst_method1=1
else return scalar error_rmst_method1=0

if (e(converged)>0) {
		gen tt1 = 1 in 1
		standsurv, at1(trt 0) at2(trt 1) ci se timevar(tt1) contrast(difference) rmst atvars(rmst_trt0
rmst_trt1) contrastvar(rmst_diff)
		summ rmst_diff, meanonly
		return scalar rmstdiff_method1= r(mean)
		summ rmst_diff_se, meanonly
		return scalar rmstdiff_se_method1= r(mean)
		summ rmst_diff_lci, meanonly
		return scalar rmstdiff_lci_method1= r(mean)
		summ rmst_diff_uci, meanonly
		return scalar rmstdiff_uci_method1= r(mean)
		summ rmst_trt0, meanonly
		return scalar rmst0_method1= r(mean)
		summ rmst_trt0_se, meanonly
		return scalar rmst_trt0_se_method1=r(mean)
		summ rmst_trt0_lci, meanonly
		return scalar rmst0_lci_method1= r(mean)
		summ rmst_trt0_uci, meanonly
		return scalar rmst0_uci_method1= r(mean)
		summ rmst_trt1, meanonly
		return scalar rmst1_method1= r(mean)
		summ rmst_trt1_se, meanonly
		return scalar rmst_trt1_se_method1=r(mean)
		summ rmst_trt1_lci, meanonly
		return scalar rmst1_lci_method1= r(mean)
		summ rmst_trt1_uci, meanonly
		return scalar rmst1_uci_method1= r(mean)
}
else {
		return scalar rmstdiff_method1=.
		return scalar rmstdiff_se_method1=.
		return scalar rmstdiff_lci_method1=.
		return scalar rmstdiff_uci_method1=.
		return scalar rmst0_method1=.
		return scalar rmst_trt0_se_method1=.
		return scalar rmst0_lci_method1=.
		return scalar rmst0_uci_method1=.
		return scalar rmst1_method1=.
		return scalar rmst_trt1_se_method1=.
		return scalar rmst1_lci_method1=.
		return scalar rmst1_uci_method1=.
}

*****************************************
* Method 2: PP
*****************************************
restore, preserve
gen eventPP = event
gen stimePP = stime
gen ctime=.
replace ctime= float(0.6666667) if MNA2==1
replace eventPP=0 if MNA2==1
replace stimePP=ctime if MNA2==1
```

```
replace ctime= float(0.3333333) if MNA1==1
replace eventPP=0 if MNA1==1
replace stimePP=ctime if MNA1==1
replace ctime= float(0.0000001) if MNA0==1
replace eventPP=0 if MNA0==1
replace stimePP=ctime if MNA0==1
stset stimePP, failure(eventPP) id(id)

capture stcox trt age
if (e(converged)>0) return scalar conv_hr_method2=1
else return scalar conv_hr_method2= 0
if (_rc>0) return scalar error_hr_method2=1
else return scalar error_hr_method2=0

if (e(converged)>0) {
      return scalar hr_method2=exp(_b[trt])
                        return scalar hr_se_method2=exp(_b[trt])*_se[trt]
}
else {
      return scalar hr_method2=.
                        return scalar hr_se_method2=.
}

capture stpm2 trt age, scale(h) df(2) lininit nolog eform vce(robust)
if (e(converged)>0) return scalar conv_rmst_method2=1
else return scalar conv_rmst_method2= 0
if (_rc>0) return scalar error_rmst_method2=1
else return scalar error_rmst_method2=0

if (e(converged)>0) {
                        gen tt1 = 1 in 1
                        standsurv, at1(trt 0) at2(trt 1) ci se timevar(tt1) contrast(difference) rmst atvars(rmst_trt0
rmst_trt1) contrastvar(rmst_diff)
                        summ rmst_diff, meanonly
                        return scalar rmstdiff_method2= r(mean)
                        summ rmst_diff_se, meanonly
                        return scalar rmstdiff_se_method2= r(mean)
                        summ rmst_diff_lci, meanonly
                        return scalar rmstdiff_lci_method2= r(mean)
                        summ rmst_diff_uci, meanonly
                        return scalar rmstdiff_uci_method2= r(mean)
                        summ rmst_trt0, meanonly
                        return scalar rmst0_method2= r(mean)
                        summ rmst_trt0_se, meanonly
                        return scalar rmst_trt0_se_method2=r(mean)
                        summ rmst_trt0_lci, meanonly
                        return scalar rmst0_lci_method2= r(mean)
                        summ rmst_trt0_uci, meanonly
                        return scalar rmst0_uci_method2= r(mean)
                        summ rmst_trt1, meanonly
                        return scalar rmst1_method2= r(mean)
                        summ rmst_trt1_se, meanonly
                        return scalar rmst_trt1_se_method2=r(mean)
                        summ rmst_trt1_lci, meanonly
                        return scalar rmst1_lci_method2= r(mean)
                        summ rmst_trt1_uci, meanonly
                        return scalar rmst1_uci_method2= r(mean)
}
else {
                        return scalar rmstdiff_method2=.
                        return scalar rmstdiff_se_method2=.
                        return scalar rmstdiff_lci_method2=.
                        return scalar rmstdiff_uci_method2=.
                        return scalar rmst0_method2=.
                        return scalar rmst_trt0_se_method2=.
                        return scalar rmst0_lci_method2=.
                        return scalar rmst0_uci_method2=.
                        return scalar rmst1_method2=.
                        return scalar rmst_trt1_se_method2=.
                        return scalar rmst1_lci_method2=.
                        return scalar rmst1_uci_method2=.
```

```
}

****************************************
* Method 3: IPCW
****************************************
restore, preserve

*Reshape data from wide format to long format
reshape long hBMI MNA, i(id) j(time)

*** Create time-dependent non-adherence for this interval
gen visit=time
replace time=float(0.3333333) if time==1
replace time=float(0.6666667) if time==2
drop if time>stime
gen timeend = min(stime,time+0.3333333)

*IPCW Step 1: Censor observations for non-adherence and reformat the data
**********************************************************************

* Generate Non-adherence indicator and time of non-adherence (in years)
gen xoind=MNA
gen xotime=.
replace xotime= 0 if xoind==1 & visit==0
replace xotime= float(0.3333333) if xoind==1 & visit==1
replace xotime= float(0.6666667) if xoind==1 & visit==2
by id: egen xotime1 = min(xotime)
replace xotime=xotime1
drop xotime1

gen xotdo=0
replace xotdo= 1 if (xotime>=time) & (xotime<time+float(0.3333333)) & (xoind==1)
replace xotdo= . if (xotime<time) & (xoind==1)

*** Create time-dependent outcome for graft loss (txlosstdo) in each interval (0-4 months, 4-8 months, 8-12 months)
gen txlosstdo = 0
replace txlosstdo = 1 if event==1 & timeend==float(stime)

*** Stset the data
stset timeend, time0(time) failure(txlosstdo)

* Create dummies for being the first and last observation per patient
by id: gen firstobs = _n==1
by id: gen lastobs  = _n==_N

***Note, things change over time as case mix of patients changes.
gen t1 = 0
replace t1=1 if visit==0
gen t2 = 0
replace t2=1 if visit==1
gen t3 = 0
replace t3=1 if visit==2

***Note that the impact of hBMI depends on age, and vice versa. So need interactions
by id: gen cat1 = 0
by id: replace cat1 = 1 if (age==0 & hBMI==0)
by id: gen cat2 = 0
by id: replace cat2 = 1 if (age==1 & hBMI==0)
by id: gen cat3 = 0
by id: replace cat3 = 1 if (age==0 & hBMI==1)
by id: gen cat4 = 0
by id: replace cat4 = 1 if (age==1 & hBMI==1)

***Then make cat time interaction
by id: gen cat1t1 = cat1*t1
by id: gen cat2t1 = cat2*t1
by id: gen cat3t1 = cat3*t1
by id: gen cat4t1 = cat4*t1

by id: gen cat1t2 = cat1*t2
by id: gen cat2t2 = cat2*t2
```

```
by id: gen cat3t2 = cat3*t2
by id: gen cat4t2 = cat4*t2

by id: gen cat1t3 = cat1*t3
by id: gen cat2t3 = cat2*t3
by id: gen cat3t3 = cat3*t3
by id: gen cat4t3 = cat4*t3

*IPCW Step 2: Model the probability of being censored over time
 **********************************************************************
*Use logistic regression to predict non-adherence given baseline covariates in the control arm (Non-adherence Model 1) &
*Estimate the probability of non-adherence for each patient-observation included in the regression
logistic xotdo age time if trt==0
if (e(converged)>0) return scalar conv_mna1_method3=1
else return scalar conv_mna1_method3= 0
if (_rc>0) return scalar error_mna1_method3=1
else return scalar error_mna1_method3=0
if (e(converged)>0) {
predict pn_mna if e(sample), pr
}

*Use logistic regression to predict non-adherence given the interaction between baseline and time-updated covariates in the control arm (Non-adherence Model 2) &
*Estimate the probability of non-adherence for each patient-observation included in the regression
logistic xotdo cat1t1 cat2t1 cat3t1 cat4t1 cat1t2 cat2t2 cat3t2 cat4t2 cat1t3 cat2t3 cat3t3 cat4t3 time if trt==0
if (e(converged)>0) return scalar conv_mna2_method3=1
else return scalar conv_mna2_method3= 0
if (_rc>0) return scalar error_mna2_method3=1
else return scalar error_mna2_method3=0
if (e(converged)>0) {
predict pd_mna if e(sample), pr
}

*Use logistic regression to predict non-adherence given baseline covariates in the experimental arm (Non-adherence Model 3)
*Estimate the probability of non-adherence for each patient-observation included in the regression
logistic xotdo age time if trt==1
if (e(converged)>0) return scalar conv_mna3_method3=1
else return scalar conv_mna3_method3= 0
if (_rc>0) return scalar error_mna3_method3=1
else return scalar error_mna3_method3=0
if (e(converged)>0) {
predict pn1_mna if e(sample), pr
}

*Use logistic regression to predict non-adherence given the interaction between baseline and time-updated covariates in the experimental arm (Non-adherence Model 4)
*Estimate the probability of non-adherence for each patient-observation included in the regression
logistic xotdo cat1t1 cat2t1 cat3t1 cat4t1 cat1t2 cat2t2 cat3t2 cat4t2 cat1t3 cat2t3 cat3t3 cat4t3 time if trt==1
if (e(converged)>0) return scalar conv_mna4_method3=1
else return scalar conv_mna4_method3= 0
if (_rc>0) return scalar error_mna4_method3=1
else return scalar error_mna4_method3=0
if (e(converged)>0) {
predict pd1_mna if e(sample), pr
}

replace pn_mna = pn1_mna if trt==1
replace pd_mna = pd1_mna if trt==1
drop pn1_mna pd1_mna

*IPCW Step 3: For each individual at each time, compute the inverse probability of remaining uncensored
**********************************************************************************************************
*Estimate the probabilities of remaining uncensored 'adhered' and the IPCW weights
sort id time
gen num = 1-pn_mna if firstobs
replace num = num[_n-1] * (1-pn_mna) if !firstobs

gen denom = 1-pd_mna if firstobs
```

```
replace denom = num[_n-1] * (1-pd_mna) if !firstobs

gen weight = 1/denom
gen sweight = num/denom

**Decalre the data as survival data with stablised weight incorporated and time0 specified for clustering
stset timeend txlosstdo if xotdo==0 [iw=sweight], time0(time)

*IPCW Step 4: Obtain IPCW RMST & HR estimates

**********************************************************************************************************
** Obtain HR using Cox model
capture stcox trt age
if (e(converged)>0) return scalar conv_hr_method3=1
else return scalar conv_hr_method3= 0
if (_rc>0) return scalar error_hr_method3=1
else return scalar error_hr_method3=0

if (e(converged)>0) {
      return scalar hr_method3=exp(_b[trt])
                        return scalar hr_se_method3=exp(_b[trt])*_se[trt]
}
else {
      return scalar hr_method3=.
                        return scalar hr_se_method3=.
}

** Obtain IPCW RMST estimates using stpm2 model
capture stpm2 trt age, scale(h) df(2) lininit nolog eform vce(robust)
if (e(converged)>0) return scalar conv_rmst_method3=1
else return scalar conv_rmst_method3= 0
if (_rc>0) return scalar error_rmst_method3=1
else return scalar error_rmst_method3=0

if (e(converged)>0) {
                        gen tt1 = 1 in 1
                        standsurv, at1(trt 0) at2(trt 1) ci se timevar(tt1) contrast(difference) rmst atvars(rmst_trt0
rmst_trt1) contrastvar(rmst_diff)
                        summ rmst_diff, meanonly
                        return scalar rmstdiff_method3= r(mean)
                        summ rmst_diff_se, meanonly
                        return scalar rmstdiff_se_method3= r(mean)
                        summ rmst_diff_lci, meanonly
                        return scalar rmstdiff_lci_method3= r(mean)
                        summ rmst_diff_uci, meanonly
                        return scalar rmstdiff_uci_method3= r(mean)
                        summ rmst_trt0, meanonly
                        return scalar rmst0_method3= r(mean)
                        summ rmst_trt0_se, meanonly
                        return scalar rmst_trt0_se_method3=r(mean)
                        summ rmst_trt0_lci, meanonly
                        return scalar rmst0_lci_method3= r(mean)
                        summ rmst_trt0_uci, meanonly
                        return scalar rmst0_uci_method3= r(mean)
                        summ rmst_trt1, meanonly
                        return scalar rmst1_method3= r(mean)
                        summ rmst_trt1_se, meanonly
                        return scalar rmst_trt1_se_method3=r(mean)
                        summ rmst_trt1_lci, meanonly
                        return scalar rmst1_lci_method3= r(mean)
                        summ rmst_trt1_uci, meanonly
                        return scalar rmst1_uci_method3= r(mean)
}
else {
                        return scalar rmstdiff_method3=.
                        return scalar rmstdiff_se_method3=.
                        return scalar rmstdiff_lci_method3=.
                        return scalar rmstdiff_uci_method3=.
                        return scalar rmst0_method3=.
                        return scalar rmst_trt0_se_method3=.
                        return scalar rmst0_lci_method3=.
```

```
                return scalar rmst0_uci_method3=.
                return scalar rmst1_method3=.
                return scalar rmst_trt1_se_method3=.
                return scalar rmst1_lci_method3=.
                return scalar rmst1_uci_method3=.
}

************************************
*Method 4: SNFTM with G-estimation
************************************
restore
*** Reshape data from wide format to long format
reshape long hBMI MNA, i(id) j(time)

gen visit=time
replace time=float(0.3333333) if time==1
replace time=float(0.6666667) if time==2
drop if time>stime
gen timeend = min(stime,time+0.3333333)

*** Create vist variable to go into the stgest3 model
replace visit=3 if visit==2
replace visit=2 if visit==1
replace visit=1 if visit==0

*** Generate Non-adherence indicator and time of first non-adherence event
gen xoind=MNA
gen xotime=.
replace xotime= 0 if xoind==1 & visit==1
replace xotime= float(0.3333333) if xoind==1 & visit==2
replace xotime= float(0.6666667) if xoind==1 & visit==3
by id: egen xotime1 = min(xotime)
replace xotime=xotime1
drop xotime1

*** Create time-dependent outcome for graft loss (txlosstdo) in each interval (0-4 months, 4-8 months, 8-12 months)
gen txlosstdo = 0
replace txlosstdo = 1 if event==1 & timeend==float(stime)

*** stset the data
stset timeend, failure(txlosstdo) id(id)

by id: gen adlag = xoind[_n-1]
replace adlag=0 if adlag==.

*** Estimate the Acceleration Factor as the effect of time-dependent non-adherence (MNA) on survival time
*** This should be done for each arm separately using the interaction between baseline and time-dependent
covariates

***Note, things change over time as casemix of patients changes.
gen t1 = 0
replace t1=1 if visit==0
gen t2 = 0
replace t2=1 if visit==1
gen t3 = 0
replace t3=1 if visit==2

*** Note that the impact of hBMI depends on age, and vice versa. So need interaction
by id: gen cat1 = 0
by id: replace cat1 = 1 if (age==0 & hBMI==0)
by id: gen cat2 = 0
by id: replace cat2 = 1 if (age==1 & hBMI==0)
by id: gen cat3 = 0
by id: replace cat3 = 1 if (age==0 & hBMI==1)
by id: gen cat4 = 0
by id: replace cat4 = 1 if (age==1 & hBMI==1)

*** Then make cat time interaction
by id: gen cat1t1 = cat1*t1
by id: gen cat2t1 = cat2*t1
by id: gen cat3t1 = cat3*t1
```

```
by id: gen cat4t1 = cat4*t1

by id: gen cat1t2 = cat1*t2
by id: gen cat2t2 = cat2*t2
by id: gen cat3t2 = cat3*t2
by id: gen cat4t2 = cat4*t2

by id: gen cat1t3 = cat1*t3
by id: gen cat2t3 = cat2*t3
by id: gen cat3t3 = cat3*t3
by id: gen cat4t3 = cat4*t3

*** Estimate Accelaration Factor as the effect of time-dependent non-aderehcne on survivial time
*** This should be done for each arm seperatley and calulate admin censoring
preserve
drop if trt==0

***Run the g-estimation on the Exp group
capture stgest3 xoind cat1t1 cat2t1 cat3t1 cat4t1 cat1t2 cat2t2 cat3t2 cat4t2 cat1t3 cat2t3 cat3t3 cat4t3 adlag,
visit(visit) lasttime(admin) range (-5 5) model(all) outcome(mgale) test(cluster) nograph nolist nocheckobs replace
if (e(converged)>0) return scalar conv_stgest1_method4=1
else return scalar conv_stgest1_method4= 0
if (_rc>0) return scalar error_stgest1_method4=1
else return scalar error_stgest1_method4=0
if (e(converged)>0) {
di r(trcaus)
scalar af1 = r(trcaus)
scalar survadminc1 = admin/af1
}

restore, preserve
drop if trt==1

***Run the g-estimation on the Control group
stgest3 xoind cat1t1 cat2t1 cat3t1 cat4t1 cat1t2 cat2t2 cat3t2 cat4t2 cat1t3 cat2t3 cat3t3 cat4t3 adlag, visit(visit)
lasttime(admin) range (-5 5) model(all) outcome(mgale) test(cluster) nograph nolist nocheckobs replace
if (e(converged)>0) return scalar conv_stgest0_method4=1
else return scalar conv_stgest0_method4= 0
if (_rc>0) return scalar error_stgest0_method4=1
else return scalar error_stgest0_method4=0
if (e(converged)>0) {
di r(trcaus)
scalar af0 = r(trcaus)
scalar survadminc0 = admin/af0
}

** Adjust stime and event usnig the AF generated from the g-estimation**
restore
sort id
collapse (max) trt age stime event admin xotime xoind, by(id)

*** Control group
gen cfact = (xotime + ((stime-xotime)/(af0))) if (trt==0 & xoind==1)
replace cfact = stime if (trt==0 & xoind==0)
gen dcfact = event if trt==0
replace dcfact=0 if (cfact>admin & trt==0)
replace cfact = admin if (cfact>admin & trt==0)

*** Exp group
replace cfact = (xotime + ((stime-xotime)/(af1))) if (trt==1 & xoind==1)
replace cfact = stime if (trt==1 & xoind==0)
replace dcfact = event if trt==1
replace dcfact=0 if (cfact>admin & trt==1)
replace cfact = admin if (cfact>admin & trt==1)

***Stset the data specifying the exit time to 1 year (follow-up time)
stset cfact, failure(dcfact) id(id)

capture stcox trt age
if (e(converged)>0) return scalar conv_hr_method4=1
else return scalar conv_hr_method4= 0
```

```
		if (_rc>0) return scalar error_hr_method4=1
		else return scalar error_hr_method4=0

		if (e(converged)>0) {
			return scalar hr_method4=exp(_b[trt])
				return scalar hr_se_method4=exp(_b[trt])*_se[trt]
		}
		else {
			return scalar hr_method4=.
				return scalar hr_se_method4=.
		}

		capture stpm2 trt age, scale(h) df(2) lininit nolog eform vce(robust)
		if (e(converged)>0) return scalar conv_rmst_method4=1
		else return scalar conv_rmst_method4= 0
		if (_rc>0) return scalar error_rmst_method4=1
		else return scalar error_rmst_method4=0

		if (e(converged)>0) {
				gen tt1 = 1 in 1
				standsurv, at1(trt 0) at2(trt 1) ci se timevar(tt1) contrast(difference) rmst atvars(rmst_trt0
rmst_trt1) contrastvar(rmst_diff)
				summ rmst_diff, meanonly
				return scalar rmstdiff_method4= r(mean)
				summ rmst_diff_se, meanonly
				return scalar rmstdiff_se_method4= r(mean)
				summ rmst_diff_lci, meanonly
				return scalar rmstdiff_lci_method4= r(mean)
				summ rmst_diff_uci, meanonly
				return scalar rmstdiff_uci_method4= r(mean)
				summ rmst_trt0, meanonly
				return scalar rmst0_method4= r(mean)
				summ rmst_trt0_se, meanonly
				return scalar rmst_trt0_se_method4=r(mean)
				summ rmst_trt0_lci, meanonly
				return scalar rmst0_lci_method4= r(mean)
				summ rmst_trt0_uci, meanonly
				return scalar rmst0_uci_method4= r(mean)
				summ rmst_trt1, meanonly
				return scalar rmst1_method4= r(mean)
				summ rmst_trt0_se, meanonly
				return scalar rmst_trt1_se_method4=r(mean)
				summ rmst_trt1_lci, meanonly
				return scalar rmst1_lci_method4= r(mean)
				summ rmst_trt1_uci, meanonly
				return scalar rmst1_uci_method4= r(mean)
		}
		else {
				return scalar rmstdiff_method4=.
				return scalar rmstdiff_se_method4=.
				return scalar rmstdiff_lci_method4=.
				return scalar rmstdiff_uci_method4=.
				return scalar rmst0_method4=.
				return scalar rmst_trt0_se_method4=.
				return scalar rmst0_lci_method4=.
				return scalar rmst0_uci_method4=.
				return scalar rmst1_method4=.
				return scalar rmst_trt1_se_method4=.
				return scalar rmst1_lci_method4=.
				return scalar rmst1_uci_method4=.
		}
}
end

set rng mt64s          // set the  stream 64-bit Mersenne Twister
set rngstream 11       //set the stream of rng
simulate hr_method1=r(hr_method1)  ///
hr_se_method1=r(hr_se_method1) ///
conv_hr_method1=r(conv_hr_method1) ///
error_hr_method1=r(error_hr_method1) ///
rmstdiff_method1= r(rmstdiff_method1) ///
```

rmstdiff_se_method1=r(rmstdiff_se_method1) ///
rmstdiff_lci_method1= r(rmstdiff_lci_method1) ///
rmstdiff_uci_method1= r(rmstdiff_uci_method1) ///
rmst0_method1= r(rmst0_method1) ///
rmst_trt0_se_method1= r(rmst_trt0_se_method1) ///
rmst0_lci_method1= r(rmst0_lci_method1) ///
rmst0_uci_method1= r(rmst0_uci_method1) ///
rmst1_method1= r(rmst1_method1) ///
rmst_trt1_se_method1= r(rmst_trt1_se_method1) ///
rmst1_lci_method1= r(rmst1_lci_method1) ///
rmst1_uci_method1= r(rmst1_uci_method1) ///
conv_rmst_method1=r(conv_rmst_method1) ///
error_rmst_method1=r(error_rmst_method1) ///
hr_method2=r(hr_method2) ///
hr_se_method2=r(hr_se_method2) ///
conv_hr_method2=r(conv_hr_method2) ///
error_hr_method2=r(error_hr_method2) ///
rmstdiff_method2= r(rmstdiff_method2) ///
rmstdiff_se_method2=r(rmstdiff_se_method2) ///
rmstdiff_lci_method2= r(rmstdiff_lci_method2) ///
rmstdiff_uci_method2= r(rmstdiff_uci_method2) ///
rmst0_method2= r(rmst0_method2) ///
rmst_trt0_se_method2= r(rmst_trt0_se_method2) ///
rmst0_lci_method2= r(rmst0_lci_method2) ///
rmst0_uci_method2= r(rmst0_uci_method2) ///
rmst1_method2= r(rmst1_method2) ///
rmst_trt1_se_method2= r(rmst_trt1_se_method2) ///
rmst1_lci_method2= r(rmst1_lci_method2) ///
rmst1_uci_method2= r(rmst1_uci_method2) ///
conv_rmst_method2=r(conv_rmst_method2) ///
error_rmst_method2=r(error_rmst_method2) ///
conv_mna1_method3=r(conv_mna1_method3) ///
error_mna1_method3=r(error_mna1_method3) ///
conv_mna2_method3=r(conv_mna2_method3) ///
error_mna2_method3=r(error_mna2_method3) ///
conv_mna3_method3=r(conv_mna3_method3) ///
error_mna3_method3=r(error_mna3_method3) ///
conv_mna4_method3=r(conv_mna4_method3) ///
error_mna4_method3=r(error_mna4_method3) ///
hr_method3=r(hr_method3) ///
hr_se_method3=r(hr_se_method3) ///
conv_hr_method3=r(conv_hr_method3) ///
error_hr_method3=r(error_hr_method3) ///
rmstdiff_method3= r(rmstdiff_method3) ///
rmstdiff_se_method3=r(rmstdiff_se_method3) ///
rmstdiff_lci_method3= r(rmstdiff_lci_method3) ///
rmstdiff_uci_method3= r(rmstdiff_uci_method3) ///
rmst0_method3= r(rmst0_method3) ///
rmst_trt0_se_method3= r(rmst_trt0_se_method3) ///
rmst0_lci_method3= r(rmst0_lci_method3) ///
rmst0_uci_method3= r(rmst0_uci_method3) ///
rmst1_method3= r(rmst1_method3) ///
rmst_trt1_se_method3= r(rmst_trt1_se_method3) ///
rmst1_lci_method3= r(rmst1_lci_method3) ///
rmst1_uci_method3= r(rmst1_uci_method3) ///
conv_rmst_method3=r(conv_rmst_method3) ///
error_rmst_method3=r(error_rmst_method3) ///
conv_stgest1_method4=r(conv_stgest1_method4) ///
error_stgest1_method4=r(error_stgest1_method4) ///
conv_stgest0_method4=r(conv_stgest0_method4) ///
error_stgest0_method4=r(error_stgest0_method4) ///
hr_method4=r(hr_method4) ///
hr_se_method4=r(hr_se_method4) ///
conv_hr_method4=r(conv_hr_method4) ///
error_hr_method4=r(error_hr_method4) ///
rmstdiff_method4= r(rmstdiff_method4) ///
rmstdiff_se_method4=r(rmstdiff_se_method4) ///
rmstdiff_lci_method4= r(rmstdiff_lci_method4) ///
rmstdiff_uci_method4= r(rmstdiff_uci_method4) ///
rmst0_method4= r(rmst0_method4) ///
rmst_trt0_se_method4= r(rmst_trt0_se_method4) ///

```
rmst0_lci_method4= r(rmst0_lci_method4) ///
rmst0_uci_method4= r(rmst0_uci_method4) ///
rmst1_method4= r(rmst1_method4) ///
rmst_trt1_se_method4= r(rmst_trt1_se_method4) ///
rmst1_lci_method4= r(rmst1_lci_method4) ///
rmst1_uci_method4= r(rmst1_uci_method4) ///
conv_rmst_method4=r(conv_rmst_method4) ///
error_rmst_method4=r(error_rmst_method4), ///
reps(1900) seed(13183) saving(estimatesc1, replace): mysimc1

use estimatesc1, clear
gen idrep= _n   // generate idrep number
order idrep, first

***reshape estimates data to long format
reshape long hr_method hr_se_method conv_hr_method error_hr_method ///
rmstdiff_method rmstdiff_se_method rmstdiff_lci_method rmstdiff_uci_method rmst0_method rmst_trt0_se_method ///
rmst0_lci_method rmst0_uci_method rmst1_method rmst_trt1_se_method rmst1_lci_method rmst1_uci_method ///
conv_mna1_method error_mna1_method conv_mna2_method error_mna2_method conv_mna3_method ///
error_mna3_method conv_mna4_method error_mna4_method conv_stgest1_method error_stgest1_method ///
conv_stgest0_method error_stgest0_method conv_rmst_method error_rmst_method, i(idrep) j(method)

***rename variable names to sensible names
rename hr_method hr
rename hr_se_method hr_se
rename conv_hr_method conv_hr
rename error_hr_method error_hr
rename rmstdiff_method rmstdiff
rename rmstdiff_se_method rmstdiff_se
rename rmstdiff_lci_method rmstdiff_lci
rename rmstdiff_uci_method rmstdiff_uci
rename rmst0_method rmst0
rename rmst_trt0_se_method rmst0_se
rename rmst0_lci_method rmst0_lci
rename rmst0_uci_method rmst0_uci
rename rmst1_method rmst1
rename rmst_trt1_se_method rmst1_se
rename rmst1_lci_method rmst1_lci
rename rmst1_uci_method rmst1_uci
rename conv_mna1_method conv_mna1
rename error_mna1_method error_mna1
rename conv_mna2_method conv_mna2
rename error_mna2_method error_mna2
rename conv_mna3_method conv_mna3
rename error_mna3_method error_mna3
rename conv_mna4_method conv_mna4
rename error_mna4_method error_mna4
rename conv_stgest1_method conv_stgest1
rename error_stgest1_method error_stgest1
rename conv_stgest0_method conv_stgest0
rename error_stgest0_method error_stgest0
rename conv_rmst_method conv_rmst
rename error_rmst_method error_rmst

*** Order vars
order rmstdiff_se, after(rmstdiff)
order hr, after(rmst1_uci)
order hr_se, after(hr)
order conv_stgest1 error_stgest1 conv_stgest0 error_stgest0, after(error_mna4)
order conv_hr, after(conv_rmst)
order error_hr, after(conv_hr)

*** Label variables and values
label variable idrep "Rep num"
label variable method "Method"
label variable hr "Hazard Ratio"
label variable hr_se "Standard Error of Hazard Ratio"
label variable rmstdiff "Difference in RMST"
label variable rmstdiff_se "Standard Error of the Difference in RMST"
label variable rmstdiff_lci "RMST 95% CI: Upper bound"
label variable rmstdiff_uci "RMST 95% CI: Lower bound"
```

```
label variable rmst0 "Restricted Mean Survival Time 'Control Group'"
label variable rmst0_se "Standard Error of RMST 'Control Group'"
label variable rmst0_lci "RMST 95% CI: Lower bound 'Control Group'"
label variable rmst0_uci "RMST 95% CI: Upper bound 'Control Group'"
label variable rmst1 "Restricted Mean Survival Time 'Exp Group'"
label variable rmst1_se "Standard Error of RMST 'Exp Group'"
label variable rmst1_lci "RMST 95% CI: Lower bound 'Exp Group'"
label variable rmst1_uci "RMST 95% CI: Upper bound 'Exp Group'"
label variable conv_hr "HR model converged"
label variable error_hr "Error - HR model"
label variable conv_rmst "RMST model converged"
label variable error_rmst "Error - RMST model"
label variable conv_mna1 "Non-adherence model(1) converged"
label variable error_mna1 "Error - Non-adherence model(1)"
label variable conv_mna2 "Non-adherence model(2) converged"
label variable error_mna2 "Error - Non-adherence model(2)"
label variable conv_mna3 "Non-adherence model(3) converged"
label variable error_mna3 "Error - Non-adherence model(3)"
label variable conv_mna4 "Non-adherence model(4) converged"
label variable error_mna4 "Error - Non-adherence model(4)"
label variable conv_stgest1 "G-estimation converged 'Exp Group'"
label variable error_stgest1 "Error - G-estimation 'Exp Group'"
label variable conv_stgest0 "G-estimation converged 'Control Group'"
label variable error_stgest0 "Error - G-estimation 'Control Group'"
label define nylab 0 "No" 1 "Yes"
label values conv_hr conv_rmst error_hr error_rmst nylab
label values conv_mna1 error_mna1 conv_mna2 error_mna2 conv_mna3 error_mna3 nylab
label values conv_mna4 error_mna4 conv_stgest1 error_stgest1 conv_stgest0 error_stgest0 nylab
label define methodlab 1 "ITT" 2 "PP" 3 "IPCW"  4 "SNFTM"
label values method methodlab

***Save labelled estimates dataset
save adjusted1, replace
```

Appendix H: Search strategies used to identify existing methodological frameworks

The literature search strategy used to identify any existing frameworks for addressing non-adherence in the context of HTA is provided in this section. The search was run on OVID MEDLINE(R) and Web of Science databases. The search used the following terms and yielded the number of recorded reported in the last column. This was complemented with an author search via Web of Science on Grutters JP who published a relevant conference abstract in Value in Health.

Ovid MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations and Daily 1946 to June 21, 2021

#	Terms	Results
1	(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).m_titl.	157078
2	framework.m_titl.	40921
3	methodology.m_titl	21597
4	guidance.m_titl.	18391
5	recommendations.m.titl	38489
6	methods.m.titl	165392
7	2 or 3 or 4 or 5 or 6	283329
8	1 and 7	4433
9	(cost-effectiveness or economic evaluation or health technology assessment).m_titl.	28300
10	8 and 9	12

Web of Science 1964 to 2021

#	Terms	Results
1	TI=(compliance or adherence or pharmacoadherence or persistence or persistency or concordance or initiation or implementation or noncompliance or	513,851

	nonadherence or nonpersistence or discontinuation or pharmionics or therapeutic alliance or patient irregularity or treatment refusal).m_titl.	
2	TI=(framework or methodology or guidance or recommendations or methods)	8,790,201
3	#2 AND #1	42,623
4	TI=(cost-effectiveness or economic evaluation or health technology assessment)	59,719
5	#4 AND #3	29

Author search (Web of Science), June 21, 2021

#	Terms	Results
1	AU=(Grutters JP)	14

Appendix I: Case study supplementary tables

Table 43 provides the coefficients from the IPCW weighing models. This covers 18 non-adherence (logistic) models across the three treatment arms with 6 models per arm (2 models for each interval one incorporated baseline confounders only and the other incorporated both baseline and time-dependent confounders. The predicted probabilities generated from these models were used to generate the stabilised weights for the IPCW adjusted analysis.

Table 43: The Coefficients from the non-adherence models by treatment arm and time interval

Model No.	Interval	Covariate	Coefficient	SE	[95% Confidence interval	
Group A: Standard-dose cyclosporine						
Non-adherence models with baseline confounders only						
1	0-3 months	Age	1.013	0.0172	0.980	1.048
1	0-3 months	Gender	2.375	1.2364	0.856	6.589
2	3-6 months	Age	1.021	0.0174	0.988	1.056
2	3-6 months	Gender	0.939	0.4047	0.404	2.186
3	6-12 months	Age	0.978	0.0093	0.960	0.996
3	6-12 months	Gender	0.894	0.2224	0.549	1.455
Group A: Standard-dose cyclosporine						
Non-adherence models with baseline and time-dependent confounders						
4	0-3 months	Age	1.010	0.0181	0.975	1.046
4	0-3 months	Gender	2.392	1.2471	0.861	6.646
4	0-3 months	BMI	1.039	0.0602	0.928	1.164
4	0-3 months	Acute rejection	1.000	-	-	-
5	3-6 months	Age	1.019	0.0177	0.985	1.055
5	3-6 months	Gender	0.956	0.4140	0.409	2.234
5	3-6 months	BMI	1.033	0.0574	0.926	1.151
5	3-6 months	Acute rejection	0.732	0.3797	0.265	2.023
6	6-12 months	Age	0.978	0.0094	0.960	0.997
6	6-12 months	Gender	0.882	0.2205	0.540	1.440
6	6-12 months	BMI	0.975	0.0299	0.918	1.035
6	6-12 months	Acute rejection	0.958	0.5663	0.301	3.052
Group B: Low-dose cyclosporine						
Non-adherence models with baseline confounders only						

7	0-3 months	Age	1.004	0.0163	0.973	1.036
7	0-3 months	Gender	1.742	0.8331	0.683	4.448
8	3-6 months	Age	1.005	0.0172	0.972	1.040
8	3-6 months	Gender	0.827	0.3661	0.347	1.969
9	6-12 months	Age	0.997	0.0091	0.979	1.015
9	6-12 months	Gender	1.388	0.3426	0.856	2.252
Group B: Low-dose cyclosporine						
Non-adherence models with baseline and time-dependent confounders						
10	0-3 months	Age	1.010	0.0169	0.977	1.044
10	0-3 months	Gender	1.818	0.8756	0.707	4.672
10	0-3 months	BMI	0.932	0.0555	0.829	1.047
10	0-3 months	Acute rejection	1.000	-	-	-
11	3-6 months	Age	1.008	0.0174	0.975	1.043
11	3-6 months	Gender	0.845	0.3770	0.353	2.026
11	3-6 months	BMI	0.950	0.0559	0.847	1.066
11	3-6 months	Acute rejection	0.178	0.1842	0.024	1.349
12	6-12 months	Age	0.995	0.0093	0.977	1.014
12	6-12 months	Gender	1.360	0.3379	0.836	2.213
12	6-12 months	BMI	1.020	0.0287	0.965	1.078
12	6-12 months	Acute rejection	1.448	0.6145	0.630	3.326
Group C: Tacrolimus						
Non-adherence models with baseline confounders only						
13	0-3 months	Age	1.009	0.0145	0.981	1.038
13	0-3 months	Gender	1.225	0.5065	0.545	2.755
14	3-6 months	Age	1.014	0.0149	0.985	1.044
14	3-6 months	Gender	0.846	0.3392	0.385	1.856
15	6-12 months	Age	0.992	0.0084	0.975	1.008
15	6-12 months	Gender	0.692	0.1625	0.436	1.096
Group C: Tacrolimus						
Non -adherence models with baseline and time-dependent confounders						
16	0-3 months	Age	1.004	0.0152	0.975	1.035
16	0-3 months	Gender	1.222	0.5057	0.543	2.750
16	0-3 months	BMI	1.045	0.0491	0.953	1.146
16	0-3 months	Acute rejection	1.000	-	-	-
17	3-6 months	Age	1.012	0.0155	0.982	1.043

17	3-6 months	Gender	0.838	0.3368	0.382	1.842
17	3-6 months	BMI	1.020	0.0518	0.923	1.127
17	3-6 months	Acute rejection	0.727	0.4597	0.210	2.511
18	6-12 months	Age	0.991	0.0086	0.974	1.008
18	6-12 months	Gender	0.689	0.1620	0.434	1.092
18	6-12 months	BMI	1.012	0.0281	0.959	1.069
18	6-12 months	Acute rejection	0.760	0.4790	0.221	2.614

Note: the coefficient for acute rejection at for the interval 0-3 months was automatically omitted from the model as no acute rejection was recorded at baseline (day zero of kidney transplantation)

Table 44-46 provides the estimates of graft survivor functions from the IPCW adjusted sensitivity analyses including the SEs and 95% confidence intervals.

Table 44: Graft survivor function from sensitivity analysis - Differential real-world non-adherence levels

	Time	Beg. Total	Fail	Survivor Function	SE	95% Confidence Interval
Standard-dose cyclosporine (Control)						
	Baseline	390	0	1	-	-
	Month 3	329.91	29.0616	0.9203	0.0142	0.8874 0.9439
	Month 6	307.94	7.0263	0.9	0.0158	0.8641 0.9264
	Month 9	230.01	0	0.9	0.0158	0.8641 0.9268
	Month 12	220.00	2.98	0.8882	0.017	0.8499 0.9173
Low-dose cyclosporine						
	Baseline	399	0	1	-	-
	Month 3	352.15	14.9974	0.9595	0.0102	0.9337 0.9754
	Month 6	334.07	4.9516	0.9456	0.0119	0.9169 0.9646
	Month 9	248.66	3.23	0.9335	0.0135	0.9014 0.9554
	Month 12	231.75	2.99	0.9221	0.0148	0.8873 0.9465
Low-dose tacrolimus						
	Baseline	401	0	1	-	-
	Month 3	346.84	17.1325	0.9534	0.011	0.9262 0.9707
	Month 6	296.29	0.9995	0.9502	0.0114	0.9222 0.9683
	Month 9	255.15	0	0.9502	0.0114	0.9222 0.9683
	Month 12	241.87	0	0.9502	0.0114	0.9222 0.9683

Table 45: Graft survivor function from IPCW sensitivity analysis - Assuming perfect adherence between 0- 3 months

	Time	Beg. Total	Fail	Survivor Function	SE	95% Confidence Interval
Standard-dose cyclosporine (Control)						
	Baseline	390	0	1	-	-
	Month 3	330	29	0.9205	0.0142	0.8876 0.944
	Month 6	294.1	6.96	0.8995	0.0159	0.8633 0.9265
	Month 9	192	0	0.8995	0.0159	0.8633 0.9265
	Month 12	181.9	2.99	0.8853	0.0177	0.8454 0.9154
Low-dose cyclosporine						
	Baseline	399	0	1	-	-
	Month 3	352	15	0.9595	0.0102	0.9337 0.9754
	Month 6	321.1	4.82	0.9454	0.0119	0.9164 0.9645
	Month 9	210.6	3.24	0.9311	0.0141	0.8973 0.9541
	Month 12	193.7	2.99	0.9177	0.0159	0.8801 0.9438
Low-dose tacrolimus						
	Baseline	401	0	1	-	-
	Month 3	347	17	0.9538	0.011	0.9267 0.971
	Month 6	313.1	2.96	0.9449	0.012	0.9159 0.9641
	Month 9	193	0	0.9449	0.012	0.9159 0.9641
	Month 12	182.9	0	0.9449	0.012	0.9159 0.9641

Table 46: Graft survivor functions from sensitivity analysis - Adjusting for non-adherence between 6-12 months only

	Time	Beg. Total	Fail	Survivor Function	SE	95% Confidence Interval
Standard-dose cyclosporine						
	Baseline	390	0	1	-	-
	Month 3	330	29	0.9205	0.0142	0.8876 0.944
	Month 6	294	7	0.8993	0.0159	0.8631 0.9264
	Month 9	197.08	0	0.8993	0.0159	0.8631 0.9264
	Month 12	186.98	2.99	0.8855	0.0176	0.8458 0.9155
Low-dose cyclosporine						
	Baseline	399	0	1	-	-
	Month 3	352	15	0.9595	0.0102	0.9337 0.9754
	Month 6	321	5	0.9449	0.012	0.9158 0.9641
	Month 9	217.79	3.24	0.9311	0.0141	0.8976 0.9539
	Month 12	200.91	2.99	0.9181	0.0157	0.881 0.9539
Tacrolimus						
	Baseline	401	0	1	-	-
	Month 3	347	17	0.9538	0.011	0.9267 0.971
	Month 6	313	3	0.9448	0.012	0.9157 0.9264
	Month 9	202.95	0	0.9448	0.012	0.9157 0.964
	Month 12	193.85	0	0.9448	0.012	0.9157 0.964

Table 47: Causal parameter psi and causal survival time ratio from the g-estimation by treatment group

Treatment group	Causal parameter psi	Causal survival time ratio
Standard-dose cyclosporine	-2.57	13.04
Low-dose cyclosporine	- 2.85	17.29
Tacrolimus	-4.11	60.82