Addressing treatment switching in Health Technology Assessment

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Rebecca Hannah Boucher

Department of Health Sciences

University of Leicester

25th September 2018

Addressing treatment switching in Health Technology Assessment

Rebecca Hannah Boucher

Abstract

Background
Sometimes individual patient level (IPD) must be reconstructed data from summary information when treatment switching has occurred (i.e. proportion of patients changed treatment arms during the course of a randomised control trial) to facilitate re-analysis when the IPD is unavailable. However, to re-analyse overall survival (OS), information is needed on the time to treatment switching; this can usually be approximated by time to progression (TTP). Therefore, the reconstructed data must include TTP and time to death for patients, estimated using an illness-death modelling framework.

Methods
Here it is assumed only summary information of Progression-free survival (PFS) and OS are available. Using coordinates extracted from the Kaplan-Meier curves, the survival distributions are modelled. These are then combined with the PFS and OS risk tables, models for TTP, and estimates of the censoring distributions and post-progression survival (PPS) The data are then simulated and combined to obtain the underlying survival data. The correct proportion of treatment switchers is selected from those experiencing disease progression and the dataset analysed using a Rank Preserving Structural Failure Time Model (RPSFTM) to account for treatment switching. Multiple datasets are created from these models; each is analysed separately and the results averaged over to obtain a final point estimate.

Results
The simulated data are, on average, broadly representative of the original IPD, both in terms of the reported summary statistics and the RPSFTM analysis.

Conclusions
This application demonstrates the success with which this method can be used to reconstruct the data, and achieve an appropriate re-analysis for treatment switching, fulfilling a fundamental gap in the research.

Acknowledgements

I would like to thank my supervisors Professor Keith Abrams and Professor Paul Lambert for their support, advice and enthusiasm during my study.

Many thanks go to all the members of the Advisory Panel: Dr Nicholas Latimer, Dr Ian White, Professor Nicky Welton, Dr Elisabeth George, Maud Pacou, Joshua Ray, Elaine Wright, Eric Low and Professor Anne Thomas, for their guidance. I have really appreciated the friendly, supportive environment of the Biostatistics research group, and a big thank you to all of those who kindly took part in the Reproducibility Study.

Gratitude is extended to the National Institute for Health Research, who funded the work involved in producing this thesis, and have played a significant role in my career development to date. I have been very fortunate with all the opportunities that this funding entitled me to and the training I received throughout my studies.

I wish to thank members of the Mathematics department at Keele University; in particular Dr David Bedford for inspiring me to study Mathematics at University, and Dr Maria Heckl for first suggesting I would be capable of studying for a PhD.

My thanks also go to my colleagues at the Cancer Research Clinical Trials Unit at Birmingham for their interest, support and encouragement particularly in the last few weeks.

A huge thank you must go to my family and fiancé for their continued encouragement, support, hugs and confidence in me and my ability. Thank you to my fiancé for his reassurance, my Nan for her enthusiasm and my Dad, for encouraging me to continue learning. Many thanks to my mum, especially for her belief in me, and for helping me to develop a love of Maths.

Table of contents

List of Tables

List of Figures

List of Abbreviations

A	Anastrazole (Monotherapy)
AIC	Akaike information criterion
BIC	Bayesian Information Criterion
BSC	Best Supportive Care
CEA	Cost-Effectiveness Analysis
Cens.	Censored Observations
Chemo.	Chemotherapy
CI	Confidence Interval
CONSORT	Consolidating Standards for Reporting Trials
CTMP	The Center for Medical Technology Policy
Df	Degrees of freedom
EGFR	Epidermal Growth Factor Receptor
EGFR +ve	Epidermal Growth Factor Receptor Positive Mutation
EMA	European Medicines Agency
ERG	Evidence Review Group
FDA	Food and Drug Administration
HR	Hazard Ratio
HTA	Health Technology Assessment
IC	Indirect Comparison
ICER	Incremental Cost-Effectiveness Ratio
IPCW	Inverse Probability of Censoring Weighting
IPD	Individual Patient Data
IPE	Iterative Parameter Estimation
IPLD	Individual Patient Level Data
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
ITT	Intention-to-treat
LCH	Log Cumulative Hazard
LTFU	Lost to Follow-up
MTC	Mixed Treatment Comparison
NAR	‘Numbers at Risk’

NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NMA	Network Meta-Analysis
No.	Number
NSCLC	Non-Small-Cell Lung Cancer
ORR	Objective Response Rate
OS	Overall Survival
PD	Transition from Disease Progression to Death
PFS	Progression-free Survival
PH	Proportional Hazards
PP	Per Protocol
PPC	Post-progression Censoring
PPS	Post-progression Survival
Prob.	Probability
QALY	Quality-Adjusted Life Year
RCC	Renal Cell Carcinoma
RCS	Restricted Cubic Splines
RCT	Randomised Control Trial
RMST	Restricted Mean Survival
RPSFTM	Rank Preserving Structural Failure Time Models
SD	Transition from Stable to Death
SE	Standard Error
SNM	Structural Nested Models
SP	Transition from Stable to Disease Progression
STA	Single Technology Appraisal
T+A	Trastuzumab and Anastrozole (Combination therapy)
TA	Technology Appraisal
TKI	Tyrosine-Kinase Inhibitor
Trt	Treatment
TSD	Technical Support Document
TTP	Time to Progression

Chapter 1: Why treatment switching is an important issue in Health Technology Assessments

1.1 Treatment switching within Health Technology Assessment

1.1.1 Treatment switching in Randomised Control Trials

Within randomised control trials (RCTs) the term 'non-compliance' is widely used to refer to a variety of issues; since it describes any deviation, in which patients depart from taking the medication as specified in the protocol. Hence, 'non-compliance' encompasses patients who discontinue treatments or never receive their randomised intervention. Treatment switching, also known as crossover, is a specific form of non-compliance; as, in the context of this research, it is when a patient switches to an alternative therapy (often that of the other treatment arm) from that which they were randomised to. This change in treatment arms could happen during any stage of the trial. Whilst it is common across many disease areas for treatment switching to occur before the patient starts on their randomised treatment, this thesis concentrates on the alternative situation; that is, where treatment switching happens after the randomised treatment regimen has already started. Recent research (Morden, 2009, Latimer, 2012) has shown that permitting patients to switch treatment arms after the start of the first intervention is a complex methodological issue, and highlighted further areas of work which this thesis seeks to address.

Conventional analyses such as intention-to-treat (ITT) and per protocol (PP) approaches are frequently implemented to account for non-compliance. ITT is an approach whereby a patient is analysed according to the intervention group they were randomised to, regardless of whether they actually received that treatment. This gives a pragmatic estimate, representing what is likely to happen in practice as not all patients would be suitable to receive that particular treatment or take the treatments according to the protocol. PP, therefore, provides a better estimate of the efficacy, as it purely analyses patients according to the intervention received (and also providing that they complied with the treatment protocol). This means that patients who do not follow the treatment regimen as specified in the protocol are completely excluded. However, under this approach, patients who switch treatment groups before the administration of any treatment (and once again providing they follow the regimen for this new treatment

according to the protocol) are analysed as part of the treatment group to which they switch.

As this research concentrates on treatment switching or crossover where the switch does not occur at the start of treatment, patients will have experienced more than one of the interventions being compared in the trial. This type of treatment switching in RCTs is particularly common in trials for advanced or metastatic cancer or heart disease. In these trials, the outcome is typically time-to-event.

Treatment switching in data assessed for a time-to-event outcome can have a profound effect, especially when the treatment to which the patients are switching is considerably more effective. Thus, crossing over to the alternative treatment can result in an increased survival time. Under these circumstances adopting an ITT approach leads to the treatment effect being underestimated (Morden, 2009). Simulation studies have shown that the greater the proportion of patients switching, and the true treatment effect, the greater the bias (Morden, 2009, Latimer, 2012).

Treatment switching or crossover in a RCT should not be confused with 'crossover trials'. It differs considerably from 'crossover trials' in which crossover is part of the overall design. In a 'crossover trial' design, all patients should undergo both interventions. In this way each patient can act as their own control. These studies are, hence, restricted to chronic conditions as a patient must start the subsequent interventions in the same disease state as they were for the first. Where this cannot be achieved, such as with curative treatments or worsening conditions, crossover trials are infeasible. In contrast, for RCTs with crossover (in terms of a treatment switch), crossover is not an inherent part of the design (the analyst is not necessarily interested in how the patient responded under two different treatment regimens), not all patients may switch and it is not intended to use the patient as their own control. Moreover, it is typically permitted for ethical or practical reasons, and given the conditions it occurs in, the patient is likely to be at a more severe level of the disease when starting the second treatment.

1.1.2 Purpose of Health Technology Assessment

Health Technology Assessment (HTA) plays a vital role in today's society. In the United Kingdom, an intervention can only be prescribed by the National Health Service (NHS)

if it has been reimbursed, following a recommendation by the National Institute for Health and Care Excellence (NICE). Once an intervention is licensed, a manufacturer may submit documentary evidence, known as the manufacturer's submission, to the NICE demonstrating the clinical- and cost-effectiveness of their intervention. Once this has been submitted it is sent to an Evidence Review Group (ERG), typically an academic body who review the evidence, conduct reanalysis where necessary and provide feedback to the NICE committee. Following this, the evidence from the manufacturer and the ERG is considered by the NICE committee, and a decision is made on whether to recommend the treatment for reimbursement or not. The clinical and cost-effectiveness is usually based on one or more phase three RCTs. These trials are known as the pivotal evidence.

1.1.3 Available methodology for use with data where treatment switching has been permitted

1.1.3.1 *Simple methods*

1.1.3.1.1 Intention-to-treat

Intention-to-treat analysis (ITT) is a routinely used approach within the field of medical statistics, and widely considered as the 'gold standard' for analysing RCTs (Gupta, 2011). In an ITT analysis, all randomised subjects are included within the analysis, with patients being analysed in the groups to which they were assigned at randomisation; regardless of medical adherence, subsequent treatment withdrawal or protocol deviation (Fisher, 1990). As the ITT approach includes patients who may not have complied with the protocol within the treatment group, the true treatment difference will be attenuated. However, this attenuation is typically accepted as it gives more pragmatic results. These are considered reflective of the treatment effect that could be seen in the 'real world' as, in practice, not every patient can be expected to strictly comply with protocol conditions.

1.1.3.1.2 Per Protocol

The Per Protocol (PP) method is another popular analysis used within RCTs, often conducted as a secondary or sensitivity analysis to ITT. PP is the converse of an ITT analysis. Whereas the ITT approach endeavours to give a pragmatic view, PP analysis is purely concerned with measuring the efficacy of the treatment. Therefore, patients are analysed according to the treatment they received, rather than that to which they were initially randomised. In terms of adjusting for treatment switching in time-to-event data,

this can be conducted in one of two ways; patients who ultimately switched treatments could be excluded entirely from the analysis, or alternatively their follow-up could be included in the analysis up until the time that they switch, at which point they are censored.

In theory, this approach could resolve the issues surrounding treatment switching in HTA as it would only include the follow-up relevant to the decision problem (e.g. by excluding patients who have unusual treatment pathways, or by ignoring the follow-up for any alternative treatments). However, the complication is that patients do not switch at random.

For example, one typical reason for control group patients switching is to allow them the potential benefit of experimental intervention as a second-line / rescue medication. In these circumstances the patients who do switch treatments will be those with a more severe level of disease. Another common reason is to allow control patients the experimental intervention after treatment un-blinding, if this has been demonstrated to be superior. In contrast to the patient population who switched on progression, these patients changing treatments after un-blinding are likely to be the stronger patients of the control group, since they have already survived for a substantial part of the follow-up. This introduces selection bias, breaking the randomisation, and causing misleading results. Depending on the percentage of treatment switching, the selection bias can have profound implications for the power of the study and, thus the level of uncertainty in the decision problem. (Latimer, 2012). In addition, the approach by which treatment switchers are excluded from the analysis also results in bias because it conditions on future events.

1.1.3.1.3 Treatment as Time Varying Covariate

This is a simple extension of the semi-parametric proportional hazards (PH) Cox model (Cox, 1972), in which covariates are allowed to vary over time and is practically implemented by partitioning each patient's survival time into intervals based on which treatment they received at that timepoint (Cox, 1984). Since this method extends a commonly used survival model, it is easy to implement and understand. The treatment covariate is recorded as a binary variable, with zero typically representing the control intervention and one, the experimental treatment regime. This covariate is a function of time, allowing patients to change from one treatment to another.

The model can be written in the form,

$$\lambda_i(t) = \lambda_0(t) \exp\left(\beta\, x_i(t)\right) \quad (1\text{-}1)$$

Where λ represents the hazard, λ_0 the baseline hazard and $x_i(t)$, the binary variable for intervention as described above. It should be noted that $x_i(t)$ is a function of time and thus allows the individual to change from one treatment to the other.

To estimate the effect of the treatment, each patient's survival time is entered in according to the duration they had spent on that particular treatment. This approach, however, suffers similarly to PP when applied in a treatment switching context, if crossover is related to the patient's prognosis. Once again, this relationship violates randomisation and introduces often considerable selection bias (Morden, 2009).

1.1.3.1.4 Summary of the Simple Methods

The simple approaches are all popular, regularly implemented techniques within the analysis of RCTs. Hence, the concepts and findings can be easily understood and interpreted by a variety of people, including those in the pharmaceutical industry and decision-makers. However, in the presence of treatment switching from the control group to the intervention arm, the ITT analysis does not give the comparison required for decision-making. The other two approaches (PP and Treatment as a Time Varying Covariate) 'adjust' for changes in intervention, and therefore allow for the comparison of current versus potential (i.e. if the experimental intervention is introduced) NHS practice. Nevertheless, because of the nature of treatment switching (it does not occur randomly) randomisation is typically violated and results are considerably biased. As previously mentioned, this bias compromises the statistical power, and has severe implications on the uncertainty in the decision problem. Recent research (Latimer, 2012) has concluded that these methods, despite being used most commonly, are not appropriate analyses for inclusion in HTA submissions when treatment switching has occurred.

1.1.3.2 *Complex Methods*

This section discusses some more complex methods developed for addressing non-compliance within studies, that were also considered potential approaches for adjusting for treatment switching (Morden, 2009, Latimer, 2012). Of particular interest are some randomisation-based techniques which have been especially developed in order to

preserve the randomisation of the trial, two observational methods, originally designed to be used on observational datasets, and a two-stage method.

1.1.3.2.1 Adjusted Cox Model (Law and Kaldor, 1996)

Law and Kaldor (1996) proposed the Adjusted Cox model, another extension of the popular Cox model (Cox, 1972). A benefit of this method is that switching from either group can be modelled. Initially, from a non-statistical point of view, this method appears very intuitive as patients are grouped according to which treatment they received first and whether they switch treatments. This results in four groups: those who remain on the control intervention through the trial; those who only receive the experimental intervention; those who switch from the control intervention to the experimental sometime throughout the trial; and those who switch from the experimental group to the control. The previously mentioned Cox model with treatment as a time-varying covariate is then fitted. Given that the Cox model is a PH model, the PH assumption must hold. The underlying hazards of switchers and non-switchers are assumed to be multiplicative factors.

Whilst the Adjusted Cox Model seems intuitive, there are statistical flaws; in particular that by grouping switchers, a treatment switcher's hazard of dying is zero until their time of switch. This is because a patient cannot switch once they have died; therefore, the model prohibits them from dying until they have switched. However, this is not appropriate as the patient would always have been at risk of dying. Another fundamental assumption of survival models, which it violates, is that which states that stratification or conditioning cannot be based on future events as, otherwise, it leads to immortal time bias (Lévesque, 2010).

1.1.3.2.2 Causal Proportional Hazards Estimator (Loyes and Goethebeur, 2003)

The Causal Proportional Hazards Estimator (Loyes and Goethebeur, 2003) is restricted to scenarios with all-or-nothing compliance. 'All-or-nothing compliance' means that a patient's switch is said to occur at time zero, and so a patient can only receive one treatment. This is because the method divides the treatment group into compliers (those who adhered to their allocated treatment regimen) and non-compliers (those who changed interventions). In addition, the Causal Proportional Hazards Estimator can only adjust for

switching in one trial arm. This latter point should not affect the decision problem too adversely as it is only strictly necessary to account for 'crossover' in the control arm to ensure that this reflects current NHS practice. The 'all-or-nothing' compliance does, however, impede this method being used in practice; as in the context of this research, patients switch at a time later than the start of the study (e.g. disease progression) which violates this assumption, and thus leads to bias.

1.1.3.3 *Randomisation Based Methods*

1.1.3.3.1 Rank-Preserving Structural Failure Time Models (RPSFTM)

Robins and Tsiatis (1991) published a paper discussing a class of models, known as 'Rank Preserving Structural Failure Time Models (RPSFTM)'. The principle benefit for their use in a treatment switching context is that they endeavour to adjust for crossover, whilst preserving the randomisation of the trial. A RPSFTM is a particular type of Accelerated Failure Time (AFT) model. With an AFT model, the covariate of interest is assumed to have a multiplicative effect on the underlying survival time, rather than assuming PH. In other words,

$$S_1(t) = S_2(\alpha t) \tag{1-2}$$

Where the multiplicative effect, α is referred to as an acceleration factor. The acceleration factor is interpreted as the extent to which a patient's life is accelerated by the covariate of interest.

The principle benefit of using a RPSFTM in a treatment switching context is that they adjust for crossover, whilst preserving the randomisation of the trial. The RPSFTMs aim to model the patients underlying survival time assuming they received no (or more often, the control) treatment. These underlying survival times are referred to as the 'counterfactual' times. A patient's observed survival time (be it through death or censoring) will be denoted as, *T*, and their counterfactual time, as *U.* It should be noted that *T* is known as this comes directly from the data, whereas *U* is unknown. *T* and *U* are then related in the way described below.

For patients who have only experienced the control treatment during the entire study, *T=U* (e.g. the time they would have lived had they only received the control intervention is what was actually observed).

Provided that the counterfactual times can be assumed independent of randomisation, the observed time for a patient is expressed as the sum of the time the patient spent on the control treatment, T_C and the time spent on the new treatment, T_N (shown in Equation 1-3).

$$T = T_C + T_N \tag{1-3}$$

Typically, treatment switching research only considers treatment switching mechanisms where only the patients in the control group can switch, this means that for patients randomised to the new intervention, their time on the control treatment is zero.

The following causal model is used to relate the observed and counterfactual times.

$$U = T_C + \exp(-\psi)\, T_N \tag{1-4}$$

The factor, $\exp(-\psi)$, referred to as the acceleration time, can be interpreted such that values less than one indicate a protective effect, whilst those above one, signify a harmful effect.

A binary process $X_i(t)$ is defined, which takes the value one when a patient receives the experimental intervention and zero otherwise. The equation for the causal model can be written in the form (given in Equation 1-5):

$$U_i = \int_0^{T_i} \exp[\, \psi\, X_i(t)]\, dt \tag{1-5}$$

The method uses a test-based approach, whereby plausible values for ψ are tested. Initially the counterfactual times are calculated for each patient based on the current estimate for ψ. The test statistic, $Z(\psi)$ for that value of ψ using a specified statistical model is then computed. Potential tests include the log rank test, Cox, exponential or Weibull models. This process is repeated using different values of ψ until a value of ψ is found such that $Z(\psi) = 0$ and hence balances the counterfactual times between the two trial arms.

A fundamental and largely untestable assumption of the RPSFTMs, which gives the method its name, is that two patients receiving the same treatment regimen must follow

the same pattern if both had received the alternative treatment. In other words, in a pair of patients receiving the same treatment, the patient who died first would always follow this pattern, regardless of which of treatment both patients had received.

Another more vital assumption of these models is that that they assume a 'common treatment effect'. In other words, patients experience the same treatment effect, regardless of when they start receiving the treatment. Therefore, patients switching to a treatment benefit as much as those initially randomised to that intervention. This assumption may well be violated in the context of this research, as often patients switch on disease progression. Therefore, treatment switchers will start on the new treatment at a more severe level of disease than patients initially randomised, and are consequently less likely to benefit as much. Another consideration is whether treatment switching is related to prognosis (as it often is in the case in the context of this research), since in these circumstances RPSFTM findings are subject to bias unless the data is re-censored (White, 1999).

As an AFT modelling approach is used, an acceleration factor is obtained, rather than the more commonly presented hazard ratio (HR). However, particularly when submitting the results as part of a NICE TA, it is more usual to convert this acceleration factor to a HR. This is often achieved by calculating the counterfactual dataset using the estimated acceleration factor, and then fitting a standard Cox (1972) (or Weibull PH (Collett, 2003)) model to the data. Whilst this produces a satisfactory point estimate, the standard error (SE) obtained from the model would be too precise. Therefore, standard practice is to calculate the SE, by 'preserving the p-value'. This essentially means that the p-value and, thus test statistic, are retained from the ITT analysis, and the SE calculated from these and the point estimate. The practical calculation needed is described in Section 4.8.2.4.

1.1.3.3.2 Iterative Parameter Estimation Algorithm (IPE)

Whilst this method is a distinct method in its own right, in practice the Iterative Parameter Estimation (IPE) Algorithm is often grouped with the RPSFTMs (Branson and Whitehead, 2002). This is because RPSFTM and IPE use exactly the same underlying theory; the key difference is the estimation process. Rather than using a test-based approach, the acceleration factor is computed using a likelihood-based technique.

A parametric failure time model, such as a Weibull, log-logistic, log-normal or gamma distribution, is fitted to the data to compare treatment arms, similarly as would be done in an ITT analysis. This estimate can be used as the initial value for the acceleration factor, e^{ψ}. This factor is then applied to the time estimates, and the same chosen model fitted once again, giving a revised estimate for ψ. The time estimates are adjusted by this revised estimate, the model refitted, and a new estimate for ψ obtained. This process is repeated until at last the estimate for ψ converges. Re-censoring is important particularly in the model fitting process (e.g. when adjusting the observed survival time estimates). Whilst bootstrapping is recommended for obtaining the SE, it can be computationally time consuming, and so alternatively the SE could be calculated from the final regression model. Computing the SE this last way will result in a smaller SE.

Alongside the 'common treatment effect' and 'rank preserving' assumptions described in Section 1.1.3.3.1 for the RPSFTM, the IPE has the additional condition, that the parametric form must be appropriate for the survival estimates. This last condition is testable, and thus it is essential to check the parametric model is a suitable form for the data.

1.1.3.3.3 Parametric Randomisation-Based Methods

Like the IPE, this approach (Walker, 2004) uses a parametric distribution, but for this method three models are now used. These include a causal model which relates a patient's counterfactual time, *U*, to their observed failure time; a model for the association between *U*, the counterfactual times, and the test statistic *Z* (this is typically a bivariate frailty model, either positive stable or gamma); and a marginal cumulative hazards model. Whilst a maximum likelihood estimate approach could be used in this method, it is extremely sensitive, and hence augmented models are recommended to maintain the randomisation balance between groups. Additionally, these augmented models are more robust if the parametric model has been mis-specified (Morden, 2011).

1.1.3.4 *Observational Methods*

Once prognosis-related treatment switching occurs, randomisation has been violated and thus the trial becomes more like an observational study. Thus, certain methods designed specifically for observational studies were suggested, provided that the RCT data contains

the necessary information (i.e. satisfactory for confounding control). These approaches may be particularly useful when the 'common treatment effect' assumption does not hold. That is to say, patients who switch receive a different (possibly attenuated) treatment effect to those randomised to that treatment.

1.1.3.4.1 Structural Nested Models (Robins, 1998)

Structural nested models (SNM) are causal models which have been designed to estimate a time-dependent treatment effect for a survival endpoint when time-dependent confounding is present (Robins, 1998). The underlying theory is that, by conditioning on covariates and previous treatment history, the treatment becomes randomly assigned. As with the RPSFTM and IPE, an AFT structure is used, alongside counterfactual survival times, and with exposure to treatment accelerating the time-to-event by the factor, $e^{-\psi}$. One advantage of SNMs is that time-varying covariates can be included. This method also requires the specification of when a patient becomes at risk of switching treatments. Two key assumptions of this model are that: the counterfactuals are independent of exposure to treatment; and that there are 'no unmeasured confounders' (the idea that all possible factors leading to treatment switching are included within the dataset, and can, therefore, be conditioned on). This last assumption may be a limitation of using SNMs in RCTs, as the datasets are often considerably smaller than observational studies, and thus it is difficult to assess the suitability of this assumption using the observed data. This can also be particularly problematic if the confounders change over time, as they would need to be recorded. As with the RPSFTM, g-estimation is typically used to predict the acceleration factor, by determining the value of ψ for which the counterfactual time and treatment exposure become independent.

1.1.3.4.2 Inverse Probability of Censoring Weighting (Robins, 2000)

In contrast to SNM, Inverse Probability of Censoring Weighting (IPCW) utilises a PH modelling approach, rather than an AFT model (Robins, 2000). It is useful in accounting for informative censoring, as uncensored observations are up-weighted based on similarity of their covariate values those of the censored patients. This weighting aims to remove selection bias, by up-weighting uncensored patients. In the context of treatment switching, patients who switch tream, are artificially censored at the time of the treatment switch. The weights are then included in a standard analysis such as the Cox model, with

treatment as the exposure, to obtain the estimate of the treatment effect, adjusted for crossover.

The IPCW also relies on the 'no unmeasured confounders assumption' in order to appropriately calculate the weights. However, this is generally untestable in the data that are available. In addition, given that this approach was originally designed for observational studies which are often considerably larger than RCTs, there may be problems when calculating the weights; this often occurs when there is an especially large proportion of patients switching treatments, or rare covariate values. In these cases, the weights often become very large and potentially unstable.

1.1.3.4.3 Two Stage Method (Latimer, 2012)

This approach (Latimer, 2012) is the most recent method proposed for treatment switching, and is reliant on the existence of a relationship between treatment switching and disease progression. As apparent by its name, it comprises of two stages; the first stage treats the data as a randomised trial, whilst the second stage, analyses the data as if it were an observational study.

It is imperative that a secondary baseline time, namely disease progression, exists. For the patients in the control group, the difference in survival time from this secondary baseline between switchers and non-switchers is modelled, using an AFT model. Once this difference has been obtained, the survival times for treatment switchers (post-secondary baseline) is adjusted accordingly. When added to the pre-secondary baseline time, this gives a revised survival time for treatment switchers. Using this adjusted dataset, a standard survival model can then be fitted to obtain an estimate for the treatment effect, having accounted for crossover.

The 'no unmeasured confounders' assumption must hold at the secondary baseline (usually at disease progression), and patients must switch soon after this baseline in order to avoid time-dependent confounding. For disease areas such as advanced or metastatic cancer, this is often the case; e.g. in practice patients will switch treatment soon after disease progression or not change at all. It also requires that data are available at the time of switching, and ideally post-switch, for the second stage of the method. This can be problematic in RCTs, when follow-up post-progression is more limited

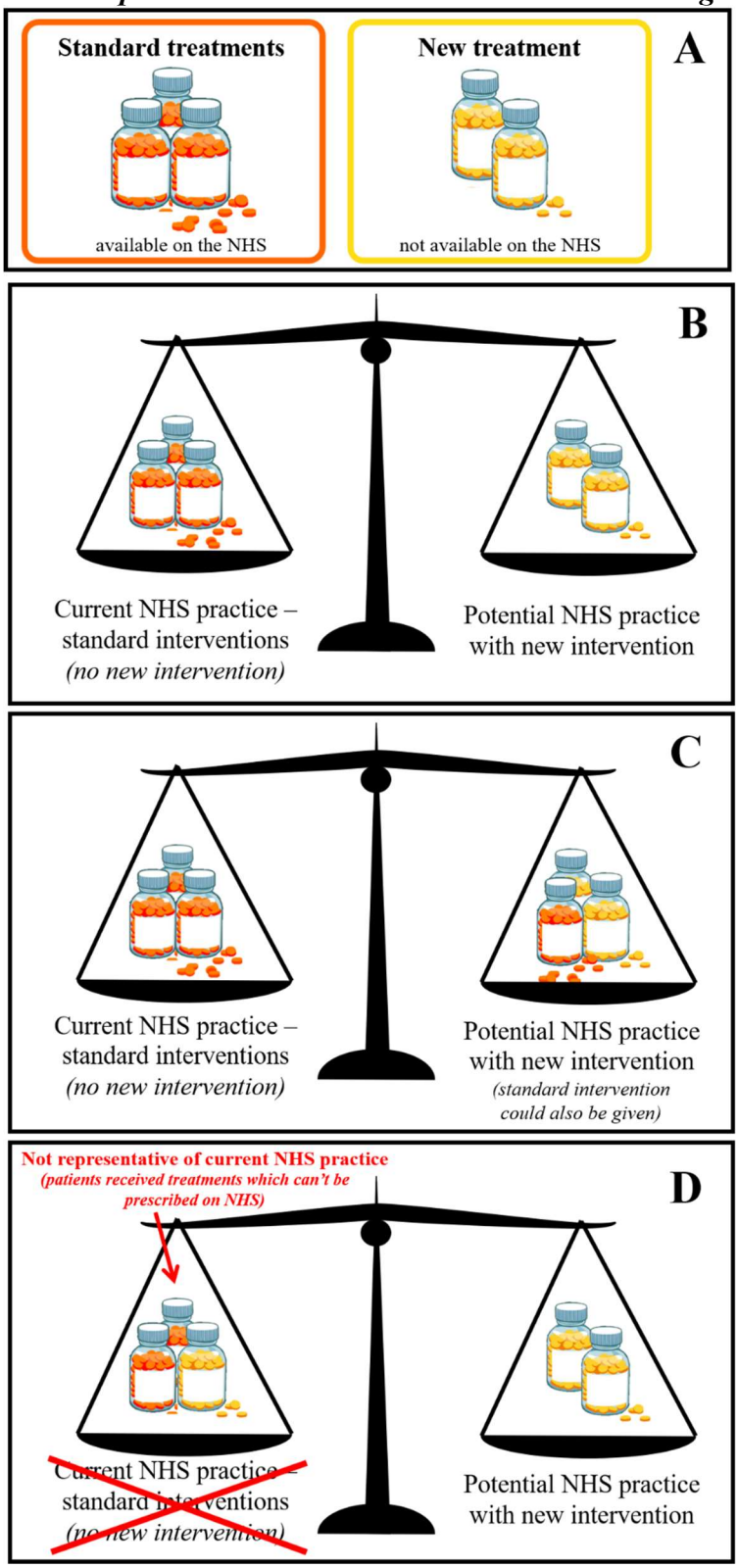
1.1.4 Impact of the methodology for treatment switching within Health Technology Assessment

1.1.4.1 *Effect of treatment switching on Health Technology Assessment submissions*

Treatment switching has considerably less impact for regulatory bodies, such as the Food and Drug Administration (FDA) or European Medicines Agency (EMA) than for reimbursement agencies, like NICE. This is because the regulatory bodies' main concern is that an intervention is safe, and hence they consider estimates of progression-free survival (PFS) are satisfactory evidence. Whereas, reimbursement agencies are principally interested in two related areas: the intervention's efficacy over a lifetime horizon (i.e. an average patient's lifetime) – estimated from overall survival (OS) – and its cost-effectiveness. Both of these are affected by treatment switching (Jonsson, 2014). As explained earlier, if patients switch to a superior treatment at disease progression, the true efficacy is underestimated. Including this underestimate within cost-effectiveness analysis (CEA) can lead to exceptionally high and very sensitive estimates for CEA. A clear example of this occurred in one Technology Appraisal (TA), TA169 (NICE, 2009a), submitted to NICE in Renal Cell Carcinoma (RCC) where the Incremental Cost-Effectiveness Ratio (ICER; a measure of the extra cost in pounds gained for every extra quality-adjusted life year (QALY)) ranged between £29,440 – within NICEs £30,000 per QALY willingness to pay threshold – to £104,715. This variation was largely caused by the way in which the OS data was used in the CEA model. The ERG group's estimate of £104,715 was obtained only using the ITT estimates for OS; whereas the manufacturer applied the same adjustments used to improve PFS model fit, to the OS data, which gave the lower estimate. Other methods (including PP) resulted in estimates around £71,000 per QALY (NICE, 2009a).

The immense uncertainty about the cost-effectiveness, due to treatment switching, has ramifications for decision-making. Decision-makers must either refuse to recommend the intervention or, more often, request further evidence; thereby delaying the decision until more evidence can be provided on which they can make a more certain decision.
Therefore, this impacts on the wider society as it means patients are either denied or required to wait longer for a new potentially cost-effective intervention. In particular, in oncology, these are patients with advanced or metastatic cancer, who have a poor prognosis and who would benefit from receiving these interventions as soon as practically possible.

Figure 1-1: Real-world problem with and without treatment switching



A) treatments compared in the real-world problem; B) the real-world decision problem of interest (without any treatment switching); C) the real-world decision problem of interest, remains unaffected despite treatment switching from the 'new intervention' group; D) real-world problem as it stands with treatment switching in the control group, this no longer represents the relevant decision problem.

Historically, there has been some doubt as to why an ITT approach is unacceptable for decision-making; primarily because it is the traditional approach employed to account for non-compliance or drop-outs, and believed to give pragmatic estimates. Although the true treatment effect is naturally attenuated, this attenuation is seen as representative of practice (as not all patients would strictly adhere to the treatment regimen set out in the protocol). However, the key issue is that the ITT analysis does not address the decision-problem in question; which is to directly compare current NHS practice (where the intervention being assessed cannot be prescribed to the general public), to the scenario where it could be prescribed. Using estimates for the standard treatment, where some patients have switched to the intervention being assessed violates this. Therefore, it is not an appropriate comparison, nor the basis for the reimbursement decision as the decision to not recommend is potentially based on the improvement the patients in the control group have received by switching to the experimental intervention. This is shown diagrammatically in Figure 1-1.

1.1.4.2 *Evaluation of methodology and current recommendations within the United Kingdom*

Given the variety of methods, and the differing levels of complexity, this issue was prioritised by NICE, and consequently a simulation study was undertaken to compare a range of approaches (Morden, 2009). This original simulation compared nine different methods (ITT; PP: excluding and censoring switchers; treatment as a time-varying covariate; adjusted Cox model; Causal proportional hazards estimator; RPSFTM: log-rank, Cox, Exponential and Weibull; IPE algorithm; and Parametric randomisation-based methods) over sixty-four scenarios. The findings indicated that the simple techniques performed extremely poorly. Whilst the ITT approach grossly underestimated the true treatment effect; PP methods and including treatment as a time-varying covariate proved exceptionally biased. Of the more complex methodology, which method proved the most accurate still remained unclear; although the IPE was tentatively suggested as perhaps the most reliable.

A second simulation study (Latimer, 2012) was conducted with the aim of expanding the previous investigation. This further research had several differences to that of Morden (2009). In particular, the study included a more complex data generation technique, and encompassed more methods for addressing treatment switching and scenarios. The

simulated data allowed for the patients' survival time to be time-dependant and related to patient's treatment and prognosis. This change in technique improved the realism of the simulated data, and hence rendered the conclusions more generalizable to current practice. A change in the data also permitted the investigation of more observational methods, such as SNM, IPCW and the two-stage method. A follow-up simulation study to this second one has since been performed to explore further scenarios and increase the evidence base (Latimer, 2017).

The conclusions formed by Latimer (2012) were broadly similar to the Morden study in that simple methods yield considerable bias, and no one method proves itself to be the best-performing approach over all scenarios. This simulation study demonstrated that there are a core of the more complex methods that work consistently well, providing that their assumptions or criteria are fulfilled. These methods consist of the RPSFTM (or their parametric equivalent the IPE), the IPCW and the two-stage method; all of which are very different in approach and data requirements. For example, to use the RPSFTM, the 'common treatment effect' assumption must hold; whilst for the IPCW, the 'no unmeasured confounders' assumption must be valid. These findings were reaffirmed by Latimer's second simulation study (Latimer, 2018c).

As discussed, the recommended approaches greatly vary in their methodology; however, they also differ considerably in the estimates they give. An illustration of this occurs with TA215 for Pazopanib as a treatment for RCC (NICE, 2011). The HR for the IPCW was 0.642 (0.266, 1.248), compared with 0.310 (0.073, 1.715) and 0.501 (0.136, 2.348) for the unweighted adjusted and weighted unadjusted RPSFTMs respectively.

The considerable differences, between the estimates, fuel the reservations concerning the application of the methods and their validity in a specific context (Latimer, 2016). NICE have, more recently, begun to advocate methods for addressing treatment switching (Van Engen, 2014), where necessary. A Technical Support Document (TSD) has been published (Latimer, 2014), to provide advice on addressing treatment switching, and the theoretical and practical application of the recommended methods. These recommendations directly follow from the former simulation studies. In addition, the TSD stresses choosing a method whose criteria agree with the data, testing all potentially appropriate recommended approaches, and the importance of providing valid justification

for the final choice of statistical model. Latimer (2015) describes how analysis with recommended methods may still be dismissed by NICE in the event that it is not deemed appropriate. Discussion with stakeholders has reinforced the need for appropriate and adequate justification for the choice of methodology, as a measure to improve transparency and acceptability to decision-makers. Consequently, this has been a key area of recent research for Latimer et al. (Bell, 2014, Bell, 2015, Watkins, 2016, Latimer, 2018a, Latimer, 2018b).

1.2 Objectives and structure of the thesis

1.2.1 Objectives

The key aims of this thesis are to:

1) Determine the impact that previous research has had on methods used to analyse data with treatment switching in NICE TAs;
2) Assess changes with regard to the use of secondary analysis, namely indirect comparisons (IC) or mixed treatment comparisons (MTC), and in the type of treatment switching occurring;
3) Evaluate the impact the inclusion of biased estimates might have on an IC;
4) Develop methodology to address treatment switching when only summary data are available;
5) Investigate the effectiveness of these methods, particularly for alternative secondary analysis such as proving PFS as a surrogate for OS.

1.2.2 Thesis structure

Chapter 1 aims to introduce the background to treatment switching, and its effect on HTA. It highlights the variety of methods that can be used and discusses previous research in this field.

Chapter 2 addresses the first two objectives (the impact of previous research on NICE TAs, and frequency of ICs and MTCs in NICE TAs) set out in section 1.2.1, by updating and extending previous reviews of NICE TAs of interventions for advanced or metastatic cancer. It primarily looks at the prevalence of treatment switching, and the methodology used to analysis the data, also stratifying into key time periods.

Chapter 3 sets out key background as to why including inappropriately analysed methods in secondary analysis might be particularly hazardous; and evaluates the impact of conducting ICs using adjusted and / or unadjusted estimates (Objective 3). It then describes previous approaches that have been taken to adjust for treatment switching, when only summary data are available, and the limitations of these.

A substantial amount of methodological development was required to address these limitations, and is explained across Chapters 4 and 5. This was achieved by generating simulation techniques for reconstructing individual patient level data (IPLD). Initially, in Chapter 4, the approach was developed and evaluated to produce IPLD for one outcome. This was then extended considerably more (Chapter 5) so as to create paired data using an 'illness-death' modelling approach. To accomplish this aim, it was necessary to establish how the method could be implemented depending on the summary information available. This novel methodology can then be used to address Objective 4 (accounting for treatment switching appropriately when only summary data are available) of the project.

The penultimate chapter (Chapter 6) focuses on the last aim of the thesis, and considers the impact treatment switching has on proving surrogacy. Consequently, it uses a case-study in Non-small-cell lung cancer, to show the impact using revised ITT estimates, which account for treatment switching appropriately. It also highlights issues analysts can face due to differential reporting, and gives suggestions for how to overcome them, also detailing the final modifications and extensions that must be made to the approach.
The final chapter (Chapter 7) is divided into two sections. The first describes recommendations for features that should be consistently reported in time-to-event trials. This is stratified further depending on the trial characteristics, e.g. data with treatment switching, secondary analysis including studies with treatment switching, or just survival analysis in general. These recommendations draw heavily on material from Chapters 2 and 6. This initial section also describes the process of identifying the important information about treatment switching from publication, how to determine whether it is possible, and if so, how to adjust for treatment switching based on the information available. The second part of the chapter summarises the research presented throughout the thesis. It discusses the uses, strengths and limitations of the methodology, their

implications on current practice and places it in context. Finally, suggestions for future work are described.

Chapter 2: The effect of treatment switching in practice and the reporting of studies with treatment switching

2.1 Chapter overview

This chapter starts by examining the prevalence of treatment switching in NICE TAs, and describes how the methodology has changed, particularly since the publication of former research (Morden, 2009, Latimer, 2012) by reviewing cancer TAs submitted to NICE. It updates previous reviews, and includes all relevant TAs published before January 2017. It also considers how prevalent ICs and MTCs are, and how these characteristics (treatment switching and comparisons) affect the recommendation. The second part of this chapter considers the evidence that is available with regard to treatment switching studies, by further examining a subset of the TAs identified in the first part. It concentrates on understanding what information is presented at different levels of evidence (e.g. TA summary report, Manufacturers Submission, trial publication etc.). In particular, the focus was on determining what information was routinely reported relating to the reasons for and proportion of treatment switching.

2.2 Changes in practice with regard to methodology for studies with treatment switching

2.2.1 Review of National Institute of Health and Care Excellence Technology Appraisals

2.2.1.1 *Purpose of the review*

2.2.1.1.1 Background

As part of his research, Latimer (2012) reviewed TAs submitted to NICE of non-screening or non-surgical interventions for advanced or metastatic cancer or cancer of all severities, which had been submitted to NICE between 2000 and 2010. This consisted of forty-five appraisals (thirteen not having fulfilled the inclusion criteria). Of these, twenty-five were identified as containing treatment switching. Further investigation found that the general approach was to: (1) ignore treatment switching completely; (2) to identify it as an issue but not adjust; (3) use an ITT approach; (4) employ PP methods (either excluding switchers completely from the analysis or censoring switchers at the time of switch); or (5) on rare occasions use Ad Hoc methods such as case-mix, using reference

data for the control arm; or exploiting PFS. Only one TA published within this time-frame applied methods that were later recommended which included a RPSFTM and the IPCW (the two-stage method having been proposed subsequent to this time). Following the completion of this review, two further appraisals were published which had included analysis using an RPSFTM. This slight change followed the publication of the Morden simulation study (Morden, 2011), and could imply an increase in awareness of the methods for treatment switching, and the relevant merits of each.

Another review was undertaken to update the Latimer review up to May 2013 (Boucher, 2013b). Of the thirty-six TAs published after 2010 which met the inclusion criteria, there were ten appraisals with treatment switching. As previously mentioned two had used a RPSFTM in addition to simple methods of adjustment, and one further appraisal published in 2011 had explored using both an RPSFTM and IPCW. However, both Latimer's conclusions (Latimer, 2012) and TSD recommendations (Latimer, 2014) insist that adequate justification is given for which method is used. This is vital in view of the widely differing results the methods may yield. Nevertheless, the justification for the final choice of method relies on the RSPFTM having been viewed as suitable in an earlier appraisal rather than the methodological requirements having been fulfilled. If, in this case the common treatment effect did not hold, then even though a currently recommended method had been employed, the results from this analysis could still remain biased.

In addition, this second review (Boucher, 2013b) had a further focus; to identify the effect of secondary analysis in studies where the pivotal evidence (that is the trial or trials used to provide evidence of the intervention's clinical and cost-effectiveness, and on which decisions are primarily made) contained treatment switching. The motivation for this was that including estimates within secondary analysis can give misleading results, if treatment switching has not been appropriately accounted for. Given the importance of adjusting estimates for treatment switching, and the lack of using appropriate methodology, this situation was highly likely to have occurred. For the purpose of this review, the 'secondary analysis' was restricted to the use of IC. Since an IC compares two or more interventions, if at least one trial for each intervention is biased due to inappropriately adjusted treatment switching, the final estimates for the IC will be

inaccurate. However, this inaccuracy will become even more pronounced if one of the trials has been adjusted, whilst the other studies have not.

This review found that, over thirteen and half years, thirteen appraisals where crossover was an issue in the pivotal evidence had also conducted an IC. Of the eleven appraisals for which information about the additional studies included in the IC could be found, all contained at least one trial where treatment switching had been permitted. Of these trials, for which only the reported analysis was available for inclusion in the IC, no recommended method had been used. Where methods for treatment switching had been employed, these were either ITT or PP approaches. In at least one circumstance, the method that was used was dictated by the data collection (Steineck, 1990), as the paper stated that no information was recorded on patients once they had switched. Due to the lack of adjustment in HTA submissions, in general, the pivotal evidence had not been adjusted appropriately either.

As already stated only TAs where treatment switching occurred within the pivotal evidence were investigated further for ICs, and consequently for the inclusion of summary data with treatment switching. The motivation for scrutinising TAs with treatment switching, was that if the pivotal evidence contained crossover, it was highly likely that some of the other trials incorporated through the IC would also have allowed treatment switching. However, this might well underestimate the actual number of TAs with ICs including summary data with unadjusted crossover.

2.2.1.1.2 Aims of this review

This review aims to assess:

1) How common it is for pivotal evidence in NICE appraisals in interventions for advanced or metastatic cancer to contain treatment switching, and whether this has changed since 2010.
2) Which methods are tried (i.e. tested on the data, but not necessarily chosen as the final method) in practice on data with treatment switching, particularly for appraisals published since 2010.
3) The impact treatment switching has on the recommendation and if this varies by adjustment method.

4) How common it is for IC / MTC to be conducted as part of a NICE appraisal in interventions for advanced or metastatic cancer.
5) Whether conducting an IC / MTC when the pivotal evidence contains treatment switching potentially impacts on the recommendation.

2.2.1.2 *Inclusion / exclusion criteria*

For the review, the studies included will be identified using the following inclusion / exclusion criteria:

Inclusion

- TAs published by NICE between January 2000 and December 2016
- Complete appraisals
- Appraisals assessing interventions for treating cancer patients.

Exclusion

- TAs labelled as 'terminated'.
- Appraisals in disease areas other than cancer
- Appraisals for surgical or screening interventions.
- Appraisals for treatments exclusively aimed at adjuvant or early cancer patients.

2.2.1.3 *Review findings*

In addition to the 81 TAs identified in previous similar reviews (covering TAs published between January 2000 and May 2013), a further 53 TAs were found; bringing the total to 134 eligible TAs. TAs are reviewed periodically, and the guidance replaced. Therefore, for several TAs, particularly between 2000 and 2003, it was not possible to obtain the original reports and hence relevant information on these (unless previously described in the Latimer review (Latimer, 2012)) These replaced or subsequently withdrawn TAs are given in Table 2-1.

There are a considerably high number of TAs, 30, published in 2016 (as can be seen in Figure 2-1). However, several of these were re-evaluating and reviewing interventions from previously published TAs.

Table 2-1: TAs which have been replaced or withdrawn by year

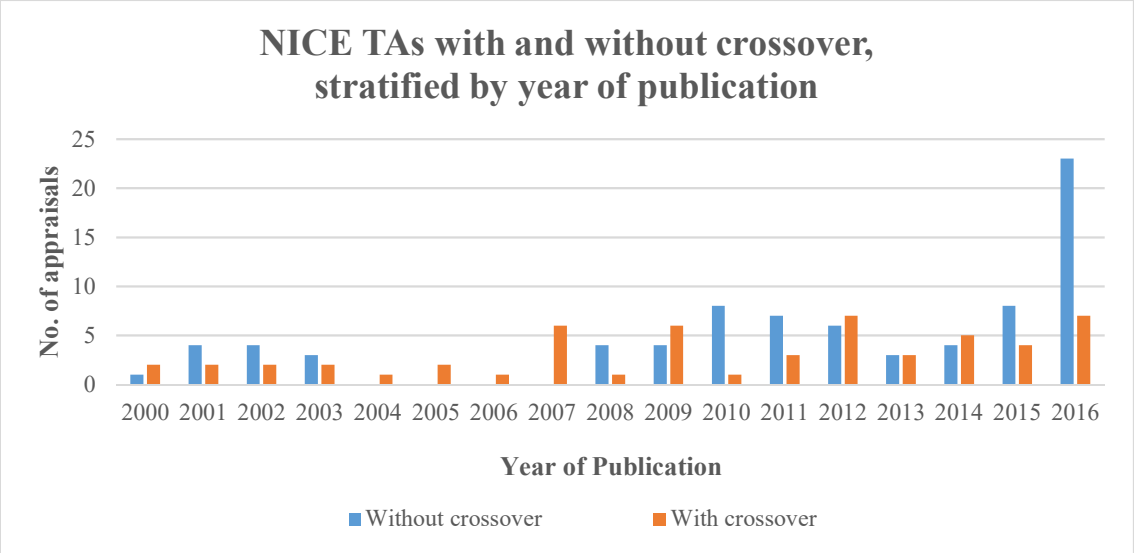
Year of Publication	TAs which have since been replaced or withdrawn
2000	TA17
2001	TA26
2002	TA37 TA45 TA50 TA54
2003	TA62 TA65
2008	TA147
2012	TA241
2013	TA296
2016	TA376

A list of the TAs included in the review can be found in Appendix A.

2.2.1.3.1 Prevalence of treatment switching in pivotal evidence

Of the 134 TAs, 55 (41%) used trials with treatment switching for the pivotal evidence. Interestingly, the proportion of TAs with treatment switching has decreased slightly over each of the reviews (55.6% for TAs between 2000 and 2010 (Latimer, 2012), 36.1% between January 2010 and May 2013 (Boucher, 2013b), and 31.5% for May 2013 to end of December 2016).

Figure 2-1: NICE TAs with and without crossover, stratified by year



Whilst the overall proportion of treatment switching may have decreased slightly, Figure 2-1 shows clearly that treatment switching is a consistently occurring phenomenon with at least one TA each year containing pivotal evidence with crossover.

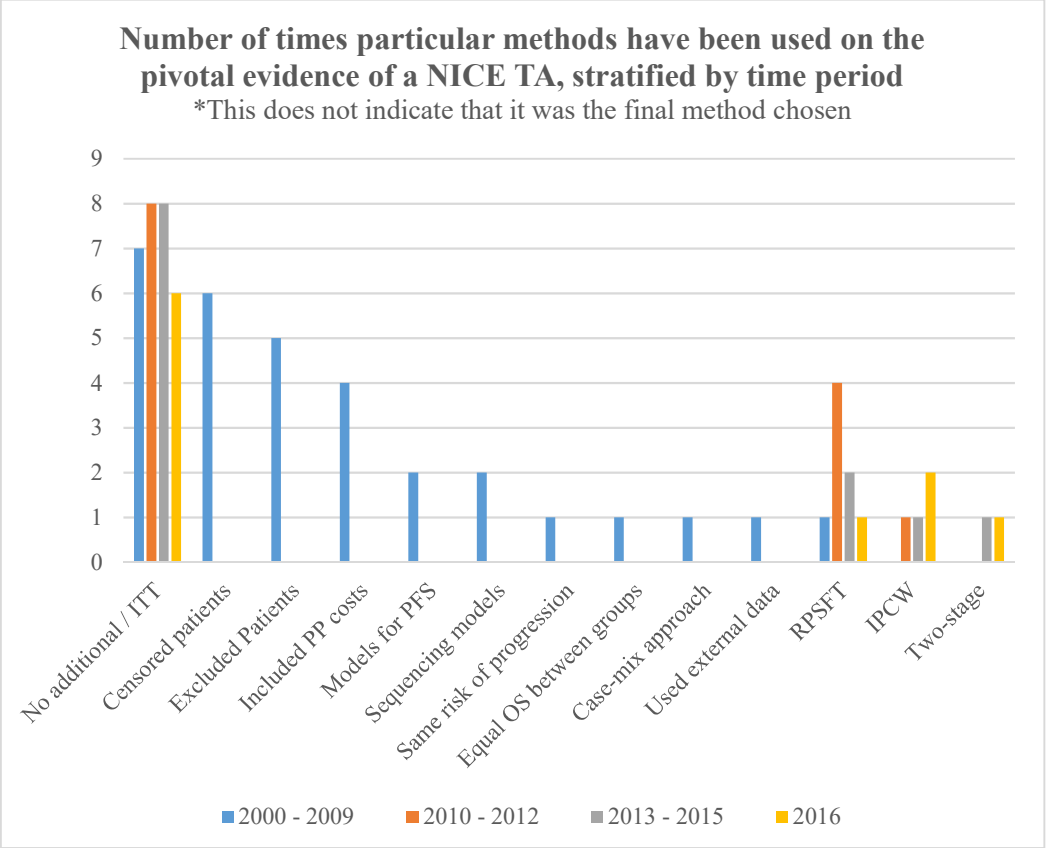
2.2.1.3.2 Methodology used on treatment switching data in pivotal evidence

To understand the impact the previous research has had on the methodology and for ease of reporting, the timescale has been divided into four periods: 2000 – 2009; 2010 – 2012; 2013 – 2015; and 2016. The reasons for these are presented in Table 2-2.

Table 2-2: Choice of cut-points for the timescale

Time period	Motivation for cut-point
2000 – 2009	Time period prior to research on appropriate statistical methods for adjusting for treatment switching.
2010 – 2012	Initial research on treatment switching methodology; discouraging the use of simple approaches for treatment switching.
2013 – 2015	Subsequent research on treatment switching methodology; establishing recommended methods, NICE and TSD guidance, and publicising them.
2016	Guidance now established and publicised.

Figure 2-2: Treatment switching methods in NICE TAs



In the first time period (2000 – 2009), a number of different methods were employed, though rarely any of those currently recommended. Since 2009, the methods fall into two categories: (1) the use of an ITT approach or no additional adjustment for treatment switching (not advocated to account for crossover); or (2) one of the recommended methods (RPSFTM, IPCW or two-stage). For the recommended methods, there is some

suggestion of a trend over time. Initially, between 2010–2012, the RPSFTM was the most popular of the three recommended methods (three examples compared to, one IPCW and no two-stage). However, more recently, applications have become more balanced across all three methods. This also highlights how manufacturers are testing different methods, as well potentially reflecting a deeper level of understanding about treatment switching issues. In the 2010–2012 time period, the choice of the RPSFTM was often justified by saying that this method had been deemed appropriate by NICE in other TAs (e.g. TA215 (NICE, 2011)), rather than checking the different model assumptions and data requirements, where possible. Nevertheless, there is some evidence that the choice of model is now becoming a more informed decision, based on relevant assumptions and data requirements.

2.2.1.3.3 Use of secondary analyses in pivotal evidence

The secondary analyses that often appear in TAs are ICs or MTC / network meta-analyses (NMAs). Figure 2-3 shows how their usage has increased over time since 2006 (when they first appeared), to the extent that in the last two years, there were more TAs which reported comparisons than those which did not. Figure 2-4 demonstrates the breakdown, per year, of comparisons into ICs and NMA. Until 2011 and between 2013 and 2014, all comparisons were ICs; in 2011 and 2012, the comparisons were evenly split between ICs and NMAs. The years 2015 and 2016 show the greatest number of NMAs, and also a high number of ICs.

Figure 2-3: Comparisons (indirect or mixed treatment) in NICE TAs

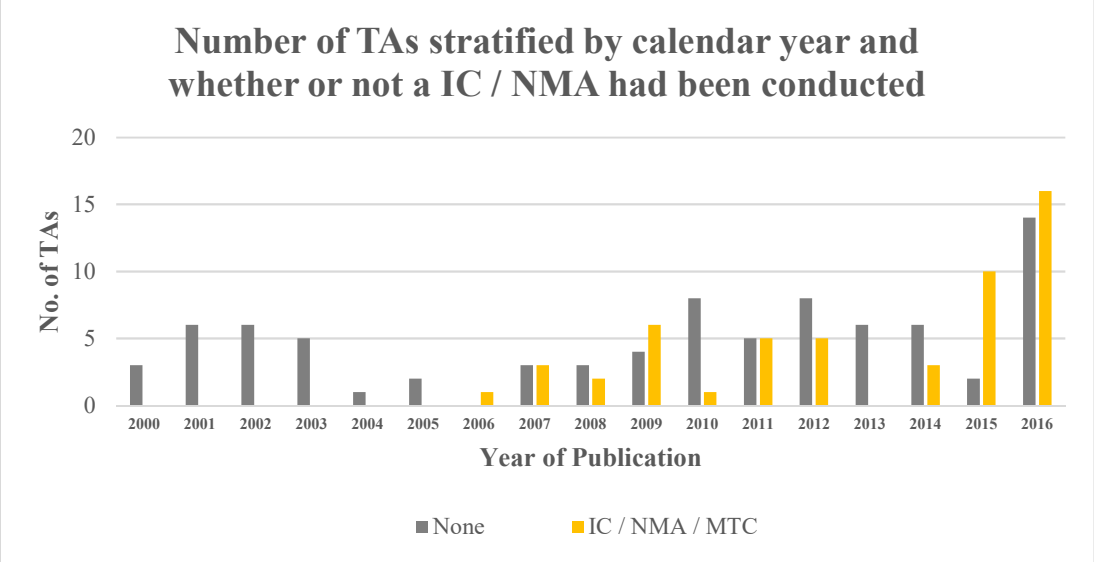
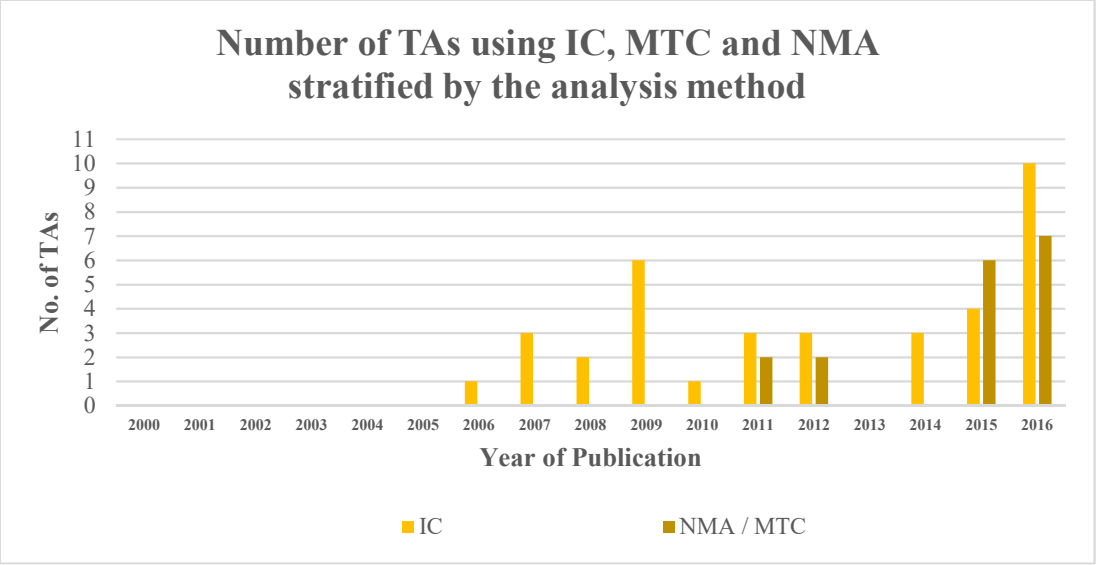


Figure 2-4: Type of comparison used in NICE TAs



2.2.1.3.4 *Notable example (TA417)*

TA417 is a particularly informative example (NICE, 2016). This TA evaluated the use of Nivolumab for previously treated advanced RCC patients. Although the pivotal evidence, the CheckMate 025 trial (Cella, 2016), did not permit treatment switching, the NMA the manufacturer conducted contained several trials which had (e.g. TARGET and RECORD-1). In this TA, the manufacturer wished to also present a crossover-adjusted NMA. For some trials, estimates accounting for treatment switching had been reported, e.g. RECORD-1 which presented an RPSFTM analysis. However, no such similar appropriate analysis existed for the TARGET trial (Escudier, 2009b), and thus the manufacturer settled on using the immature OS data as this was known to not be affected by treatment switching. Nevertheless, the use of this data has been criticised, because of its immaturity.

A vital point to this example is the manufacturer's awareness of the issues of treatment switching, and potential dangers in including unadjusted results in the analysis. It is important that they have tried to account for treatment switching, but evident that there are limitations to performing 'crossover-adjusted' comparisons (e.g. ICs, NMAs) when summary data estimates have not been reported using appropriate recommended methods.

2.2.1.3.5 Recommendations by NICE

For the purposes of the previous review (Boucher, 2013b), the type of recommendation was classified into two categories: (1) positive; or (2) negative. The definitions were as follows:

- Positive recommendation: any recommendation that the intervention be reimbursed and thus used in the NHS (regardless of any additional access schemes or restrictions to the patient population, e.g. to those with a specific biomarker)
- Negative recommendation: the intervention was not recommended for reimbursement

Initially, it was anticipated that the same definition would be employed again. However, on reflection it was decided that the above definition could lose vital information for the purpose of this review. That is, that, potentially, positive recommendations for TAs with treatment switching in the pivotal evidence may be subject to conditions more often than those without. Therefore, the definition for positive recommendation was revised and the new category of partial recommendation introduced. These were defined as follows:

- Positive recommendation: for this category the intervention must have been recommended without any conditions as a first choice or option;
- Partial recommendation: the intervention has been recommended for reimbursement subject to conditions, e.g. for a specific subgroup only, with a patient access scheme, with a discount agreement, etc.

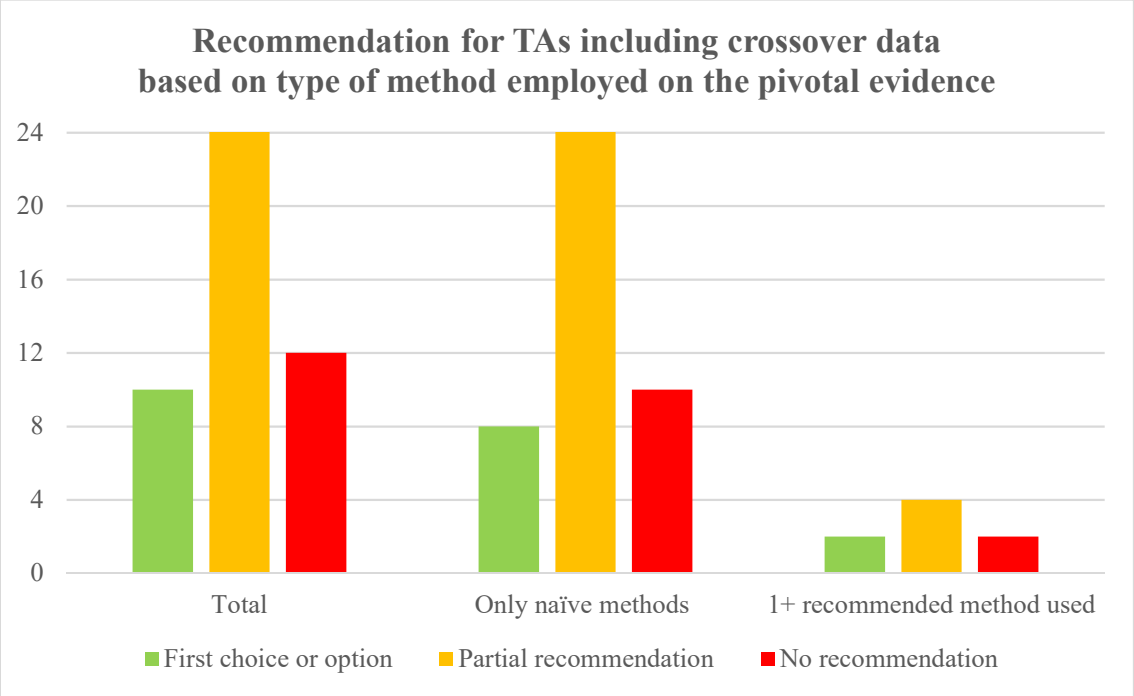
The definition for negative recommendation remained unchanged. Since this decision was made at a later date, for many of the TAs given in Table 2-1, the information needed to distinguish between positive and partial recommendations was unavailable. These appraisals were consequently excluded from the analysis. It can be noted that for two of the TAs (in Table 2-1), some basic information had previously been recorded: TA241 had not been recommended and TA296 had been recommended (but no information about whether this recommendation was subject to conditions was available).

2.2.1.3.6 For TAs with treatment switching in the pivotal evidence

This section principally concerns itself with how the method might influence the recommendation. Figure 2-5 shows the recommendations stratified by the method type. Partial recommendations are the most popular type of recommendation. There are some

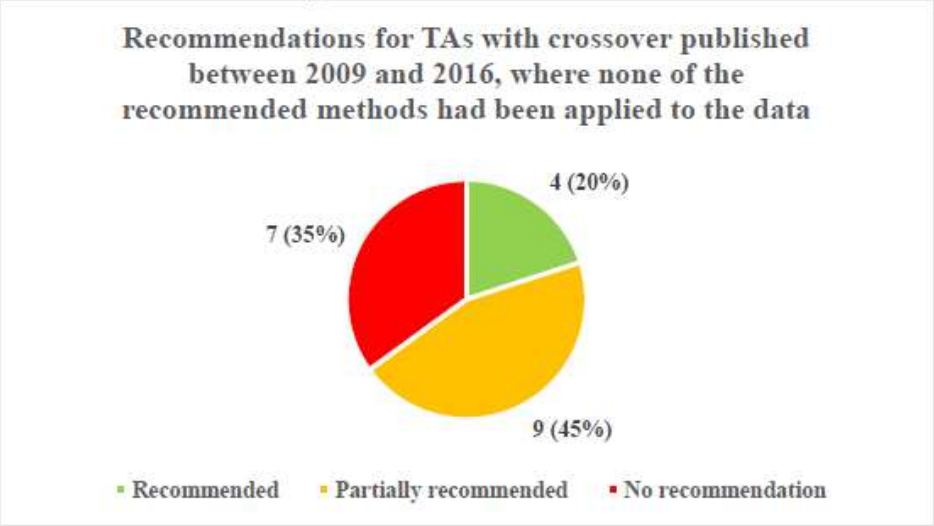
slight differences in terms of receiving a 'positive recommendation' (first choice or option – without conditions) and 'no recommendation' depending on the method used, with slightly more TAs having 'no recommendation' than a positive result when using naïve approaches, whilst there are the same number for those TAs using more appropriate methods. It should be noted though, that very few appraisals have used appropriate methods, 10 in total.

Figure 2-5: Recommendations stratified by type of crossover method



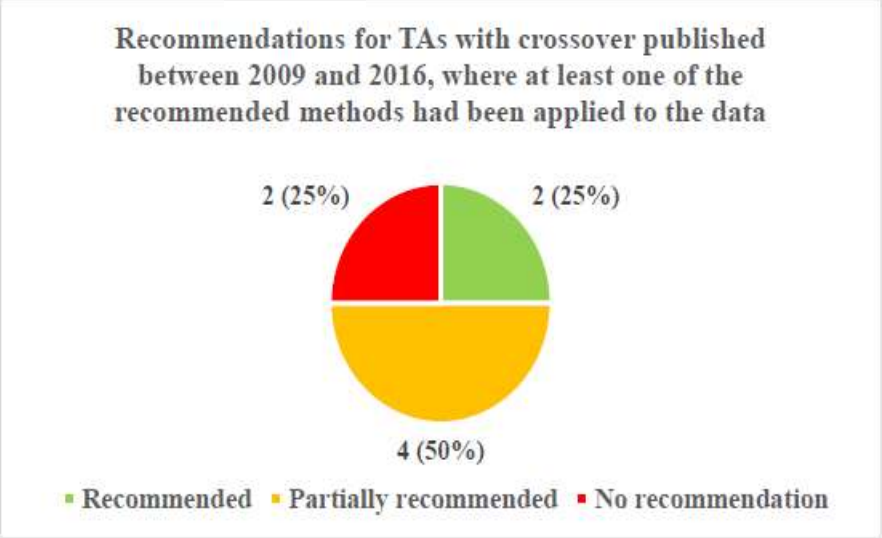
To assess the impact of the method on the recommendation further, the timescale has been restricted to between 2009 and 2016. The motivation for this was that first research into appropriate methods for treatment switching was published in 2009, thus establishing that naïve approaches should not be employed. This restriction does, however, result in very small numbers, but it is interesting to see from Figure 2-6 and Figure 2-7 that there could be some suggestion that 35% of the TAs using naïve methods to analyse their pivotal evidence do not get recommended, compared to 25% which use more advanced methods.

Figure 2-6: Recommendations for TAs with crossover – no recommended methods



The number (%) for each recommendation can be found adjacent to the corresponding section of the pie chart.

Figure 2-7: Recommendations for TAs with crossover - recommended methods

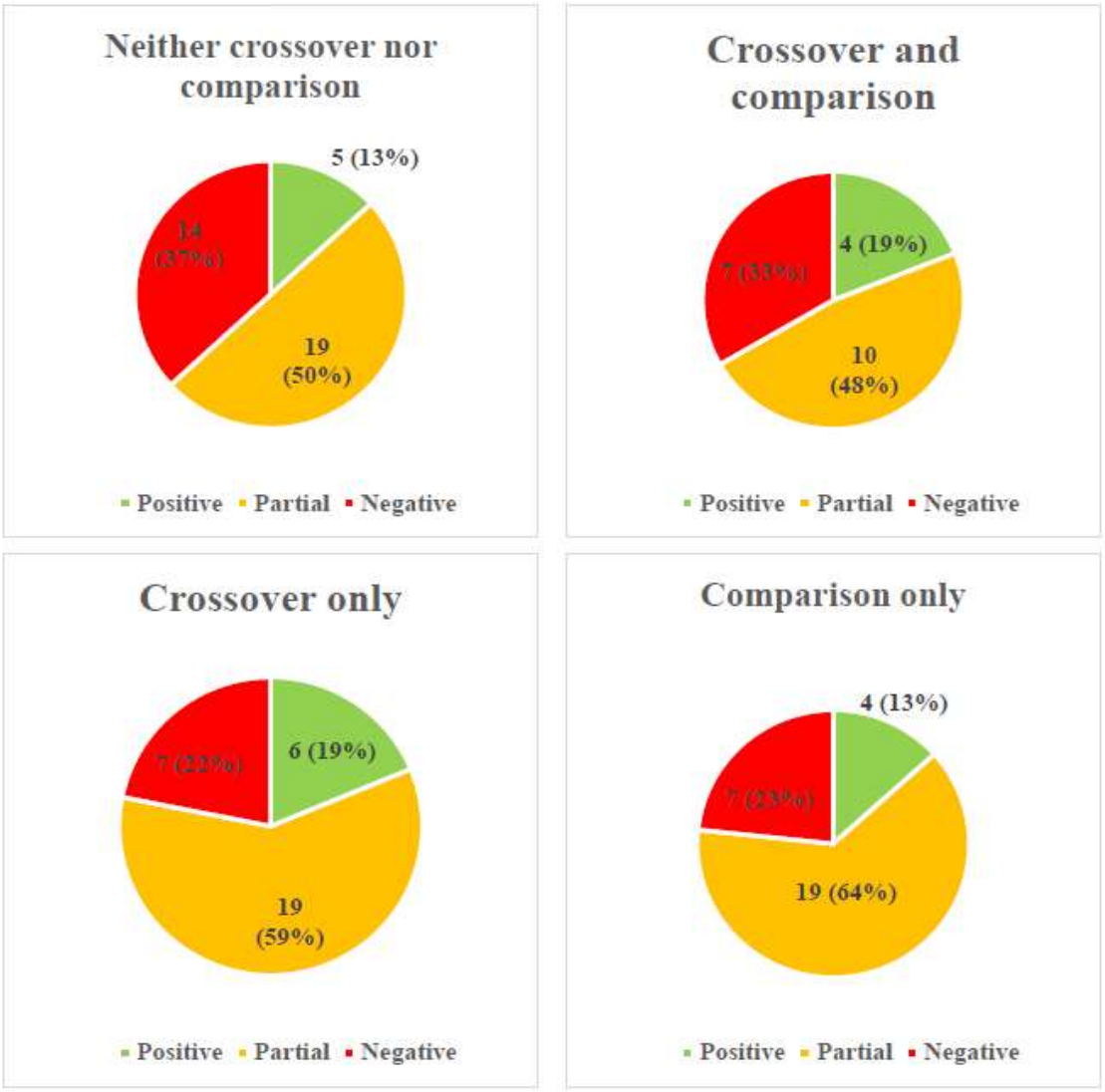


The number (%) for each recommendation can be found adjacent to the corresponding section of the pie chart.

2.2.1.3.7 Stratified by characteristics

From Figure 2-8, which shows the breakdown of recommendation based on the characteristics (e.g. neither crossover and comparison, both crossover and comparison, only crossover, only comparison), the groups with either both crossover and comparison or neither crossover nor comparison, seem to have a considerably larger proportion of negative recommendations than where only one occurred (approximately 35% – 40%

Figure 2-8: Recommendations based on TA characteristics

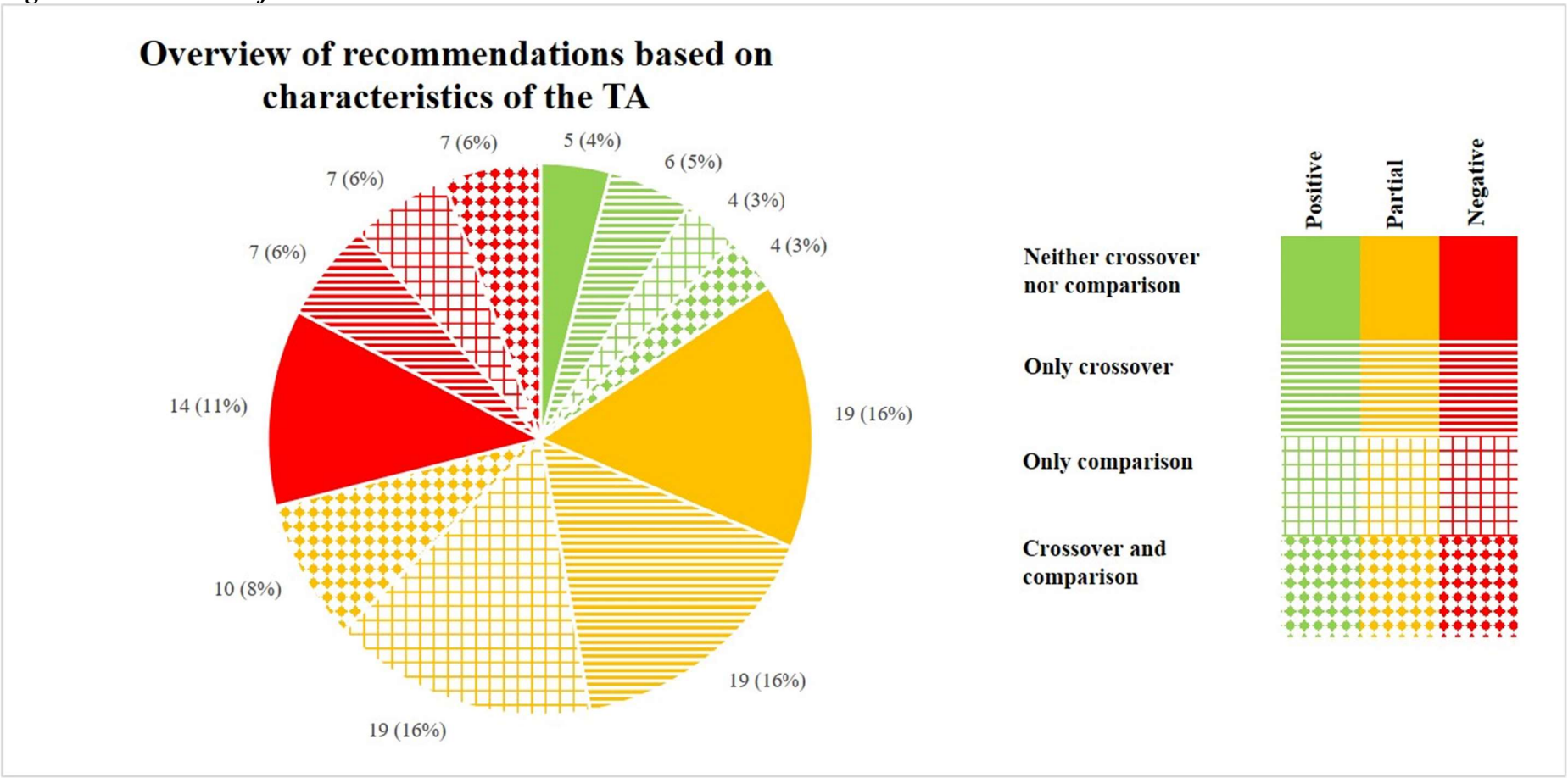


The number (%) for each recommendation can be found on/adjacent to the corresponding section of the pie chart.

compared to 20% – 25%). There are similar proportions of positive recommendations across all groups.

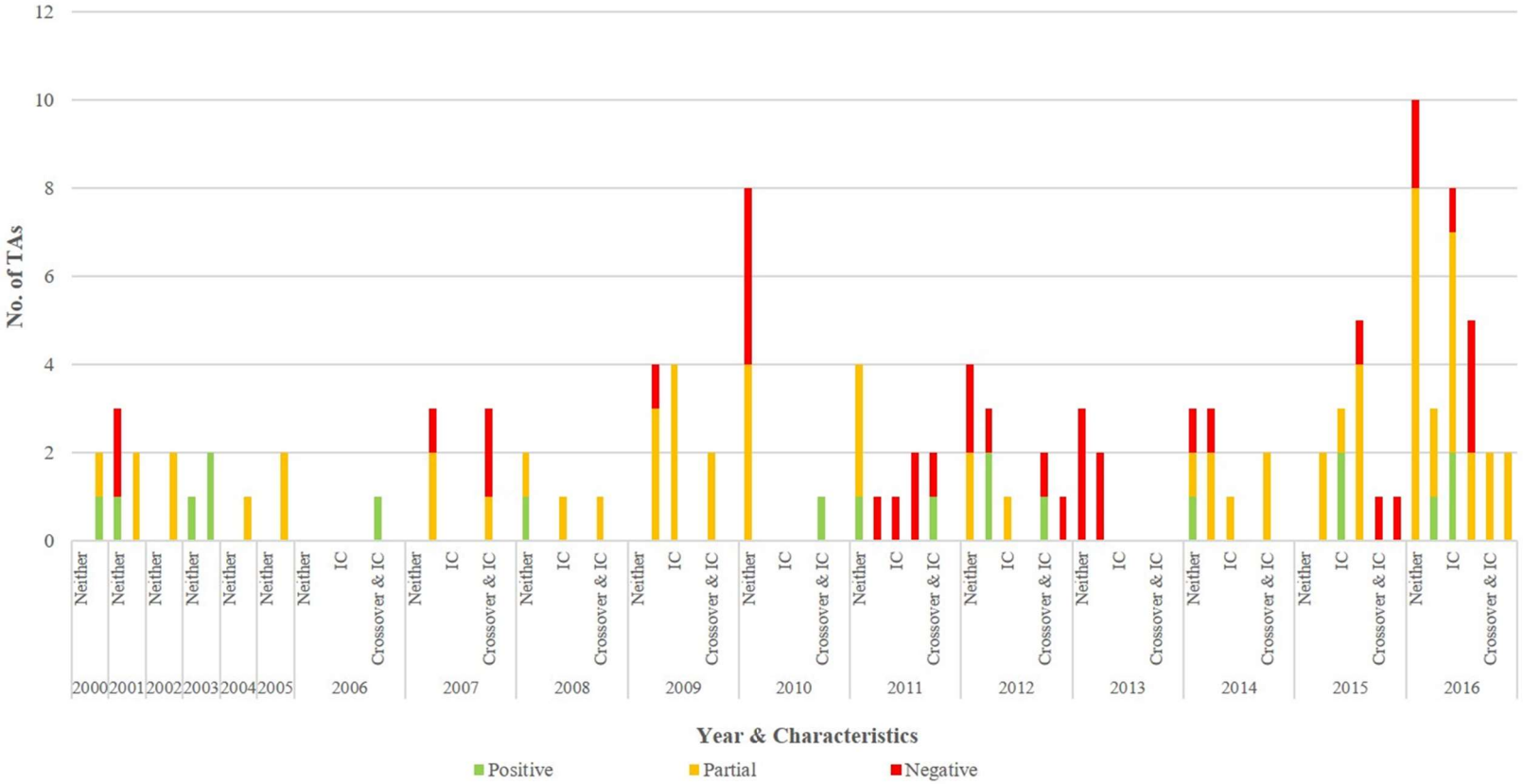
From Figure 2-9, which shows a breakdown of TAs by recommendation and characteristics, it is clear that the most common type of recommendation is partial recommendation (e.g. positive recommendation, usually with a patient access scheme and / or discount agreement). In addition, it can be seen that of the four characteristic groups, TAs without crossover and without any comparison having been conducted are the most frequent. Looking at the recommendations by characteristics and year

Figure 2-9: Overview of recommendations



The number (%) for each recommendation can be found on/adjacent to the corresponding section of the pie chart.

Figure 2-10: Overview of recommendations, by year of publication



(Figure 2-10) did not appear to show any clear trends. Since a number of TAs between 2000 and 2003, without treatment switching, had been withdrawn and replaced, these were therefore excluded from the analysis since no recommendation information could be obtained on them. To investigate whether this did potentially lead to bias, the analyses used for Figure 2-8 and Figure 2-9 were rerun restricting the time period to 2004 – 2016. These figures can be found in the appendices, but showed little difference to those presented here. Stratification based on whether the treatment switching was adjusted using recommended methods, and whether an IC or MTC was conducted did not show any significant findings and the results can be found in Appendix B.

2.2.1.4 *Conclusions*

Treatment switching continues to be an issue in NICE TAs, and whilst the prevalence has declined slightly (40.1%) since the last review (55.6%), examples have arisen each year. A particular strength of this work is how it has highlighted a considerable change in the methodology used within NICE TAs, particularly over the last six years, and how the issues and approaches to adjust for treatment switching have been better understood and accepted. Rather than the wide variety of methods used originally to account for treatment switching, the approaches now fall into two categories: those where appropriate methods have been used, and those where treatment switching has effectively been ignored. There is also some evidence that guidance provided by NICE and in the TSD has been effective, in that more of the recommended methods are being tested on the data, and more recently that appropriate choices and justifications are now being given.

In terms of the recommendation, it is possibly still too early to ascertain what effect the methods for treatment switching are having on the recommendation, since there are only ten appraisals which use recommended methods, and not all of these may have chosen appropriate models. One aspect that this review does not take account of is the time of the appraisal process. It could be that there are differences in the length of time taken to provide a recommendation for different methods, e.g. if the NICE panel, perhaps, ask for more information for TAs where naïve methods have been used, which may be less necessary if more appropriate analyses have already been submitted.

It is clear to see that ICs and NMA are becoming more frequently reported in TAs, and this leaves scope for potential problems with including treatment switching data in these,

depending on whether or not it has been appropriately analysed. Where it has not, it is debatable what approach should be taken. This uncertainty in approach is due, in part, to manufacturers rarely having access to the IPLD for a competitor's product. As a result, they are often unsure of the action to take, if these reported estimates do not adequately account for treatment switching (available evidence is explored in section 2.3). In addition, it is not clear what effect including a comparison (whether or not treatment switching occurred in the pivotal evidence) has on the recommendation, and if the inclusion of this provides greater support or not.

2.2.1.4.1 Views on accounting for treatment switching and the impact this has on the uptake of appropriate methodology

This review provides some evidence in highlighting the effect of previous research into treatment switching, the establishment of guidelines and the publicity of this former work, by examining how submissions have changed. However, underpinning these changes are the perspectives of stakeholders (notably analysts in the pharmaceutical industry, decision makers). This section seeks to put the review evidence into some of the context surrounding it. Investigations have continuously highlighted the difficulty in understanding and acceptability of adjustment methods for stakeholders (Maervoat, 2014, Henshall, 2016, Latimer, 2015, Latimer, 2016, Zhang, 2016). The European Annual Conference of the International Society for Pharmacoeconomic and Outcomes Research (ISPOR) has provided a key platform for the discussion of and education on treatment switching matters. In particular, during the sixteenth and seventeenth (2013 and 2014) there were several workshops (Maervoat, 2014, Van Engen, 2014) and talks covering treatment switching related topics, and these and the questions which followed highlighted the variation in expertise when applying or interpreting results from the appropriate methods. Moreover, the discussion emphasised that many stakeholders did not feel confident enough to apply the methods, or did not understand how the approaches worked.

One common theme was that, whilst different HTA bodies appreciate the issues caused by treatment switching, they take different stances on the methodology considered appropriate. In particular, the German HTA body, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), whilst now accepting treatment switching is an issue, seemed fervently against any of the adjustment methods; and based

on a workshop (IQWiG, 2014) specifically stated that they would not use adjusted estimates in the decision-making process. The German position seems a consistent one with Zhang (2016) highlighting that a case-study of submissions to the Gemeinsamer Bundesausschuss suggested that they did not accept the use of treatment switching approaches. This potentially gives conflicting advice, creating more uncertainty, and contributing to the initial reluctance in accepting results or indeed using this methodology. This might also provide some reasoning as to why there are still appraisals which choose to ignore treatment switching, rather than address it; given that there is no real uniformity across HTA bodies compelling them to use these complex methods. In addition, little difference in recommendation in the United Kingdom is seen depending on whether adjustment has been made.

2.3 Routinely reported and available information for treatment switching trials

To understand the information that is typically reported for trials with treatment switching, it was decided to examine the evidence sources further. Given time constraints, the TAs were restricted to those appraising a single technology where treatment switching had occurred and a comparison (either IC or MTC) had been performed. This particular subset was chosen because it gave the most scope for showcasing the variety of evidence for different levels of information e.g. information on pivotal evidence and on supporting studies only included for the comparison. The subset comprised of 15 appraisals (TA34, TA91, TA101, TA116, TA124, TA162, TA171, TA214, TA215, TA258, TA319, TA321, TA338, TA377, TA422), of which 3 were subsequently excluded; 2 of these (TA91 and TA338) because the guidance has since been replaced, meaning that the original documents were no longer available; and one (TA422) since it differed in evidence base and motivation for TA. TA422 was predominately an update from a previous submission, and was being reviewed in preparation for the end of the Cancer Drugs Fund. In addition to this, no information was available in the TA summary report relating to evidence appraised, due to a change in the format of the TA report structure. Instead the reader was referred to the committee papers.

This investigation started by reviewing all of the TA documents, examining the summary of ‘evidence’, for details including:

- Trial names and sample sizes
- Outcomes reported
- Statistics presented
- Details about data cut-off and follow-up lengths
- Information about treatment switching
- Information about the comparison (either IC or MTC / NMA) and any studies included in it

To showcase a variety of evidence sources, once the TA summary had been examined, for some of the included TAs, a copy of the ERG report and / or manufacturers submission was obtained, as were further publications that related to some or all of the included studies. It should be noted that the review focussed on the key trials used in the pivotal evidence and / or the comparison. Additional supporting trials, and in particular, observational studies were not investigated further (e.g. to publication level).

2.3.1 Findings

A total of 12 TAs, 3 manufacturers submissions, 3 ERG reports and 39 trial publications from peer-reviewed journals were examined. The findings are stratified by evidence source. A list of the manufacturers submissions, ERG reports and trial publications are available in Appendix C.

2.3.1.1 *TA evidence summary*

The evidence collected from the TAs fell into two broad categories: those features which were related to the TA (e.g. number of key RCTs, reporting of trial names etc.), reported in Table 2-3, and those that were specific to each trial included in the TA (and could not easily be summarised over the TA, e.g. whether treatment switching had occurred in that trial, whether the data cut-off date was stated, etc.), given in Table 2-4.

The majority of TAs rely on one key RCT, however, one third of these provided additional evidence as well. Many TAs reported trial names and the primary endpoint. However, considerably less information is given for the trials included in the IC, with only a quarter giving all the names of the trials. This becomes even more evident when the trial level data are examined. It is not surprising that little, in terms of the IC trials, has been

reported, as these are not directly relevant to the decision problem. However, treatment switching occurring in these trials could impact on any findings from the IC, and thus would benefit from being reported.

Table 2-3: TA level details obtained from the TA evidence summary

	No. of TAs (n = 12)	(%)
No. of key RCTs		
1	9	(75.0)
2	3	(25.0)
Additional evidence used	4	(33.3)
Trial names reported		
None	1	(8.3)
Some	1	(8.3)
All	6	(50.0)
Sample size of key trials reported	4	(33.3)
Primary endpoint for the trials	9	(75.0)
Reason for the IC		
To improve comparisons / scope of the decision problem	7	(58.3)
For reasons associated with CEA	1	(8.3)
As 'sensitivity analysis'	1	(8.3)
No. of additional trials		
Unclear	1	(8.3)
1	4	(33.3)
2	3	(25.0)
3	2	(16.7)
4	1	(8.3)
5	0	(0.0)
6	1	(8.3)
Trial names reported		
None	7	(58.3)
Some	2	(16.7)
All	3	(25.0)

The trial information typically reported for the pivotal evidence consisted of the size of the trial, its impact on TTP or PFS, and its effect on OS. The effect estimates were mostly in terms of median survival and HRs. Information about treatment switching is much more variable, although for a large proportion of trials, there was some clear indication that crossover did occur in the text. In addition, typically a reason was provided for why treatment switching occurred and the treatment arms affected were given.

These findings suggested that the manufacturers submissions contain more detail than the ERG report, as might be expected. Kaplan-Meier curves were occasionally reported for the pivotal evidence for the Manufacturers Submission, but more commonly, for OS, in the ERG report. Once again, there was relatively little information about the studies in the IC, although here the names of the trials were reported, and a little more detail is

provided about treatment switching. This particularly concerns whether treatment switching did occur, why it was permitted and which treatment arm it occurred in.

Table 2-4: Trial level details obtained from the TA evidence summary
The total number of trials and related percentages have not been reported as some TAs were unclear about the number of trials included and thus calculating these statistics would not necessarily be an accurate reflection of the truth.

		Number of trials	
		For the pivotal evidence	**For the IC**
Sample size			
Per group		4	2
Overall		6	2
Reported the primary endpoint		10	2
Primary endpoint was:			
TTP		3	0
PFS		4	0
OS		3	0
Joint PFS and OS		1	2
Duration of response as a secondary endpoint		4	0
Data cut-off date		1	1
Maximum or median length of follow-up		4	1
Statistics:			
Median with CI:	TTP	2	0
	PFS	6	0
	OS	10	0
Median, point estimate only:	TTP	2	1
	PFS	1	1
	OS	1	1
HR:	TTP	2*	0
	PFS	6	2
	OS	10	2*
Events:	PFS (per group)	1	0
	OS (per group)	2	0
	OS (overall)	1	0
Whether treatment switching occurred clearly reported		7	1
Reason for treatment switching specified		8	0
Treatment switching occurs in			
One treatment arm		5	1
Both treatment arms		4	0
Treatment switching proportions reported		3	1
Recommended methodology used		3	1
Justification for the choice of methodology		0	0

** indicates a pooled estimate*

2.3.1.2 *Manufacturers Submissions and ERG Reports*

Of the 14 TAs, 3 manufacturers submissions (TA116, TA214, TA215) and 3 ERG reports (TA101, TA214, TA215) were identified. The findings are given in Table 2-5.

Table 2-5: Evidence in the manufacturer's submission or ERG report

Number of trials based on information from the:	Manufacturers submission		ERG reports	
Trial detail	**Pivotal evidence**	**IC**	**Pivotal evidence**	**IC**
Sample size				
Per group	3	15	4	5
Overall	0	0	2	0
Primary endpoint was:				
TTP	0	1	0	0
PFS	2	1	1	1
OS	1	7	2	0
Joint PFS and OS	0	0	0	0
Joint TTP and OS	0	1	0	0
Other	0	1	0	1
Kaplan-Meier curves				
PFS K-M with 'risk table'	1	0	0	0
PFS K-M without 'risk table'	1	0	0	0
OS K-M with 'risk table'	1	0	0	0
OS K-M without 'risk table'	1	0	2	0
Events				
TTP or PFS events (per group)	1	0	1	2
OS events (per group)	2	0	3	3
OS events (overall)	0	0	0	1
Data cut-off date	1	2	2	0
Maximum or median length of follow-up	2	8	0	0
Statistics:				
Median TTP	1	5	1	7
TTP HR	1	3	0	5
Median PFS	2	9	3	6
PFS HR	2	7	2	8
Median OS	3	8	3	8
OS HR	3	4	4	8
Whether treatment switching occurred				
Possible or definite, according to the authors	3	6	3	1
Unclear in publication	0	3	0	0
Prohibited	0	0	0	2
Reason for treatment switching specified	2	4	2	1
Treatment switching occurs in				
One treatment arm	2	4	1	1
Both treatment arms	0	2	2	0
Treatment switching proportions reported	2	0	2	1
Recommended methodology used	1	0	1	0
Justification for the choice of methodology	1	0	0	0

2.3.1.3 *Individual publications*

Of the 12 included TAs, all the RCTs in the TA were identified and reviewed for 4 TAs (TA171, TA258, TA377) and some of those for TA101 (1 RCT of 7), and TA319 (2 of 4 RCTs). For TA214, the 4 studies used by the manufacturer were examined, as was the RIBBON-1 trial added by the ERG. 12 of the 13 trials documented in the Manufacturers Submission for TA215 were appraised. For several of the trials reviewed, the results were spread across two separate publications; the initial paper tended to report TTP / PFS

outcomes, occasionally in conjunction with immature OS data, whilst the follow-up paper then reported the mature OS data (sometimes referring to the previously published or partially updated TTP / PFS results). In total, 39 papers were examined; 9 of which were 'follow-up' papers.

2.3.1.3.1 General information

Table 2-6: Primary endpoints as reported in the publication

	Reported in original paper		Reported in follow-up paper		Total papers reporting	
	n = 30	**(%)**	**n = 9**	**(%)**	**n = 39**	**(%)**
TTP	3	(10.0)	0	(0.0)	3	(7.7)
PFS	10	(33.3)	2	(22.2)	12	(30.8)
OS	8	(26.7)	0	(0.0)	8	(20.5)
Joint PFS and OS	3	(10.0)	1	(11.1)	4	(10.3)
Other	3	(10.0)	0	(0.0)	3	(7.7)

Table 2-7: Commonly reported information about enrolment and follow-up

	Reported in original paper		Reported in follow-up paper		Total papers reporting	
	n = 30	**(%)**	**n = 9**	**(%)**	**n = 39**	**(%)**
Dates of Recruitment	21	(70.0)	4	(44.4)	25	(64.1)
Date of data cut-off for analysis	9	(30.0)	5	(55.6)	14	(35.9)
Median or maximum follow-up length	5	(16.7)	3	(33.3)	8	(20.5)

Table 2-8: Kaplan-Meier curves reported in the trial publication

		Reported in original paper		Reported in follow-up paper		Total papers reporting	
		n = 30	**(%)**	**n = 9**	**(%)**	**n = 39**	**(%)**
Without 'risk table'	TTP	2	(6.7)	1	(11.1)	3	(7.7)
	PFS	6*	(20.0)	1	(11.1)	7	(17.9)
	OS	7	(23.3)	1	(11.1)	8	(20.5)
With 'risk table'	TTP	3	(10.0)	0	(0.0)	3	(7.7)
	PFS	14	(46.7)	1	(11.1)	15	(38.5)
	OS	14	(46.7)	7	(77.8)	21	(53.8)

** One of Kaplan-Meier curves included in the 'PFS' results was actually described as 'event-free survival'*

Table 2-9: Number of events

Events		Reported in original paper		Reported in follow-up paper		Total papers reporting	
		n = 30	**(%)**	**n = 9**	**(%)**	**n = 39**	**(%)**
TTP or PFS	Per group	3	(10.0)	2	(22.2)	5	(12.8)
	Overall	1	(3.3)	0	(0.0)	1	(2.6)
OS	Per group	11	(36.7)	5	(55.6)	16	(41.0)
	Overall	5	(16.7)	0	(0.0)	5	(12.8)

Table 2-10: Effect estimates routinely used

		Reported in original paper		Reported in follow-up paper		Total papers reporting	
		n = 30	(%)	n = 9	(%)	n = 39	(%)
Median survival	TTP	6	(20.0)	2	(22.2)	8	(20.5)
	PFS	18	(60.0)	3	(33.3)	21	(53.8)
	OS	15	(50.0)	7	(77.8)	22	(56.4)
HR	TTP	6	(20.0)	1	(11.1)	7	(17.9)
	PFS	16	(53.3)	2	(22.2)	18	(46.2)
	OS	16	(53.3)	8	(88.9)	24	(61.5)

Table 2-6 to 2-10 tabulate the details about the general survival analysis details (e.g. endpoints, enrolment, Kaplan-Meier curves, etc). Clearly, from Table 2-6 the most common endpoint was PFS and this was primarily reported in the original paper and only occasionally eluded to in the follow-up publication. In terms of information about enrolment and follow-up (Table 2-7), in the majority of papers, the dates of recruitment were given. However, information about either the average (or maximum) length of follow-up or the data cut-off date, were less frequently reported. The majority of papers published at least one Kaplan-Meier curve. In the case of the follow-up publication, this was almost always for OS, and rarely any other endpoint. Whilst the majority of Kaplan-Meier curves (Table 2-8) were accompanied by a risk table there was still a noticeable proportion that did not (almost a quarter – 6 / 29 for PFS; 7 / 29 for OS). In terms of events (Table 2-9), OS was the most frequently reported outcome, generally detailing the number of deaths for each treatment group. Events for other endpoints were occasionally given, although in some cases these were only for the overall population, rather than in each group. Median survival and HRs continued to be routinely used (Table 2-10).

2.3.1.3.2 Crossover-specific information

Table 2-11 documents the information concerning treatment switching that was found in the trial publications. In the majority of the papers (23/39), little information could be obtained on the existence of treatment switching in that study. In some papers, however, treatment switching, although not directly stated was implied in other ways (e.g. through the discussion of post-study treatments). Only 2 papers clearly stated that treatment switching was prohibited. Overall, there were similar numbers of trials which permitted treatment switching in all arms, to those where it was only permitted in some (e.g. for a two arm trial, only allowed for control group, for a three arm trial, patients in two of the three arms were allowed to crossover). Reasons were given in several of the papers, and

the most popular were: to allow as a rescue treatment following disease progression, or to administer the superior treatment after treatment un-blinding (or an amendment to the protocol). In some of the publications, patients could switch for either of these reasons. Treatment switching proportion was sometimes reported, mostly in the follow-up paper.

Table 2-11: Commonly available information on treatment switching

	Reported in original paper		Reported in follow-up paper		Total papers reporting	
	n = 30	(%)	n = 9	(%)	n = 39	(%)
Treatment switching:						
Was prohibited	2	(6.7)	0	(0.0)	2	(5,1)
Was not reported on	16	(53.3)	7	(77.8)	23	(59.0)
Occurred in one / some arm(s)	7	(23.3)	2	(22.2)	9	(23.1)
Occurred in all arms	5	(16.7)	2	(22.2)	7	(17.9)
Reason for treatment switching:						
As post-study treatment	2	(6.7)	1	(11.1)	3	(7.7)
Following disease progression	9	(30.0)	3	(33.3)	12	(30.8)
Available after un-blinding / protocol amendment	7	(23.3)	5	(55.6)	12	(30.8)
Treatment switching proportion	6	(20.0)	5	(55.6)	11	(28.2)

2.3.1.3.3 Additional notes

There are a couple of additional findings that are noteworthy; another outcome that is commonly reported in TA summaries as one of the secondary analyses is 'duration of response'. Studies from the more recent TA, sometimes report OS in several different ways; using the ITT approach; and stratifying by 'as treated', dividing the patients into groups depending on treatment given (either two or three groups, e.g. control treatment only, experimental treatment only, control and experimental treatment). A few papers also reported median time to switch. Motzer (2009) chose to report additional information for the 'risk table'. Alongside the 'number at risk' the number of deaths were reported. Several of the other trials for which the papers were not reviewed were reported only available in terms of Abstracts.

2.3.2 Discussion

In terms of obtaining the relevant information, the principle difficulties often related to the identification of the trials. It was difficult to obtain copies of either the Manufacturers Submission or ERG report. Without these, which included references to the actual publications used, identifying individual trials was exceptionally challenging as not all publications report the trial name. Thus, this evidence cannot be considered as an

exhaustive list of evidence for all the TAs considered. However, it gives some indication of the level of information routinely available from different evidence sources.

The main aim of identifying individual papers was to determine whether more detailed information (e.g. number of events, Kaplan-Meier curve, 'numbers at risk' table) regarding the survival distributions was available, alongside details about treatment switching. This was clearly fulfilled. In most cases a Kaplan-Meier curve, with 'at risk' table was available for a given trial, and in many cases so was a PFS K-M curve. However, in terms of considering the most up-to-date data, PFS and OS may not both be reported with the same length of follow-up. This, in particular, has a great bearing on the methods developed and discussed in future chapters (Chapter 5 and Chapter 6).

In terms of treatment switching, relatively little information tends to be clearly reported, which means finding relevant 'crossover' information can be difficult. Sometimes it is only eluded to in terms of post-study therapies (often presented as a table in the supplementary appendix); whilst in other cases it can be clearly documented in the main text. The reasons for treatment switching, and, in particular, when it occurs are rarely stated. Of the few papers that do document when treatment switching has been permitted, the majority are 'on disease progression' as second-line therapy or to control group patients on 'un-blinding' or following interim analyses when one treatment has been demonstrated to be superior to the other. Most papers indicate the extent of treatment switching (e.g. proportion receiving as 'post-study' or switching). However, where treatment switching occurs for different reasons, the extent to which it has occurred for each reason is almost never stated.

The most popular reported measures of efficacy for all outcomes (TTP, PFS and OS) are median survival and HR, being reported in the majority of trials. It is clear that, for most trials, PFS is used as the primary outcome; this means that even if crossover is not part of the study design, there is huge potential for treatment switching to have occurred following disease progression in terms of a second or subsequent line therapy. The trial level tables only capture whether the Kaplan-Meier was accompanied by a risk table or not. However, where risk tables were presented these varied wildly in terms of the number of intervals (e.g. times at which the 'number at risk' were reported).

The publication for updated OS in the MM-009 and MM-010 trials (Dimopoulos, 2009) should be praised for the detail in its paper relating to treatment switching. The authors clearly specified the reasons for treatment switching, and how many patients had switched for each reason. This would be particularly useful when forming a decision about how much a study was affected by treatment switching.

There is no doubt that across all types of evidence, the reporting is very varied. There is, perhaps, slightly more consistency for publications, most likely due to the existing reporting guidelines. However, these still vary in terms of the outcomes they report, and in particular the information on treatment switching. In terms of the TA summaries, these vary vastly with some such as TA34 reporting mostly on the findings, and very little on the trials used (e.g. trial names); and those such as TA258 reporting very detailed summary information. Given that these methods are still relatively new, it is important to continue monitoring how and when they are used.

Chapter 3: Secondary analysis using former studies with treatment switching

3.1 Chapter overview

This chapter principally explores and evaluates how findings from an IC are affected by the inclusion of studies with treatment switching. In particular how these estimates and statistical significance change depending on whether adequate adjustment for treatment switching has been implemented. Three examples are shown, two of which use individual simulated datasets, before a full simulation study is undertaken. This simulation study is divided into two parts: an initial study in which a variety of different scenarios (11 in total) were tried; and a second more 'systematic' study where all distinct combinations of a subset of covariates used in the initial study were chosen (136 scenarios in total). The initial simulation study focussed on examining HRs or treatment effects that had been observed in case studies, or which were perhaps more extreme, in order to assess the sensitivity of ICs to treatment switching. The final section of this chapter discusses former research undertaken to address this issue.

3.2 Impact of appropriateness of methodology on secondary analysis

Given the recent research (Morden, 2011, Latimer, 2012) in the field of treatment switching, and in particular the guidelines (Latimer, 2014) that have been released; it is to be expected that, in the future, the recommended methods will be implemented more regularly to account for treatment switching. In addition, appropriate justification for the final choices will be provided alongside. Nevertheless, this future improvement will not address issues within previously published appraisals. However, this is problematic as conducting secondary analysis, such as ICs, within TAs is becoming increasingly required.

An IC is conducted where no direct head-to-head comparison exists for treatment A and B, say; but where other trials exist in which treatments A and B have been compared to a common comparator (Bucher, 1997). Figure 3-1 describes this diagrammatically for when there are only three possible treatments and two RCTs. The two novel treatments are labelled 'A' and 'B', and the study which compares this novel treatment to the common comparator 'X', denoted Study 'A' and Study 'B' for the respective treatments. Treatment

X represents a regimen such as with Placebo or best supportive care. If there are no head-to-head trials comparing 'A' and 'B' then as indicated in the Figure 3-1 by the dashed line these are compared indirectly. Figure 3-1 highlights the 'network' of treatments e.g. connection of treatments through trials.

Figure 3-1: Indirect Comparison of two treatments

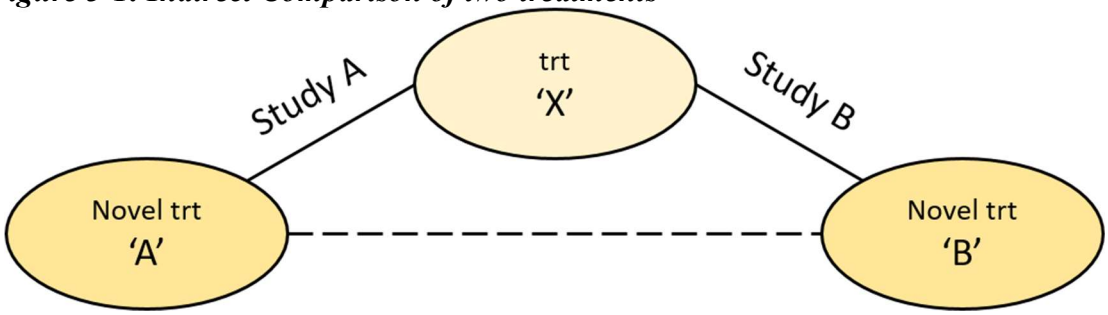
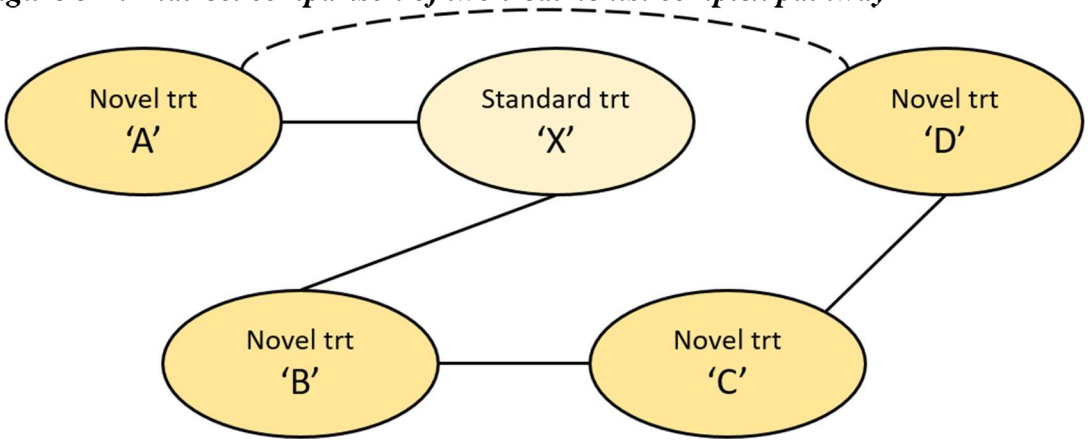


Diagram showing an indirect comparison of treatment A and B, where there are two studies A and B, showing the efficacy of each of the treatments to the comparator X. The solid lines represent the direct evidence, and the dashed line, the indirect evidence.

This can be extended to encompass a number of treatments, provided each intervention is linked by another, as shown in Figure 3-2.

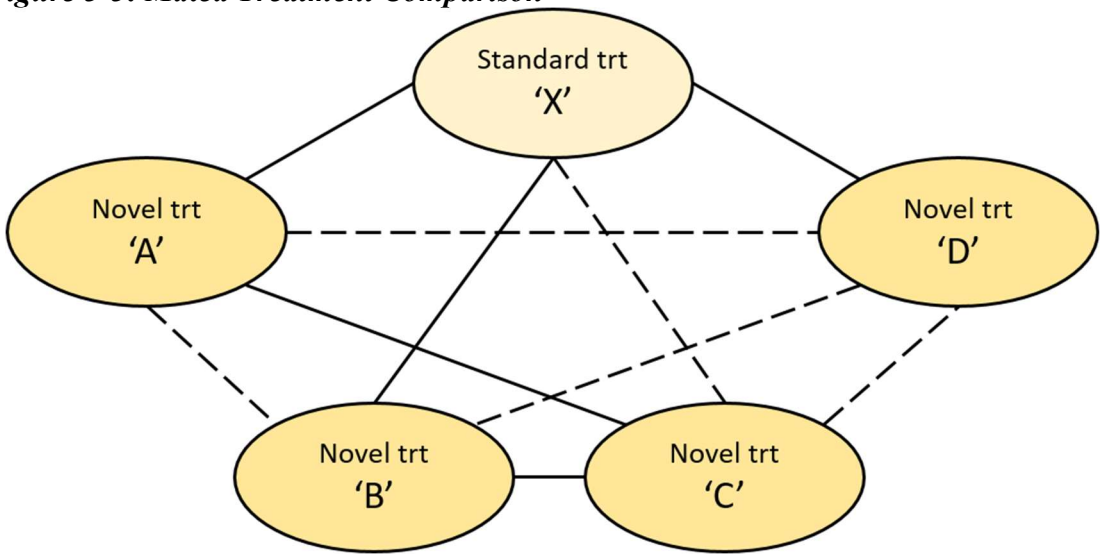
Figure 3-2: Indirect comparison of two treatments: complex pathway



Treatments 'A' and 'D' are compared indirectly (highlighted by the dotted line), by using the studies connecting A and D through treatments X, B, C. Direct head-to-head studies are shown by the solid black lines.

Methods exist whereby all treatments can be compared with one another in one analysis, and which take into account both direct and indirect evidence; these methods are known as 'Mixed Treatment Comparison' (MTC) or Network Meta-Analysis (NMA). An example of a network for a NMA is given in Figure 3-3 (Lumley, 2002, Lu, 2004, Caldwell, 2005).

Figure 3-3: Mixed Treatment Comparison



Since there is at least one study (solid lines) connecting each treatment to another, all pairwise treatment comparisons can be calculated - either using the direct evidence (solid lines) or indirectly (dashed lines).

These methods can be especially useful in HTA as they provide comparisons of the intervention under review with those routinely used in practice for that disease or alternative potentially 'gold standard' treatments. One example of this is presented in TA215 (NICE, 2011) which assessed pazopanib for treating metastatic RCC. The pivotal evidence consisted of a RCT (Sternberg, 2010) which compared pazopanib with placebo, both of which were administered in conjunction with best supportive care (defined as monitoring of progression, symptom control and palliative care without active treatment). However, sunitinib is a treatment previously recommended as a first-line option for patients with this condition and hence, a comparison between sunitinib and pazopanib was desirable. This comparison was achieved using an IC, which included other treatments such as interferon-alpha, vinblastine and medroxyprogesterone in order to link the network. This IC was in a form similar to that shown in Figure 3-2.

The work conducted by Latimer (2012) not only demonstrated the need to adjust for treatment switching, but also that the more appropriate methods were rarely used in practice. Therefore, examples, such as TA171 Multiple Myeloma – lenalidomide, have caused concern (NICE, 2009b). In this appraisal, the manufacturer indirectly compared lenalidomide with bortezomib. The evidence for bortezomib came from the APEX RCT (Richardson, 2005), which had been used as evidence of bortezomib's clinical- and cost-

effectiveness in TA129 (NICE, 2007), and in which all patients receiving the control intervention of high dose dexamethasone were offered bortezomib after the interim analyses showed bortezomib to be far superior. No adjustment for treatment switching was considered necessary in the final analysis, since the results showed a significantly lower HR for both time to progression (TTP) and OS. This unadjusted estimate was then included within the IC. Given that a considerable number of control group patients switched treatments, the actual effectiveness of bortezomib will have been underestimated. Treatment switching also occurred in the MM-009 and MM-010 RCTs used to provide the evidence for lenalidomide (Weber, 2007, Dimopoulos, 2007, Dimopoulos, 2009), but in this case some effort had been taken to adjust for the crossover, though not using one of the currently recommended methods (external data had been used). Especially, since both treatments show a statistically significant effect on survival (Bortezomib: TTP HR 0.55 (0.44, 0.69), OS HR 0.57 (0.4, 0.81); Lenalidomide: TTP HR 0.35 (0.29, 0.43), OS HR 0.66 (0.45, 0.96), and because of the effect of treatment switching, there will be considerable uncertainty about the treatment effect between lenalidomide and bortezomib.

In order to have more certainty in situations such as these, it is necessary to have estimates where an appropriate method of analysis (from those recommended if treatment switching has occurred) has been employed. To achieve this aim, the data will likely need to be reanalysed. However, currently whilst manufacturers will have access to their own individual patient data (IPD), and can therefore adjust it; for competitors' products, they are often solely reliant on published summary information. At present, all methods for treatment switching, and in particular those that have been recommended, require IPD.

3.3 Illustrative examples of impact on an Indirect Comparison

To increase understanding of the impact that using 'unadjusted' and 'adjusted' estimates within ICs has, three examples are described below. For the first two examples, the data (for the two included trials) has been simulated, based on plausible characteristics of TAs (e.g. HR, survival proportion, treatment switching proportion etc.). There were two principle motivations for this: (1) due to the minority of examples which contain treatment switching in both trials, where ITT analyses and 'adjusted' estimates are available for both trials (it permitted a range of HRs and treatment switching proportions to be

explored); (2) to ensure that the treatment switching has been appropriately accounted for by the adjustment method chosen (this is untestable in trial data; although justification can be based on method assumptions and data requirements, it is not infallible). The third example (BRIM-3 and BREAK-3) presents one of the rare existing ICs where the manufacturer reported both an 'unadjusted' and 'adjusted' estimate (NICE, 2014). For this example, it also demonstrates what might have happened if they had conducted the same IC, but only had one 'adjusted' estimate rather than two.

3.3.1 Indirect Comparison

Only a simple IC (i.e. of the form shown in Figure 3-1) (Bucher, 1997) was used, where:

$$\ln(HR_{AB}) = \ln(HR_{AC}) - \ln(HR_{BC}) \quad (3\text{-}1)$$

$$SE(\ln(HR_{AB})) = \sqrt{SE(\ln(HR_{AC}))^2 + SE(\ln(HR_{BC}))^2} \quad (3\text{-}2)$$

Four different comparisons were considered which correspond to the following situations:

I. An IC of the ITT HRs. Where only ITT HRs are presented, that is to say no adjusted HRs are available, the ITT results might be used. Alternatively, if some studies have reported an adjusted HR but not others, the ITT results might be chosen to provide consistency.

II. An IC using the adjusted HR. This is the ideal solution in which all studies will have been adjusted for treatment switching before inclusion in an IC.

III./IV. An IC in which adjusted HRs are used where available and ITT used otherwise. This is potentially the most likely situation for the future. Typically manufacturers have their own IPD and so can analyse this any way they choose. Since guidelines have now been produced, this should ensure that some means of adjustment will be made. However, the comparator interventions come from previously published trials and hence are unlikely to have been appropriately adjusted for treatment switching.

3.3.2 Simulation of the data

As specified in section 3.3.1, the examples described in Section 3.3.2 and 3.3.4 comprise of an IC with two trials. The data for each trial have been simulated as follows.

Both studies had a sample size of 500, with 1:1 randomisation. The underlying survival was simulated from a Weibull distribution, with shape parameter of 0.5, and a scale parameter of 1.322, which means that, on the control treatment, at three years, approximately 90% of the patients will have died. To allow for additional variation between the patients, 70% of patients were assumed to have a more severe disease on enrolment, and therefore, their survival was reduced by 25%. For this, a Bernoulli distribution with probability of 0.70 was used, and patients for whom this was 1 had their survival time multiplied by 0.75.

Only administrative censoring was assumed, and this was achieved using a uniform distribution between 730 and 1095, designed to represent a maximum follow-up time of between 2 and 3 years. Therefore, patients were recruited at a constant rate over the space of a year, and the time of analysis was exactly three years into the study.

Whilst the treatment effect and proportion of crossover varied over the two studies; both studies had a level of treatment switching, and were designed to have a clear treatment effect. A Bernoulli distribution was used to allocate switchers, with different probabilities for those with different disease severities (as shown in Table 3-1); patients with a higher disease severity being more likely to switch. Once a patient was assigned to switch, their switch time was generated from a uniform distribution, implying that a patient who switched was likely to switch at any point throughout the study. Treatment switchers received the same treatment effect as patients randomised to the intervention.

3.3.3 Simulated example with differential treatment effects and treatment switching proportions

3.3.3.1 *Methods and Results*

For both studies, an ITT HR was estimated, as was an adjusted HR. To obtain an adjusted HR, a RPSFTM was fitted (as described in Section 1.1.3.3.1). The final acceleration factor was used to obtain the counterfactual dataset, then the original analysis (such as a Cox

model) conducted to give the final 'adjusted HR'. The SE was calculated by preserving the p-value (given in Section 1.1.3.3.1).

Table 3-1: Study specific simulation information - Example 1
Summary of the information on treatment switching and efficacy used for the simulation of each of the studies

	Study A:	**Study B:**
Treatments:		
Novel treatment reference	A	B
Standard treatment reference	X	X
Level of treatment switching:	Low	High
Switching proportion for patients with:		
Severe disease	25%	95%
Moderate disease	5%	10%
Underlying HR	0.70	0.31

Table 3-2: Study-specific information for the single simulated dataset - Example 1
*Summary of the information on treatment switching and efficacy for the simulated dataset representing each study. NOTE: Only <u>one</u> dataset has been simulated per study. * CI: Confidence Interval*

	Study A:	**Study B:**
No. of patients switching from standard to novel treatment, (%):	45 (18.0%)	183 (73.2%)
ITT HR, (95% CI*; p-value):	0.797 (0.665, 0.956; p = 0.013)	0.560 (0.456, 0.688; p <0.001)
RPSFTM-adjusted HR, (95% CI*; p-value):	0.738 (0.581, 0.939; p = 0.013)	0.334 (0.228, 0.488; p <0.001)

Using the simulation parameters, within the IC formula, the underlying difference between the two treatments should be a HR of 2.258. Hence, from the IC, the underlying mortality rate for patients on treatment A is more than twice that of patients on treatment B. Given that only a single dataset has been simulated for each trial, the actual HR will only be comparable (not exact).

So, starting with the most preferable scenario, that is adjusted HRs are available for all groups, a HR of 2.210 (95% CI: 1.408, 3.641) is achieved. This suggests that there is a

statistically significant difference between the two treatments, and that the mortality rate for those receiving treatment A is 2.210 times that of those patients on treatment B.

Table 3-3: Comparison of HRs calculated from an IC - Example 1

Study A	Study B	HR (95% CI)	Statistically significant?
Unadjusted	Unadjusted	1.423 (95% CI: 1.080, 1.875)	Yes
Unadjusted	Adjusted	2.386 (95% CI: 1.562, 3.467)	Yes
Adjusted	Unadjusted	1.318 (95% CI: 0.961, 1.806)	No
Adjusted	Adjusted	2.210 (95% CI: 1.408, 3.641)	Yes

This HR is very similar to the underlying rate but it should be noticed that there is a wide confidence interval (CI).

Consider next what happens when no adjustment has been made in either case. Here, the HR: 1.423 (95% CI: 1.080, 1.875) still shows a statistically significant difference between the two treatments. Even though it is statistically significant, the HR is considerably lower, only suggesting the hazard rate is 1.423 times higher. In addition, the true underlying value of 2.258 does not lie within the CI. The CI is affected by the estimates for the HR and in addition, without adjusting for treatment switching there is less uncertainty incorporated in the SE and hence this results in a narrower CI for the IC.

Now, examine the third scenario, where only some (one of the two) HRs have been adjusted for treatment switching. First, assume both analyses, adjusted and unadjusted are available for Study B. Therefore, the study where there was the largest amount of treatment switching, and greatest treatment effect has been adjusted for. Using this combination achieves a HR of 2.386 (95% CI: 1.562, 3.467), higher than the previous two ICs. This time there is a smaller CI than the case where both HRs had previously been adjusted, but this still incorporates the true HR (unlike the scenario with both unadjusted).

Looking at the final example, where Study A has been adjusted for treatment switching and not Study B, it can be seen that the HR is 1.318 (95% CI: 0.961, 1.806). This time, in contrast to the other scenarios, a statistically significant difference cannot be seen. Similarly, to the other cases where at least one adjusted HR has been used, there is an increase in the uncertainty compared to the unadjusted IC. Like the unadjusted analysis,

the true value is not contained within the CI. In summary, there appears to be potential bias unless both studies have accounted for treatment switching.

3.3.4 Simulated example with the same treatment effect and differential treatment switching proportions

From the first example, it could be concluded that, if the adjusted analyses for all treatment groups are not available, the ITT analyses should be used for consistency. However, this second example highlights the potential downfall of adopting that stance.

The survival data were simulated in a similar way to the previous example (and as set out in section 3.3.2); hence, using the same underlying survival, randomisation ratio, sample size and censoring distribution. In this scenario, both treatment effects compared with the control treatment were set to the same; but, the amount of treatment switching for each study was different.

It should be noted, that based on the simulation parameters, the underlying difference between the two treatments should be a HR of 1. Hence, the mortality rate for patients on treatment C was the same as that for patients on treatment D.

Table 3-4: Study specific simulation information - Example 2
Summary of the information on treatment switching and efficacy used for the simulation of each of the studies

	Study C:	**Study D:**
Treatments:		
Novel treatment reference	C	D
Standard treatment reference	X	X
Level of treatment switching:	High	Low
Switching proportion for patients with:		
Severe disease	95%	25%
Moderate disease	10%	5%
Underlying HR	0.50	0.50

Table 3-5: Study-specific information for the single simulated dataset - Example 2
Summary of the information on treatment switching and efficacy for the simulated dataset representing each study. NOTE: Only one dataset has been simulated per study.

	Study C:	Study D:
No. of patients switching from standard to novel treatment, (%):	192 (76.8%)	46 (18.4%)
ITT HR, (95% CI*; p-value):	0.693 (0.573, 0.839; p <0.001)	0.484 (0.398, 0.590; p <0.001)
RPSFTM-adjusted HR, (95% CI*; p-value):	0.530 (0.383, 0.735; p <0.001)	0.465 (0.380, 0.570; p <0.001)

Table 3-6 gives the results of conducting an IC on both the unadjusted HR, both adjusted HRs and where one of the HR has been adjusted and the other not.

Table 3-6: Comparison of HRs calculated from an IC - Example 2

Study C	Study D	HR (95% CI)	Statistically significant?
Unadjusted	Unadjusted	1.432 (1.090, 1.881)	Yes
Unadjusted	Adjusted	1.490 (1.129, 1.967)	No
Adjusted	Unadjusted	1.095 (0.749, 1.600)	No
Adjusted	Adjusted	1.140 (0.778, 1.671)	No

Most importantly, the HR has changed considerably over the different scenarios. This is useful to examine as the HR is not affected by sample size, unlike the statistical significance. However, this example shows that it could be possible that statistical significance of the treatment effect between intervention C and D is different if the unadjusted HRs are used, compared to any other pairing. As the underlying treatment effects, are in fact the same, the first comparison, with both unadjusted estimates, gives a misleading conclusion. Although no example which included such a large treatment effect combined with such high proportion of treatment switching has been identified, an example exists for which the treatment switching proportion was 98% (ITT HR: 0.83) and another separate example (in a different type of cancer) where an ITT (unadjusted for treatment switching) HR estimate is reported as 0.55 (treatment switching was 62%). In theory, there could be such an example which combined both of these estimates in the future, potentially resulting in similar findings to this example.

These two examples highlighted the potential variation in results from an IC, depending on how treatment switching was addressed in the trials. Having simulated the data, there was the advantage of confirming that the 'adjustment' method has performed well, and

in knowing how biased the IC was for each scenario. The original motivation of simulating the datasets was the lack of examples. However, one example does exist, and this is presented in Section 3.3.5.

It is hard to draw firm conclusions having only looked at two different situations, and one single comparison for each. To explore these suppositions conclusively, a full simulation study would be required.

3.3.5 The BRIM-3 and BREAK-3 trials

This final example demonstrates an actual case presented to NICE, and is quite remarkable as it is one of the only examples where both of the trials used in the IC reported ITT and RPSFTM estimates for the treatment effect (NICE, 2014). Whilst the TA reported the results for scenarios I and II (both unadjusted and both adjusted estimates); here, those for scenarios III and IV demonstrate how the results may have been affected if both the adjusted analyses had not been available.

TA321 considered whether Dabrafenib was cost-effective for treating patients with unresectable or metastatic BRAF V600 mutation-positive melanoma (NICE, 2014). The pivotal evidence came from the BREAK-3 trial, which compared Dabrafenib with Dacarbazine (Hauschild, 2014). However, the manufacturer also chose to conduct an IC between Dabrafenib and Vemurafenib, a treatment recommended by NICE, subject to the patient access / discount scheme, in 2012. The evidence for Vemurafenib came from the BRIM-3 trial (where the control group received Dacarbazine treatment) (McArthur, 2014).

Information on treatment switching and treatment efficacy for both of the studies is given in Table 3-7. This highlights that, whilst the ITT HR (for OS) is the same for both trials, the treatment switching proportion is noticeably different (34% vs. 57%).

Table 3-8 shows the results from using different combinations of the unadjusted and adjusted estimates. Whilst the point estimate and SE clearly change, the statistical significance does not, which is a reflection of the small sample size.

Table 3-7: Study-specific information for the BRIM-3 and BREAK-3 trials
Summary of the information on treatment switching and efficacy for each study

	BRIM-3	BREAK-3
Treatments:		
Novel treatment	Vemurafenib	Dabrafenib
Standard treatment	Dacarbazine	Dacarbazine
No. of patients switching from standard to novel treatment, (%):	115 (34%)	36 (57%)
ITT HR, (95% CI):	0.76 (0.63, 0.93)	0.76 (0.48, 1.21)
RPSFTM-adjusted HR, (95% CI):	0.64 (0.53, 0.78)	0.55 (0.21, 1.43)

Table 3-8: Comparison of HRs calculated from an IC – BRIM-3 and BREAK-3

BRIM-3	BREAK-3	HR (95% CI)	Statistically significant?
Unadjusted	Unadjusted	1.00 (0.61, 1.64)	No
Unadjusted	Adjusted	0.72 (0.27, 1.93)	No
Adjusted	Unadjusted	1.19 (0.72, 1.95)	No
Adjusted	Adjusted	0.86 (0.32, 2.29)	No

3.4 Initial simulation study – part 1: Specific scenarios

3.4.1 Background

Whilst the case study examples indicated that there were issues relating to ICs in trials with treatment switching, there was always the possibility that these findings could be in the extreme, and hence rarely occur in practice. Therefore, a small simulation study was conducted to assess the impact of crossover in a slightly wider context. The aim was to assess, if the same circumstances were to stand, how variable the difference between ICs using ITT and RPSFTM-adjusted estimates were, and any changes on the statistical significance that might occur.

3.4.2 Methods

3.4.2.1 *Simulation of the data*

The underlying survival data was generated in the same way as for the simulated examples and hence the details of this can be found in Section 3.3.1. However, the underlying HR and switching proportions varied across scenarios (and also study). In generating the scenarios, the key aim was to explore many different possible

combinations relating to the trial characteristics (thus representing different examples seen in practice) but using a small number of scenarios. As a consequence, there was no systematic method to the assignment of crossover proportion and treatment effect.

Throughout the review of the TAs, it was noticed that HRs for OS ranged between 0.30 and 1. So assuming that assessors and manufacturers would only be interested in comparing treatments indirectly where there was some suggestion of a protective effect when compared to the standard treatment; the underlying HR was restricted to a selection of values between 0.31 (strong protective effect) and 0.95 (weak protective effect).

It was also of interest to consider how the results were affected, both depending on how high the switch proportion was, and based on which type of prognosis switched more. Although the choices for p_1, p_2 (the probability of switching for patients with a 'poor' and 'good' prognosis respectively) were chosen arbitrarily, the proportion of switchers overall primarily lay between 18% and 80% (once again in line with examples in the literature). However, there were a few exceptions which were specifically designed to have very little treatment switching (< 5%).

The table below shows the structure of the scenarios regarding the treatment effect, and the probability of switching for patients depending on prognosis (p_1 for those with a 'poor' prognosis and p_2 for those with a 'good' prognosis).

These encompass a wide range of possible situations, exploring those occasions:

- Where one study may have the stronger underlying treatment effect and the higher proportion of treatment switching (scenarios 1,2,3,5,8);
- Where the studies have the same underlying treatment effect but different proportions of patients switching (scenarios 4,6);
- Where one study had the more effective treatment, but the other allowed more treatment switching (scenario 7);
- How the selection process (choosing which type of prognosis switched more) affected the situation (scenarios 9,10,13);
- Where the studies had different underlying treatment effects but the same proportion of treatment switching (scenario 11);

- Where the studies were exactly the same in terms of underlying HR and treatment switching proportion (mechanism) (scenario 12).

Table 3-9: Scenario information
The underlying true HR and probability of switching (depending on disease severity) for each of the simulated studies in the IC comparing treatment 'A' – assessed in 'Study A' and treatment 'B' – assessed in 'Study B'.

Scenario	Study A			Study B		
	HR	p_1	p_2	HR	p_1	p_2
1	0.70	0.70	0.1	0.50	0.90	0.10
2	0.70	0.25	0.05	0.31	0.95	0.10
3	0.50	0.70	0.1	0.70	0.25	0.05
4	0.70	0.95	0.4	0.70	0.25	0.05
5	0.70	0.90	0.2	0.95	0.25	0.05
6	0.95	0.95	0.4	0.95	0.25	0.05
7	0.70	0.90	0.4	0.55	0.45	0.05
8	0.40	0.95	0.4	0.60	0.01	0.01
9	0.70	0.10	0.7	0.50	0.10	0.90
10	0.70	0.05	0.25	0.31	0.10	0.95
11	0.70	0.70	0.25	0.50	0.70	0.25
12	0.70	0.70	0.25	0.70	0.70	0.25
13	0.70	0.70	0.25	0.70	0.50	0.25

For each scenario considered, 1000 replications were carried out, and thus the results reported are an average over these 1000.

3.4.2.2 *Statistical analyses undertaken*

For each dataset, regardless of scenario, the proportion of treatment switching, ITT HR using a Cox PH model, RP SFTM-adjusted HR were calculated for each 'study'. Once these estimates had been obtained, four ICs were conducted:

1) Using both ITT HR for the studies
2) Using the adjusted HR for study A, and the ITT HR for study B
3) Using the ITT HR for study A, and the adjusted HR for study B
4) Using both of the adjusted HRs

As well as computing the point estimates and the uncertainty, whether the comparison was statistically significant or not (i.e. whether the value 1 was contained within the 95% CI for the IC) was also recorded.

3.4.2.2.1 Performance Measures

For both adjusted and unadjusted of the study-specific HRs and each of the ICs, several performance measures were calculated. These included the bias, mean squared error (MSE) and the coverage. These were defined as follows (Boucher, 2013b):

Bias, δ

$$\delta = \hat{\beta} - \beta \tag{3-3}$$

Where $\hat{\beta}$ is the estimate obtained from that specific analysis and β the true underlying value.

Mean squared error (MSE)

$$MSE = \left(\bar{\hat{\beta}} - \beta\right)^2 + \left(SE\left(\bar{\hat{\beta}}\right)\right) \tag{3-4}$$

Where $\bar{\hat{\beta}}$ is the mean value of all the estimates ($\hat{\beta}$) for a specific analysis, and once again β is the true underlying value.

Since this method takes account of both the bias and the variability of the estimates, it provides an expedient overall assessment of a particular analysis method.

Coverage

Proportion of simulations where the true underlying value is contained within the 95% CI. For a 95% CI, the coverage should be approximately 95%.

3.4.3 Results

3.4.3.1 *Point estimates*

Table 3-10 details the estimates both for the studies, in terms of average adjusted and unadjusted HR and treatment switching proportion. As expected, where the treatment switching proportion is very low and / or the HR is close to 1, the ITT and RPSFTM-adjusted results are very similar.

3.4.3.2 *Comparison of statistical significance across ICs*

Table 3-11 gives the values of the HR for the IC using every combination of adjusted and unadjusted estimates, contrasted with the true underlying HR. Using both ITT estimates tends to underestimate the treatment effect, sometimes drastically, whilst only using one adjusted and one unadjusted tends to lead to an over- or under-estimate depending on which has the stronger treatment effect and / or high treatment switching proportion. Using both adjusted estimates always performs well, and the majority of the time, it gives the results most similar to the underlying value.

Table 3-12 shows the proportion of the 1000 simulations which were statistically significant for each of the four ICs and each of the scenarios. There are quite noticeable differences in the proportions, depending on whether neither, one or both have been adjusted. Whether the proportion of statically significant observations increases or decreases, when comparing the IC with both ITT estimates to that with both adjusted, varies from scenario to scenario. However, care must be taken in interpreting these findings, as highly biased estimates may be wrongly significant. Essentially, these are just a measure for the power of the analysis.

Table 3-10: Initial simulation - averaged scenario-study-specific information
The scenario-specific details for each study, then used in the IC, averaged over all 1000 datasets with the exception of the true HR – this is the underlying scenario value.

Scenario	Study A				Study B			
	Crossover proportion	True HR	ITT HR	Adjusted HR	Crossover proportion	True HR	ITT HR	Adjusted HR
1	51.9%	0.70	0.766	0.695	66.1%	0.50	0.663	0.511
2	19.0%	0.70	0.727	0.697	69.5%	0.31	0.554	0.315
3	52.1%	0.50	0.628	0.508	19.0%	0.70	0.728	0.699
4	78.4%	0.70	0.806	0.695	19.1%	0.70	0.727	0.699
5	69.1%	0.70	0.789	0.693	19.0%	0.95	0.949	0.945
6	78.5%	0.95	0.962	0.947	18.9%	0.95	0.949	0.945
7	75.0%	0.70	0.802	0.697	32.9%	0.55	0.621	0.564
8	78.5%	0.40	0.650	0.411	1.0%	0.60	0.601	0.612
9	28.1%	0.70	0.753	0.698	34.1%	0.50	0.611	0.516
10	11.0%	0.70	0.725	0.700	35.4%	0.31	0.460	0.317
11	56.3%	0.70	0.777	0.697	56.7%	0.50	0.644	0.512
12	56.6%	0.70	0.779	0.699	56.8%	0.70	0.776	0.695
13	71.5%	0.70	0.807	0.697	35.5%	0.50	0.593	0.512

Table 3-11: Initial simulation - Average IC HR depending on analysis method
The HR estimates calculated from an IC on each of the 1000 scenario-specific estimates and then averaged over on the log-HR scale. The estimates have then been converted back to the HR scale to make them easier to interpret.

Scenario	HR obtained from the indirect comparison of Study A versus Study B:				
	Using the true underlying values	**ITT estimates for A & B**	**Adjusted for A ITT for B**	**ITT for A Adjusted for B**	**Adjusted for A & B**
1	1.397	1.155	1.048	1.498	1.359
2	2.258	1.311	1.258	2.310	2.216
3	0.714	0.862	0.698	0.897	0.727
4	1.000	1.107	0.955	1.152	0.994
5	0.737	0.832	0.731	0.835	0.734
6	1.000	1.014	0.998	1.019	1.002
7	1.273	1.291	1.122	1.423	1.237
8	0.667	1.082	0.684	1.063	0.672
9	1.397	1.232	1.142	1.460	1.353
10	2.258	1.576	1.522	2.287	2.209
11	1.400	1.207	1.083	1.517	1.361
12	1.000	1.004	0.901	1.121	1.006
13	1.400	1.363	1.177	1.576	1.362

Table 3-12: Initial simulation study - statistical significance depending on method
The proportion of the 1000 simulations for which the IC HR was a statistically significant result.

Scenario	Proportion of simulations for which the indirect comparison of Study A versus Study B was statistically significant:			
	ITT estimates for A & B	**Adjusted for A ITT for B**	**ITT for A Adjusted for B**	**Adjusted for A & B**
1	4.7%	0.2%	29.5%	8.7%
2	21.6%	9.3%	86.6%	79.5%
3	4.6%	28.1%	1.8%	16.0%
4	2.0%	0.2%	3.4%	0.2%
5	6.9%	16.6%	5.2%	12.3%
6	0.1%	0.5%	0.1%	0.3%
7	19.0%	0.8%	38.0%	2.5%
8	1.4%	15.0%	1.4%	18.9%
9	11.5%	1.8%	36.6%	12.4%
10	67.9%	55.0%	97.3%	95.3%
11	9.1%	0.7%	38.1%	8.3%
12	0.2%	1.2%	1.3%	0.1%
13	32.0%	1.3%	60.2%	6.6%

3.4.3.3 *Performance measures*

There is a considerable difference over the scenarios in terms of the bias (both absolute and percentage), and the coverage for study A, as shown in Table 3-13. Ignoring the direction (i.e. positive or negative values), the bias of the estimate ranges between 0.9% and 38.2% for the ITT estimate, and reduces to between 0.4% and 1.8% for the adjusted estimates. The MSE for the ITT was largely comparable across all scenarios and between the ITT and RPSFTM except for scenarios 6 and 8. The coverage is exceptionally variable for the ITT, it ranged from 0.1% to 94.9%. In two of the scenarios, the coverage was close to what is expected; a further nine ranged between 60% and 93%. The remaining scenarios all had poor coverage (less than 40%). For the adjusted estimates, apart from one scenario the coverage was either around the expected 95% or higher. Even if the SE is correct, where there is high bias, the coverage will not be appropriate, and thus, will not necessarily be useful.

Table 3-13: Initial simulation – Performance measures for Study A
Performance measure (bias, MSE and coverage) for the simulated 'Study A'.

Scenario	ITT				RPSFTM-adjusted			
	Absolute bias	Prop. Bias	MSE	Coverage	Absolute bias	Prop. Bias	MSE	Coverage
1	0.070	8.2%	0.077	83.7%	-0.001	-1.4%	0.080	96.8%
2	0.030	3.2%	0.070	92.6%	0.000	-0.9%	0.066	96.5%
3	0.130	20.0%	0.077	36.2%	0.014	0.6%	0.075	92.8%
4	0.109	12.7%	0.087	67.7%	0.001	-1.7%	0.099	97.8%
5	0.092	10.9%	0.082	76.1%	-0.001	-1.8%	0.090	96.7%
6	0.016	0.9%	0.088	94.9%	0.008	-1.4%	0.147	95.2%
7	0.106	12.3%	0.088	65.8%	0.004	-1.3%	0.099	97.4%
8	0.253	38.2%	0.129	0.1%	0.018	0.9%	0.079	96.2%
9	0.056	6.6%	0.076	86.2%	0.002	-0.8%	0.073	97.2%
10	0.028	3.0%	0.068	94.3%	0.003	-0.4%	0.063	97.2%
11	0.081	9.5%	0.080	79.2%	0.002	-1.1%	0.083	97.4%
12	0.082	9.7%	0.080	78.4%	0.004	-0.8%	0.084	97.8%
13	0.111	12.9%	0.088	67.1%	0.004	-1.3%	0.097	97.4%

Based on the results from Table 3-14, study B follows a similar trend to that for study A. The bias typically reduces, although in estimates where there was initially low bias (<0.5%), the adjusted estimate does contain slightly more bias (1.1%. to 1.4%). Once again, for the ITT estimates the coverage varies substantially – 0% to 95%. Here five studies have values of the coverage close to 95% (93.1% to 95.4%). The adjusted results

were acceptably close to 95% (between 92% and 97%) in most scenarios. The MSE was comparable for all scenarios except scenario 2.

Table 3-14: Initial simulation - Performance measures for Study B
Performance measure (bias, MSE and coverage) for the simulated 'Study B'.

Scenario	ITT				RPSFTM-adjusted			
	Absolute bias	Prop. Bias	MSE	Coverage	Absolute bias	Prop. Bias	MSE	Coverage
1	0.165	24.1%	0.092	17.4%	0.017	0.8%	0.082	94.6%
2	0.247	43.7%	0.121	0.0%	0.011	-0.7%	0.067	93.7%
3	0.031	3.5%	0.068	93.1%	0.002	-0.5%	0.065	98.2%
4	0.031	3.3%	0.071	93.5%	0.002	-0.6%	0.068	96.8%
5	0.003	-0.5%	0.087	95.4%	0.000	-1.1%	0.097	94.8%
6	0.003	-0.5%	0.088	94.5%	-0.001	-1.1%	0.097	94.6%
7	0.074	11.1%	0.065	74.6%	0.018	1.7%	0.068	93.0%
8	0.004	-0.3%	0.058	94.5%	0.015	1.4%	0.060	93.1%
9	0.113	17.6%	0.074	46.0%	0.020	2.0%	0.071	92.2%
10	0.152	32.2%	0.072	3.9%	0.011	0.8%	0.052	92.1%
11	0.147	22.0%	0.086	26.3%	0.018	1.3%	0.078	93.1%
12	0.079	9.3%	0.079	78.6%	0.000	-1.4%	0.083	97.3%
13	0.095	15.2%	0.067	56.8%	0.016	1.6%	0.067	92.9%

Table 3-15: Initial simulation - IC performance measures (part 1)
The performance measure (bias, MSE and coverage) for the simulated ICs for two of the IC-scenarios – using both ITT estimates; and using Study A RPSFTM-adjusted and Study B ITT estimates.

Scenario	ITTA versus ITT B				adjusted A versus ITT B			
	Absolute bias	Prop. Bias	MSE	Coverage	Absolute bias	Prop. Bias	MSE	Coverage
1	-0.232	-22.0%	0.211	91.3%	-0.338	-34.8%	0.274	83.3%
2	-0.934	-73.9%	1.059	15.0%	-0.988	-81.3%	1.156	12.7%
3	0.155	16.4%	0.141	92.4%	-0.005	-3.9%	0.124	99.0%
4	0.117	8.9%	0.161	98.0%	-0.032	-6.2%	0.163	99.8%
5	0.102	10.7%	0.117	97.4%	0.003	-2.0%	0.115	99.3%
6	0.022	0.6%	0.133	99.9%	0.013	-1.8%	0.182	99.5%
7	0.031	0.5%	0.181	99.7%	-0.134	-15.0%	0.215	98.5%
8	0.426	37.8%	0.332	23.5%	0.033	0.4%	0.150	99.0%
9	-0.153	-14.5%	0.196	96.6%	-0.243	-23.6%	0.228	92.4%
10	-0.667	-44.7%	0.667	58.8%	-0.722	-49.7%	0.732	54.2%
11	-0.182	-17.0%	0.200	95.2%	-0.304	-30.8%	0.264	89.6%
12	0.013	-0.4%	0.136	99.8%	-0.088	-12.2%	0.147	98.8%
13	-0.025	-3.7%	0.188	98.9%	-0.207	-20.6%	0.243	96.9%

Table 3-16: Initial simulation - IC performance measures (part 2)
The performance measure (bias, MSE and coverage) for the simulated ICs for two of the IC-scenarios – using Study A RPSFTM-adjusted and Study B ITT estimates; and using both ITT estimates.

Scenario	ITT A versus adjusted B				adjusted A versus adjusted B			
	Absolute bias	Prop. Bias	MSE	Coverage	Absolute bias	Prop. Bias	MSE	Coverage
1	0.126	5.2%	0.294	99.1%	-0.013	-4.7%	0.268	99.8%
2	0.114	-0.3%	0.570	99.2%	0.017	-4.6%	0.536	99.7%
3	0.191	19.7%	0.157	90.4%	0.024	0.3%	0.130	99.4%
4	0.162	12.5%	0.180	96.6%	0.007	-2.0%	0.169	99.8%
5	0.106	11.0%	0.125	97.4%	0.006	-1.7%	0.120	99.4%
6	0.028	1.0%	0.141	99.9%	0.019	-1.4%	0.188	99.7%
7	0.167	9.5%	0.253	97.8%	-0.015	-4.7%	0.236	99.4%
8	0.407	36.7%	0.315	24.9%	0.020	-1.4%	0.147	98.9%
9	0.083	3.0%	0.255	98.3%	-0.024	-4.8%	0.240	99.1%
10	0.070	-0.5%	0.454	99.0%	-0.009	-4.0%	0.433	99.1%
11	0.141	6.3%	0.298	98.7%	-0.014	-4.8%	0.272	99.3%
12	0.134	9.8%	0.190	98.7%	0.020	-0.8%	0.172	99.9%
13	0.197	10.0%	0.297	97.3%	-0.014	-4.6%	0.263	99.1%

From Table 3-15, Table 3-16, it can be seen that the proportion of bias varies considerably, and as this example shows the bias can go in either direction. Starting with Table 3-15 and looking across all scenarios, the bias could be as little as 0.4% to 73.9% for the comparison using both ITT estimates and 0.4% to 81.3% for the adjusted estimate for study A and ITT estimate for study B. The MSE is a lot higher than for the individual estimates, especially for the second scenario. The coverage is also very variable (ranging from 15% to 99.9% when using both the ITT estimates and 12.7% to 99.8% for the adjusted A and ITT B comparison).

3.4.4 Conclusions

The data for this example was simulated in such a way that the RPSFTM was expected to perform well in adjusting for the treatment switching, based on previous simulation studies (Morden 2009, Latimer 2012). In practice, it could not be guaranteed how successful any adjustment would be at obtaining the actual treatment effect, and thus there may be additional bias in the adjusted estimates used in the 'adjusted' IC analysis. Yet, having ensured that the adjustment method in the simulation study was appropriate highlighted that in these circumstances the adjusted estimates should be similar to the true

values, and hence the IC using both adjusted estimates would perform well, albeit with additional uncertainty.

The data presented here are by far a simplification of those that could be obtained in practice. For example, whilst the proportion of treatment switchers was related to the severity level, the switching time was not controlled by any time dependent covariates (such as severity, age, etc.). In addition, the underlying survival data for both studies was the same. This meant that exactly the same population was assumed for both studies, which may not occur in practice. The initial simulation results are, therefore, very important, as they have shown that even in these simple examples where additional external biases may be reduced, the data remain sensitive and whether treatment switching was accounted for has a clear impact.

The changes in the proportion of simulations where the result was statistically significant is perhaps surprising. When adjusting a single estimate for treatment switching using the RPSFTM, the uncertainty is typically calculated by preserving the p-value (section 1.1.3.3.1). This approach merely increases the uncertainty, and does not alter the significance. To some extent then, it might have been hypothesized that, in general, for the IC, either:

1) The significance would not change – i.e. the proportion of simulations showing a statistically significant effect should remain the same.
2) A statistically significant result would be less likely, given the increase in uncertainty – i.e. the proportion of simulations showing a statistically significant effect will be less for comparisons using at least one adjusted estimate, compared to the analysis using both ITT estimates.

However, neither of these hypotheses were justified by the data. Indeed, in some circumstances the proportion of statistically significant effects increased substantially. Whilst these findings may give some insight into the effect of treatment switching in ICs, it is difficult to see how this could be used in practice. As the actual underlying treatment effect is unknown, and the level of bias varies by this unknown quantity, there will always be a high level of uncertainty.

In terms of the comparison using the adjusted and unadjusted estimates, similar approaches were followed as for the case studies. These estimates were highly sensitive and very much influenced by scenarios. For example, where the estimate with the higher treatment switching proportion and which showed the greater treatment effect was adjusted for crossover, and the other one not, the IC HR more closely resembled the adjusted analysis; albeit that this was an overestimate of the true effect, and there might also have been a higher probability of having a statistically significant result. In contrast where the estimate for the study with less treatment switching, and showing less benefit was adjusted, the results resembled the ITT analysis more.

It should also be noted that the analyses both in terms of the IC HR and probability of being statistically significant were very sensitive to the data used.

3.5 Simulation study – part 2: Systematic selection of scenarios

3.5.1 Background

The initial simulation study examined different potential scenarios. It highlighted a number of exceptionally valuable points, depending on which estimates (unadjusted or adjusted) were used. However, due to the nature of the simulation scenarios (a wide variety of characteristics were spread across a very limited number of scenarios), it was difficult to compare across situations and examine the effect that all the different parameters had. It did not seem feasible due to computation time to investigate all combinations of the values of the HR, and probabilities of switching which differed depending on prognosis for each study, as this would have led to at least 1,728 scenarios. Nevertheless, a more systematic approach did seem warranted to give a greater understanding of the effect in terms of bias, MSE and coverage.

3.5.2 Methods

3.5.2.1 *Simulation of the data*

As before, for the case studies and the initial simulation study, the underlying survival data was simulated according to section 3.3.1. However, here the treatment effect and treatment switching proportion were restricted to four possible values, for both of the studies. These are given below.

Table 3-17: 'Systematic simulation' - Characteristics of the simulated datasets

Characteristic	Values used in the simulation
Treatment effect	0.31; 0.50; 0.70; 0.95
Treatment switching proportion	25% poor prognosis; 5% good prognosis
	50% poor prognosis; 25% good prognosis
	70% poor prognosis; 25% good prognosis
	90% poor prognosis; 40% good prognosis

This resulted in 256 potential scenarios. It was then noticed that since the same set of values were being used for study A and B, this would lead to some duplicate findings. Therefore, to improve computation time, essentially 'duplicate' scenarios were ignored. An example of a duplicate scenario is given in Table 3-18, since essentially Scenario X is the same as Scenario Y, just with the study letters reversed. Therefore, the findings should be almost identical (some variation is likely to occur naturally as part of the simulation), only with the study names reversed.

Table 3-18: Example of 'duplicate' scenarios

Study	Parameter	Scenario X	Scenario Y
Study A	Treatment effect	0.31	0.95
	Crossover proportion	25% poor prognosis; 5% good prognosis	70% poor prognosis; 25% good prognosis
Study B	Treatment effect	0.95	0.31
	Crossover proportion	70% poor prognosis; 25% good prognosis	25% poor prognosis; 5% good prognosis

This reduced the number of scenarios to 136, which are illustrated in Figure 3-4.

3.5.2.2 *Analyses*

For each dataset, the HRs and corresponding SE were calculated using a Cox model (Cox, 1972) and a RPSFTM (Robins and Tsiatis, 1991). An IC was then performed, as described in sections 3.3.1. and 3.3.3.1, using each of the estimates (both ITT; one adjusted and one ITT estimate; both adjusted). The main aim of this research was to assess the various performance measures, which are reported in Section 3.5.3. The performance measures of interest were the bias (both absolute and percentage), MSE and coverage, as defined in Section 3.4.2.2.1.

Since it is difficult to summarise all of the 136 scenarios in a clear way, the scenarios have been grouped into five categories; those with:

(1) the same treatment effect and crossover proportion;

(2) with less effective novel treatment in study B and lower proportion of crossover in study B;

(3) less effective novel treatment in study B and the same proportion of crossover;

(4) less effective novel treatment in study B and higher proportion of crossover in study B;

(5) the same treatment effect and a higher proportion of crossover in study B.

This is illustrated in Figure 3-5.

Figure 3-4: 'Systematic simulation' scenarios

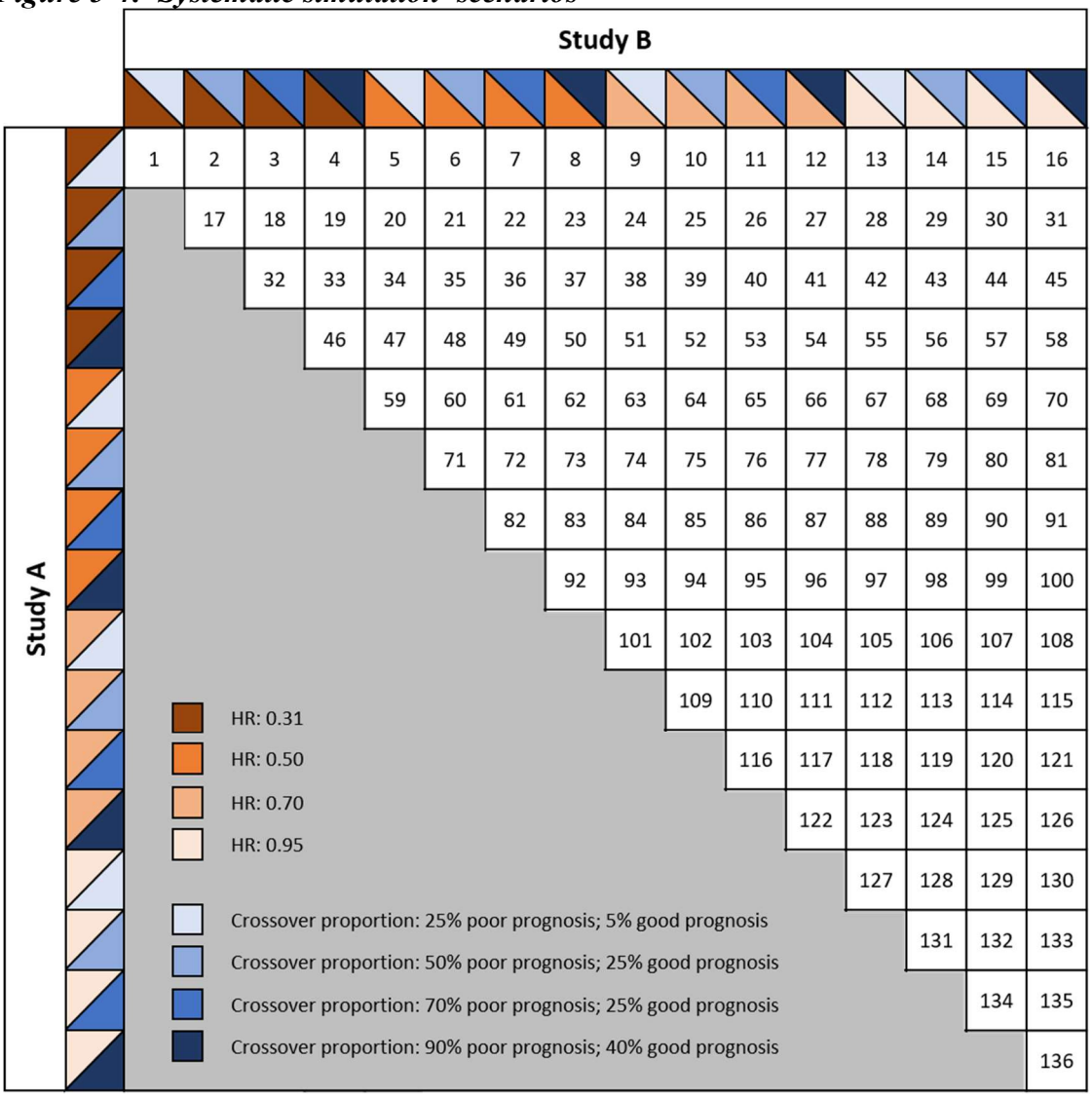


Illustration of the HR and treatment switching proportions in each of the 136 simulation scenarios of the 'Systematic simulation study'.

3.5.3 Results

Detailed results about the bias (absolute and proportional), MSE and coverage for each of the 136 scenarios are available in the Appendix D.

Table 3-19 gives a summary over all scenarios and also across each of the five groups. Where the underlying treatment effect for both studies and the crossover proportion are equal (Group 1), the ICs using both ITT estimates produce very little bias. Consequently, little was gained by using the adjusted estimates, and in fact the bias potentially increased when using the RPSFTM-adjusted (range in comparison to the ITT estimates). For the remaining groups, the comparisons using both RPSFTM-adjusted estimates showed a marked decrease in bias than the both ITT IC, reducing from -54.9% to 46.6% for the both ITT to -4.8% to 3.1% for the both RPSFTIM-adjusted. Similar figures for the MSE and Coverage are available in Figure 3-11 to Figure 3-15 and Figure 3-16 to Figure 3-20 respectively. Table 3-20 gives the summarised results for MSE and coverage. The MSE remains relatively consistent across groups for each of the IC-scenarios. This is particularly interesting since the MSE is a measure which balances the precision against the bias, thus the fact that they are consistent across ICs means that any decrease in bias, is being consistently matched by a loss in precision. Therefore, there is a trade-off to be made and it may be worth accepting a small amount of bias to improve the precision. It should be noted, though, that the MSE is only one way in which to contrast the bias and precision. The coverage was exceptionally variable (3.6% to 100%); almost all of the both adjusted ICs having 99% or 100% coverage, and similar results for the ITT estimates when the crossover proportion and HR were the same in both studies. Coverage has been presented for completeness, but given the highly biased nature of estimates does not prove very useful.

Figure 3-5: 'Systematic simulation' - Grouped simulation scenarios

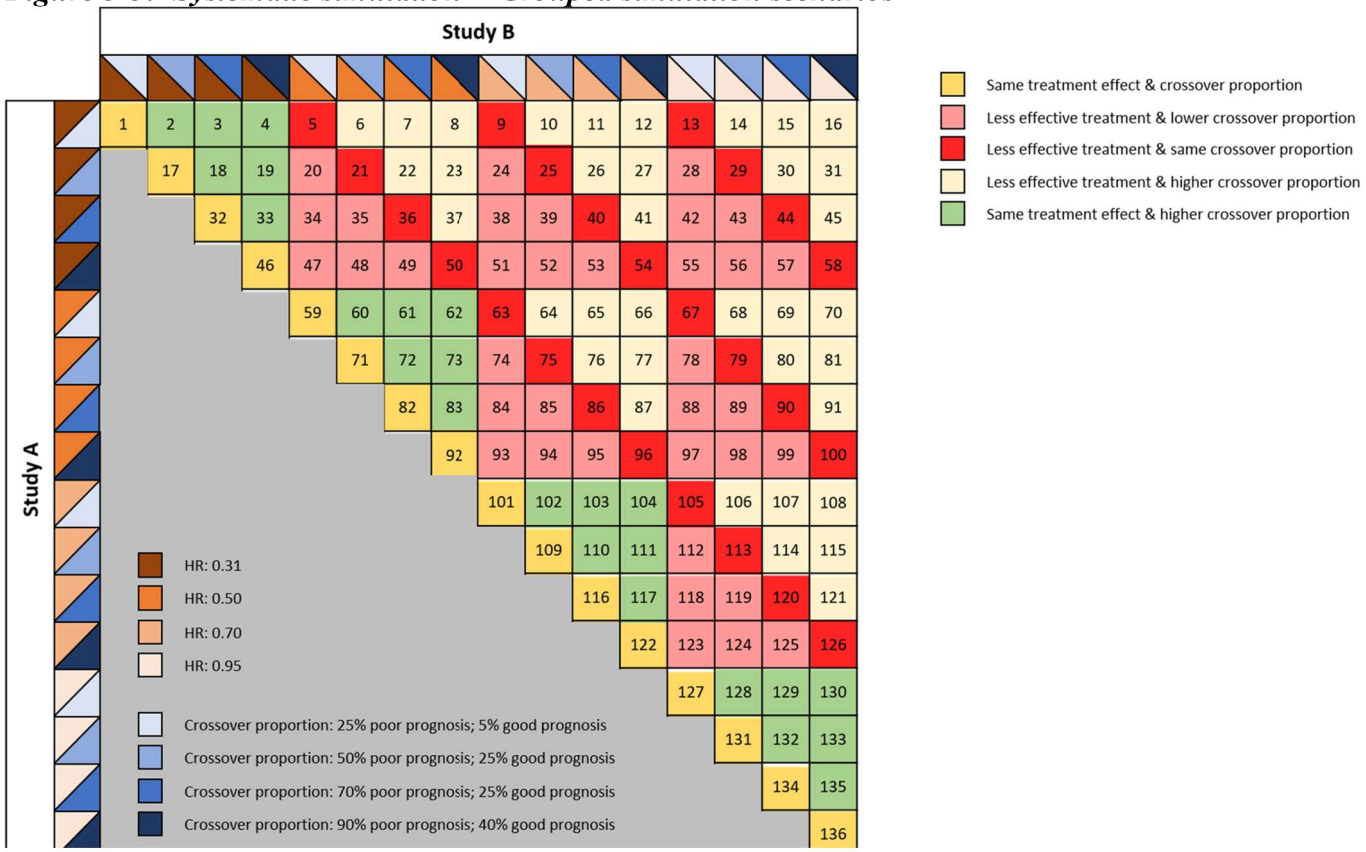


Figure 3-6: 'Systematic simulation' - Absolute bias (Group 1)

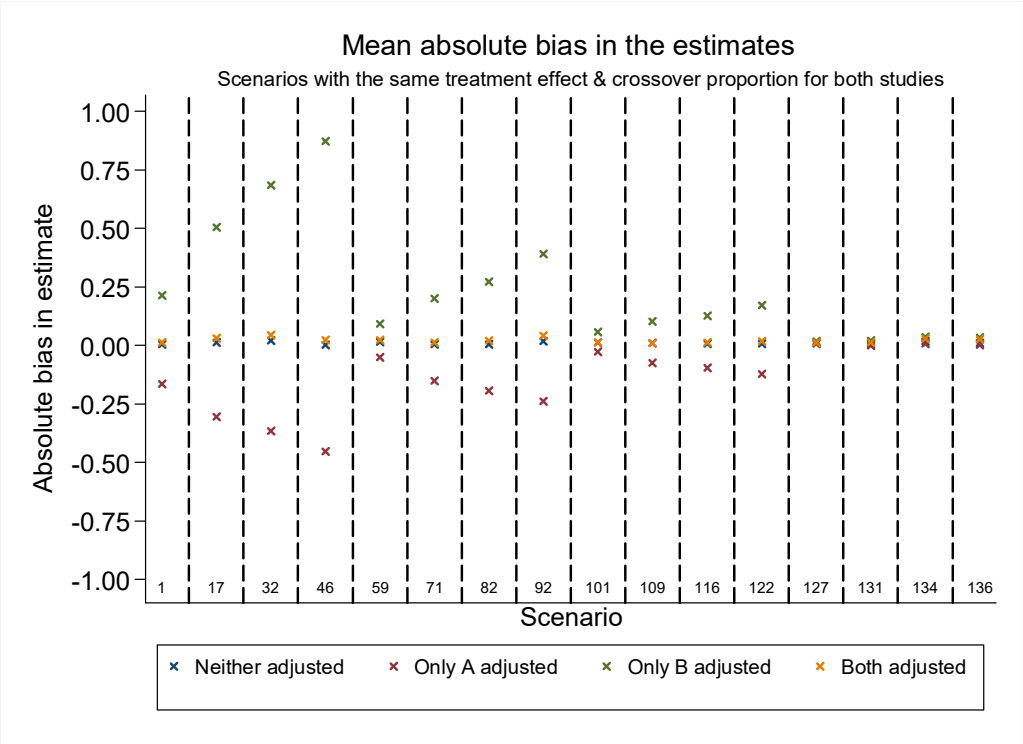


Illustration of the mean value of the absolute bias in the estimate for each scenario where both novels treatments were equally as effective and had an equal proportion of crossover.

Figure 3-7: 'Systematic simulation' - Absolute bias (Group 2)

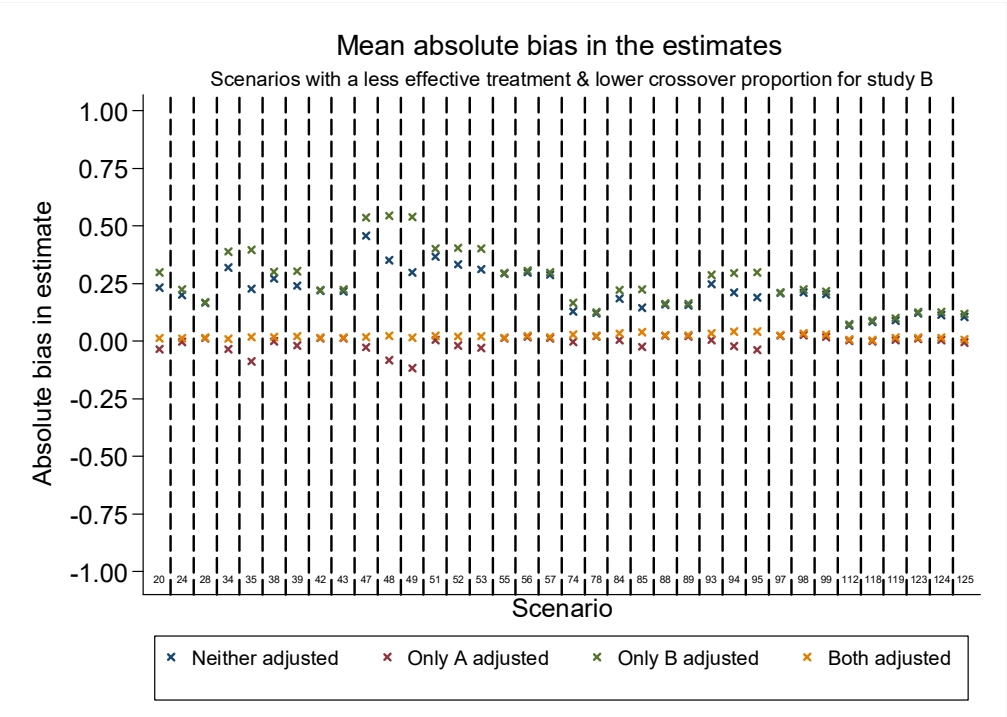


Illustration of the mean value of the absolute bias in the estimate for each scenario where study B had a less effective treatment novel treatment and a lower proportion of crossover.

Figure 3-8: 'Systematic simulation' - Absolute bias (Group 3)

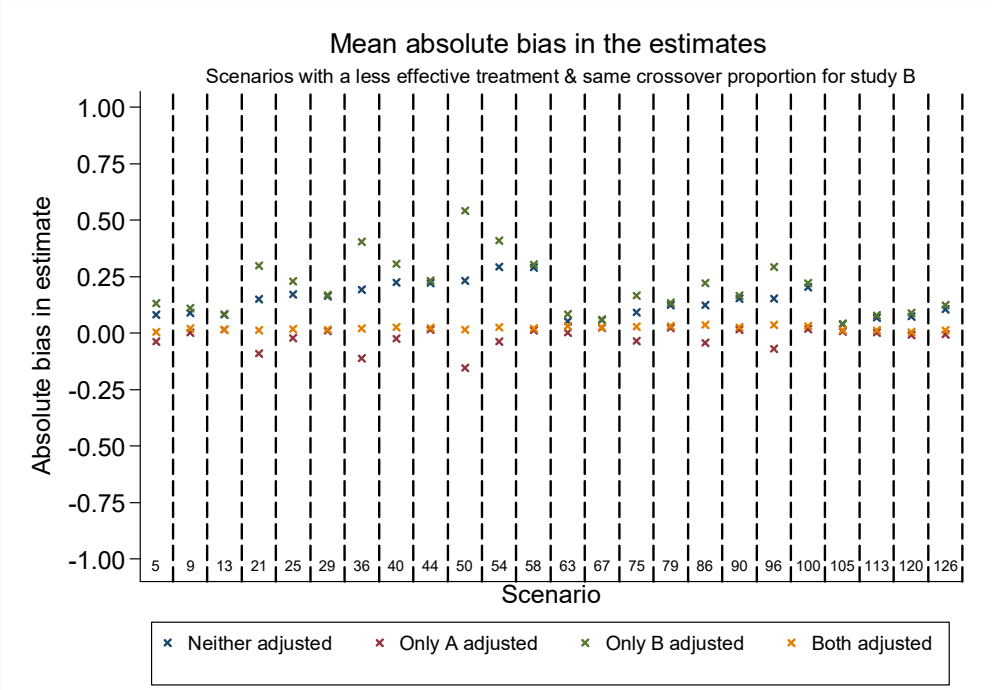


Illustration of the mean value of the absolute bias in the estimate for each scenario where study B had a less effective treatment novel treatment but both studies had the same proportion of crossover.

Figure 3-9: 'Systematic simulation' - Absolute bias (Group 4)

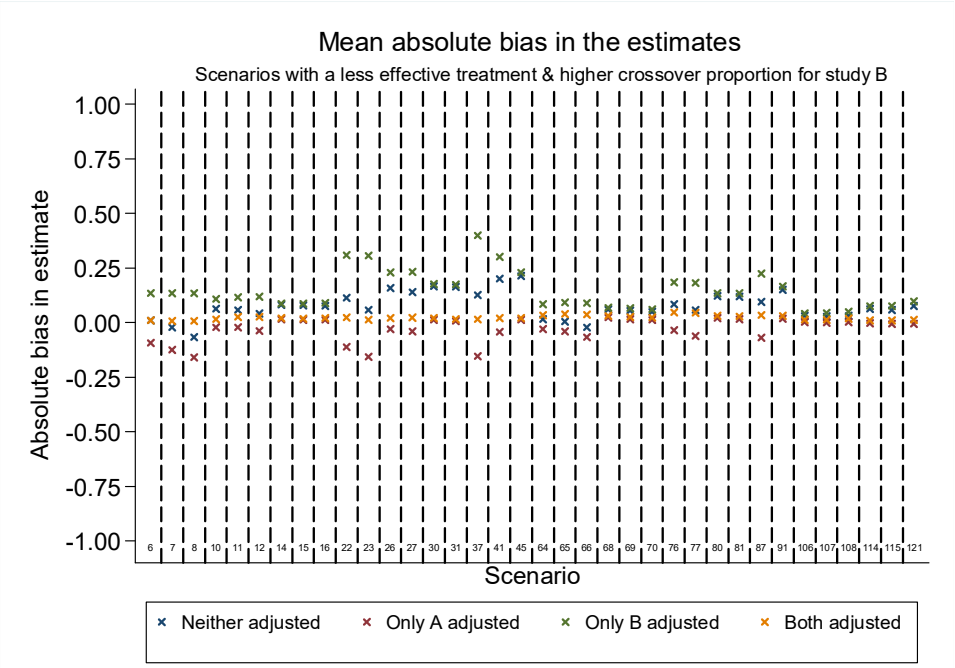


Illustration of the mean value of the absolute bias in the estimate for each scenario where study B had a less effective treatment novel treatment but a higher proportion of crossover.

Figure 3-10: 'Systematic simulation' - Absolute bias (Group 5)

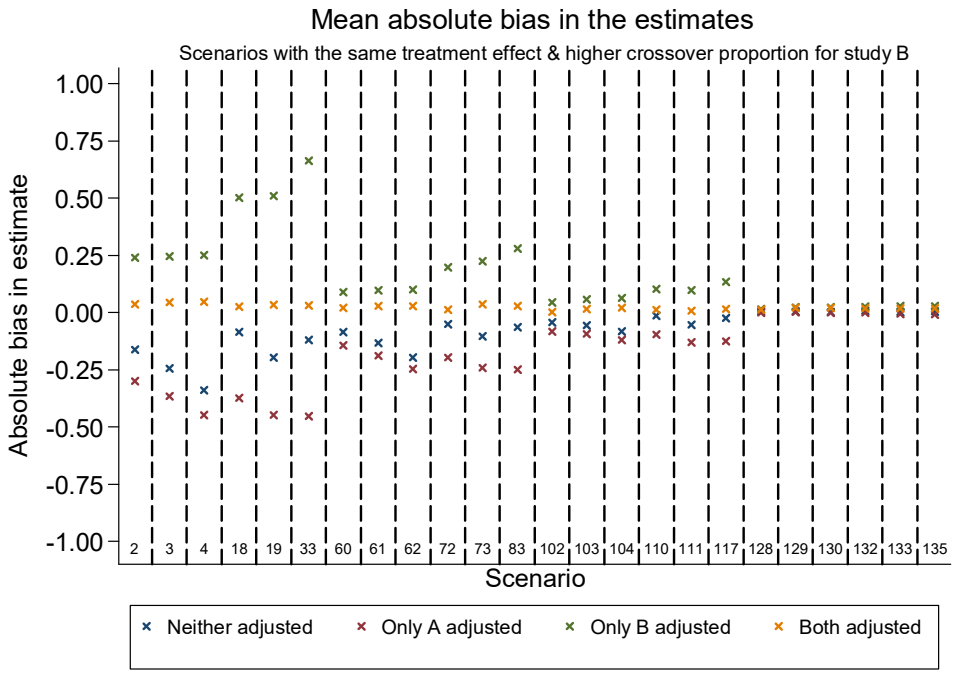


Illustration of the mean value of the absolute bias in the estimate for each scenario where both novel treatments are equally as effective but Study B has a higher proportion of crossover.

Figure 3-11: 'Systematic simulation' - MSE (Group 1)

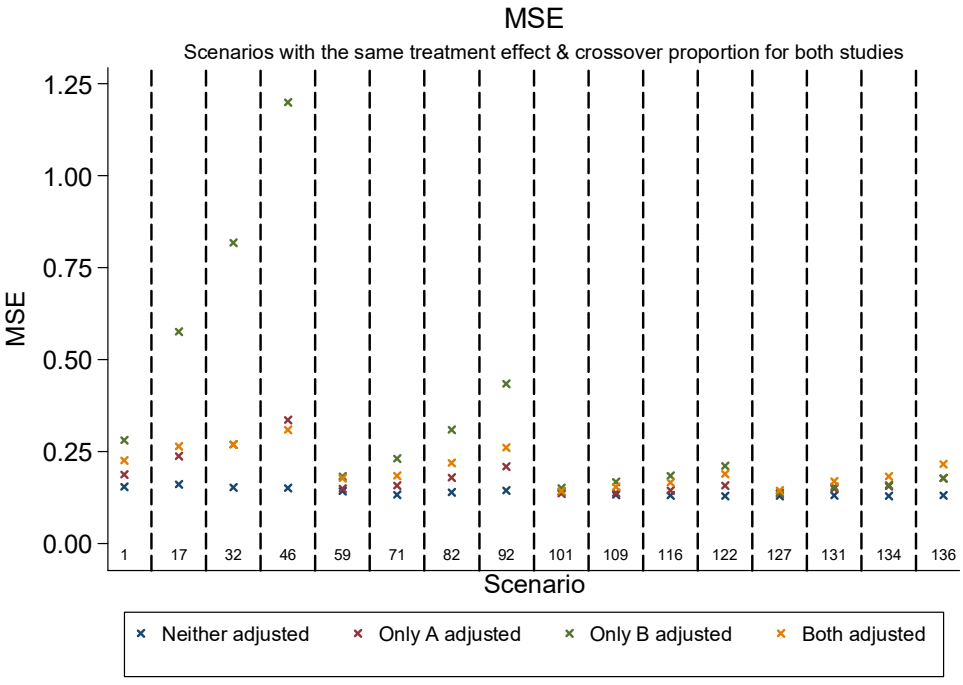


Illustration of the mean value of the MSE in the estimate for each scenario where both novels treatments were equally as effective and had an equal proportion of crossover.

Figure 3-12: 'Systematic simulation' - MSE (Group 2)

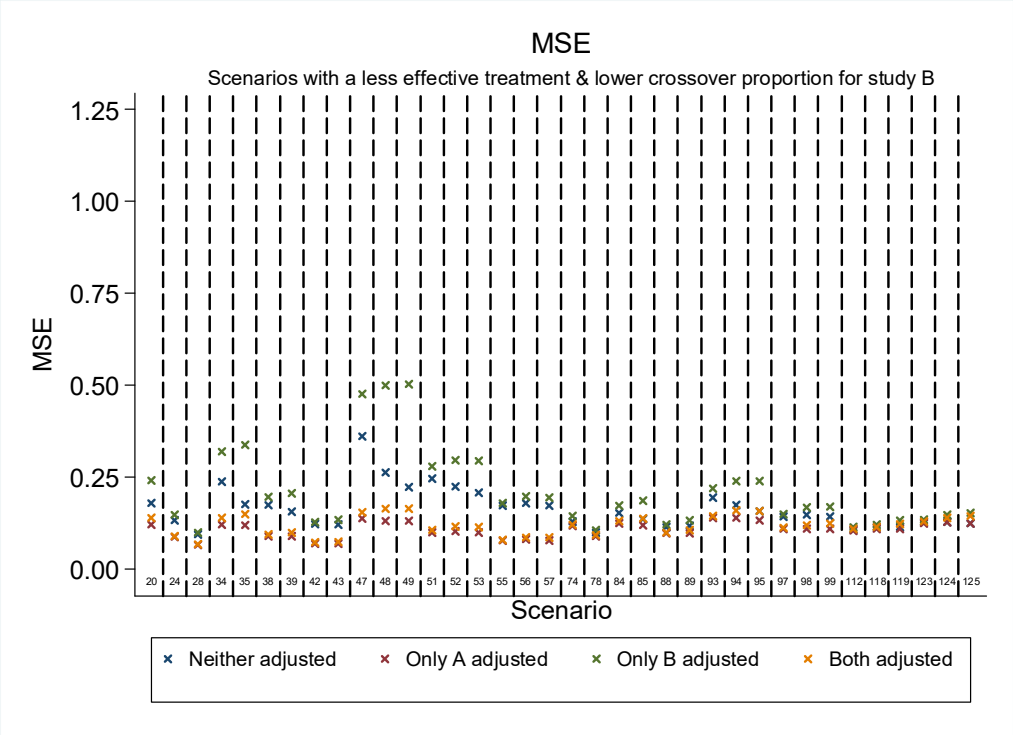


Illustration of the mean value of the MSE in the estimate for each scenario where study B had a less effective treatment novel treatment and a lower proportion of crossover.

Figure 3-13: 'Systematic simulation' - MSE (Group 3)

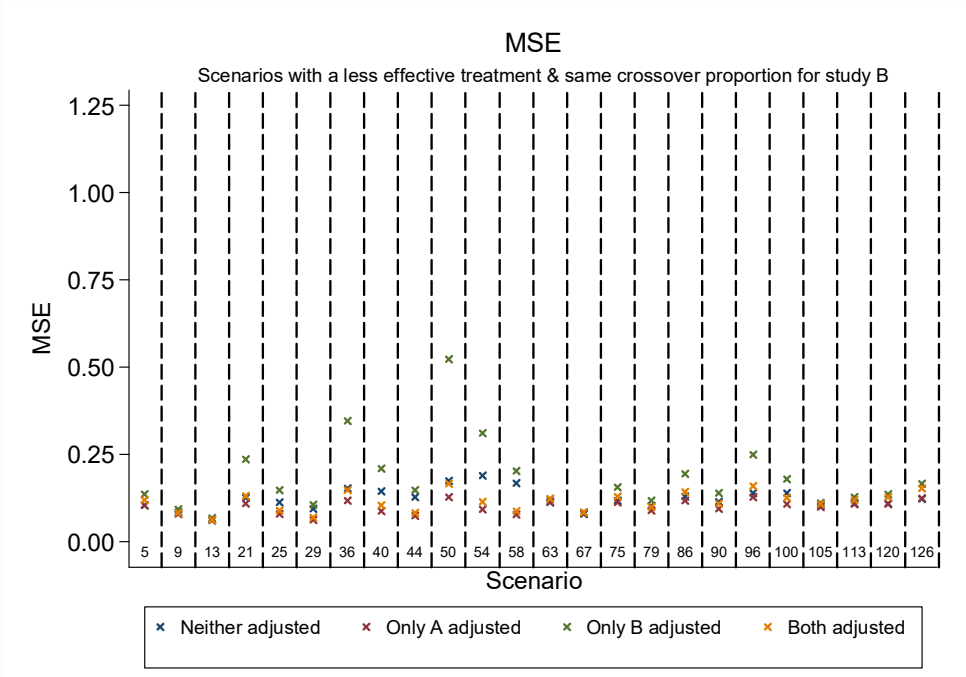


Illustration of the mean of the MSE in the estimate for each scenario where study B had a less effective treatment novel treatment but both studies had the same proportion of crossover.

Figure 3-14: 'Systematic simulation' - MSE (Group 4)

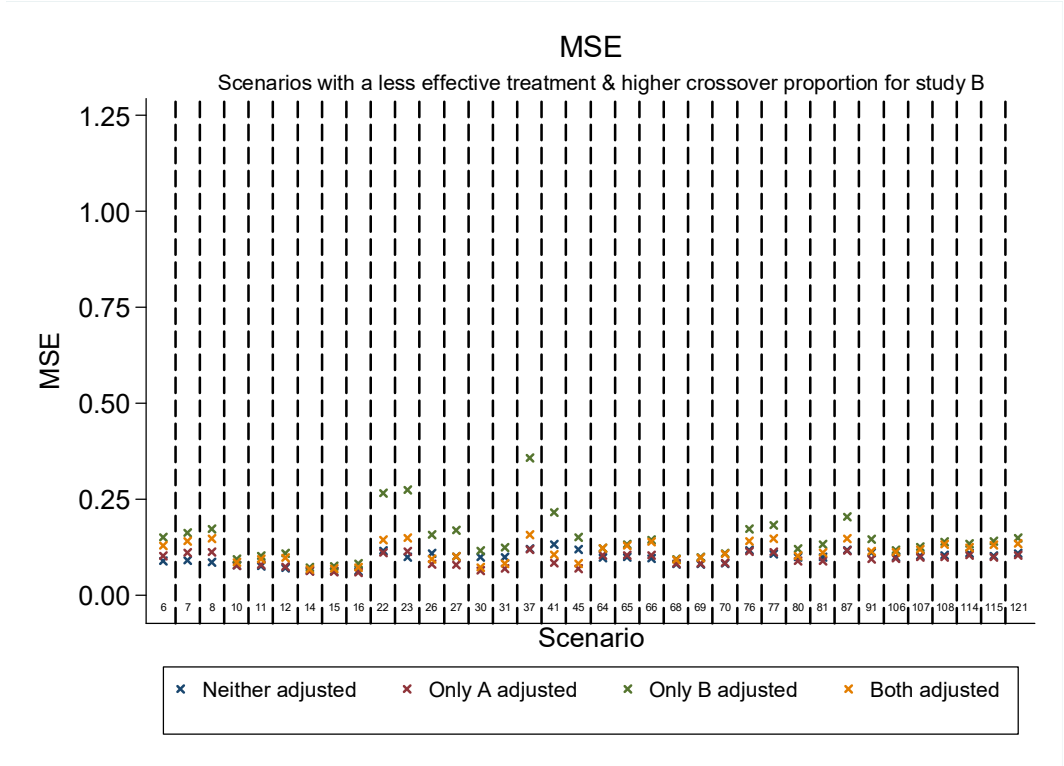


Illustration of the mean value of the MSE in the estimate for each scenario where study B had a less effective treatment novel treatment but a higher proportion of crossover.

Figure 3-15: 'Systematic simulation' - MSE (Group 5)

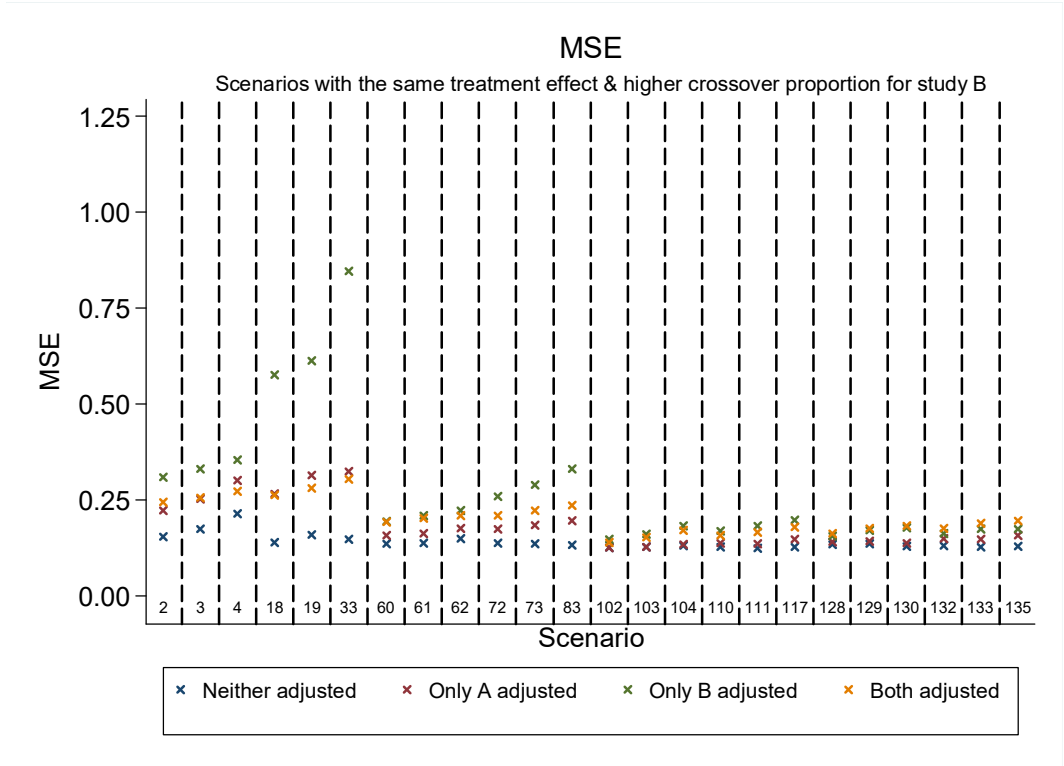


Illustration of the mean value of the MSE in the estimate for each scenario where both novel treatments are equally as effective but Study B has a higher proportion of crossover.

Figure 3-16: 'Systematic simulation' - Coverage (Group 1)

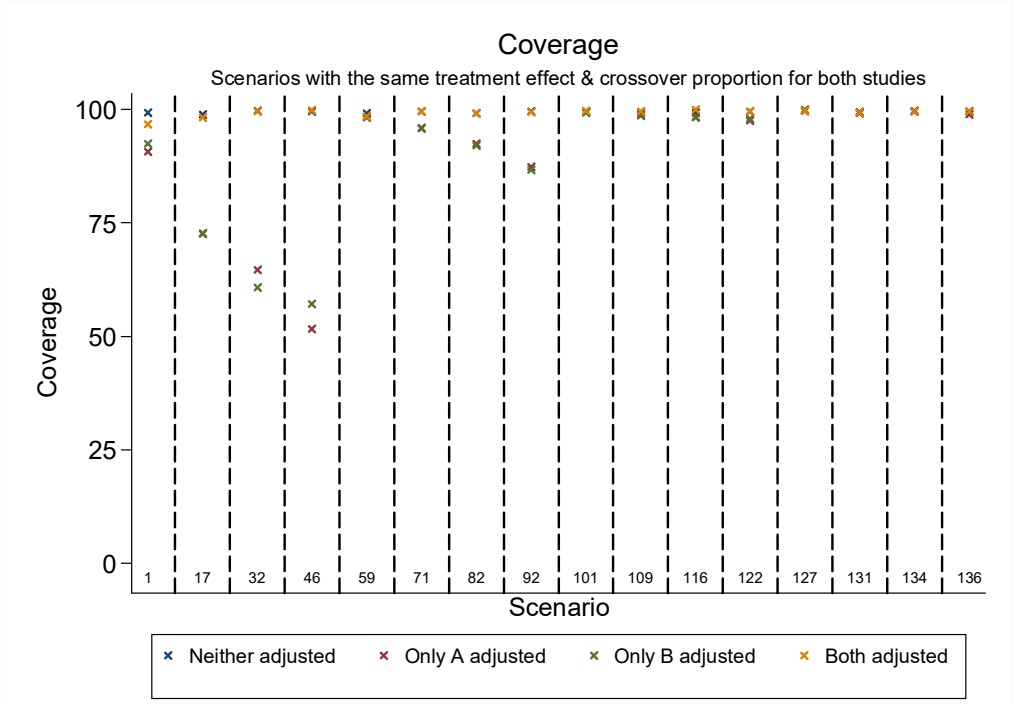


Illustration of the mean value of the coverage in the estimate for each scenario where both novels treatments were equally as effective and had an equal proportion of crossover.

Figure 3-17: 'Systematic simulation' - Coverage (Group 2)

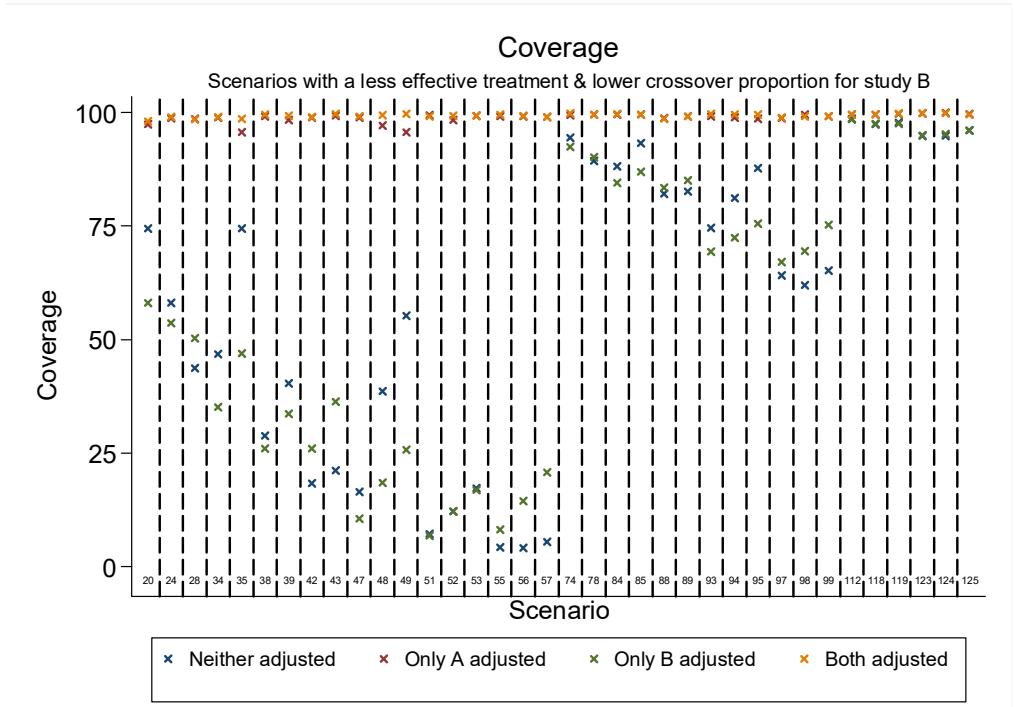


Illustration of the mean value of the coverage in the estimate for each scenario where study B had a less effective treatment novel treatment and a lower proportion of crossover.

Figure 3-18: 'Systematic simulation' - Coverage (Group 3)

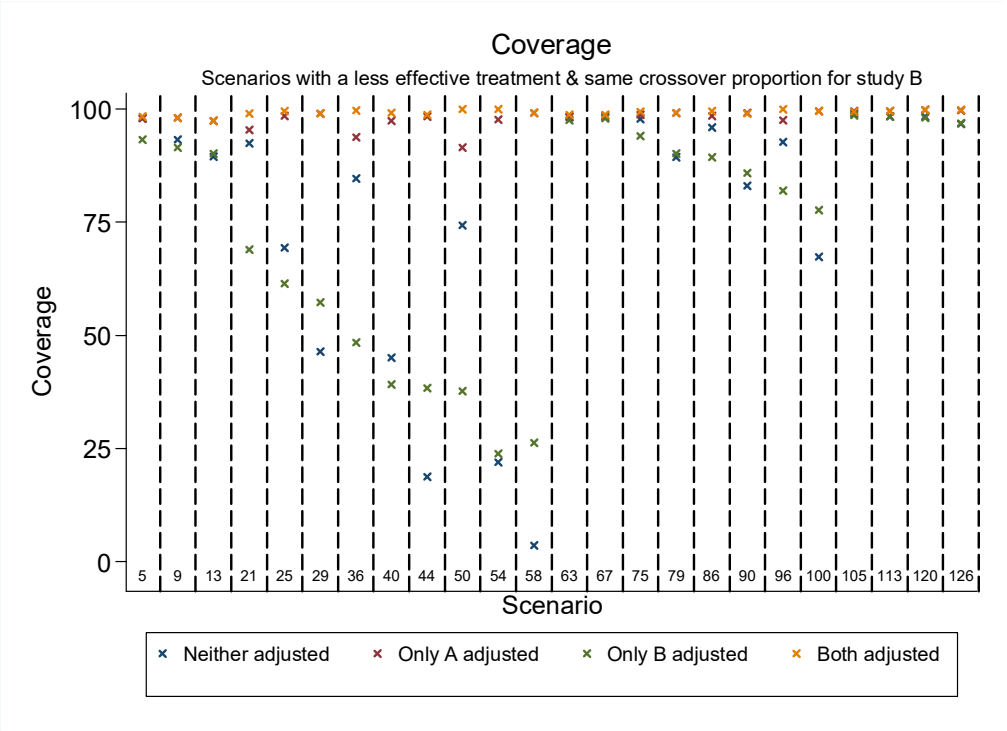


Illustration of the mean coverage in the estimate for each scenario where study B had a less effective treatment novel treatment but both studies had the same proportion of crossover.

Figure 3-19: 'Systematic simulation' - Coverage (Group 4)

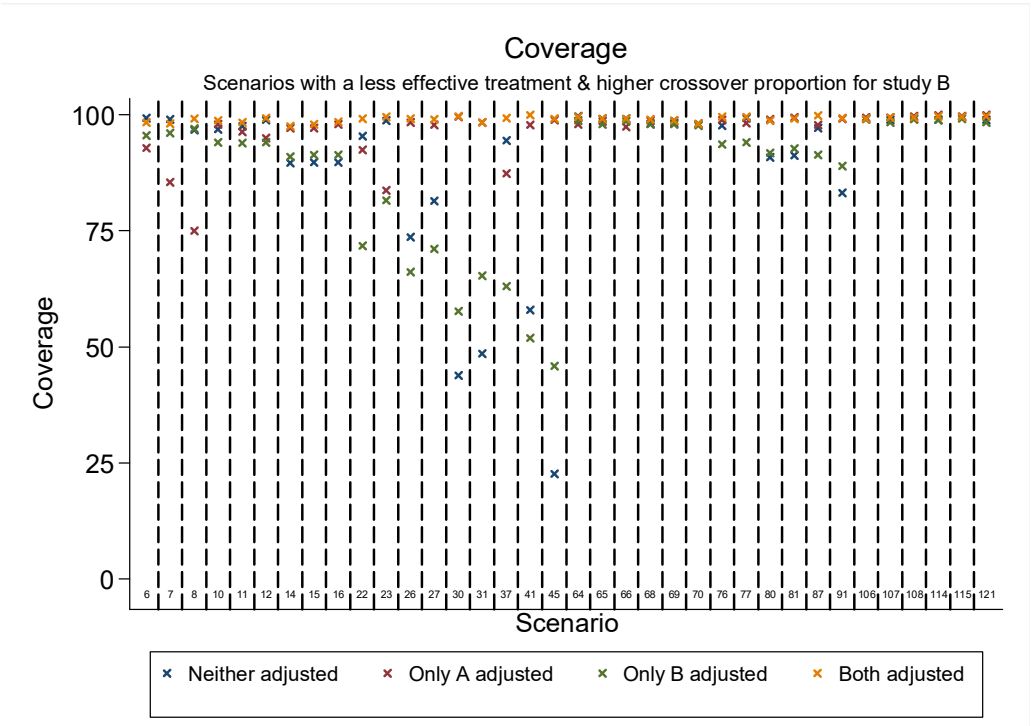


Illustration of the mean value of the coverage in the estimate for each scenario where study B had a less effective treatment novel treatment but a higher proportion of crossover.

Figure 3-20: 'Systematic simulation' - Coverage (Group 5)

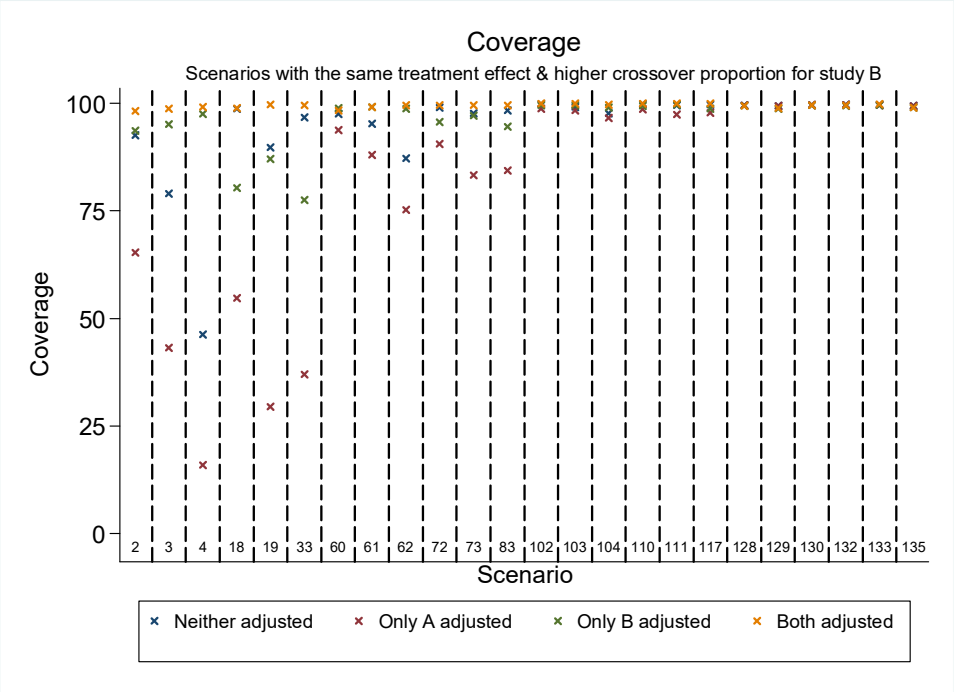


Illustration of the mean value of the coverage in the estimate for each scenario where both novel treatments are equally as effective but Study B has a higher proportion of crossover.

Table 3-19: 'Systematic simulation' - Grouped results for bias

Average, minimum and maximum bias over all the scenarios and the grouped simulation scenarios for each of the IC-scenarios.

Grouping	Statistic	Neither adjusted		Only A adjusted		Only B adjusted		Both adjusted	
		Absolute	%	Absolute	%	Absolute	%	Absolute	%
Overall	Mean	0.089	11.3%	-0.063	-13.4%	0.202	20.7%	0.021	-0.7%
	Minimum	-0.340	-54.9%	-0.454	-93.5%	0.016	-0.7%	0.001	-6.8%
	Maximum	0.456	46.6%	0.026	2.0%	0.873	47.1%	0.046	3.1%
Same treatment effect & crossover proportion	Mean	0.010	-0.9%	-0.139	-23.7%	0.237	14.1%	0.021	-2.0%
	Min	0.002	-2.1%	-0.454	-93.5%	0.017	-0.2%	0.010	-6.8%
	Max	0.020	0.1%	0.008	-1.1%	0.873	43.7%	0.044	-0.4%
Less effective treatment & less crossover	Mean	0.217	27.2%	-0.009	-5.5%	0.263	30.4%	0.021	-0.5%
	Min	0.069	7.0%	-0.117	-29.9%	0.073	7.4%	0.004	-4.2%
	Max	0.456	46.6%	0.026	1.4%	0.544	47.1%	0.042	2.1%
Less effective treatment & same crossover	Mean	0.143	19.6%	-0.021	-7.4%	0.204	24.4%	0.021	-0.1%
	Min	0.038	3.2%	-0.155	-40.0%	0.042	3.7%	0.003	-4.5%
	Max	0.292	46.0%	0.023	2.0%	0.541	46.7%	0.036	2.6%
Less effective treatment & more crossover	Mean	0.076	11.1%	-0.032	-9.5%	0.145	18.5%	0.022	0.1%
	Min	-0.068	-14.6%	-0.160	-39.3%	0.042	3.3%	0.006	-4.1%
	Max	0.213	38.4%	0.022	2.0%	0.399	39.3%	0.046	3.1%
Same treatment eff & more crossover	Mean	-0.085	-12.5%	-0.180	-30.4%	0.168	10.0%	0.023	-1.8%
	Min	-0.340	-54.9%	-0.453	-91.5%	0.016	-0.7%	0.001	-4.8%
	Max	0.008	-0.8%	0.001	-1.9%	0.663	36.5%	0.046	-0.7%

Table 3-20: 'Systematic simulation' - Grouped results for MSE and Coverage
Average, minimum and maximum results for the MSE and coverage over all the scenarios and the grouped simulation scenarios for each of the IC-scenarios.

Grouping	Statistic	MSE				Coverage			
		Neither adjusted	Only A adjusted	Only B adjusted	Both adjusted	Neither adjusted	Only A adjusted	Only B adjusted	Both adjusted
Overall	Mean	0.131	0.123	0.214	0.141	81.2%	93.9%	78.8%	99.2%
	Minimum	0.062	0.059	0.066	0.062	3.6%	15.9%	6.8%	96.7%
	Maximum	0.361	0.335	1.200	0.309	99.8%	100.0%	99.9%	100.0%
Same treatment effect & crossover proportion	Mean	0.139	0.182	0.335	0.204	99.4%	90.3%	90.5%	99.2%
	Min	0.128	0.135	0.137	0.141	98.9%	51.6%	57.1%	96.7%
	Max	0.160	0.335	1.200	0.309	99.7%	99.7%	99.9%	100.0%
Less effective treatment & less crossover	Mean	0.162	0.106	0.210	0.117	58.6%	98.9%	56.4%	99.3%
	Min	0.094	0.065	0.099	0.067	4.1%	95.6%	6.8%	98.1%
	Max	0.361	0.139	0.502	0.163	98.6%	99.9%	98.5%	99.8%
Less effective treatment & same crossover	Mean	0.121	0.097	0.179	0.113	77.2%	98.0%	74.3%	99.2%
	Min	0.062	0.060	0.066	0.062	3.6%	91.5%	23.9%	97.4%
	Max	0.188	0.127	0.522	0.165	99.0%	99.8%	98.6%	99.9%
Less effective treatment & more crossover	Mean	0.096	0.091	0.147	0.113	89.5%	96.4%	87.6%	99.1%
	Min	0.062	0.059	0.072	0.067	22.7%	75.0%	45.9%	97.6%
	Max	0.131	0.118	0.357	0.156	99.7%	100.0%	99.1%	99.9%
Same treatment eff & more crossover	Mean	0.139	0.183	0.274	0.205	94.6%	81.1%	96.1%	99.4%
	Min	0.123	0.125	0.147	0.139	46.3%	15.9%	77.6%a	98.2%
	Max	0.214	0.324	0.846	0.303	99.8%	99.7%	99.7%	100.0%

3.5.4 Conclusions

This follows a similar trend to the initial simulation. In scenarios where there is the same treatment switching effect and switching proportion for both studies, the IC using both ITT estimates gives very little bias, and extremely high coverage. However, in other scenarios the bias and coverage range quite considerably, even if the proportion of treatment switching is the same in both groups using the ITT estimates. There are also differences in the variability depending on which estimates are used in the IC. Once again, the IC using both adjusted estimates performs consistently well at representing the true HR between the two treatments of interest. As before, this will be due to the adjustment method being extremely appropriate for the data requirements; had the common treatment effect assumption been violated the bias would most likely have been greater.

3.6 Overall conclusions from the simulation studies

The key findings of this, and the previous study, are (1) that the impact of treatment switching is substantial depending on whether or not the estimates have or have not been appropriately adjusted; and (2) that the bias and coverage is highly related to the underlying treatment effect and the proportion of bias.

The simulation studies show that, theoretically, there is less cause for concern if only ITT analyses are available when the proportion of crossover and the underlying HR for each of the new treatments (compared with the common comparator) are the same. However, in reality the true HRs are unknown (and any ITT analysis will be biased to an unknown degree). Therefore, it would be very difficult to determine in practice whether an example complied with these conditions. It could actually be that the true treatment effect for one study is less than the other, and thus the bias could be considerable.

The estimates for the coverage caused the most concern, given that when both adjusted estimates were included within the IC, the coverage remained at above 97%. This occurred both in the initial and the 'systematic' simulation study findings. From the initial simulation study where the study-specific performance measures were calculated, it was possible to see some additional uncertainty than expected (e.g. coverage values of 96% or higher) which could be contributing. An alternative suggestion is that some of this could be contributable to the exclusions of the 'severity' variable in the model. Exploration has shown that the inclusion of this variable in the ITT Cox model analysis reduces the SE, potentially accounting for some of the poor coverage, the rest being due to bias. Since the p-value is preserved in the RPSFTM analysis, some of the additional variation will undoubtedly carry across to the adjusted estimates. This means that although the point estimate should be unbiased, the SE will be increased. Whilst this will have contributed towards the poor coverage using both adjusted estimates, it does not fully explain it. Greater uncertainty around the estimate would also impact on decision making, if these results are used as inputs for a probabilistic economic decision model.

At the most basic level, the findings from these simulation studies emphasise the importance of examining any additional summary data for treatment switching, when conducting an IC. It is vital to know: whether treatment switching has occurred, the

proportion of crossover, and if any adjustment has been made as these factors clearly affect the results. In light of the findings from this study, where a non-negligible proportion of treatment switching has occurred and no adjustment has been made, any results must be considered with caution.

In situations, when only the unadjusted HRs are reported, even if the amount of treatment switching is known, it is hard to draw any conclusions of how affected this estimate may be. Previous simulation studies (Morden, 2009, Latimer, 2012) have indicated that the more treatment switching there is, and the greater the true treatment effect is, the more bias toward the null hypothesis (i.e. the closer the HR is to one); but in a clinical example where the true treatment effect will not be known, the full impact of conducting an IC cannot be fully assessed. Using appropriately adjusted estimates for treatment switching, therefore, is highly recommended, even though this will result in a higher level of uncertainty around the estimate.

When adjusted HRs are available for some but not all treatments, using a mixture of adjusted and unadjusted HRs is likely to lead to a great discrepancy, and possibly high bias. The simulation studies, presented in this Chapter have highlighted the wide-ranging consequences of adopting such an approach; in particular, the treatment effect is underestimated, or overestimated depending on whether it is the more or less effective treatment that has been adjusted for.

However, as has been demonstrated, using both the unadjusted estimates is not a fair representation either, and hence, ICs should be conducted using adjusted estimates for all RCTs with treatment switching. Given the past practice in NICE TAs (discussed in Chapter 2), this would mean re-analysis of RCTs with treatment switching must occur before their inclusion into an IC. But, in order to conduct this re-analysis, the data must be available at individual patient level. For manufacturers conducting an IC with respect to a competitor's product, this is unlikely to be feasible under the STA process. Therefore, at present they would be impelled to either exclude the study or include the ITT analysis, neither of which is ideal. Research was therefore undertaken to produce a method whereby some form of adjustment could be performed only using routinely reported summary statistics.

3.7 Potential solutions

Although no formal evaluation (such as in Sections 3.3, 3.4 and 3.5) of ICs including summary data affected by treatment switching had taken place, the issue had already been raised and thus, prior to the research presented here (Chapters 4 – 6), work had been undertaken to develop a method which could be used to adjust appropriately for treatment switching which only used routinely reported information (Boucher, 2013b). For this purpose, two broad approaches were suggested: the first, to use the simulation studies (Morden, 2009, Latimer, 2012) to estimate the associated level of bias, and then adjust by this amount; the second, to reconstruct the IPD to permit existing recommended methods to be used. The methods (Chapter 4 and 5) outlined in this thesis drastically extend the earlier work described in section 3.7.2.

3.7.1 Directly adjusting the summary data

The first method, known as the 'adjustment factor method', calculated the 'adjustment factor' – one minus the proportion of bias (given the trial characteristics) – and then multiplied the HR by this factor. However, the proportion of bias could be computed in one of two ways. The first of which required the trial characteristics to exactly match one of the scenarios, then the proportion of bias for that scenario was used. The second approach fitted a linear regression to the simulations and trial characteristics, and aimed to predict the level of bias for a given set of characteristics. These methods were theoretically and computationally easy to use, and a simulation study demonstrated that the level of bias was generally reduced when compared with an ITT approach (Boucher, 2013a, Boucher, 2013b). However, the characteristics used to define the simulation scenarios are rarely known in practice, and hence made the method difficult to use routinely. Overall, the 'adjustment factors' are only useful as tools for sensitivity analysis.

3.7.2 Reconstructing individual patient level data

The alternative proposed approach was to reconstruct the IPD. A method to reconstruct individual patient survival times using coordinates extracted from a Kaplan-Meier curve had been proposed and subsequently advocated (Guyot, 2012). This method, referred to as the 'Guyot method' (described in more detail in Section 4.3.4) was, therefore, decided upon. The method effectively back transforms the time coordinates; calculates the number of events at a given time, and when the 'numbers at risk' table is presented, the number

of censorings over an interval; and finally distributes the censorings evenly over the interval or time span (if the 'numbers at risk' information is not given).

However, further difficulties remain after the IPD survival times have been generated. Although the analyst now has survival times for individual patients, there is still no information on which of these patients switched and when they switched. Consequently, this must be estimated. Essentially bootstrapping without replacement was used to allot switchers and non-switchers; thus patients are selected at random from the control group, and assigned as switchers, until the number that was reported as crossing over has been reached. This bootstrapping process is repeated multiple times, for computational purposes 100 repetitions was decided upon, in order to incorporate the uncertainty around who switches.

Approximate switch times for patients were calculated by multiplying their OS time by the ratio of median times for PFS and OS. This mechanism was also used as part of the methods in Chapter 4 and thus, the calculation of the switch time is described in more detail in Section 4.8.2.3. The data, for each of the 100 repetitions, was analysed using a recommended method, and the result recorded. All the results were then averaged over to give one estimate of the treatment effect. The simulation study showed that, whilst the bootstrapping mechanism and switch time estimation worked relatively well, reducing the bias from an ITT estimate when the percentage of crossover was high and effectiveness of the treatment strong, there were difficulties using the Guyot method (Boucher, 2013b).

The Guyot method was found to be harder to implement than expected, and examples comparing the reconstructed data with the original IPD showed many discrepancies. In addition, summary statistics calculated from the reconstructed data were at times vastly different to those reported. This caused concern for implementing the method in practice, and the key conclusions of the research relating to this approach were that other techniques for reconstructing data should be explored or developed (Boucher, 2013b).

Consequently, although the reconstructing IPD approach had tremendous potential, as it stood, none of the methods produced from this project could be effectively and

confidently used in practice to reduce the bias due to treatment switching in summary statistics. Further research was required to develop this reconstruction approach further.

3.7.2.1 *Generating the survival times*

Whilst the Guyot method (Guyot, 2012) is largely recommended, in this context, problems existed with this method, emerging largely through digitizing the curve (using a software package to digitally read and record coordinates from an uploaded copy of the Kaplan-Meier curve). This highlighted how dependent the method is on having accurate coordinates. The principle difficulty related to the type and quality of the graph. Poor quality graphs clearly impact, however, other elements such as dotted or dashed lines, exceptionally thin or thick lines make it challenging to take clearly take points. Although there is a line recognition component of the software, DigitizeIt (2013), this seldom works effectually, and it is almost always necessary to take points manually by clicking on the location the coordinate is to be taken at. However, this also induces error. In particular thick lines cause difficulties, and can lead to conflicting coordinates. It is not possible for the survival proportion to increase, nevertheless, this can easily occur within the coordinates; for example, having a 0.871 probability of survival at 25 days but a 0.872 probability at 29 days. One of these must be incorrect but in practice it is almost impossible to tell which. It could be that the cursor had been slightly to low on the first, or too high on the second. Sometimes a reasonable guess can be made depending on the coordinates either side but more often than not, it becomes an arbitrary choice made by the analyst, and thus highly subjective. It was also concerning that a different choice of coordinate could lead to a quite different dataset, as the time coordinates are back-transformed to give the survival times, and potentially result in noticeably different results. Given the sensitivity of the data it was felt that the uncertainty surrounding the reconstructed data was reflected in the final results.

3.7.2.2 *Constructing and analysing the treatment switching information*

The general approach by which treatment switching information was reconstructed is described in sections 4.8.2.2 and 4.8.2.3, as this was also utilised during the methodological development which informs that Chapter. This principle difference with the methods employed in Boucher (2013b), was the replicating of the treatment switching information 100 times, the individual analysis for each of these replications using one of

the recommended methods (namely a variation of the RPSFTM) and the averaging of the results. In addition, this project ensured that the median times from the group where treatment switching has not occurred were used to avoid bias.

Chapter 4: Reconstructing Individual Patient Level Data for Overall Survival

4.1 Chapter overview

This chapter discusses the current methods available for reconstructing IPLD for one outcome (namely OS), which is the initial step towards reconstructing and reanalysing treatment switching data. It continues by proposing a novel simulation approach, detailed in Section 4.4. This approach is illustrated by its application to a dataset for Neutron therapy, before being assessed in a reproducibility study. During the illustrative example and reproducibility study, the simulation technique is contrasted with the Guyot method (described in Section 4.3.4). Extensions needed in order to apply treatment switching methods are given and two case studies presented.

4.2 Introduction

In previous research (Boucher, 2013a, Boucher, 2013b), the approach by which IPLD was reconstructed and re-analysed indicated the greatest potential for use in practice (more details are available in Section 3.7.2.2). However, in order to promote this, the principle limitations in terms of the accuracy of the IPLD must be resolved. One solution was to consider other methods for generating survival times, which might lead to more representative data. More importantly, however, it was felt necessary to be able to reflect some of the uncertainty in the reconstruction process within the final analysis. This could be achieved by generating multiple datasets, although the most recommended methods, which include the Guyot (2012) and Hoyle and Henley (2011) methods, for data reconstruction do not particularly lend themselves to this.

It should be noted that reconstructing IPLD is not restricted purely to aiming to address treatment switching, but for other reasons as well. For example, it may be necessary to obtain various summary statistics, such as restricted mean survival time (RMST), or CEA, the primary reason which has fuelled research in this area. Alternatively, it might be that the analysis method employed already is inappropriate, perhaps if a PH model, such as the Cox (1972) or Weibull (Collett, 2003**)** model, has been applied when the PH assumption is clearly violated.

4.3 Evaluation of current methodology

A variety of approaches already exist, however, each of these have limitations, principally how they accommodate uncertainty about the data reconstruction process. They typically assume that the dataset they produce is equivalent to the IPD.

4.3.1 Notation

Before describing the methods, some notation used in this section (4.3) will be introduced. The 'numbers at risk' table is a key component for both of the more complex methods. The period of time between two values of the risk-set will be referred to as the interval, and denoted *I*. The value of the risk-set at the start of the interval will be written as n_I and the survival proportion at the start of the interval as S_I. The number of censorings and number of events occurring over the course of the interval are represented by c_I and d_I respectively.

4.3.2 Naïve approaches

For calculating statistics such as RMST, naïve approaches (Wan, 2015) are often taken. These involve extracting pairs of survival and time coordinates from the Kaplan-Meier curve, and then using a 'least squares' or a 'graphical' approach to estimate parameter value for simple survival parametric distributions, such as exponential or Weibull. If a 'least squares' approach is employed, the sum of the squares of the residuals (the difference between the model estimates and the actual data) is minimised to obtain the parameter values. If, instead, graphical methods are used, the data is transformed onto a scale, such as the log cumulative hazard (LCH) and log times scales, and then a linear regression model fitted. For values such as the RMST, the model obtained can just be integrated between the relevant limits (e.g. for the RMST at five years, integrating the survival function between zero and five). However, no particular attention is given to censoring for these methods, and they are heavily reliant on the simple survival models chosen, representing the data well. A key criticism of these approaches has been that they also do not account for uncertainty.

4.3.3 Hoyle and Henley approach

This method takes extracted coordinates at the same time points at which the numbers at risk are reported, as well as a quarter, half and three-quarters of the way through an

interval. It aims to essentially calculate the number of events and censored observations, and number at risk over each quarter of an interval. One of its key assumptions is that the censorings are distributed evenly throughout the interval, thus, if say eight censoring had occurred throughout the course of a four month interval, two would have occurred each month.

The first stage of the method is to calculate the number of deaths and number of censorings. The survival at the start of the next interval is essentially made up of the survival at the beginning of the current interval, multiplied by the probability of surviving the interval, calculated as it would be for life-tables.

$$S_{I+1} = S_I\left(1 - \frac{d_I}{n_i - \frac{1}{2}c_i}\right) \tag{4-1}$$

In addition, the number at risk at the end of the interval is essentially the number at risk at the start of the interval, minus the number of events and censorings over the interval.

$$n_{I+1} = n_I - c_I - d_I \tag{4-2}$$

Using this information, and rearranging it, the number of deaths and censorings can be computed as follows.

$$d_I = \frac{(n_I + n_{I+1})(S_I - S_{I+1})}{S_I + S_{I+1}} \tag{4-3}$$

$$c_I = \frac{2\,(S_{I+1}n_I - S_I\,n_{I+1})}{S_I + S_{I+1}} \tag{4-4}$$

These estimates will be rounded to the nearest integer.

Once these estimates have been obtained, the method then goes on to split each interval in half, calculating the same parameters as before. For this, the additional notation has been introduced: $I_{(x,y)}$ to indicate that the survival / numbers at risk at the start of, and the number of censorings and deaths during the time period from x^{th} proportion of the interval to the y^{th} proportion is being calculated. For example, $I_{\left(0,\frac{1}{2}\right)}$ means the time period from the start of interval *I* until halfway through the interval; $I_{\left(\frac{3}{4},1\right)}$ would indicate the period

from three-quarters of the way through until the end of the interval. Where the *x,y* subscript is omitted, this means that the whole interval is being considered. Therefore, the survival proportion and the number at risk can be calculated using the equations below.

$$S_{I_{\left(\frac{1}{2},1\right)}} = S_I \left(1 - \frac{d_{I_{\left(0,\frac{1}{2}\right)}}}{n_I - \frac{1}{4} c_I} \right) \tag{4-5}$$

$$S_{I+1} = S_{I_{\left(\frac{1}{2},1\right)}} \left(1 - \frac{\frac{1}{4} c_I + d_{I_{\left(0,\frac{1}{2}\right)}} + d_{I_{\left(\frac{1}{2},1\right)}}}{n_I - \frac{1}{4} c_I} \right) \tag{4-6}$$

$$n_{I+1} = n_I - c_I - d_{I_{\left(0,\frac{1}{2}\right)}} - d_{I_{\left(\frac{1}{2},1\right)}} \tag{4-7}$$

The numbers of death over these smaller time periods, can be calculated as follows:

$$d_{I_{\left(\frac{1}{2},1\right)}} = \left(S_{I_{\left(\frac{1}{2},1\right)}} - S_{I+1} \right) \left(\frac{\left(S_{I_{\left(\frac{1}{2},1\right)}} n_I + S_{I_{\left(\frac{1}{2},1\right)}} n_{I+1} \right) + 2\, S_I\, n_{I+1}}{S_{I_{\left(\frac{1}{2},1\right)}} S_I + S_{I_{\left(\frac{1}{2},1\right)}} S_{I+1} + 2\, S_I\, S_{I+1}} \right) \tag{4-8}$$

$$d_{I_{\left(0,\frac{1}{2}\right)}} = n_I + 3 n_{I+1} - \left(3\, S_{I+1} + S_{I_{\left(\frac{1}{2},1\right)}} \right) \left(\frac{\left(S_{I_{\left(\frac{1}{2},1\right)}} n_I + S_{I_{\left(\frac{1}{2},1\right)}} n_{I+1} \right) + 2\, S_I\, n_{I+1}}{S_{I_{\left(\frac{1}{2},1\right)}} S_I + S_{I_{\left(\frac{1}{2},1\right)}} S_{I+1} + 2\, S_I\, S_{I+1}} \right) \tag{4-9}$$

Based on all of the above:

$$c_I = n_I - n_{I+1} - d_{I_{\left(0,\frac{1}{2}\right)}} - d_{I_{\left(\frac{1}{2},1\right)}} \tag{4-10}$$

And,

$$n_{I_{\left(\frac{1}{2},1\right)}} = n_I - d_{I_{\left(0,\frac{1}{2}\right)}} - \frac{1}{2} c_I \tag{4-11}$$

Having achieved this, for greater accuracy each half-interval is split again. By the same principles used to estimate expressions 4-5 to 4-11 this means that,

$$d_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} = \left(S_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} - S_{I_{\left(0,\frac{1}{2}\right)}}\right)\left(\frac{S_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} n_I + S_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} n_{I\left(\frac{1}{2},1\right)} + 2\, S_{I_{\left(\frac{1}{2},1\right)}} n_{I+1}}{S_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} S_I + S_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} S_{I_{\left(\frac{1}{2},1\right)}} + 2\, S_I\, S_{I_{\left(\frac{1}{2},1\right)}}}\right) \quad (4\text{-}12)$$

$$d_{I_{\left(\frac{3}{4},1\right)}} = \left(S_{I_{\left(\frac{3}{4},1\right)}} - S_{I+1}\right)\left(\frac{S_{I_{\left(\frac{3}{4},1\right)}} n_{I_{\left(\frac{1}{2},1\right)}} + S_{I_{\left(\frac{3}{4},1\right)}} n_{I+1} + 2\, S_{I_{\left(\frac{1}{2},1\right)}} n_{I+1}}{S_{I_{\left(\frac{3}{4},1\right)}} S_{I_{\left(\frac{1}{2},1\right)}} + S_{I_{\left(\frac{3}{4},1\right)}} S_{I+1} + 2\, S_{I_{\left(\frac{1}{2},1\right)}} S_{I+1}}\right) \quad (4\text{-}13)$$

For simplicity, deaths in the first quarter, and third quarter are calculated as follows.

$$d_{I_{\left(0,\frac{1}{4}\right)}} = d_{I_{\left(0,\frac{1}{2}\right)}} - d_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} \quad (4\text{-}14)$$

$$d_{I_{\left(\frac{1}{2},\frac{3}{4}\right)}} = d_{I_{\left(\frac{1}{2},1\right)}} - d_{I_{\left(\frac{3}{4},1\right)}} \quad (4\text{-}15)$$

Based on the assumption mentioned earlier about evenly distributing censored observations, the following statement holds.

$$c_{I_{\left(0,\frac{1}{4}\right)}} = c_{I_{\left(\frac{1}{4},\frac{1}{2}\right)}} = c_{I_{\left(\frac{1}{2},\frac{3}{4}\right)}} = c_{I_{\left(\frac{3}{4},1\right)}} = \frac{c_I}{4} \quad (4\text{-}16)$$

In other words, the censoring for each quarter is exactly the same e.g. the total censorings over the interval, divided by 4.

When these estimates have been obtained, a model is fitted to the data, with the parameters estimated by maximising the likelihood. Suggested models include exponential, Weibull, logistic, log-normal and log-logistic distributions.

Tierney (Williamson, 2002) has proposed an amended version of the method for use if the 'numbers at risk' table has not been reported.

4.3.4 Guyot approach

This approach predominantly uses the theory behind the calculation of the Kaplan-Meier curve. It involves extracting pairs of survival and time coordinates at each distinct event

time. The coordinates are typically extracted by uploading a scanned copy of the Kaplan-Meier curve into a digitizing software package, such as DigitizeIt (mentioned in Section 3.7.2.1). The software permits the analyst to click on the location of an 'event' (e.g. each step of the curve), and easily record its coordinates. This list of coordinates can then be transported into a statistical software package, and manipulated to reconstruct IPLD. Since, the Kaplan-Meier curve, up to the k^{th} distinct event time, is calculated as follows:

$$S(t_k) = \prod_{z=0}^{k} p_z\,, \qquad p_z = 1 - \frac{d_z}{n_z} \tag{4-17}$$

Where d_z, n_z are the number of events and numbers at risk at time t_z; then, the number of events calculated at a given event time is:

$$d_k = n_k\left(1 - \frac{S(t_k)}{S(t_{k-1})}\right) \tag{4-18}$$

As it currently stands, censoring is being ignored (or assumed to occur at the very end of the timescale). However, in reality, censoring is likely to have occurred throughout the whole timescale, and as such will need to be accounted for. Where possible, if the numbers at risk are available, these should be used to calculate the number of censored observations throughout the intervals. This is obtained using equation (4-19):

$$c_I = \left(\frac{S_{I+1}}{S_I}\right) n_I - n_{I+1} \tag{4-19}$$

This value is rounded to the nearest integer.

The censorings are assumed to happen evenly throughout the interval, and so for the m^{th} person of M people censored in interval I, the censoring time occurs at:

$$t_{cens_{Im}} = t_I + m\left(\frac{t_{I+1} - t_I}{M+1}\right) \tag{4-20}$$

Once the censoring times have been allocated, these are then used in the 'number at risk' needed to back-transform the survival estimate, and calculate the number of events at that specific time point. When the number of events have been calculated over the entire interval, the 'number at risk' for the reconstructed dataset is computed and compared with that of the IPD. If there are differences, this difference is added to the number of

censorings over the interval, and the process of reconstructing the data for that interval repeated. The estimates for intervals are then summed, and compared with the actual number of censorings (if the number of events has been reported), and adjusted accordingly. Where the number of events is not available, the initial estimates are assumed to be accurate, and no re-estimation is done. Without the number of events and 'numbers at risk' table, this method assumes that all patients experienced the event of interest.

4.3.5 Limitations of current approaches

The three broad approaches for facilitating reanalysis of time-to-event data vary considerably in their theory and implementation. Recent studies that have compared these methods have concluded that the Hoyle and Henley and the Guyot approaches typically perform better than naïve techniques and result in lower levels of bias (Wan, 2015). However, there are still limitations with these, principally in terms of the censoring distribution. Both methods assume evenly distributed censored observations which may or may not be valid. In addition, the IPLD created by these approaches is treated as being equivalent to the original IPD; in that the results obtained from this data are assumed equivalent to those from the IPD. However, there is the potential for error in the dataset. The censoring assumptions may not be appropriate, the coordinates may well contain measurement error, to mention some potential issues. This is particularly true for the Guyot method, since this approach relies on obtaining accurate values of the coordinates at every distinct event time. As previously described in detail in Section 3.7.2.1, (based on Boucher (2013b)), this is difficult to achieve in practice, and is impacted on by the quality of the scanned survival curve. In large sample sizes with many events, event times can be difficult to isolate since the step function can appear more of a smooth continuous curve. Also, dotted line styles can make it hard to determine the event time. Thick line styles complicate the data extraction process as there is more room for variation and error. Since the survival function is monotonic, there must not be any inconsistencies within the extracted coordinates. E.g.,

- the survival proportion cannot exceed any previous estimate of the survival;
- there should not be two different estimates of the survival proportion at the same time;
- the survival proportion must be bounded between zero and one;

- the time coordinates must be greater than zero.

Nevertheless, it is difficult to ensure that all these conditions are met. Thus, if any are violated, coordinates are typically removed subjectively (e.g. the analyst removes the estimates they believe to be the inaccurate ones). For these reasons, it seemed important that uncertainty within the data reconstruction process is taken into account during the analysis.

4.3.6 Rationale for a simulation approach

To combat the current problems that had been associated with reconstructing IPLD (in particular extracting exact event times, and relying on a single dataset), a new 'simulation approach' was developed. This novel technique draws on some of the methodology behind the naïve methods and the Guyot approach. Like naïve model fitting approaches, this method models the survival, using coordinates extracted from the survival curve (as is used for both the Guyot and Hoyle and Henley approaches). An important extension is also ensuring flexible models are used (e.g. no restriction to exponential or Weibull models). Also, in contrast to the naïve methods, a censoring distribution is modelled in addition to the survival curve. For this novel simulation approach, a similar technique to that applied in the Guyot method is used to generate the censoring distribution. Survival times and censoring times are then simulated from the respective models, before being combined to produce a single complete dataset. This method deviates most from other approaches; since, rather than only producing a single dataset, multiple datasets are simulated from the models and analysed individually. To summarise the datasets overall, an average is taken across all the individual results. A key difference between the simulation method and the Hoyle and Henley or Guyot approaches, therefore, is that this method cannot claim to create a single dataset almost identical to the original IPD. This is because the simulated survival and censoring times come from models. However, long-term the average point estimate over multiple simulated datasets will tend towards the true underlying value, provided that the models are a reasonable fit.

Should this method prove successful, this could also provide a solution to generating the treatment switching information. If the OS could be reconstructed effectively, then so long as the PFS Kaplan-Meier was given, PFS could also be constructed. However, OS and PFS would need to be related to each other, during the simulation. Given that this could be developed, this might also address issues surrounding who switched, as

assuming patients only crossover on progression, there would be more information on patients who did (or did not progress), which could be used to indicate prognosis. These combined solutions would strengthen the model considerably. The aim therefore became to first develop a technique whereby IPLD could be simulated using coordinates extracted off a Kaplan-Meier curve, and then once this had been established, extending this method to simultaneously simulate PFS and OS (explained in Chapter 5).

4.4 Simulation approach

4.4.1 Outline of method

The process of simulating IPLD comprises of ten key stages. These stages are illustrated graphically in Figure 4-1, and listed below.

1. Extract coordinates for the survival proportion and time from the scanned survival curve for each treatment arm;
2. Transform the extracted coordinates onto the LCH and log time scales;
3. Model the survival distribution for each treatment arm, using restricted cubic splines (RCS);
4. Construct a censoring distribution based on the information reported;
5. Simulate survival times for each patient from the survival model;
6. Simulate censoring times for each patient from the censoring distribution;
7. Define the observed survival time and status of each patient; by taking the minimum of the times obtained in steps 5 and 6.
8. Analyse the dataset using the analysis method of choice, recording the results;
9. Repeat stages 5 – 8 to produce multiple datasets, as this incorporates uncertainty around the reconstruction process;
10. Calculate an average of the recorded results to obtain a final point estimate.

Figure 4-1: Simulation Technique Reconstruction Process

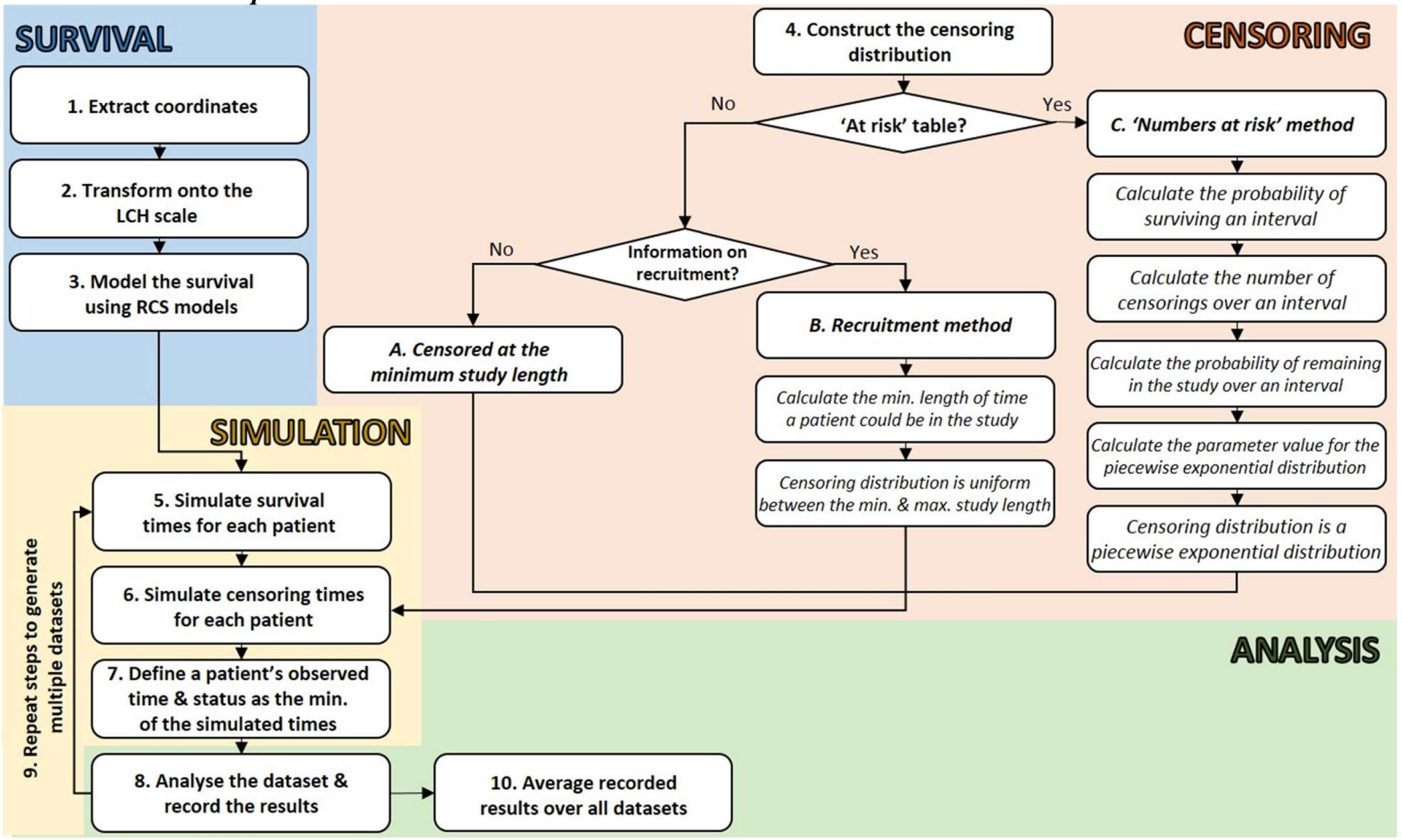


Diagram of the process of reconstructing the IPLD using a simulation approach, showing the four key phases: creating the survival distribution; creating the censoring distribution; simulating the dataset; and the analysis.

4.4.1.1 *Stage 1: Extracting the data*

As with the Guyot (2012) and the Hoyle and Henley (2011) methods, this approach begins by obtaining estimates of the survival proportion at a number of different time points. To do this effectively, scanned Kaplan-Meier curves should be transported into a digitizing software package. Unlike the other data reconstruction approaches, rather than at exact event times, here the aim is only to capture the shape of the survival curve, so that the survival distribution can be modelled. This does not require exact event times. The reason for this choice is that, in practice, it can be exceptionally hard to accurately extract coordinates at each distinct event time. This is particularly true for large trials with many events (resulting in a smooth-looking function instead of a clear step-function), having many event times close together, and line styles, such as thick or dotted lines.

4.4.1.2 *Stage 2: Transformation of the data on to the log cumulative hazard scale*

The extracted coordinates are then transformed, in preparation for the modelling process, and to ensure that the subsequent model has an intuitive interpretation. The survival coordinates are transformed onto the log cumulative hazard (LCH) scale, whilst the values for the time are transformed onto the log scale. This choice of transformation means that, if a linear function is fitted, the resulting survival distribution is the commonly used Weibull model. Consequently, if this simple linear model is extended to encompass more complex terms, these additional terms can be interpreted as merely modelling the departure from linearity.

4.4.1.3 *Stage 3: Modelling the survival distribution*

The choice of model is crucial to obtaining representative survival data. As discussed in Section 4.4.1.2, using a linear model on the transformed (LCH) scale, would give a Weibull model. However, the data needing be reconstructed often have quite complex non-linear shapes, which require more complex models. Using polynomial terms is not necessarily appropriate, since in practice these often contain turning points, whereas the survival distribution must be strictly monotonic. Therefore, the application of RCS for log time within the model has been selected. Whilst these models are not restricted to monotonic functions, usually given the nature of the data, (e.g. the true survival function will never exhibit a turning point), they remain monotonic. On the rare occasion that it is non-monotonic, astute choices for knot locations and / or use of additional coordinates

can correct this. The splines functions ensure a flexible model, and since the LCH scale has been used, this model is equivalent to the flexible parametric survival model (Royston and Parmar, 2002).

In terms of the practical implementation, RCS functions are computed from the log time variable. These are then included within an ordinary least squares regression, with the LCH variable as the outcome, and the splines functions as the exposure. The resulting model can then be written in the form:

$$\ln(H) = S(\ln t \mid k, \alpha) \tag{4-21}$$

Where H is the hazard function, t time, k the knots and α the coefficients.

4.4.1.4 *Stage 4: Modelling the censoring distribution*

This stage differs quite considerably from the alternative data reconstruction methods. It should be remarked that the censoring distribution is more difficult to construct than the survival function, since the majority of the information about censoring is given implicitly, and its level varies between studies. Three different techniques by which a censoring distribution can be constructed have been proposed, based on the data available. The information these methods use are:

A) the maximum study length;
B) recruitment and follow-up times; and
C) 'Numbers at risk' table.

Of these three, method C (using the 'Numbers at risk' table) is the most preferable.

4.4.1.4.1 Approach A: Maximum study length

This technique relies on the smallest information possible, and can be applied in the absence of any other information than the Kaplan-Meier curve. The maximum study length can either be determined from the publication, if perhaps the survival is reported for a fixed length of follow-up (e.g. 6-month, 1-year survival, etc.), or estimated as the maximum value of the Kaplan-Meier curve, regardless of whether it is a censored observation or an event). This estimated maximum time is used to censor any patients whose simulated survival time exceeds this value. One important assumption of this method is, however, that patients cannot be censored earlier than this time point, e.g. if withdrawing from the study or being lost to follow-up (LTFU).

4.4.1.4.2 Approach B: Using recruitment and follow-up times

This second approach requires the dates of recruitment and the data cut-off date to have been reported in the publication. This technique makes two key assumptions: that of administrative censoring only (patients must remain in the study, other than leaving due to an event, for a minimum length of time which will be described below; e.g. they cannot leave because of withdrawal of consent, or being LTFU); and secondly by assuming that patients enrol in the study uniformly throughout the recruitment period.

This method works by calculating the minimum length of study time any patients who do not experience an event must have remained in the trial: this is essentially the length of time between end of recruitment and the data cut-off dates. To create the variation within the censoring times, the maximum study length also needs to be determined: this is the difference between the data cut-off date and the start date of recruitment. Once these have been ascertained, censoring times can then be simulated uniformly between these values. In other words, denoting the censoring distribution as C, and t_{max}, t_{end} as the time between the start of the randomisation and the data cut-off, and the time between the start and end of the randomisation period respectively; then $C \sim \text{Uniform}(t_{\text{max}} - t_{\text{end}},\ t_{\text{max}})$.

Once again, this method imposes a distribution for administratively censoring patients. Approach A (maximum study length) censoring is essentially a special case of this method, in which all patients were randomised at the same time (recruitment / randomisation for all patients was one day or less), such that $t_{\text{end}} = 0$.

4.4.1.4.3 Approach C: Using the 'Numbers at risk' table

This approach employs a similar technique to that used for constructing the censoring distribution in the Guyot method; that is to estimate the number of patients who were censored during a given interval. In order to do this, however, it is imperative that the 'number at risk' table has been reported (e.g. as illustrated in Figure 4-7). The advantage of this approach is that it can take account of other types of censoring e.g. LTFU, withdrawing consent, rather than just administrative.

To calculate the number of censored patients for a particular interval, the following formula, used to construct lifetables, is relied on:

$$p_i = 1 - \frac{d_i}{n_i - \frac{1}{2}c_i} \tag{4-22}$$

Where d_i, c_i denote the number of events and censorings over a particular interval, *I*, respectively and n_i , the number at risk at the start of the I^{th} interval.

Sometimes, the denominator shown in equation (4-22), is referred to as the 'effective number at risk'; this essentially is the 'number at risk' taking account of those patients who leave over the course of the interval due to censoring. This formula specifically assumes that censoring occurs evenly throughout the study, by multiplying the number of censorings by a half; this means that censored patients only actually contribute towards the risk-set for half the interval.

Equation (4.22) can be re-arranged to obtain an expression for c_i

$$c_i = 2\left(n_i - \frac{d_i}{1 - p_i}\right) \tag{4-23}$$

However, expressions for d_i or p_i do not yet exist, and thus, these must be estimated.

For the method to perform well, accurate estimates for the survival at the 'numbers at risk' time points are required. Therefore, it should be ensured that the survival coordinate nearest to these time points is an appropriate estimate. If this is the case, then:

$$p_i = \frac{S(t_i)}{S(t_{i-1})} \tag{4-24}$$

In terms of d_i, consideration must be given to how the difference in the 'number at risk' (denoted y_i) over the interval is composed. Since there are only two ways in which a patient can leave the risk set, by either having an event or being censored, then:

$$y_i = d_i + c_i \tag{4-25}$$

Since y_i can be estimated directly from the data, and *c* is the parameter of interest, then

$$c_i = 2\left(n_i - \left(\frac{y_i - c_i}{1 - p_i}\right)\right) \tag{4-26}$$

Re-arranging this, estimates for the number of censored observations, $\hat{c}_i$, can be obtained using the following expression:

$$\hat{c}_i = \frac{2\big(y_i - n_i(1-p_i)\big)}{p_i + 1} \tag{4-27}$$

Nevertheless, assuming the total number of censored patients (over the whole time period) is reported or can be calculated, which is represented by *C*. Then, there may be occasions where:

$$C \neq \sum_{i=1}^{I} \hat{c}_i \tag{4-28}$$

And *I* is the last interval.

In these cases, to ensure the simulated data should contain approximately the right number of censorings on average, the estimates for each interval are scaled accordingly based on equation (4-29).

$$\hat{c}_i{}' = \frac{C}{\sum_1^I \hat{c}_i}\ \hat{c}_i \tag{4-29}$$

Where $\hat{c}, \hat{c}'$ are the initial and revised estimates for the number of censored observations respectively.

This is perhaps where generating the censoring distribution using a simulation approach deviates most from other methods. Instead of allocating patients to censoring times evenly spaced throughout the timescale, a piecewise exponential model has been chosen, where each interval has its own hazard rate. This hazard rate essentially depends on the probability of remaining in the study and the length of the interval.

$$\lambda_i = -\frac{\ln\left(\dfrac{\hat{c}_i{}'}{n_i - \dfrac{1}{2}(y_i - \hat{c}_i{}')}\right)}{l_i} \tag{4-30}$$

Where l_i is the length of the i^{th} interval.

By using the piecewise exponential models, the assumption that the hazard of being censored remains constant over the course of an interval is being made, but can vary between intervals.

4.4.1.5 *Stage 5: Simulating from the survival distribution*

The next stage of the method is to generate a time for each patient; this can be achieved by simulating from the survival model. But, first, the survival distribution model (given in Equation (4-21) must be rearranged in terms of *t*. This is solved for different values of *S(t)*, to give 'simulated' survival times for each patient. If simple survival models, such as an exponential or Weibull distribution, have been chosen, this rearrangement could be computed analytically. If, however, a more complex model, such as the flexible parametric distribution, is used, the process is considerably more complex, and it is no longer possible to calculate the function analytically. Therefore, the method becomes reliant on user written software packages, which employ root solving techniques to solve the expression and produce the simulated survival times. One example of a package that can be used is the stsurvsim command in Stata (Royston, 2012).

4.4.1.6 *Stage 6: Simulating from the censoring distribution*

Having simulated a survival time estimate for each patient, their censoring time needs to be generated. Once again this will depend on the choice of censoring distribution. If the 'maximum time censoring' approach is used, then each individual's censoring time is the value of the maximum time. If instead, the 'recruitment times censoring' approach has been chosen to define the censoring distribution, then times are simulated from the uniform distribution outlined in Section 4.4.1.4.2. Finally, should a 'numbers at risk' table approach be adopted, the simulation model is a piecewise exponential distribution. In practice this is implemented by simulating an interval-specific time for each interval and each patient. This is estimated, as follows, where for patient *x* their censoring time, for a particular interval, can be calculated as:

$$t_{x_i} = \begin{cases} 0 & , \quad \text{if } t_{x_{i-1}} \leq t_{\max_{i-1}} \\ t_i & , \quad \text{if } t_i \leq t_{\max_i} \\ t_{\max_i} & , \quad \text{if } t_i > t_{\max_i} \end{cases} \tag{4-31}$$

Where,

$$t_i = -\frac{\ln(S)}{\lambda_i} \tag{4-32}$$

And $S \sim \text{Uniform}(0,1)$.

The total censoring time can then be calculated by summing over all the intervals, in other words,

$$t_x = \sum_{i=1}^{I} t_{x_i} \tag{4-33}$$

4.4.1.7 *Defining the patients' observed survival times and status*

Now that each patient has a simulated survival time and censoring time, their observed time and status (i.e. indicator of event or censored) must be determined. This is simply calculated by using the minimum of the two values: if the minimum value is the survival time, this patient experienced an event at the simulated survival time; if the minimum value is the censoring time, this patient was censored at their simulated censoring time.

4.4.1.8 *Stage 8: Analysing the simulated dataset*

The analysis of a single dataset can be divided into two parts: 1) replicating the reported statistics in order to confirm how representative the simulated data is to the original IPD; and 2) conducting any additional analyses.

4.4.1.8.1 Part 1: Replicating the reported statistics

This is a crucial stage of the method; its purpose is to evaluate how representative these simulated data are to the original IPD. Commonly reported statistics, such as the number of events, median survival (or survival proportion at a specific time point) and HR and, if possible, RMST, should be replicated. There will be a level of disparity between the IPLD and the IPD, but, if the IPLD is representative of the IPD, this should be small. For certain statistics, in particular the median survival, a reasonably higher level of disparity can still be acceptable. This is due in part to the fact that the median can be sensitive to reconstruction based on a step function. Therefore, having fitting a continuous model through this step function, will account for some of the difference.

4.4.1.8.2 Part 2: Conducting additional analysis

This is principally the main aim of reconstructing IPLD; to conduct a new analysis. Therefore, once it has been established that the IPLD is a reasonable proxy for the IPD, the additional analysis can be implemented on the dataset.

4.4.1.9 *Stage 9: Capturing the uncertainty by producing multiple datasets*

Since this is where this data reconstruction method differs the most from other alternative techniques, it is perhaps, the most important stage of the approach. Instead of relying on a single dataset, multiple datasets are simulated and analysed (by repeating stages 5 to 9). This means that uncertainty around the reconstruction process is specifically captured, and no one dataset is given excessive weight (e.g. considered to be exactly equivalent to the IPD).

Whilst a single reconstructed dataset may give values of the replicated statistics that differ noticeably from the IPD, in the long term, assuming the chosen model fits the data reasonably well, the results will tend towards those that could have been obtained from the IPD.

4.4.1.10 *Stage 10: Obtaining a final point estimate*

Having generated all of the datasets, and analysed these individually, for ease of interpretation or inclusion in further secondary analysis (e.g. the calculation of HR for a meta-analysis), a single point estimate would be valuable. Therefore, the average over all the datasets, is taken; similar to using parametric bootstrap.

4.5 Illustrative example of reconstructing IPD

4.5.1 Neutron Therapy trial

To illustrate how this method can be used in practice, and for an initial assessment of whether the methodology performs well, the simulation approach has been applied to the longer-term follow-up from a RCT. This RCT investigated whether treating pelvic cancer patients with Neutron therapy improved survival in contrast to Photon therapy (Errington, 1991). This specific trial was chosen as the IPD were accessible. Whilst these methods are designed to be used when the IPD are not accessible, here, it allowed a direct head-to-head comparison for the IPD and IPLD secondary analysis results. The secondary

analyses chosen were methods that accounted for non-proportional hazards, since the Kaplan-Meier curves crossed (at approximately 6 months – Figure 4-4), suggesting that the PH assumption may be violated.

4.5.2 Methods

Two methods have been used to reconstruct the data: the Guyot approach (Guyot, 2012) and the simulation method. For both methods, the effect of the level of information available was explored.

For the Guyot approach this included having:

- neither the 'numbers at risk' table or events;
- events only;
- both the events and 'at risk' table.

For the simulation approach, this meant applying all three censoring techniques.

An ITT analysis was conducted, once the data had been reconstructed using the required method. This analysis included the number of deaths, median survival time, RMST at three and a half years, and the HR estimated from a Cox model. To account for possible non-PH, two different secondary analysis methods were applied: 1) a Cox model which allowed for time-dependent effects (Bellera, 2010) by including an interaction term between log time and treatment as shown in equation (4-34)

$$h(t) = h_0(t)exp(\beta_0 + \beta_1\, trt + \beta_2\, trt\, \times log(t)) \qquad (4\text{-}34)$$

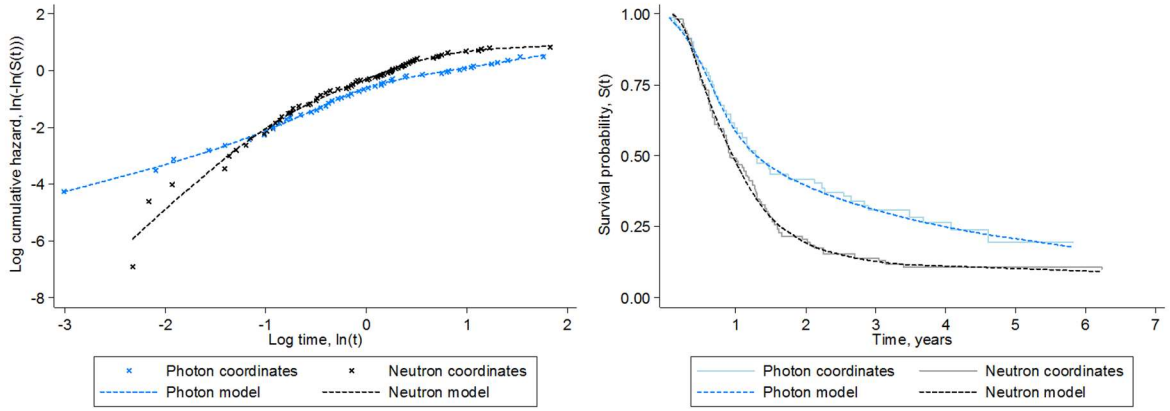
and 2) piecewise Cox models, which effectively estimate different HR, for pre-specified time periods, as shown in equation (4-35).

$$h(t) = h_0(t)\exp\left(\beta_j x_{ij}\right), \qquad where\ x_{ij} = trt_i \ \text{if}\ t_{j-1} \leq t < t_j \qquad (4\text{-}35)$$

4.5.3 Results

4.5.3.1 *Model fitting*

Figure 4-2: Fitted models compared to the coordinates – Neutron therapy example



The chosen models compared to the extracted coordinates: (left) on the LCH scale – as fitted during the modelling process; (right) transformed back to the survival scale.

For the simulation method, models had to be determined for each treatment arm. The simulation models chosen for both treatment arms used four degrees of freedom (df). These decisions were based on:

- a visual inspection of the curve with the extracted coordinates;
- the Akaike Information Criterion (AIC) statistic;
- the Bayesian Information Criterion (BIC) statistic;

compared over models using different degrees of freedom, from 3 df to 9 df.

Once the degrees of freedom had been chosen, four different knot locations (for that number of degrees of freedom) were tried. These models were plotted and showed very little difference on visual inspection. Hence, the knot location was judged not to have been influential on the results. More information, including the curves and model fitting criteria results, is provided in the Appendix E.

The results for the initial ITT analysis, contrasting IPD and IPLD are shown in Table 4-1. For the simulation method, the results have been averaged over the two thousand generated datasets. Also shown in Table 4-1 is the effect of the information available to develop the censoring distribution. The findings presented here show how much poorer the data reconstruction methods are when little information is available on censoring (e.g. when the maximum time simulation method or Guyot approach without the 'risk table'

Table 4-1: Initial analysis for the IPD and IPLD – Neutron therapy example

Comparison of the initial analysis over the IPD and IPLD (averaged over 2000 datasets); accounting for both the Guyot and simulation approach, and the different levels of information for the developing the censoring distribution.

Dataset	Results from Initial Analysis											
	Number of events				Median survival time				Hazard ratio		Restricted mean survival time at 3 ½ years			
	Photon		Neutron		Photon		Neutron				Photon		Neutron	
Original IPD	47	(75.8%)	82	(89.1%)	1.27	(0.91, 2.27)	0.89	(0.74, 1.24)	1.57	(1.09, 2.25)	1.80	(1.47, 2.13)	1.29	(1.07, 1.50)
Using the Guyot approach:														
Incl. risk table & events	47	(75.8%)	82	(89.1%)	1.19	(0.86, 2.23)	0.92	(0.70, 1.27)	1.47	(1.03, 2.10)	1.78	(1.55, 2.01)	1.30	(1.15, 1.45)
Incl. no. of events	47	(75.8%)	82	(89.1%)	1.30	(0.93, 2.27)	0.92	(0.76, 1.24)	1.58	(1.10, 2.27)	1.84	(1.51, 2.17)	1.30	(1.09, 1.51)
Not incl. risk table or events	50	(80.6%)	83	(90.2%)	1.28	(0.93, 2.27)	0.92	(0.76, 1.24)	1.49	(1.05, 2.13)	1.81	(1.48, 2.14)	1.30	(1.09, 1.51)
Using the simulation approach:														
Risk table censoring	47.2	(76.1%)	82.5	(89.7%)	1.33	(0.89, 2.11)	0.95	(0.75, 1.16)	1.57	(1.09, 2.32)	1.78	(1.47, 2.11)	1.29	(1.09, 1.49)
Recruitment times censoring	47.3	(76.3%)	82.7	(89.9%)	1.33	(0.89, 2.11)	0.95	(0.74, 1.16)	1.57	(1.09, 2.29)	1.78	(1.46, 2.11)	1.29	(1.09, 1.49)
Maximum time censoring	50.9	(82.1%)	83.6	(90.9%)	1.33	(0.89, 2.11)	0.95	(0.75, 1.16)	1.51	(1.05, 2.18)	1.78	(1.46, 2.11)	1.29	(1.09, 1.49)

Table 4-2: Secondary analysis for the IPD and IPLD – Neutron therapy example

Comparisons of the IPLD results over 2000 datasets and the IPD for the secondary analyses; accounting for both the Guyot and simulation approach, and the different levels of information for the developing the censoring distribution.

Dataset	Results from Secondary Analysis									
	Coefficients from the time-dependent Cox model:				Hazard ratios from the Piecewise Cox model:					
	Main effects		Time dependent effect		0 and 6 months		6 and 12 months		12+ months	
Original IPD	0.482	(-0.115, 0.850)	0.198	(-0.220, 0.617)	1.41	(0.69, 2.90)	1.39	(0.72, 2.68)	1.79	(1.04, 3.08)
Using the Guyot approach:										
Incl. risk table & events	0.404	(0.045, 0.765)	0.171	(-0.251, 0.594)	1.23	(0.61, 2.48)	1.32	(0.68, 2.55)	1.73	(1.02, 2.93)
Incl. no. of events	0.458	(0.094, 0.822)	0.006	(-0.466, 0.479)	1.44	(0.68, 3.07)	1.47	(0.77, 2.81)	1.72	(1.02, 2.94)
Not incl. risk table or events	0.404	(0.047, 0.761)	0.052	(-0.379, 0.484)	1.29	(0.62, 2.67)	1.45	(0.75, 2.77)	1.63	(0.97, 2.74)
Using the simulation approach:										
Risk table censoring	0.489	(0.113, 0.873)	0.188	(-0.288 0.648)	1.33	(0.62, 3.07)	1.43	(0.77, 2.87)	1.87	(1.07, 3.39)
Recruitment times censoring	0.487	(0.106, 0.884)	0.184	(-0.315, 0.629)	1.33	(0.62, 3.07)	1.43	(0.77, 2.87)	1.87	(1.09, 3.36)
Maximum time censoring	0.427	(0.065, 0.802)	0.075	(-0.397, 0.503)	1.33	(0.62, 3.07)	1.43	(0.77, 2.87)	1.69	(1.10, 3.00)

Figure 4-3: Time dependent HRs

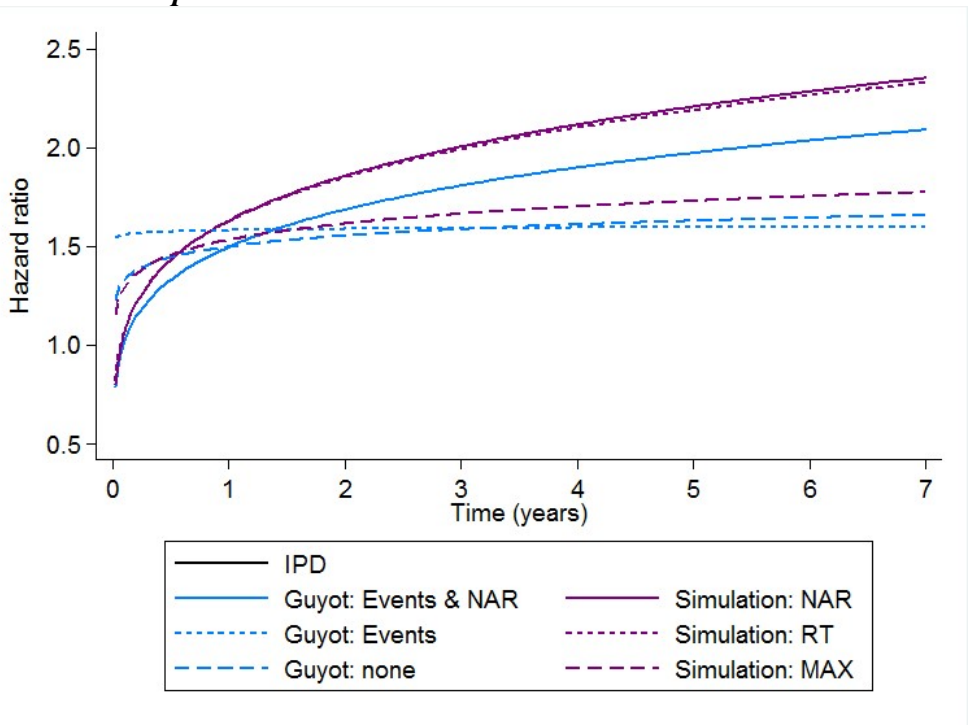


Illustration of the HR over time depending on the data being used.

and number of events are used). Interestingly, here the 'recruitment times' censoring proved almost as appropriate as the 'numbers at risk' table censoring for the simulation technique. Of the statistics reported the number of events in the photon therapy group, the HR and median survival time appeared quite sensitive across methods. As already commented upon, a slight degree of variation in the median survival time is expected. It should be noted that the three and half year RMST proved very stable over all of the simulation method results.

4.5.3.2 *Secondary Analysis*

Continuing to the secondary analysis, the results for this are presented in Table 4-2. A formal test for non-PH was non-significant, which meant that this analysis may not yield significant results. However, for illustration purposes, this was still deemed to be a good example. Examining Table 4-2, highlights, yet again, that, particularly for this example, making simplistic assumptions about censoring can lead to poorer performance in terms of obtaining IPLD results comparable to the IPD. In addition, it was also clear different levels of information available for the Guyot approach led to quite different estimates across both Cox approaches accounting for non-PH. Once again, for the simulation

technique, the ‘risk table’ and ‘recruitment times’ censoring distributions resulted in very similar findings. Further assessment of Table 4-4, showed that the IPLD findings from the time-dependent Cox model, performed exceedingly better at representing IPD than when the piecewise model had been used. This is also illustrated, for the IPD, Guyot approach (with ‘events’ and ‘number at risk’ table) and the simulation technique (using the ‘numbers at risk’ censoring method) in Figure 4-3.

4.5.4 Conclusions

This example highlights the huge impact the level of information and choice of method can have on the results. When the amount of information is reduced to the minimum (Kaplan-Meier curve without ‘at risk’ table), strong assumptions about censoring (e.g. only administrative censoring at the maximum follow-up time) are applied. This consequently often reduces the representativeness of the data to the IPD, as can be seen here (Section 4.5.3). It is, perhaps, quite surprising how well the Guyot method (including the number of events), and simulation approaches with either ‘recruitment times’ or ‘numbers at risk’ table censoring methods perform in terms of the initial analysis, and for the simulation techniques for the Cox model with time-dependent effects. None of the methods performed particularly well in relation to the piecewise model, which may suggest that this methodology is not so proficient in reconstructing IPLD which is suitable for applying piecewise models.

4.6 Reproducibility study

4.6.1 Background

Having assessed the viability of the method in practice, it was important to examine the reproducibility aspect of the approach. To ensure reliability and consistency, the simulation technique must also provide results that can be easily reproducible, especially given the simulation nature of the data.

Therefore, a small reproducibility study was undertaken. This investigated the reproducibility of both the Guyot and the simulation approach. It directly considered the effect of different participants, and image quality considerations such as line style and line thickness. To enable certain statistics to be calculated and full comparison to the IPD, the three datasets for the reproducibility study were simulated based on clinical trials

(REACT trial (Gershlick, 2005), TAnDEM trial (Kaufman, 2009) and the MRC-Focus trial (Seymour, 2007)). Information about the simulated examples can be found in Figure 4-3.

The REACT trial example was specifically chosen, as this is very representative of typical heart disease trials, and as such the Kaplan-Meier curve is a complex shape (very steep drop in the first couple of days, plateauing soon after for the remainder of the follow-up (e.g. plateau for approximately four months out of six). This shape is likely to be challenging for any method used to reconstruct the data.

4.6.2 Methods

Six participants were asked to extract coordinates twice for the three examples, once for the Guyot method and once for the simulation method. To ensure parity between the two methods, none of the participants had previously extracted coordinates for either method, nor were they experienced with the digitizing software. Each participant had a set of instructions, given in Appendix F, describing the criteria for extracting the coordinates for a particular method. In addition, the methods were labelled generically to blind participants as to the actual approach they were digitizing for. Half of the participants were required to start with the Guyot method (Method A) and then extract for the simulation technique (Method B); whilst the other half were instructed to digitize for the simulation approach (renamed Method C), before extracting for the Guyot method (now labelled Method D). By varying the order in which participants extracted the data it was hoped that the potential bias of participants improving the extraction process over the course of the example would be addressed, and thus the coordinates being extracted better for the second method.

In order to reconstruct the data using the Guyot method, the coordinates must conform to certain criteria (time coordinates are positive and distinct; monotonic survival function). As such, the following filters were applied:

- Removal of any coordinates where the survival probability was greater than one or less than zero;
- Removal of any duplicate coordinates with respect to both time and survival or just time;

- Removal of any estimate of the survival function which exceeded any of the previous estimates;
- Removal of the final estimate of survival if a patient was censored.

These filters could also be applied to the simulation technique, if it was thought it might improve the model fit.

The statistics chosen for comparison were the HR and RMST at a given time point.

To ensure fairness, models with four degrees of freedom were used for the simulation approach, for all examples, regardless of fit. Having fixed the degrees of freedom, knot locations were permitted to be changed, if the default knots led to a non-monotonic or especially ill-fitting function. One thousand datasets were generated for each example in the simulation approach.

4.6.3 Results

Results from the individual participants, both un-stratified and stratified based on extraction order, are reported in Appendix F. Table 4-5 gives the results for the RMST and HR averaged over participants for each method and example. These are discussed in more detail in sections 4.6.3.1 to 4.6.3.3.

4.6.3.1 *Guyot method compared to the IPD*

Of all three examples, the first seemed to have the most across-participant variation for the Guyot method, with HR estimates ranging between 0.497 and 0.884. This example also had the greatest difference between the results from the IPLD and IPD, (0.516 in contrast to 0.697). In comparison, for the other two examples, the point estimates for the IPLD and IPD were comparable.

It should be noted, that whilst all participants followed and typically conformed with the same instructions, for the last two examples, one person's results are markedly different. Stratifying by the order of extraction demonstrated little difference for examples 2 and 3. However, for the first example, the results did show differences: for those extracting for the Guyot approach first, the average was 0.798 (Standard Deviation: 0.141, minimum:

0.636; maximum: 0.884), and 0.596 (Standard Deviation: 0.087, minimum: 0.497; maximum: 0.657) for those who digitized for the Guyot method second.

4.6.3.2 *Simulation method compared to the IPD*

As with the Guyot approach, the first example has the greatest amount of variation over the different participants, ranging from 0.388 to 0.580. However, in terms of point estimates, all of the average results over 1000 datasets are exceptionally similar to the reported values. None of the examples showed evidence that the results depended on the ordering of the methods (e.g. Guyot first, simulation second).

4.6.3.3 *Contrasting the Guyot and simulation approach*

Both methods performed well at reconstructing the data and obtaining a point estimate for examples 2 and 3, but generally the variability was greater for the Guyot approach; partly due to the participant with different results. The first example yielded very different estimates over participants for each method and example. These are discussed in more detail in Sections 4.6.3.1 and 4.6.3.2.

4.6.4 Conclusions

Based on the above reproducibility study, it would appear that the simulation approach is an excellent alternative to the algorithm proposed by Guyot for reconstruction IPLD. The simulation method is considerably more flexible, allowing it to capture and reconstruct more complex survival data, such as that in Example 1 with greater accuracy and lower across-participant variability.

It should be commented that, even though the simulation technique did perform better for Example 1, it required a lot of effort to find suitable models to effectively capture the shape of the curve, and ensure the simulated data was truly representative of the underlying survival distribution. This was ultimately achieved by specifically chosen knot locations. It is worth noting that the Guyot method also had difficulty reconstructing the data.

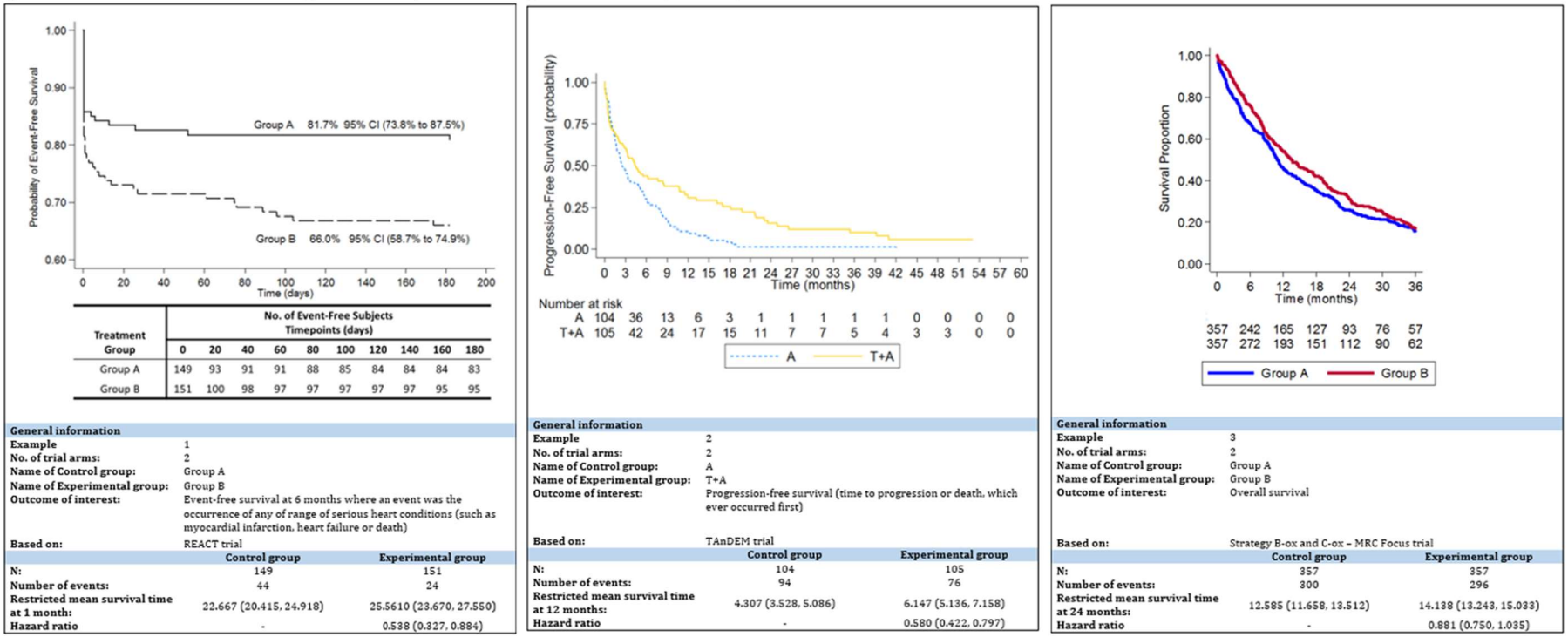
4.6.4.1 *Limitations and areas for improvement*

Should this, or a similar study be repeated, the implementation and findings have highlighted several areas that could have been further exploited to achieve the maximum potential from such an investigation. Some of the key limitations of this particular study were the small sample size both in terms of participants and examples. Both of these occurred in order to make the study feasible, as those who agreed to participate were limited on the time available to digitize the curves. Thus, the whole digitizing process was restricted to only taking one to two hours.

In addition to having more than six participants, more examples should be used; in particular examples where there were a greater sample size and many events over a short timescale (such that the Kaplan-Meier has the appearance of a smooth curve). Such examples were excluded due to the time it would take for the coordinates to be digitized. Further tests on the quality of the image could also be useful; this could include more dotted or thick lines, the choice of line colour (e.g. yellow, or pale coloured as opposed to dark lines) and those with poorer image quality.

Another factor that this study did not take account of was asking participants to digitize the same dataset more than once to facilitate an assessment of the within participant variation. The last adaptation that could have been made would be to have varied the order of the examples as well as the order of the reconstruction methods. This would have determined, for definite, if the first study (and method) a participant digitized for was distinctly different because of starting the process or due to a particular example.

Figure 4-4: Simulated examples used for the Reproducibility Study



The three simulated examples, including the Kaplan-Meier curve, risk-table, summary information and underlying case study used to generate the data from.

Table 4-3: Average HR and RMST for the IPLD over participants

Comparison of the average HR and RMST for each group over the participants for each example and method, contrasted with the IPD data

		Example	IPD	Guyot reconstructed data			Simulation approach		
				Average	SD	Range	Average	SD	Range
Hazard ratio		1	0.516	0.697	0.152	0.497 – 0.884	0.496	0.068	0.388 – 0.580
		2	0.580	0.585	0.020	0.563 – 0.622	0.576	0.013	0.552 – 0.589
		3	0.881	0.856	0.103	0.650 – 0.924	0.889	0.013	0.873 – 0.906
Restricted mean survival time	Control	1	22.667	25.747	0.213	25.535 – 26.144	22.707	0.076	22.589 – 22.775
		2	4.307	5.000	1.037	4.467 – 7.108	4.337	0.147	4.186 – 4.612
		3	12.585	13.018	0.570	12.527 – 14.098	12.506	0.084	12.344 – 12.573
	New	1	25.610	28.012	0.383	27.735 – 28.689	26.172	1.186	25.547 – 28.556
		2	6.147	6.519	0.686	6.098 – 7.909	6.049	0.140	5.891 – 6.308
		3	14.138	15.062	1.791	13.954 – 18.639	14.094	0.045	14.041 – 14.157

4.6.4.2 *Advantages of this study*

Compared to that described in the Guyot publication, the reproducibility study undertaken here had more participants, allowing the potential for variability to be assessed further. It also demonstrated how differently the guidelines for extracting coordinates could be interpreted and applied. For example, with the first study (as more points are needed than the number of events, and to avoid turning points in the curve) the instructions indicated that a quarter of the total points (in other words ten to fifteen or more), were taken at the beginning. However, these were taken in different ways. For instance, in Example 1, one participant extracted a considerable number of additional points at the time of the first event, whilst another distributed their additional points over the first 20 days. Another advantage was that by simulating the datasets, it allowed greater comparison between the IPD and IPLD. Moreover, it enabled statistics that are infrequently reported such as the RMST to be recorded, and also an appropriate cut-off time for this (suitable for the simulation technique) to be chosen. Simulating the underlying datasets also ensured that additional features, such as the dotted lines and thick line style were included in the examples, as these factors potentially impact on the results.

4.7 Discussion points

Whilst the outline of the method is easy to describe (Sections 4.4.1) and illustrate (Sections 4.5 and 4.6), there are several questions that come to light in order to use this approach in practice. For example, how many coordinates should be taken? How many datasets should be simulated? These questions, and other topics will be explored in further detail in the following sections, with reference made to the findings from the illustrative example and the reproducibility studies.

4.7.1 The location of the coordinates

One reason for choosing a modelling-based approach was to minimise the need for capturing every single event time as this can be very difficult to do in practice. However, if not digitizing at exact event times, how many points should be taken? And where should these be placed? Is it better to take as many points as possible (and throughout the whole of the time scale)? Would points extracted at evenly spaced intervals be adequate? Is it still important to try to identify and extract coordinates at exact event times (with the

proviso that it matters less if this is not obtainable)? The principle concern was whether results would differ depending on which of the above approaches was adopted.

Intuitively, some of these suggestions could be accepted or rejected. For example, taking a high proportion of points in the tail of the distribution should avoided; if not, this would give excessive weight to the estimates in the tail, which are usually less stable and often the result of small sample size. Another condition, which was inferred, was that more points should be taken where there are more events (though not necessarily at event times). Combining these two, suggests that evenly spaced intervals are not sensible, as these would give insufficient weight to the coordinates in the necessary places (e.g. more weight at the start, little weight in the tail). Finally, in order to apply appropriate models which fit well to the data, sufficient coordinates must be taken, in other words ten to twenty data points for each spline variable created.

The question still remained how many coordinates should be taken. If as many points as possible could be taken, would this improve the fit? Initial exploration demonstrated that actually in trying to obtain as many points as possible, the digitizer was likely to induce much more variation and error, as it is harder to consistently ensure monotonicity when taking so many coordinates. This substantially increased variation / error which ultimately results in more unstable models and can lead to turning points in the model fitted for the survival function.

4.7.2 The number of simulated datasets

One of the key advantages of the simulation approach is that it captures the uncertainty around the reconstruction process by producing multiple datasets. However, that leads to the question of how many datasets should be generated. In theory, it would be preferable to generate as many as possible, thus enabling more reliable and reproducible estimates to be obtained. But in practice, the more datasets that are produced, the more computationally time intensive the method becomes, not only in simulating the datasets, but also in analysing them. Ideally, one thousand or more datasets should be produced if time permits as this typically ensures reasonable reproducibility of results. Nevertheless, generating several hundred datasets can be satisfactory in terms of producing a point estimate which gives a representative average of the IPLD, and which clearly gives a huge saving in computation time. It should be noted, though, that using fewer datasets could

result in greater differences if repeating the same data reconstruction and analysis methods for a second (or third, or fourth etc.) time using the same summary data, than if several thousand datasets were produced each time. Therefore, there will always be a trade-off between feasible computation time and good reproducibility with regard to the replication of the method.

4.7.2.1 *Effect of replication using different numbers of datasets*

This section explores how the point estimate changes with regard to the number of datasets simulated and the impact of this on reproducibility. This is achieved using the information generated from the illustrative example described in Section 4.5.

Firstly, it examines how the point estimate varies depending on how many datasets are simulated. The ITT Cox model HR (on the log-scale) has been chosen to demonstrate this. Figure 4-5 highlights how this statistic changes the more datasets are produced and averaged over up to 2000 datasets, whilst Table 4-4 gives a comparison of the actual point estimate values for a selected number of datasets (e.g. 10, 500 and 2000 etc.).

Table 4-4 showed that averaging over less than 100 datasets (i.e. 10, 20 or 50) gave poorer agreement. However, based on the values presented in the table, any number of 100 datasets or more computed a relatively consistent estimate, comparable with the IPD. Nevertheless, on examining the plot in Figure 4-5: Average log-HR depending on the number of datasets simulated, which provides a continuous assessment, it can be seen that convergence actually occurred after approximately 800 datasets. Essentially, the consistent agreement at 100 or 500 datasets, were chance findings. Particularly from Figure 4-5: Average log-HR depending on the number of datasets simulated, it is clear that for small numbers of simulations, the point estimate is sensitive and so sufficient simulations must be generated to ensure convergence and reproducibility.

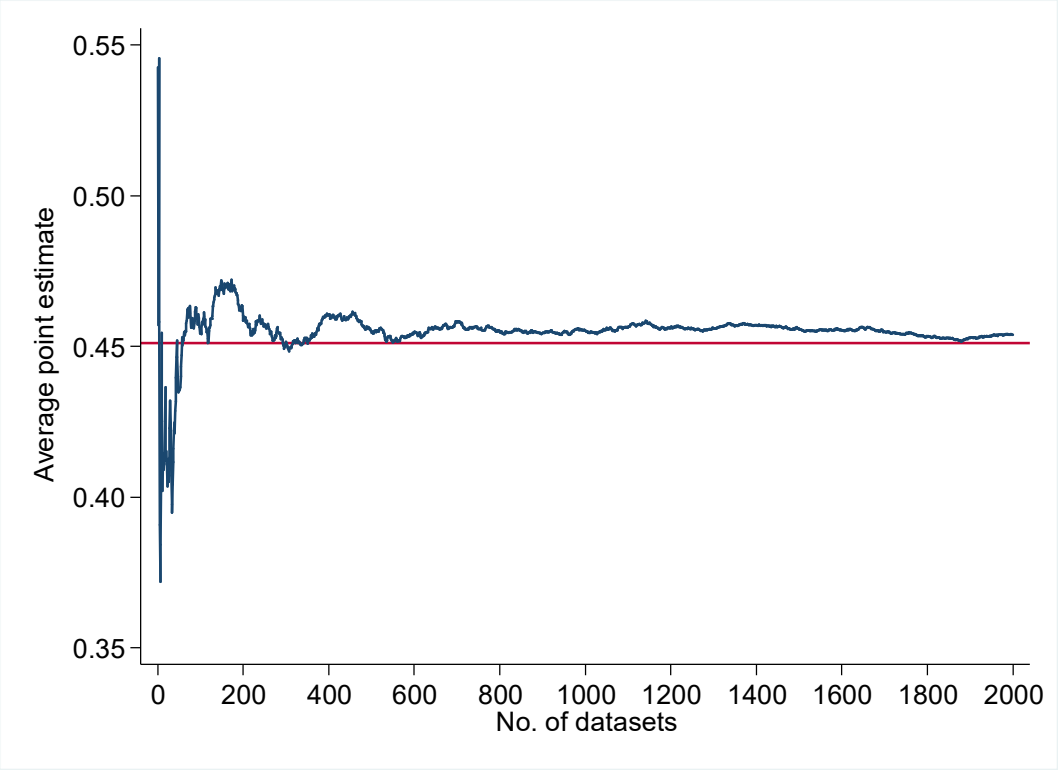
To assess reproducibility further, it was decided to explore how the ordering of the datasets impacted on the point estimate. Therefore, a ‘bootstrapping without replacement’ technique was employed to essentially ‘re-order’ the datasets. Thus, creating different subsets of datasets for a particular number of datasets e.g. 200. At the time the analysis was undertaken, the process of simulating the data was exceptionally time intensive for a large number of datasets. Consequently, it was more feasible to use the existing datasets,

than to create further simulations. Figure 4-6 and Table 4-5 illustrate how these estimates change depending on the ordering of the datasets, for smaller numbers of datasets. Since, there are only 2000 datasets, these subsets have been restricted to those up to 500 datasets, since this means that, at maximum, only a quarter of all the datasets will ever be used, allowing some variation. The estimate was calculated for five different orderings.

Table 4-4: Log-HR averaged over a given number of datasets
Change in the average value of the log-HR depending on how many simulated datasets were averaged over, contrasted with the reported estimate.

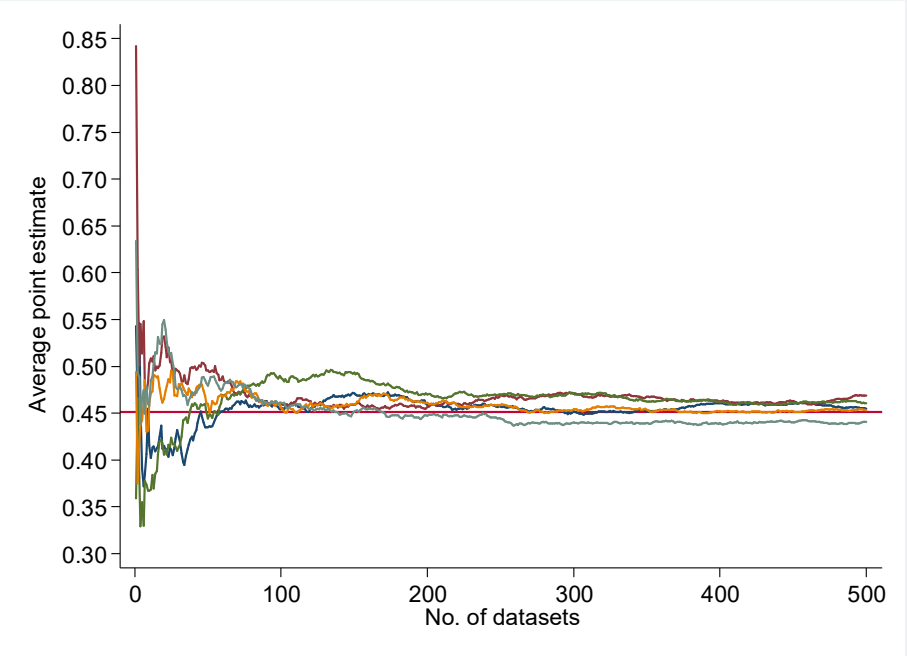
Number of datasets	Reported log-HR	Average log-HR
10	0.451	0.416
20		0.415
50		0.435
100		0.454
500		0.455
1000		0.455
1500		0.456
2000		0.454

Figure 4-5: Average log-HR depending on the number of datasets simulated



The log-HR averaged over different numbers of datasets (up to 2000) compared to the reported estimate (as shown by the red line) – results from the illustrative example in section 4.5.

Figure 4-6: Log-HR depending on the ordering of the IPLD datasets



The average log-HR for different numbers of datasets and different 'orderings' of 2000 simulated datasets, contrasted with the reported value, shown by the red line.

Table 4-5: HR depending on the ordering of the IPLD datasets
Both the average log-HR and the HR for different numbers of datasets and different 'orderings' of 2000 simulated datasets, contrasted with the reported value.

Number of datasets	Reported log-HR	Average log-HR				
		Order 1	Order 2	Order 3	Order 4	Order 5
10	0.451	0.416	0.503	0.367	0.461	0.458
20		0.415	0.532	0.406	0.465	0.549
50		0.435	0.496	0.444	0.460	0.479
100		0.454	0.461	0.488	0.456	0.461
200		0.459	0.456	0.470	0.461	0.447
500		0.455	0.468	0.461	0.453	0.441
Number of datasets	**Reported HR**	**HR**				
		Order 1	**Order 2**	**Order 3**	**Order 4**	**Order 5**
10	1.570	1.516	1.654	1.443	1.586	1.581
20		1.514	1.702	1.501	1.592	1.732
50		1.545	1.642	1.559	1.584	1.614
100		1.575	1.586	1.629	1.578	1.586
200		1.582	1.578	1.600	1.586	1.564
500		1.576	1.597	1.586	1.573	1.554

Since the datasets to be included are essentially drawn from a finite set, some observations will appear in all, a few or none of the five re-orderings.

Whilst there is not a vast difference (Table 4-5) between the values calculated there is still clear variability between the estimates depending on the re-ordering. For example, the estimates, for two hundred datasets say, vary between 1.564, reasonably close to the reported estimate, to 1.600 which might suggest the IPLD is not representing the IPD as well as would have been hoped. Given that there will be some overlap between re-orderings, which could have reduced the variability, for 200 datasets distinct subsets were assessed. However, the results from these did not differ noticeably from those reported in Table 4-5. These results from the distinct subsets (of 200 datasets) are presented in the Appendix H.

4.7.3 Does the knot location have an impact on the results?

Whether the model is sensitive to knot locations is purely attributable to the specific example. This was highlighted clearly in the examples already presented in this chapter (Sections 4.5 and 4.6). In the illustrative example (given in Section 4.5), for the chosen number of degrees of freedom, the knot location made very little difference (graphically) to the model fitted. However, in the later example in the Reproducibility Study (Section 4.6) to find a monotonic well-fitting model for the first example and several of the participants, specific knot location had to be chosen (default knot locations for the same number of degrees of freedom produced ill-fitting or non-monotonic functions). Once again, this was easily identified by plotting the fitted model against the coordinates.

There is the potential that some of this sensitivity could be accredited to the coordinates; if there is variation within the coordinates (discrepancy in the survival proportion resulting in a later estimate having a marginally higher survival proportion than at an earlier time point). For a method such as the Guyot approach, the data would have to be screened and filtered to remove conflicting estimates such as these. However, with a simulation approach the majority of the time, these small discrepancies should not affect the data too significantly, but could possibly lead to this sensitivity. Therefore, in these circumstances, also applying a screening process to these coordinates to see if this improves the model fit (and reduces the sensitivity) is advocated.

4.7.4 Acceptable limits

The last point which is difficult to determine is how to define acceptable limits, since the term ‘acceptable’ is very subjective. For example, if the publication reported that the HR was 0.70, would HR of 0.72 (0.02 difference) from the reconstructed data be acceptable? Or what about a 0.75 (difference of 0.05) HR? To some people both of these differences could be considered satisfactory, for others perhaps neither would be. And what about the number of events: what about a discrepancy of two events? In a treatment arm with six hundred participants, this might be acceptable; but in a sample size of fifty, it could be argued that this is substantially less so.

Decisions for some of the key statistics rely on other considerations, which are detailed below

- Number of events – as detailed above the difference between the reported and reconstructed data estimate also depends on the sample size of the treatment group. Larger discrepancies in a larger sample size would possibly be more acceptable.
- Median survival time – here a more lenient acceptability criteria is required, because of the underlying step function nature of the data. However, the absolute size of the discrepancy must be related to the follow-up length. E.g. 0.25-year difference in median survival for a seven-year follow-up length seems satisfactory; this same discrepancy in follow-up of a year would not be acceptable.
- RMST – although this statistic does not require such lenient boundaries as for the median survival time, the results must still be interpreted in relation to the follow-up length.
- HR – to some extent it could be argued that this statistic can be considered independent to other elements of the trial (e.g. the scale of the point estimate should not be affected by follow-up time), however, it should always be contextualised. For example, smaller sample sizes may lead to larger variation in the uncertainty, or point estimate which should be expected.

In addition to assessing the individual statistics, an overall assessment needs to be made. Intuitively, if at least one of the statistics is clearly not ‘acceptably’ close to the reported

estimate, then any findings from that reconstructed data should be treated with caution, and potentially other models considered and assessed.

4.8 Secondary analysis for treatment switching

4.8.1 Background

The main aim for developing this method of reconstructing the IPLD has always been to ultimately adjust for treatment switching using one of the currently recommended methods. Thus far, this method has only concerned itself with recreating the survival time information. However, in order to adjust for treatment switching, data related to treatment switching must also be reconstructed. Therefore, this section describes how treatment switching information can be generated. The techniques described here are similar to those used in Boucher (2013a) which was discussed in Section 3.7.2.2. In Section 5.5, a more complex and realistic method is suggested, however, this relies on having reconstructed paired data (Sections 5.4.1 and 0)

4.8.2 Methods

4.8.2.1 *Reconstruction of the treatment switching information*

Of the currently recommended methods (Latimer, 2014), the RPSFTM has been chosen for the re-analysis method. The reason for this decision is that the RPSFTM, unlike the other methods (which require all treatment switching related covariates to be available), only needs two variables:

1. indicating whether a patient switched;
2. recording time of switch, if it occurred (note, for patients in the experimental group this is denoted zero, and for non-switching control patients, this is their last observed survival time, be it through censoring or event).

Therefore, in order to apply the RPSFTM the first stage of reconstructing the treatment switching data, is to assign which patients switched. The switch time can then be calculated. These stages are detailed in Sections 4.8.2.2 and 4.8.2.3.

4.8.2.2 *Assigning treatment switchers*

Typically, information will have been reported on the number, or perhaps proportion of patients who switched treatments. There is little other information on which to base the decision about patients who switch, as it is not known how the treatment affects OS

For example, whether it is those patients with a shorter survival time who are more likely to have switched, or those with longer survival. Therefore, the choice has been made arbitrary. Patients are selected at random and assigned to switch, until the amount of treatment switchers matches that reported in the publication.

4.8.2.3 *Calculating switch times*

With regard to switch times, whilst authors may specify whether treatment switching was permitted on disease progression (NICE, 2012); after un-blinding; or following interim analysis (NICE, 2009b), this is often the only information available. Given that perhaps the most common reason for switching in advanced or metastatic cancer trials is disease progression, the method has been designed to assume that the actual switch happens at a patient's progression time. There is some justification for this argument, as although a patient might not actually switch at the time of documented disease progression, if they do switch treatments, it will be very soon after this date (Latimer, 2012).

Median PFS and OS times are routinely reported, particularly in NICE TAs. Therefore, it is assumed that the ratio of median PFS to median OS gives an approximation for the average proportion of life spent progression free. Essentially, the median PFS is treated as the average time until the patient's disease starts to progress, and the median OS as the average time until the patient dies. Therefore, given a patient's OS time, their estimated progression time is their simulated OS time multiplied by the ratio (Boucher, 2013b). It should be noted that for an exponential distribution, the ratio of two median survival times is the same as the ratio of mean survival, and also equivalent to the HR.

In the previous research addressing treatment switching using summary data (Boucher, 2013a, Boucher 2013b) (described in Section 3.7.2.2, for each dataset, the process of assigning treatment switchers was treated as a form of bootstrapping without replacement. In other words, of the patients in the control group, the process of selecting those patients who switched was repeated, effectively choosing different patients each time. Each selection were analysed separately and the results then averaged over (in a similar way to the simulation technique). However, now each of the simulated datasets just has one combination of treatment switchers. This is because uncertainty about the data reconstruction process is already being accounted for through simulating multiple datasets.

4.8.2.4 *Secondary analysis applied*

The choice of secondary analysis is relatively limited. Of the three (or four, including IPE) recommended approaches to be used to adjust for treatment switching, only the RPSFTM (or its parametric equivalent, IPE) is feasible since this approach does not rely on covariates, but does need a variable identifying the treatment switchers and the patients' switch time.

Another advantage of using the RPSFTM is that this method 'preserves the p-value'. This means that the uncertainty is purely based on the p-value of the original ITT analysis (see Section 1.1.3.3.1 for more details).

Since the test statistic, Z, is

$$Z = \frac{\ln HR}{SE(\ln HR)} \tag{4-36}$$

And as 'preserving the p-value' means that the test statistic is the same for the adjusted results as the unadjusted, then

$$Z = \frac{\ln HR_{ITT}}{SE(\ln HR_{ITT})} = \frac{\ln HR_{adj}}{SE(\ln HR_{adj})} \tag{4-37}$$

And so,

$$SE(\ln HR_{adj}) = \ln HR_{adj}\left(\frac{SE(HR_{ITT})}{\ln HR_{ITT}}\right) \tag{4-38}$$

Therefore, the secondary analysis using the RPSFTM only needs to calculate the point estimate. Hence, for this approach only, it is not necessary to calculate the RPSFTM HR estimates separately and then combine them. Instead, *n* samples can be combined as one huge dataset, and analysed together, since no SE for the adjusted HR needs to be obtained through the estimates.

It should be noted that the analysis may not give entirely appropriate estimates if the RPSFTM assumptions do not hold, e.g. it is inappropriate to assume that the treatment effect is the same between switchers and non-switchers.

4.8.3 Illustrative examples

4.8.3.1 *Pazopanib for the first-line treatment of advanced renal cell carcinoma VEG105192.*

4.8.3.1.1 Background

The first example used was taken from NICE TA215 Pazopanib for the first-line treatment of advanced RCC (NICE, 2011). The pivotal evidence for this appraisal came from the VEG105192 trial (Sternberg, 2010). This compared best supportive care (BSC), defined as the monitoring of progression, symptom control, and palliative care without active treatment, combined with either pazopanib (n = 155) or placebo (n = 78). On disease progression, patients with an ECOG status of 2 or less who had been receiving placebo, had the option of having pazopanib added to their treatment regimen. This led to 40 patients (51%) having switched treatment by the final date of analysis. To account for treatment switching reanalysis, the manufacturer applied several different methods which included a weighted RPSFTM (not currently programmed in Stata) and an unweighted RPSFTM (available in user-written package strbee (White, 2002)). Since the final analysis for the reconstructed IPLD was conducted in Stata, the secondary analysis aimed to replicate the results of the unweighted RPSFTM using a log-rank test. This, however, proved complex since, for this analysis, the g-estimation process discovered three possible values which satisfied $Z = 0$, thus giving three potential results for the acceleration factor.

Figure 4-7: Kaplan-Meier curve for OS – VEG105192 trial (Sternberg, 2010)

Image subject to copyright and so has been removed from the text. Please see the original source for image.

4.8.3.1.2 Methods

The method was performed as outlined in Section 4.4. Coordinates were extracted off the curve. The models used to simulate the data were restrictive cubic splines models with five degrees of freedom for both the Pazopanib and Placebo treatment groups. The censoring distribution was calculated using information from the 'numbers at risk' table. For this example. 2000 datasets were generated. For each simulated dataset, 40 patients from the control group were chosen randomly and assigned to switch. Given that median PFS is 2.8 months and median OS, 23.5 months; a patient's progression time, and switch time, is 12% of their OS time.

To accommodate the treatment switching, the RPSFTM method was applied to the reconstructed data. Here, the unweighted RPSFTM with the log-rank statistic was used. This was then compared with the values reported from the same analysis conducted on the IPD.

4.8.3.1.3 Results

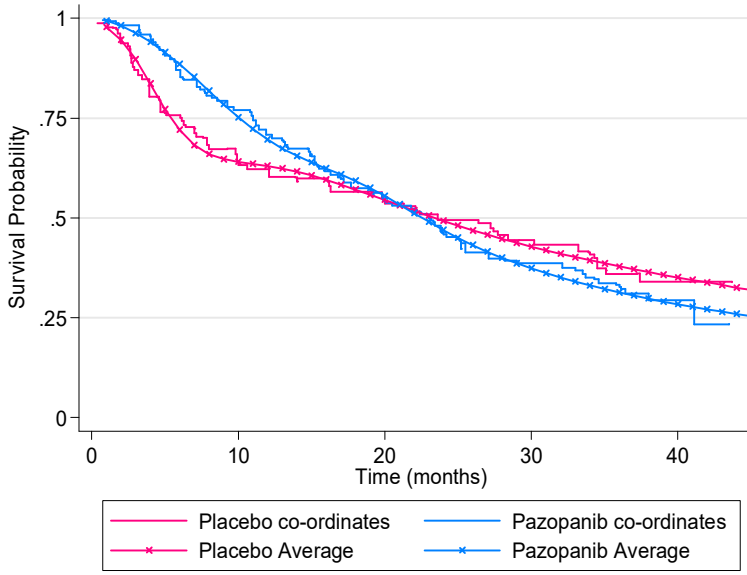
Reported summary statistics

Table 4-6: IPD and IPLD for the VEG105192 trial
Comparison of the average value over 2000 simulated datasets and the same statistics reported in the original publication

	PLACEBO				PAZOPANIB			
	Reported		**Simulated average**		**Reported**		**Simulated average**	
No. of events	49		48.8		99		102.1	
Median OS	23.5	(12.0, 34.3)	23.7	(14.3, 35.0)	22.9	(17.6, 25.4)	22.60	(18.5. 26.8)
HR	1	-	1	-	1.01	(0.72, 1.42)	1.04	(0.73, 1.51)

Initially the ITT analysis is replicated (results given in Table 4-6), in order to confirm the data are representative of the original IPD. As shown in Figure 4-8, the average survival (calculated at monthly intervals) over all 2000 datasets is comparable to the coordinates. In terms of the summary statistics, those for the Placebo group are very similar, although the CI for median OS is slightly smaller. For the pazopanib group, these differ more noticeably. On average, three more deaths are occurring, and the HR shows a 4% increase in mortality on pazopanib compared with placebo, rather than the 1% reported. The CI is also wider for the HR. Therefore, there is some difference within the reconstructed data to the original IPD. However, it is largely comparable.

Figure 4-8: Average survival compared with the coordinates – VEG105192 trial



The average survival proportion at every month compared to the extracted coordinates, for each of the treatment groups.

Reanalysis

Table 4-7: Secondary analysis for treatment switching – VEG105192 trial
Comparison of the secondary analysis, using an RPSFTM to account for treatment switching, between the original reported and the average of the simulated datasets

	Reported	Simulated average	
Acceleration factor	0.61 3.16 5.75	0.95	(0.55, 1.70)

Table 4-7. gives the three estimates for the acceleration factor (as described in Section 4.8.3.1.1), derived from the original IPD, on the left, and the value obtained from the re-analysis on the right. Rather than the multiple solutions found using the IPD, only a single value was obtained for each dataset using the IPLD.

4.8.3.1.4 Discussion

This method recreates data that are representative of the original IPD, and that can be re-analysed addressing for treatment switching. The simulation technique has the advantage that it encapsulates the uncertainty around the original Kaplan-Meier and in the reconstruction process, particularly compared to using the Guyot method (Guyot, 2012). The method was straightforward to implement in practice; the process, however, proved time-consuming, particularly since 2000 datasets were generated. This value had been

chosen to ensure robust results, i.e. should the method be performed again, the average HR would largely remain unchanged. This led to further debate on whether greater efficiency, in terms of time, or greater robustness was more important. Subsequently the number of datasets was reduced to 200.

Given that, for this specific example, a single solution was not found for the IPD, this put the method under additional pressure as there is greater potential for error and disagreement. Particularly since the method does not recreate the original dataset exactly, there will be some error and variation in the results, which should diminish once the average is taken. However, that potentially means that it is very unlikely that three solutions would have been found for every dataset. As the results highlighted, in all datasets, a single solution was obtained (but not three). In practice, this difference could not be ascertained as there would be no indication whether the IPD would produce a single, or multiple, solutions.

In conclusion, for this dataset, the simulation technique worked in terms of reconstructing the survival time IPLD, but there is less evidence that the treatment switching reanalysis has been successful. It would be, therefore, advisable to try additional analysis methods, or other examples in order to bring a greater confirmation of whether the method works. The next example should have reported, where possible, a single solution for ψ produced from a RPSFTM with the log-rank test.

4.8.3.2 *TAnDEM trial*

4.8.3.2.1 Background

TA257 (NICE 2012) assessed lapatinib (GlaxoSmithKline) or trastuzumab (Hoffman-La Roche) in combination with an aromatase inhibitor (letrozole or anastrazole) for the treatment of metastatic breast cancer. The pivotal evidence for trastuzumab came from the TAnDEM trial (Kaufman, 2009), which compared the combination treatment of trastuzumab plus anastrazole (T+A) with a treatment regimen only containing anastrazole (A). However, both the appraisal and the original publication report that, after experiencing disease progression, 73 (71%) patients randomised to receive anastrozole alone had begun to receive trastuzumab in addition. The publication only reports the ITT

analysis, but the appraisal also gives results from a RPSFTM. No information is given on what test was used to calculate the RPSFTM analysis.

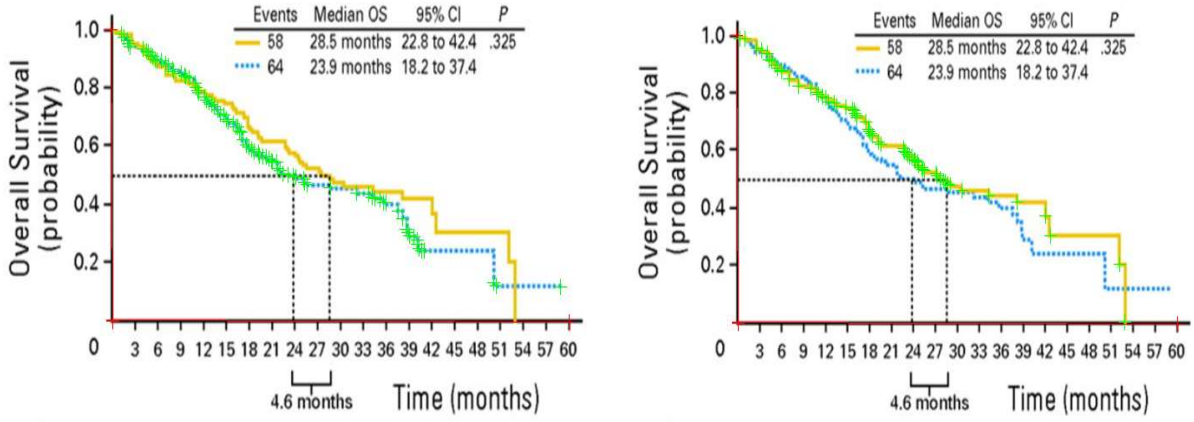
Figure 4-9: Kaplan-Meier curve for OS - TAnDEM trial

Image subject to copyright and so has been removed from the text. Please see the original source (Kaufman, 2009) for image.

4.8.3.2.2 Methods

Similarly, to the example in Section 4.7.3.1, the data are reconstructed using the process explained in Section 4.4. Figure 4-10 shows the location of the coordinates that were extracted from the Kaplan-Meier curve (given in Figure 4-9) using the digitizing software package (DigitizeIt, 2013). The left-hand side give the coordinates (denoted by the lime green plus signs) for A and on the right are the coordinates for T+A.

Figure 4-10: Location of the coordinates – TAnDEM trial



The extracted coordinates for each treatment arm; each green plus sign denoted a separate coordinate.

As afore mentioned, the coordinates from the Kaplan-Meier curve are transformed. In order to capture the shape of the curve, models with 7 and 8 degrees of freedom were fitted for A and T+A respectively. For both treatment arms the knot locations were chosen specifically (rather than using knots calculated from equally spaced percentiles), in order to maximize the fit to the curve. The censoring distribution was formed using the information from the 'numbers at risk table', in Figure 4-9, and by deducing the survival at the equivalent time points using the coordinates. A total of 200 datasets were created. This is considerably less than was used in the previous example (in Section 4.8.3.1); this change was made to reduce the computation time to a more appropriate length. Modifications to the program have since rendered this change unnecessary and more datasets (e.g. 2000) can now be generated in an appropriate amount of time. For each dataset the number of events and ITT median OS time in each of the treatment groups and the ITT (log-)HR were recorded. These were averaged over for the final point estimates.

Treatment switching information

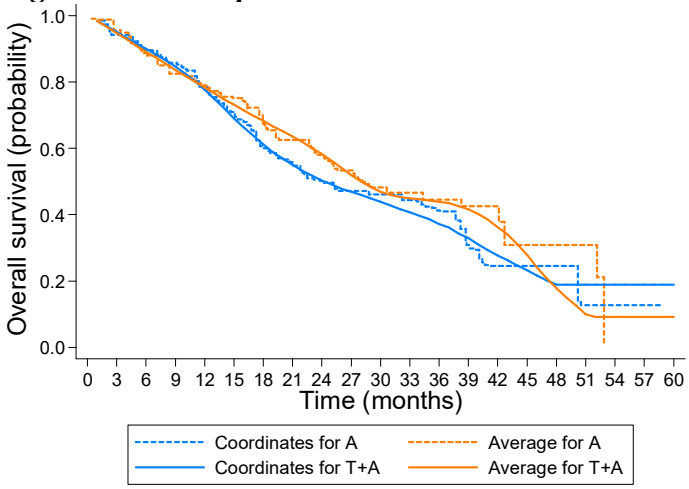
The treatment switching information was reconstructed using the process outlined in Section 4.7.2. A random subset, equalling the number of patients who switched in size, was taken from the control group and assigned to switch. Thus here, 73 of the 104 patients were chosen at random and assigned to switch. Their switch time was calculated as 10% (ratio of median survival, where median PFS was 2.4 months; and median OS, 23.9 months) of their OS. Once the treatment switching information had been reconstructed, a RPSFTM using a log rank test was fitted to the data. The value of ψ recorded. In addition, the 'adjusted' HR was recorded. Although the RPSFTM uses an AFT model, standard practice in NICE TAs is to report a HR. This statistic was hence obtained by calculating the counterfactual dataset using the estimated acceleration factor, and then fitting a standard Cox (or Weibull PH) model to the data. The ψ value and log HRs were also averaged over for the final results.

4.8.3.2.3 Results

The comparison for the original (ITT) analysis, both on the IPD and IPLD are given in Table 4-8. A comparison for the average survival is given in Table 4-8. In contrast to the previous example, the number of events was comparable for both groups. There were differences in the median survival for the ITT analysis, The HR also differed slightly –

14% decrease in mortality rate, compared with 16% decrease as reported. The fit to the survival curve was relatively good near to the start of the time scale, however when events became scarce after about 42 months, the goodness of fit deteriorated.

Figure 4-11: Average survival compared with the coordinates – TAnDEM trial



The average survival proportion at every month compared to the extracted coordinates, for each of the treatment groups.

Table 4-8: Comparison of the IPD and IPLD statistics – TAnDEM example
Comparison of the average point estimate over 200 IPLD simulated datasets to the reported estimate for the simple ITT analysis in the TAnDEM trial.

	Point estimates, (95% CI)			
	Reported		Average over 200 simulated datasets	
No. of events				
A	64		63.9	(51, 78)*
T+A	58		58.6	(44, 74)*
Median OS				
A	23.9 months	(18.2, 37.4)	24.9 months	(18.2, 35.7)
T+A	28.5 months	(22.8, 42.4)	29.2 months	(22.8, 41.7)
HR	0.84	(0.59, 1.20)	0.86	(0.58, 1.20)

The secondary analysis for treatment switching is presented in Table 4-9. For Table 4-9, the reported information was extracted from the TA. Whilst the point estimate for the reanalysis was almost identical as the one reported in the appraisal, the uncertainty is greater in the re-analysis.

Table 4-9: Secondary analysis for treatment switching – TAnDEM example
Comparison of the secondary analysis, using an RPSFTM to account for treatment switching reported in the TA and the average across 200 simulated datasets

	Point estimates, (95% CI)			
	Reported		Average over 200 simulated datasets	
Median OS				
A	21.98 months		22.25 months	
HR	0.74	(0.39, 1.38)	0.73	(0.32, 1.66)
Acceleration factor	-		1.29	

4.8.3.2.4 Discussion

The method adjusts for treatment switching well, and attains a RPSFTM analysis which mirrors that obtained using the original IPD, although with more uncertainty. Therefore, this approach enables the re-analysis of data for treatment switching where only summary data and the Kaplan-Meier curve are available. In particular, this demonstrates the success in replicating the RPSFTM analysis that could have been obtained from the original IPD. Additionally, the reduction in the number of datasets needed, improved the efficiency of the reconstruction and analysis process, without adversely affecting the robustness.

The simulation of the switch times still causes concern. At present, all treatment switchers are assumed to switch after the same proportion of their OS time, which does not seem realistic. This is particularly because no variation from this proportion is permitted. One potential solution for this is to adapt the method to simulate a pair of PFS and OS times for each patient, provided a Kaplan-Meier was available for PFS as well as OS. This would also have the ability to improve the choice of treatment switchers (as described in Section 5.5). The practicalities and methodology to achieve this were developed and are explained in Chapter 5 and Chapter 6.

4.9 Discussion and further work

This proposed simulation technique is an exceptionally useful method for reconstructing IPLD. It competently reconstructs survival time data, maintaining a good level of accuracy, even with challenging features as to curve shape or line style. Based on findings from the illustrative example and reproducibility study, it can be concluded that this new method is an excellent alternative to the Guyot reconstruction algorithm (2012).

Evidence about re-analysing reconstructed data for treatment switching is perhaps less conclusive. This approach facilitates re-analysis that would be infeasible otherwise. For example, it may provide a reasonable proxy for the results from an IPD analysis. However, the way in which the treatment switching data are generated is of most concern. As it currently stands, the switching mechanism ignores selection process, and does not permit variation in switch time.

As discussed in Section 4.7.3.2.4, the most important future development is the simulation of PFS and OS paired data. Currently, the existing methods only reconstruct one outcome at a time, so whilst PFS and OS data could be generated using these methods, they would not be paired across patients, a necessity if treatment switching is to be accounted for. In terms of PFS and OS, the principle issue that has to be contended with is that, any patient who has died prior to progression, will have the same time at PFS and OS, and thus, if a simulation approach is used, this will have to be taken into account.

Another area of research in this field to be considered is developing even more flexible censoring distributions, for example, those handling interval censoring. Interval censoring is very common for outcomes such as PFS, and therefore, likely to appear in many examples. Given the flexibility of the framework, this approach could potentially be extended further to incorporate this. Other considerations include ensuring that the model fitted is always monotonic. This could be done by using a class of RCS models that fit monotonic functions. Alternatively, it could be useful to consider weighting the coordinates to improve the model fit.

Chapter 5: Reconstructing individual patient level data with two related outcomes

5.1 Chapter overview

This chapter extends the simulation technique proposed in Chapter 4 to facilitate the production of pairs of survival times for patients, primarily for PFS and OS data. This extension involves the use of an underlying 'illness-death' modelling framework. Chapter 5 follows the development of the methodology depending on what information is available to the analyst and how this might impact on the results, in terms of assumptions made. In addition, it discusses modifications to the methods outlined in Chapter 4 to reconstruct treatment switching information. The key approaches for this chapter are, furthermore, described using illustrative examples.

5.2 Motivation and aims

In order to address an example where only summary data are available and treatment switching has not been handled appropriately, a method by which the IPD are reconstructed and re-analysed (Chapter 4) was proposed. However, whilst the survival time information is relatively straightforward to reconstruct, many assumptions are required to provide treatment switching information (Boucher, 2013a). Thus far, a purely deterministic approach had been used, whereby, provided patients switched following progression, a patient's switch time was assumed proportional to their time of death, and the ratio of the median survival times for PFS and OS (Boucher, 2015). Primarily these switch times were used both as proxy for TTP and time to treatment switch. In addition, for this initial method all patients had an equal probability of switching treatments. Whilst in some examples this gave reasonable agreement with a corresponding analysis conducted on the IPD, it clearly ignored the biological processes underlying the treatment switching mechanism; and so was a key limitation of this approach (Boucher, 2015). The aim consequently became to develop a method which allowed variability in the switch time, i.e. all patients' switch time would not necessarily be the same proportion of their OS time, and which also included some type of selection process.

Any method used to reconstruct IPD typically depends on patients switching soon after disease progression (Boucher, 2013a); this occurs for two principle reasons, the first being

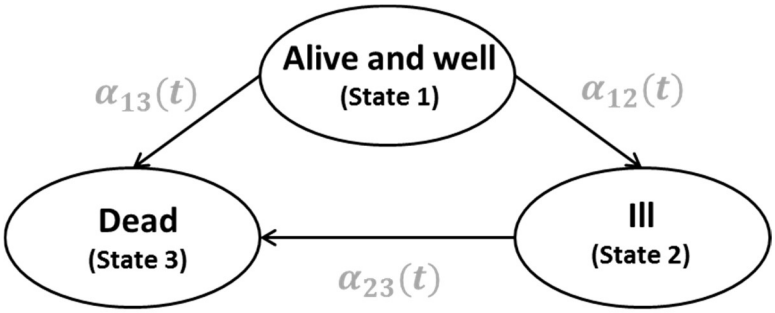
that is perhaps the most common situation, and the second because it provides a reference for constructing the switch time (Boucher, 2013, Morden, 2011, Latimer, 2013). However, thus far little use has been made of the PFS and / or TTP evidence provided, as only the median survival time for this outcome has been used. Indeed, given that many studies specifically choose PFS as the primary endpoint and OS as a secondary outcome; there is potentially a wealth of information that could be exploited further (Morden, 2011, Latimer, 2013).

To maximise the use of PFS or TTP, when the Kaplan-Meier curve for this outcome has been presented, IPLD could be generated, using the simulation approach (Chapter 4); likewise to the OS information. However, for the purpose of using this as a switch time, the simulated times for PFS and OS must be paired. Nevertheless, since PFS and OS are composite endpoints, and therefore not independent of each other, thought is required to allow for the inherent correlation. Considering the underlying structure of PFS and OS, leads to the necessity of using a framework based on an 'Illness-Death' model.

5.3 Structure of the data

5.3.1 An Illness-Death modelling structure

Figure 5-1: Standard illness-death model



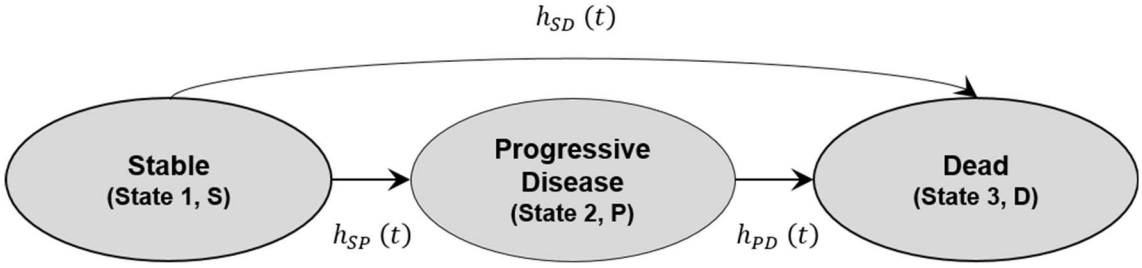
Structure of a typical three-state 'illness-death' model; with states: "Alive and well"; "Ill; and "Dead" and transition rates denoted $\alpha_{ij}(t)$ for the transition from the i^{th} to the j^{th} state.

'Illness-Death' models are a type of multistate model (Hinchliffe, 2013). 'Illness-death' models (shown in Figure 5-1) typically consist of three states: a state where patients are alive and well (state 1); a state where patients are ill (state 2); and finally, a state containing the patients who have died (state 3). Moving from one state to another is defined as a transition. There are three transitions in this model: from state 1 to state 2, that is to say, well patients become ill; from state 1 to 3, those who were alive and well

can die; and from state 2 to state 3, i.e. a patient who is ill could die. Associated with each transition is a specific hazard rate, also referred to as the transition rate. Where IPD are available, these hazard rates are estimated directly from the data. The probability of being in any one state at a given time can be calculated using the transition rates (Hinchliffe, 2013).

This thesis concentrates on studies in advanced or metastatic cancer, and thus, tend to be of a particular form (illustrated in Figure 5-2). All patients will start the study in question, with stable disease. They are followed-up to observe when, during the course of the trial they experience disease progression (e.g. an increase in tumour size). This occurs at a certain rate, denoted as: $h_{SP}(t)$ (previously referred to as $\alpha_{12}(t)$). However, some patients, often only a small proportion, will die before they reach the stage of disease progression. These are typically still included in the analysis for PFS. Assuming that the rate at which patients have died before experiencing disease progression is represented using $h_{SD}(t)$. Given that OS is included as an outcome, patients who have experienced disease progression continued to be followed-up to death. The rate at which this occurs will be indicated as $h_{PD}(t)$.

Figure 5-2: Illness-Death model with standard health states for cancer trials



Structure of a typical three-state 'illness-death' model in cancer trials; with states: "Stable (disease)"; "Progressive disease"; and "Dead" and transition rates denoted $h_{ij}(t)$ for the transition from the i^{th} to the j^{th} state.

Clearly, this is the necessary format for enabling pairs of times for progression and death for each patient in the trial to be simulated. As mentioned before, where IPD are accessible the rates $h_{SP}(t), h_{SD}(t), h_{PD}(t)$ are directly estimated. These parameters could then be used within a simulation approach. However, when only summary data are available this process is less straight-forward. In addition, the process largely depends on what level of summary data has been reported.

The survival functions for PFS and OS, denoted $S_{PFS}(t)$ and $S_{OS}(t)$ respectively are composite functions of the transition rates, and can be written as follows:

$$S_{PFS}(t) = \exp\left[-\left(H_{SP}(t) + H_{SD}(t)\right)\right] \tag{5-1}$$

$$\begin{aligned} S_{OS}(t) = \exp\left[-\left(H_{SP}(t) + H_{SD}(t)\right)\right] \\ + \exp[-H_{PD}(t)] \int_0^t h_{SP}(s) \exp\left[-\left(H_{SP}(s) + H_{SD}(s)\right.\right. \\ \left.\left. - H_{PD}(s)\right)\right] ds \end{aligned} \tag{5-2}$$

Where,

$$H_i(t) = \int_0^t h_i(u)\ du \tag{5-3}$$

Alternatively, $S_{OS}(t)$ can be written as:

$$S_{OS}(t) = S_{PFS}(t) + \exp[-H_{PD}(t)] \int_0^t h_{SP}(s) S_{PFS}(s) \exp[H_{PD}(s)]\ ds \tag{5-4}$$

Or,

$$S_{OS}(t) = S_{PFS}(t) + S_{PD}(t) \int_0^t \frac{h_{SP}(s) S_{PFS}(s)}{S_{PD}(s)}\ ds \tag{5-5}$$

And where,

$$S_i(t) = \exp[-H_i(t)] \tag{5-6}$$

This essentially means that the OS distribution is composed of complicated functions involving both PFS and time from progression to death (also known as post-progression survival (PPS)).

5.3.2 Available information

The amount of information varies considerably from study to study, with each choosing to report in a different way. Ideally, to use an underlying 'illness-death' modelling structure, information on all the transitions and the censoring distribution would need to be reported, particularly in terms of Kaplan-Meier curves. Two of the three transitions relate to commonly discussed outcomes: TTP equivalent to $S_{SP}(t)$ and PPS – essentially $S_{PD}(t)$. However, the frequency of these outcomes being reported is relatively low. TTP is sometimes reported alongside or instead of PFS, and is perhaps the most common outcome after PFS and OS. PPS is often discussed, but rarely reported. The last transition is not easily meaningful in practice, the time taken to death where death occurs before

documented disease progression, and hence is very unlikely to be reported. Similarly, time to censoring is not often of interest, and so also not reported. Given the rarity of these transitions being reported (and additionally all being reported together), it is highly improbable that, if a method could be developed using the underlying 'illness-death' modelling structure, it could be used in practice.

Instead of these transition rates, the outcome of PFS and OS are almost always reported. Since these are composite endpoints of the transition rates, in theory, and making some strong assumptions, it should be possible to extract and model the necessary parameters to enable the IPD to be reconstructed. However, determining what these assumptions should be, remained very unclear.

5.4 Exploring the levels of information available

The variety in reporting outcomes, and difficulty in identifying and constructing suitable assumptions, was key in starting to formulate this problem and methodology. Therefore, it was decided to take a gradual and systematic approach in which specific information (e.g. transitions rates and censoring) became more limited and general (e.g. composite endpoints) information became more available. This is shown in Table 5-1.

Essentially these situations translate to examples where:

1. Kaplan-Meier curves are available for all transitions, and a Kaplan-Meier for the censoring distribution (direct information on censoring)
2. Kaplan-Meier curves are available for all transitions, and risk tables (indirect information on censoring)
3. Kaplan-Meier curves are available for PFS, TTP (equivalent to the transition from stable to disease progression), and the transition from progression to death
4. Kaplan-Meier curves for PFS, TTP and OS are available and risk tables are available
5. Kaplan-Meier curves for PFS and OS

It should be noted that the 'number at risk' tables for time to death (before progression), TTP and PFS will all be identical.

The aim of using all these scenarios was to provide a variety of methods that would (1) address the different levels of information which could possibly be observed, and (2) to

obtain the following necessary stages to developing a method purely relying on PFS and OS.

The stages were:

- How to estimate the transition from stable to death when there is only PFS and TTP (scenario 3)
- How to estimate the transition from progression to death using only PFS, TTP, and OS (scenario 4)
- How to estimate the TTP survival from only PFS and OS risk tables and / or Kaplan-Meier curves (scenario 5)
- How to estimate the censoring distribution from risk tables, particularly with regard to the post-progression phase (from scenario 2 onwards)

Table 5-1: Information Scenarios

Outcome	**Statistic**	**Scenario**				
		1	**2**	**3**	**4**	**5**
Time to death, $S_{SD}(t)$ (before progression)	Kaplan-Meier curve	✓	✓	×	×	×
	Risk table	-	-	-	-	-
Time to progression, $S_{TTP}(t)$ $(S_{SP}(t))$	Kaplan-Meier curve	✓	✓	✓	✓	×
	Risk table	-	-	-	-	-
Post progression survival, $S_{PPS}(t)$ $(S_{PD}(t))$	Kaplan-Meier curve	✓	✓	✓	×	×
	Risk table	-	✓	✓	×	×
Time to censoring	Kaplan-Meier curve	✓	×	×	×	×
Progression-free survival, $S_{PFS}(t)$	Kaplan-Meier curve	-	-	✓	✓	✓
	Risk table	-	✓	✓	✓	✓
Overall survival, $S_{OS}(t)$	Kaplan-Meier curve	-	-	-	✓	✓
	Risk table	-	-	-	✓	✓

✓ : This information is available to use and necessary for this method
× : This information has not been reported
- : This information may or may not be available but is not integral to the method

5.4.1 Methods development

As described above, there is a clear dependency on the available information and the approach to be taken.

5.4.1.1 *All information (Scenario 1)*

5.4.1.1.1 Background

Whilst a situation where summary information is available on all of the transitions and the censoring is exceptionally rare, developing and evaluating an approach that could be used here is essential. Not only would it determine the format to ultimately be emulated with more limited information, but it would also provide an assessment of whether the methodology would perform satisfactorily at all. In other words, assuming this is the 'gold standard' level of summary data, if an approach with all transitions and censoring information performed poorly, there would be little point in extending this underlying methodology further to rely on less detailed information. Another reason is that manufacturers may be more prepared to produce this more detailed summary information (e.g. K-M curves for individual transitions), rather than providing direct access to the IPD. As a consequence, this could offer a suitable compromise to both parties.

5.4.1.1.2 Overview of method

For this approach, it is assumed that there is a Kaplan-Meier curve for each transition, and for the time to censoring (four Kaplan-Meier curves in total).

The method has seven key steps:

1. Extract coordinates for each outcome and each treatment group
2. Model the survival (or censoring) distribution using RCS models fitted on the log-cumulative hazard scale
3. Simulate a time from each transition / time to censoring model for that specific simulated treatment group
4. Combine the information together to define the survival outcome
    a. PFS information: take the minimum of the simulated TTP, time to death (before progression) and censoring as the observed survival time and event type (progression, death or censoring)
    b. OS information: for all patients defined as progression, take the minimum of the simulated PPS and censoring times as the observed PPS and event type. Combine this newly defined PPS information with the PFS censorings and deaths to complete the OS information
5. Analyse the dataset

6. Simulate and analyse additional datasets (using steps 3 - 5)
7. Average over all the datasets to obtain a point estimate

5.4.1.1.3 Stages 1 and 2: Extracting coordinates and modelling the survival distribution

Stages one to three are very simple extensions of the method outlined in Section 4.4. Hence, all the points relating to extracting the coordinates still hold for this method. The principal difference is that here there are more than one outcome of interest. Therefore, for every single Kaplan-Meier curve needed, and for each treatment group, coordinates must be extracted. For a two arm trial this would give eight sets of coordinates (four per treatment group). Each of these sets of coordinates has to be transformed onto the log-cumulative hazard and log time scale for the survival and time coordinates respectively. Then individually for each treatment group and outcome, RCS functions are calculated. Then, using least squares regression, a model is constructed for the survival / censoring distribution. Since it is assumed that there is a Kaplan-Meier curve reported for time to censoring, there is no need to use any of the techniques outlined in section 4.4.1.4 (e.g. 'numbers at risk' or recruitment times). Instead, it is directly modelled.

5.4.1.1.4 Potential issues with post-progression survival

The greatest complication with this method is the definition of PPS. By using an 'illness-death' modelling structure, potentially two timescales are being introduced: time from randomisation and time between transitions. However, of the three transitions only PPS survival is affected (the other transitions move from stable disease – the starting state).

Due to the difference in timescale, there are two possible definitions of PPS: (1) PPS with 'delayed entry' – the timescale starts from randomisation, but each patient enters the risk-set only after experiencing disease progression; (2) PPS with the 'reset the clock' time scale – a person's survival time is measured from the time of their progression (i.e. their progression time becomes t = 0)

The following example will describe this more clearly. Supposing a patient was enrolled on the trial, and then had documented disease progression at two months, and died six months later. In terms of the 'delayed entry' notation, this person would be denoted as

entering the risk set for PPS at time two months, and leaving the risk set (due to an event) at eight months. For the 'reset the clock' approach, this person would be recorded as entering the risk set at time zero, and leaving at six months.

The definition used has several important implications on the simulation technique employed and underlying assumptions. If the 'reset the clock' approach is adopted then essentially PPS is completely independent of any of the other transitions (i.e. it does not relate to TTP; e.g. all earlier progressors are not also all the earlier deaths). Therefore, this can be simulated in exactly the same way as the other outcomes.

On the other hand, if the 'delayed entry' technique has been employed, the theory is quite different. Here, there is conditioning on the TTP, which also needs to be built into the method. Times are still simulated from the PPS model but conditional on the simulated TTP. The principle advantage of this, is that it allows for a relationship between TTP and PPS (e.g. a longer TTP survival leads to longer PPS).

In practice, it is more likely that, if PPS is reported, the 'reset the clock' approach will be used, as this has a more meaningful practical interpretation, time from progression until death (rather than time from randomisation to death for patients who experienced disease progression). However, using the 'reset the clock' approach in this methodology is only really justified if the analyst fervently believes that there is no relationship between TTP and PPS. This is described further in Figure 5-3 and Figure 5-4.

5.4.1.1.5 Stage 3: Simulating the survival

For the TTP, time to death (before progression) and time to censoring, the times for these outcomes can be simulated in the same way as in Section 4.4.1.4.3. The only outcome which is different is the post-progression, which as explained in the section above (Section 5.4.1.1.4), depends on how it has been defined (using 'reset the clock' or 'delayed entry').

5.4.1.1.6 Stage 4: Combining the simulated data together

For the transition-specific simulated times the interpretation in the top panel is different depending on the PPS method. For delayed entry, this is the time from randomisation to event; for 'reset the clock', this is the time between transitions.

Combining the simulated times together essentially forms two key stages: combining the PFS data; and combining the OS data. For PFS, one of three things can happen to a patient: either they experienced progression; they died (before progression) or they were censored. To define which type of event a patient had, the minimum of the three simulated times is taken as the observed survival time. The choice of survival time, also defines the type of event.

OS is slightly more complicated to estimate, as some of the information carries over from PFS. Since patients that have died before progression have already been defined, only calculating those that died after progression is of interest. In addition, any patients defined as 'censored' for PFS are temporarily ignored (this imposes the assumption that patients who were censored for PFS were not followed up after this time for PPS). For all patients who progress, time from randomisation to death for patients who progress, and time to censoring are compared. Once again, the observed OS time and status, for this subgroup of the population is the minimum of the times being compared. This newly defined PPS information is then combined with the PFS deaths and censorings to complete the OS data.

As well as differences in the simulation, there are also differences when combining the information together depending on which technique was used to calculate the PPS. Figure 5-3 and Figure 5-4 illustrate the differences. Figure 5-3 shows the process when the PPS is calculated using the 'delayed entry' technique; whilst Figure 5-4 illustrates the approach when 'reset the clock' method is employed. The top panel shows example simulated transition and censoring times for four people. This example has been specially designed so that it clearly shows that there is a level of correlation between TTP and PPS for the 'delayed entry' format, with the patients with shorter

Figure 5-3: 'Delayed entry' format for PPS

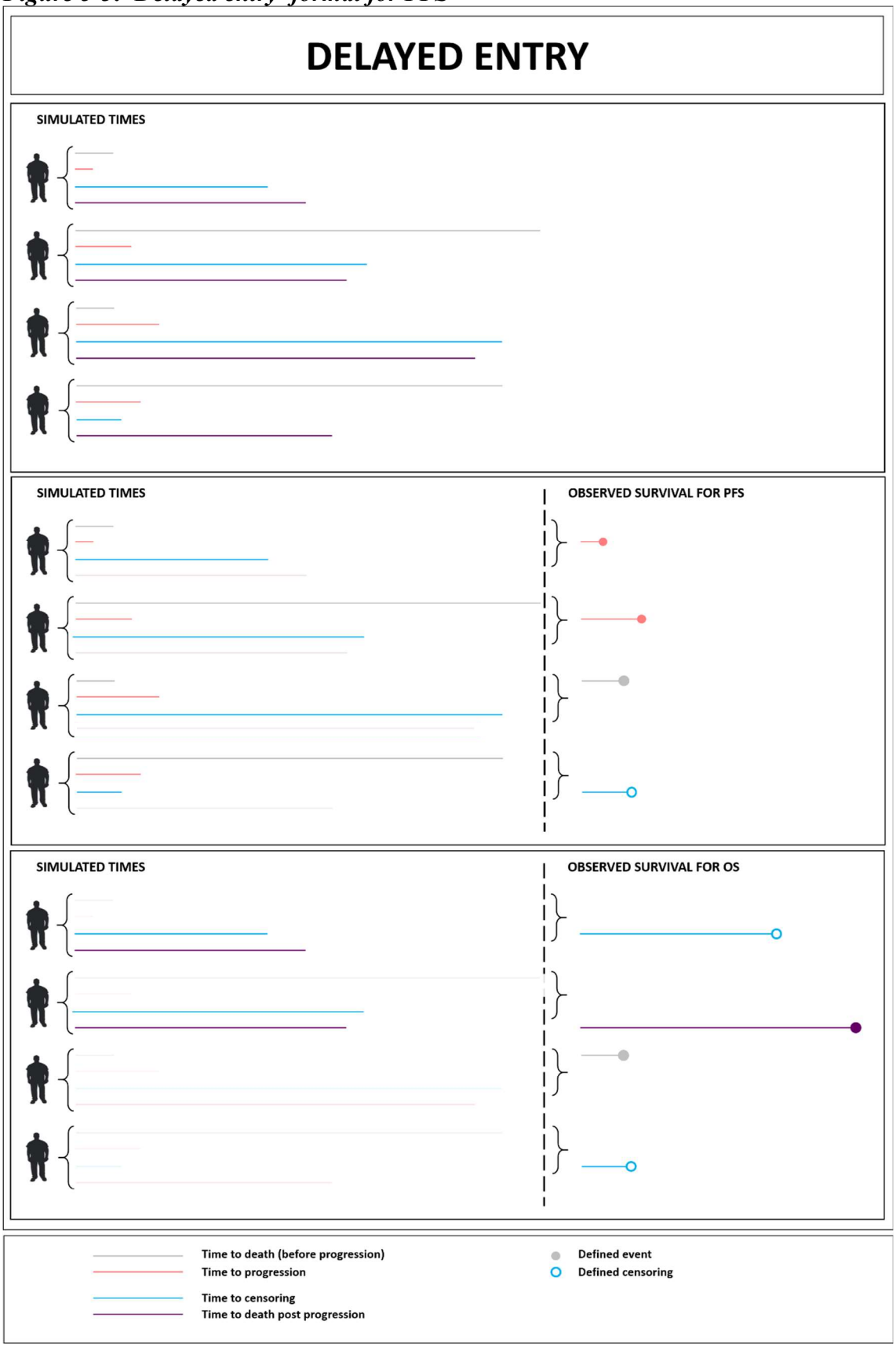


Diagram illustrating how the simulated times are combined to define the PFS and OS data if the 'delayed entry' approach is used.

Figure 5-4: 'Reset the clock' format for PPS

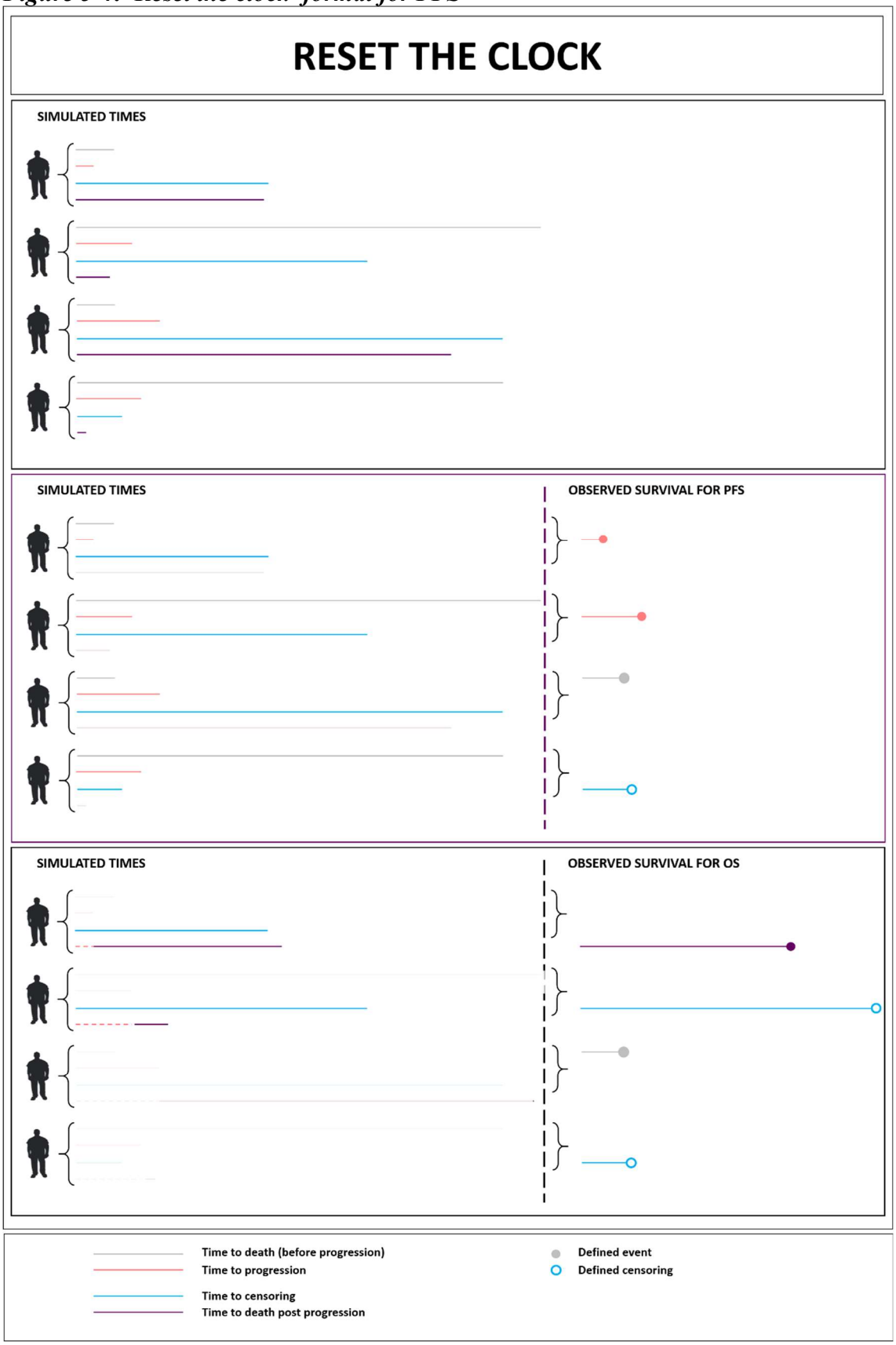


Diagram illustrating how the simulated times are combined to define the PFS and OS data if the 'reset the clock' approach is used.

simulated TTP times, also having shorter PPS times. In contrast, for the 'reset the clock' diagram, some of the patients with earlier TTP times, have a longer PPS than those with a longer TTP time.

Essentially, the middle panels are identical for both PPS approaches, since the comparison is between the times for TTP, death before progression and censoring, and thus PPS information is temporarily being ignored. It is also in the bottom panel that the differences become noticeable again.

For Figure 5-3, the last two simulated times (those for death post progression and time to censoring), are considered for patients defined as having progressed. This means, for the moment, the information on TTP or death before progression must be ignored, as is any simulated data on patients defined as censored or dead for PFS (e.g. the last two patients). The minimum of these is then taken, as the observed survival time and status. For Figure 5-4, it is slightly different. As before, the information on TTP or death before progression, and any simulated data on patients defined as censored or dead for PFS is temporarily ignored. The difference with this approach is that the simulated PPS time here is only time from progression to death, so first, the time from randomisation to death must be calculated in order to contrast it with the time to censoring. This is simply done by adding the TTP and PPS times together. Once this has been obtained, the minimum of the newly calculated PPS (from randomisation) and the time to censoring is taken as a patient's observed survival time and status.

5.4.1.2 *Estimating censoring in the presence of all transitions (Scenario 2)*

Scenario 1 examined the 'gold standard' level of summary information, in which, in addition to the summary data on the transitions, there was also a Kaplan-Meier curve for censoring. In practice, though, this would hardly, if ever, be actually reported. Therefore, censoring must be estimated in a different way. Once again, the methodology established in Chapter 4 is used to generate the censoring distribution with some minor differences. All three techniques: maximum study time; recruitment times; and 'numbers at risk' table can be used in this setting.

5.4.1.2.1 Maximum study time censoring

This extends very simply to the case where two outcomes are of interest. To use this method, the maximum study time over both outcomes for a specific treatment arm, is estimated, usually by reading it from the Kaplan-Meier curve. Then, this time is treated as the patient's 'simulated' time of censoring, and thus the data can be defined in the way outlined in Section 5.3.1.1.6.

5.4.1.2.2 Recruitment times censoring

Similarly, to the maximum study time censoring technique, the method using recruitment times also automatically extends to the reconstruction of paired data. Provided that the recruitment times and data cut-off date have been reported, a uniform distribution between the minimum and maximum length of follow-up can be determined for the censoring times distribution, where the minimum and maximum length of follow-up are defined as in Section 4.4.1.4.2. A single censoring time is then generated for each patient from this uniform distribution, and this time used for both PFS and PPS.

5.4.1.2.3 'Numbers at risk' approach

This approach is the most different of all the three possible censoring distributions when translated to the paired data setting (i.e. PFS / OS). It is heavily dependent on the information that has been reported. Under this framework, the PFS censoring must be separated from the PPS censoring.

To gauge PFS censoring, the event type (death or progression) must essentially be ignored, and hence use the PFS survival proportion. This can be achieved in one of two ways. The first is by calculating the survival proportion for both death before progression and TTP at the corresponding times to the 'numbers at risk' interval and then multiplying them. In this method, the approximate value of the survival, $\hat{S}_{PFS}(t)$ at a given time, t^* is:

$$\hat{S}_{PFS}(t^*) = \hat{S}_{TTP}(t^*)\,\hat{S}_{SD}(t^*) \tag{5-7}$$

Alternatively, if the PFS K-M curve is available, it may be better, and potentially easier to directly estimate the PFS survival proportion from that (either from visual inspection, or more preferably by taking coordinates at the interval time points). Once the PFS

survival proportion has been obtained, the estimation and censoring distribution continues in exactly the same way as it would have for a single outcome.

For the scenarios where PPS is explicitly reported, there remain differences in the methods depending on how the PPS was calculated. If the 'reset the clock' technique has been used, then calculating the PPS censoring becomes simple: it is essentially calculated using the method for a single outcome.

If the 'delayed entry' timescale has been used; this is far more complicated. In a 'delayed entry' risk-table, the number of patients at time zero will actually be zero as patients should not enter the risk-set until after progression, and progression should commence after time zero. Therefore, this risk-set, unlike most, will increase as well as decrease in number over time. However, this means that the equation used to calculate censoring is not clear, as the change in the number at risk over the i^{th} interval (denoted y_i) is not evidently defined. This is because, at present, this term has not been divided into progressions and those experiencing a PPS event, a crucial factor in determining the number of PPS censorings). This can be expressed as:

$$y_{PPSi} = d_{PPSi} + c_{PPSi} - d_{TTPi} \tag{5-8}$$

Where y_{ji}, c_{ji}, $d_{ji,}$ are the change in the number at risk, the number of censored observations and the number of events respectively over the i^{th} interval, for outcome j (e.g. TTP, PPS etc.).

It is imperative that the change in y purely due to leaving the risk set, which shall be defined as y^*, is estimated.

$$y^*_{PPSi} = d_{PPSi} + c_{PPSi} \tag{5-9}$$

In order to do this, the number of progressions must first be calculated. At this stage, essentially, this is obtained in a similar way to the number of censorings. Having already calculated the number of PFS censorings, the number of PFS events can easily be obtained.

$$d_{PFSi} = y_{PFSi} - c_{PFSi} \tag{5-10}$$

The probability of having experienced an event over the course of the i^{th} interval for outcome j, is defined as q_{ji}.

$$q_{ji} = 1 - p_{ji} \tag{5-11}$$

Where, p_{ji} is the probability of not having an event.

Without currently making any assumptions about the ordering of the deaths before progression throughout the progressions, the interval specific formula for interval, i, is:

$$q_{TTPi} = \frac{d_{TTPi}}{n_{PFSi} - \left(\frac{1}{2} c_{PFSi} + \gamma_i \, d_{SD}\right)} \tag{5-12}$$

Where n_{ji} is the number at risk for the start of the i^{th} interval for outcome j and γ_i is the average proportion of interval patients, who are censored, are at risk for.

This reduces further to

$$q_{TTPi} = \frac{d_{TTPi}}{n_{PFSi} - \left(\frac{1}{2}(y_{PFSi} - d_{PFSi}) + \gamma_i \, (d_{PFSi} - d_{TTPi}\right)} \tag{5-13}$$

$$d_{TTPi} = \frac{q_{TTPi} \left(n_{PFSi} - \frac{1}{2} y_{PFSi} + d_{PFSi} \left(\frac{1}{2} + \gamma_i\right)\right)}{1 + \gamma_i \, q_{PFSi}} \tag{5-14}$$

Once these estimates for the progression have been obtained, all that remains is their inclusion in the equation for the probability of being censored post-progression, p_{PPCi} (given in Equation (5-15)):

$$p_{PPCi} = 1 - \frac{c_{PPSi}}{(n_{PPSi} + \alpha_i \, d_{TTP}) - \frac{1}{2} d_{PPSi}} \tag{5-15}$$

Once again, returning to y^*, there is:

$$p_{PPCi} = 1 - \frac{c_{PPSi}}{(n_{PPSi} + \alpha_i \, d_{TTPi}) - \frac{1}{2}(y^*_{PPSi} - c_{PPSi})} \tag{5-16}$$

This formula has been generalised in Appendix I.

To complete the censoring distribution, this probability in conjunction with the length of interval (l_i) is included within a piecewise exponential model, such that, for each interval,

$$S_{censi} \sim Exponential(\lambda_i), \qquad \text{where } \lambda_i = -\frac{\ln p_{PPCi}}{l_i} \tag{5-17}$$

5.4.1.3 *Reconstructing the data without one of the transition rates (relating to PFS) (Scenario 3)*

5.4.1.3.1 Estimating event type for PFS when one of the relevant transitions is missing

In practice, the transition for stable to death is exceptionally unlikely to have been reported as it is not usually of particular interest since TTP, PFS and OS are often seen as more relevant outcomes. Provided that PFS data are available, though, this missing transition can be estimated, since:

$$S_{PFS}(t) = S_{TTP}(t)\, S_{SD}(t) \tag{5-18}$$

In principle then, provided that coordinates from the Kaplan-Meier curves for TTP and PFS have been extracted and modelled, the survival for time to death before progression can be simulated from:

$$S_{SD}(t) = \frac{S_{PFS}(t)}{S_{TTP}(t)} \tag{5-19}$$

However, since the ultimate aim is the simulation of PFS data, where alongside each event time is the event type (progression or death), there is an alternative approach. Beyersmann (2009) proposed a method by which data with two competing risks could be simulated using distributions for the all-cause survival and only one of the cause-specific hazards. This would be exceptionally useful as essentially this is the information that is available, as there are two competing risks (death and progression) and the all-cause survival, PFS. Using this approach, the probability for a particular event type X_T, that for which the cause-specific hazard, $\alpha_{01}(t)$, is available can be defined as:

$$P(X_T = 1 \,|T \in \mathrm{d}t, |T \geq t) = \frac{P(T \in \mathrm{d}t, X_T = 1 \,|T \geq t)}{P(T \in \mathrm{d}t \,|T \geq t} = \frac{\alpha_{01}(t)}{\alpha_{01}(t) + \alpha_{02}(t)} \tag{5-20}$$

Where $\alpha_{02}(t)$ is the cause-specific hazard for the competing risk.

Therefore, in the context of this research, the probability of the event being a progression, X_{TTP}, where X_{TTP} = 1 indicates a progression, X_{TTP} = 0 indicates a death (event due to competing risk), is:

$$(X_{TTP} = 1 \,|T \in \mathrm{d}t, |T \geq t) = \frac{h_{TTP}(t)}{h_{TTP}(t) + h_{SD}(t)} = \frac{h_{TTP}(t)}{h_{PFS}(t)} \tag{5-21}$$

Since,

$$h_{PFS}(t) = h_{TTP}(t) + h_{SD}(t) \tag{5-22}$$

Beyersmann's, (2009) method is relatively simple to implement here. The first stage is to simulate the event times from the all-cause, in this case PFS, distribution. Once an event time has been simulated for all patients in the dataset. The hazard rate at the patient's specific simulated survival time needs to be estimated for both TTP and PFS (one of the cause-specific hazards and the all-cause hazard rate). When these have been obtained, the ratio can be calculated. Probabilities close to zero suggest that these patients were more likely to have died than progressed, whilst probabilities close to one, indicate these patients most likely experienced disease progression. This patient-specific probability is then included within a Bernoulli distribution to formally define the event type.

Whilst this approach has primarily been developed, with the idea of the cause-specific information being that of TTP, this methodology would still hold in the rare case of death before progression having been reported alongside PFS, but no data for TTP. The only difference is that the simulated event type definition would be zero for a progression and one for a death.

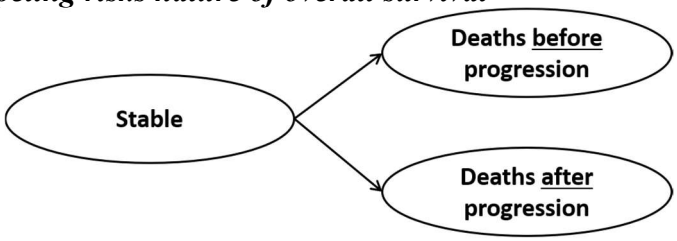
5.4.1.4 *Reconstructing the data with TTP, PFS and OS (Scenario 4)*

In this scenario, there is no longer information specifically on PPS. PPS is an outcome which appears to be seldom reported. However, obtaining information on this is integral to the method developed thus far. Particularly if the motivation is for treatment switching patients who switch on progression, then PPS becomes vital.

5.4.1.4.1 Calculating PPS from OS (Scenario 4)

This section assumes that in the absence of PPS, OS information has still been reported. Essentially, OS could be viewed as a competing risks problem, in which patients either die before progression or after progression. This is illustrated in Figure 5-5.

Figure 5-5: Competing risks nature of overall survival



In order to obtain an expression for PPS, the deaths after progression must be isolated. Therefore, the formula used to calculate OS must be examined.

The equation for OS can be written as follows:

$$S_{OS}(t) = S_{PFS}(t) + S_{PPS}(t)\int_0^t \frac{h_{TTP}(s)S_{PFS}(s)}{S_{PPS}(s)}\,ds \qquad (5\text{-}23)$$

In other words, the proportion of patients who are alive ($S_{OS}(t)$) is composed of the proportion of patients who are alive and with stable disease ($S_{PFS}(t)$), and those who are alive but with progressive disease ($S_{PPS}(t)\int_0^t \frac{h_{SP}(s)S_{PFS}(s)}{S_{PPS}(s)}\,ds$). This expression does contain the function of interest, $S_{PPS}(t)$, but is not currently in an analytical form that can be simulated from. Therefore, Equation (5-22) must be rearranged to give an expression for $S_{PPS}(t)$. To do this, however, it is first easier to obtain an expression for $h_{PPS}(t)$ because of $S_{PPS}(t)$ being explicitly involved in the integral.

Initially the Equation (5-22) is rearranged so that the integral is one side and the other terms on the other:

$$\frac{S_{OS}(t) - S_{PFS}(t)}{S_{PPS}(t)} = \int_0^t \frac{h_{TTP}(s)S_{PFS}(s)}{S_{PPS}(s)}\,ds \qquad (5\text{-}24)$$

Using the fact that $S(t) = \exp\big(-H(t)\big)$ the following is obtained:

$$\frac{\exp\big(-H_{OS}(t)\big)}{\exp\big(-H_{PPS}(t)\big)} - \frac{\exp\big(-H_{PFS}(t)\big)}{\exp\big(-H_{PPS}(t)\big)} = \int_0^t \frac{h_{TTP}(s)S_{PFS}(s)}{S_{PPS}(s)}\, ds \qquad (5\text{-}25)$$

And then using the rules of exponentials, the following expression is obtained:

$$\exp\big(H_{PPS}(t) - H_{OS}(t)\big) - \exp\big(H_{PPS}(t) - H_{PFS}(t)\big) = \int_0^t \frac{h_{TTP}(s)S_{PFS}(s)}{S_{PPS}(s)}\, ds \qquad (5\text{-}26)$$

This expression can then be differentiated with respect to t, giving the

$$\begin{aligned}\big(H'_{PPS}(t) - H'_{OS}(t)\big)&\exp\big(H_{PPS}(t) - H_{OS}(t)\big) \\ &- \big(H'_{PPS}(t) - H'_{PFS}(t)\big)\exp\big(H_{PPS}(t) - H_{PFS}(t)\big) = \frac{h_{TTP}(t)S_{PFS}(t)}{S_{PPS}(t)}\end{aligned} \qquad (5\text{-}27)$$

Since $\frac{d}{dt}\big(H(t)\big) = h(t)$ and as $S(t) = \exp\big(-H(t)\big)$, the equation below is attained:

$$\frac{\big(h_{PPS}(t) - h_{OS}(t)\big)S_{OS}(t) - \big(h_{PPS}(t) - h_{PFS}(t)\big)\, S_{PFS}(t)}{S_{PPS}(t)} = \frac{h_{TTP}(t)S_{PFS}(t)}{S_{PPS}(t)} \qquad (5\text{-}28)$$

So,

$$\big(h_{PPS}(t) - h_{OS}(t)\big)S_{OS}(t) - \big(h_{PPS}(t) - h_{PFS}(t)\big)\, S_{PFS}(t) = h_{SP}(t)S_{PFS}(t) \qquad (5\text{-}29)$$

Now that Equation (5-29) no longer contains $h_{PPS}(t)$ within an integral, the equation can be re-arranged to make $h_{PPS}(t)$ the subject. At last Equation (5-30) obtained,

$$h_{PPS}(t) = \frac{\big(h_{TTP}(t) - h_{PFS}(t)\big)S_{PFS}(t) + h_{OS}(t)S_{OS}(t)}{S_{OS}(t) - S_{PFS}(t)} \qquad (5\text{-}30)$$

Where the terms on the right-hand side of the equation can be evaluated using the models developed on the coordinates for each outcome.

This function is then integrated using integration techniques to obtain the cumulative hazard function. In practice, this was implemented using the 'integ' command in Stata. The obtained values were then transformed onto the log-cumulative hazard scale. RCS models are then used to obtain a final model for PPS.

It should be noted that here the type of PPS being used is clearly defined. Since PPS is calculated through progression, the underlying structure is 'delayed entry'; since OS is measured from randomisation. Therefore, when simulating the PPS data, the survival times are all conditional on the simulated TTP.

5.4.1.4.2 Calculating PPS censoring (Scenario 4)

The PPS censoring distribution also becomes more complex to estimate if the 'numbers at risk' table approach is used, since the information for this now needs to be extracted from the OS risk-table. This relies on making additional assumptions.

The process starts by calculating the number of censorings for PFS and OS. This is done, as if each outcome was separate (i.e. using the same approach for estimating the number of censorings if only one outcome was of interest – Section 4.4.1.4). Once these have been obtained, the number of events for each outcome can also be computed. It becomes crucial to have calculated the number of progressions. It should be noted that this approach estimates PPS, using the 'delayed entry' technique, and so issues caused by patients entering as well as leaving the risk set apply here.

Assuming d_{ji}, c_{ji} are the number of events and censorings respectively for the outcome j and interval i, then:

$$d_{SDi} = d_{PFSi} - d_{TTPi} \tag{5-31}$$

Where d_{SDi} is the number of deaths before progression for the i^{th} interval.

Since,

$$d_{OSi} = d_{SDi} + d_{PPSi} \tag{5-32}$$

And,

$$c_{OSi} = c_{PFSi} + c_{PPSi} \tag{5-33}$$

The post-progression events and censorings can easily be obtained by rearranging these expressions. These values for d_{PPSi}, c_{PPSi} are then included in the formula (Equation (5-16)), with 'delayed entry' detailed in Section 5.4.1.2.3.

5.4.1.5 *Reconstructing the data with only PFS and OS (Scenario 5)*

This last scenario differs from the last in that TTP is no longer available. This is a very crucial part of the 'illness-death' model and must therefore be indirectly estimated from

any information which is available. It is assumed that the total number of progressions for a given treatment arm has been reported.

5.4.1.5.1 Calculating TTP

Essentially this stage has two ultimate aims:

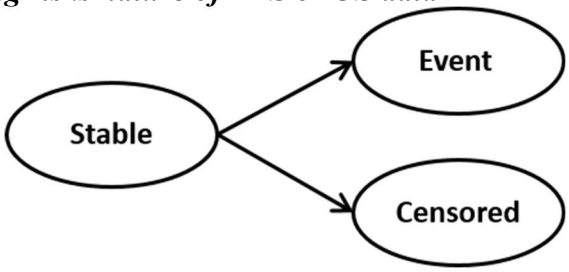
1) to estimate the number of progressions needed to calculate the post-progression censoring (PPC);
2) to obtain a model for the TTP hazard rate in order to define the event type and calculate the PPS.

The first part largely relies on knowing the number of PFS and OS events (as calculated from stage 2) since the number of events, d_{ji} for interval i and outcome j is:

$$d_{ji} = y_{ji} - c_{ji} \tag{5-34}$$

However, in order to estimate the next part of this stage, modelling the TTP hazard rate, more accurately, it is beneficial to have smaller intervals than those typically reported in the 'at risk table'. Therefore, each interval is divided into smaller intervals, known as partitions. The number of partitions is arbitrary; and so, in theory every interval could contain the same number of partitions. Nevertheless, it is more reasonable to have more partitions where there are more events. Therefore, assuming the number of coordinates to be a proxy for the number of events, the number of partitions is usually the number of coordinates in the interval, or possibly the square root of the number of coordinates (if there are many coordinates). The survival proportion must then be calculated, and / or more importantly the probability of survival for each partition is needed, so that the number of events over that partition can be calculated. Partitions will use the subscript k and intervals i.

Figure 5-6: Competing risks nature of PFS or OS data



PFS or OS data can be represented by a 'competing risks' format, as shown in Figure 5-6, where patients can only either be censored or experience an event (either PFS or OS depending on the outcome). Therefore, the probability of patients having an event or being censored can be expressed using the following formulae (Lambert, 2010):

$$p_{event_k} = \int_0^t S_{cens}(s) S_{event}(s) h_{event}(s) ds \quad (5\text{-}35)$$

$$p_{cens_k} = \int_0^t S_{cens}(s) S_{event}(s) h_{cens}(s) ds \quad (5\text{-}36)$$

Whilst, in practice, given that the models are piecewise exponential and RCS models (Durrleman, 1989), these functions cannot be integrated analytically, but can be evaluated using numerical integration techniques. Once the probabilities have been obtained for each partition, using these relevant probabilities and the number of events / censorings over the whole interval, estimates can be calculated for the number of events, number of censorings and the number at risk at the start of each partition.

Now that the number of events for both PFS and OS has been obtained, the minimum and maximum possible number of progression events per partition can be calculated. Basically, the maximum number of progressions for any partition is the number of PFS events, since essentially this means that no patients died before progression for that particular partition. The minimum number of progressions, $prog_{\mathrm{MIN}_k}$, is calculated by assuming that the maximum possible number of deaths before progression occurred, and hence is:

$$prog_{\mathrm{MIN}_k} = d_{PFS_k} - \min\left(d_{PFS_k}, d_{OS_k}\right) \quad (5\text{-}37)$$

Where, d_{PFS_i} , d_{OS_i} are the estimated number of events for PFS and OS respectively.

Assuming that the actual number of progressions, D, is between the total minimum,D_{MIN}, and the maximum D_{MAX}, number of events. Essentially the estimated number of progressions, d_{TTP_k} for the k^{th} partition is calculated as:

$$d_{TTP_k} = prog_{MIN_k} + \beta_k\left(prog_{MAX_k} - prog_{MIN_k}\right) \quad (5\text{-}38)$$

Here β_k has been chosen to be proportional to the difference between the total minimum and maximum number of progressions. Thus,

$$d_{TTP_k} = \frac{prog_{MIN_k}(D_{MAX} - D) + prog_{MAX_k}(D - D_{MIN})}{D_{MAX} - D_{MIN}} \quad (5\text{-}39)$$

Once the number of progressions has been obtained, a similar technique to that used in relative survival (Pohar Perme, 2012, Dickman, 2015) is used, and the excess hazard i.e. the difference between the hazard rate for TTP and PFS model. The process then continues by calculating the risk time over the interval,

$$\Delta_k = \left(n_k - \frac{1}{2}\left(c_{PFS_k} + d_{PFS_k}\right)\right) l_k \quad (5\text{-}40)$$

Where $\Delta_k, n_k, c_{PFS_k}, d_{PFS_k}, l_k$ are the risk time, number at risk, censorings, events and length of the k^{th} partition.

The excess hazard rate Λ_k is then calculated as:

$$\Lambda_k = \frac{\left(d_{PFS_k} - d_{TTP_k}\right)}{\Delta_k} \quad (5\text{-}41)$$

This rate is then modelled using RCS models to ensure flexible smooth curves (Durrleman, 1989), which gives a model for the excess hazard rate, $\Lambda(t)$.

$$h_{PFS}(t) = h_{TTP}(t) + \Lambda(t) \quad (5\text{-}42)$$

5.4.1.5.2 Modifications to existing methodology

Now that specific information on the TTP distribution is no longer accessible, the excess hazard (difference between the PFS and TTP hazards) is modelled rather than directly modelling the TTP hazard. However, this leads to some small modifications in some of the later formulae.

Calculation of the PPS hazard

In section 5.4.1.4.1, the PPS hazard was defined in Equation (5-30) as:

$$h_{PPS}(t) = \frac{\left(h_{TTP}(t) - h_{PFS}(t)\right)S_{PFS}(t) + h_{OS}(t)S_{OS}(t)}{S_{OS}(t) - S_{PFS}(t)} \quad (5\text{-}43)$$

Using the excess hazard notation, this now simplifies to:

$$h_{PPS}(t) = \frac{\Lambda(t)\, S_{PFS}(t) + h_{OS}(t) S_{OS}(t)}{S_{OS}(t) - S_{PFS}(t)} \tag{5-44}$$

Calculation of the probability of an event being due to progression

Considering how the excess hazard has been defined, the following equation (5.45) can be obtained:

$$h_{TTP}(t) = h_{PFS}(t) - \Lambda(t) \tag{5-45}$$

Therefore, substituting this into the expression used to calculate the probability that the event, at a particular event time, is due to progression, becomes:

$$P(X_{TTP} = 1 \,|T \in \mathrm{d}t, |T \geq t) = \frac{h_{PFS}(t) - \Lambda(t)}{h_{PFS}(t)} = 1 - \frac{\Lambda(t)}{h_{PFS}(t)} \tag{5-46}$$

Figure 5-7: Approach to be adopted depending on available information

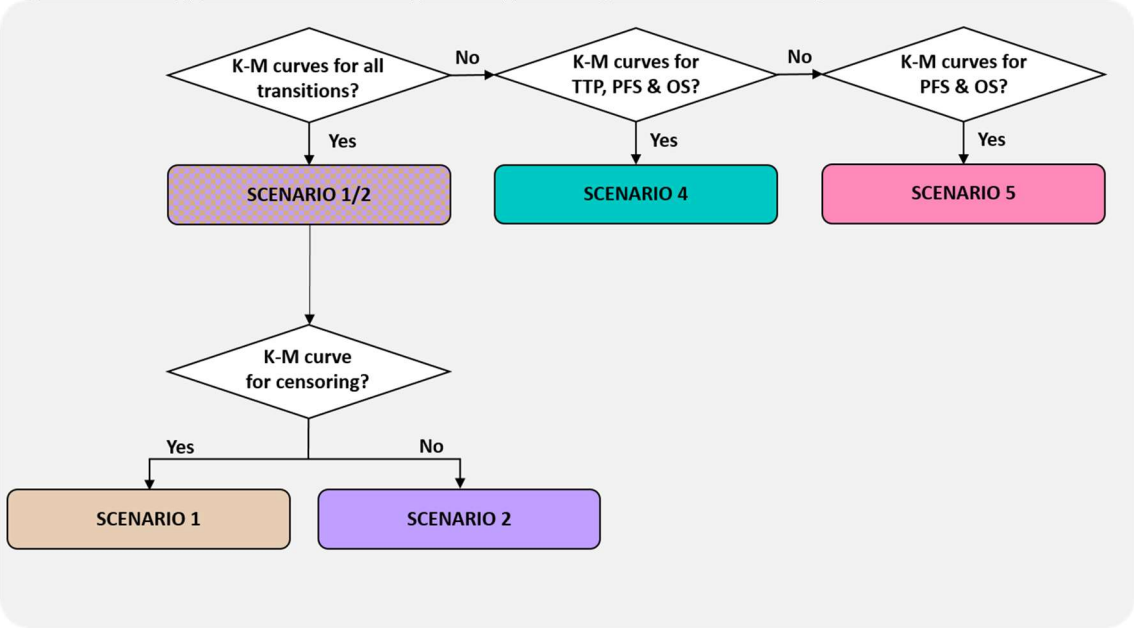
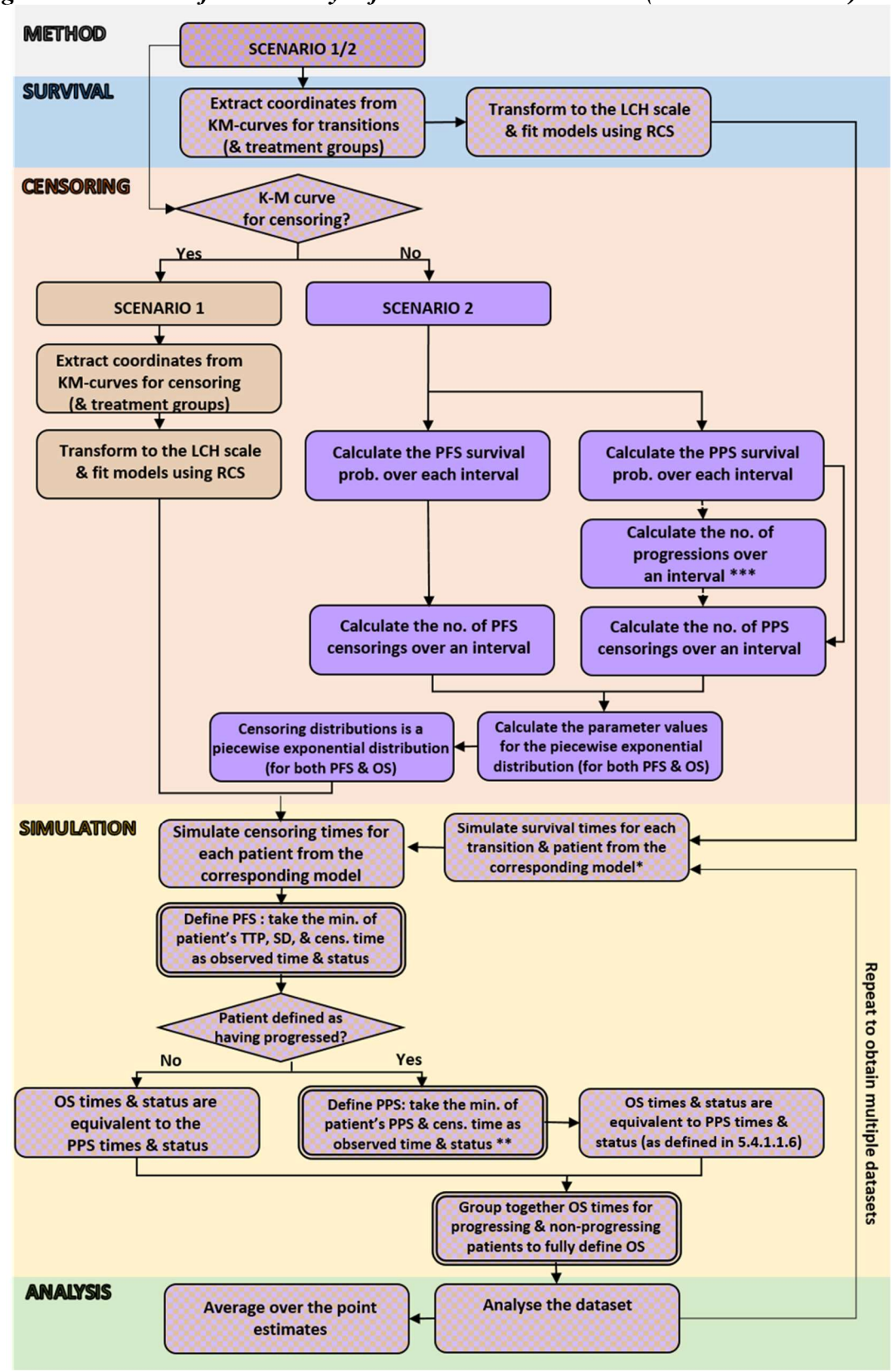


Table 5-2 outlines the stages of the method and gives a summary of the efficacy and ease of implementation for each of the key approaches; also providing a contrast between the three. Figure 5-8 to Figure 5-10 show the process for each scenario diagrammatically.

5.4.2 Overview of implementing the key methods depending on the information

The previous sections of this chapter explain the sequential adaptations of the methods. However, they give little detail of the practical implementation depending on what level

Figure 5-8: Process for summary information on transitions (Scenarios 1 and 2)

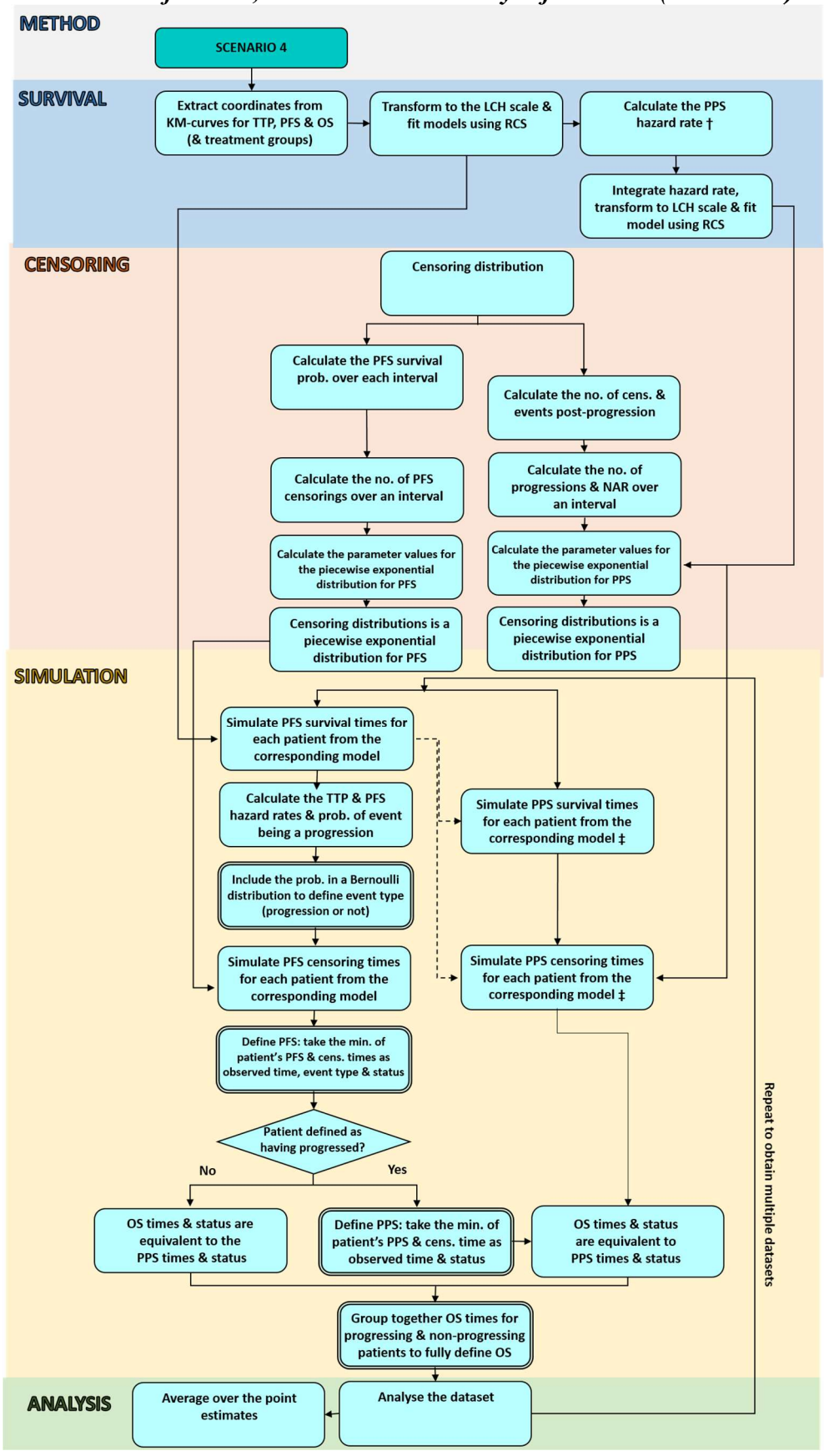


* If using a delayed entry is used, then the simulation will be conditional on the patient's simulated progression time. This does not apply to 'reset the clock' approach.

** If reset the clock, the process follows the same approach as other outcomes (e.g. TTP). If delayed entry, numbers at risk must take account of the number of progressions (e.g. number entering the risk-set)

*** If using a 'reset the clock' approach, then a patients PPS time in relation to the study duration is the simulated PPS time added to the progression time. This does not apply to 'delayed entry' approach.

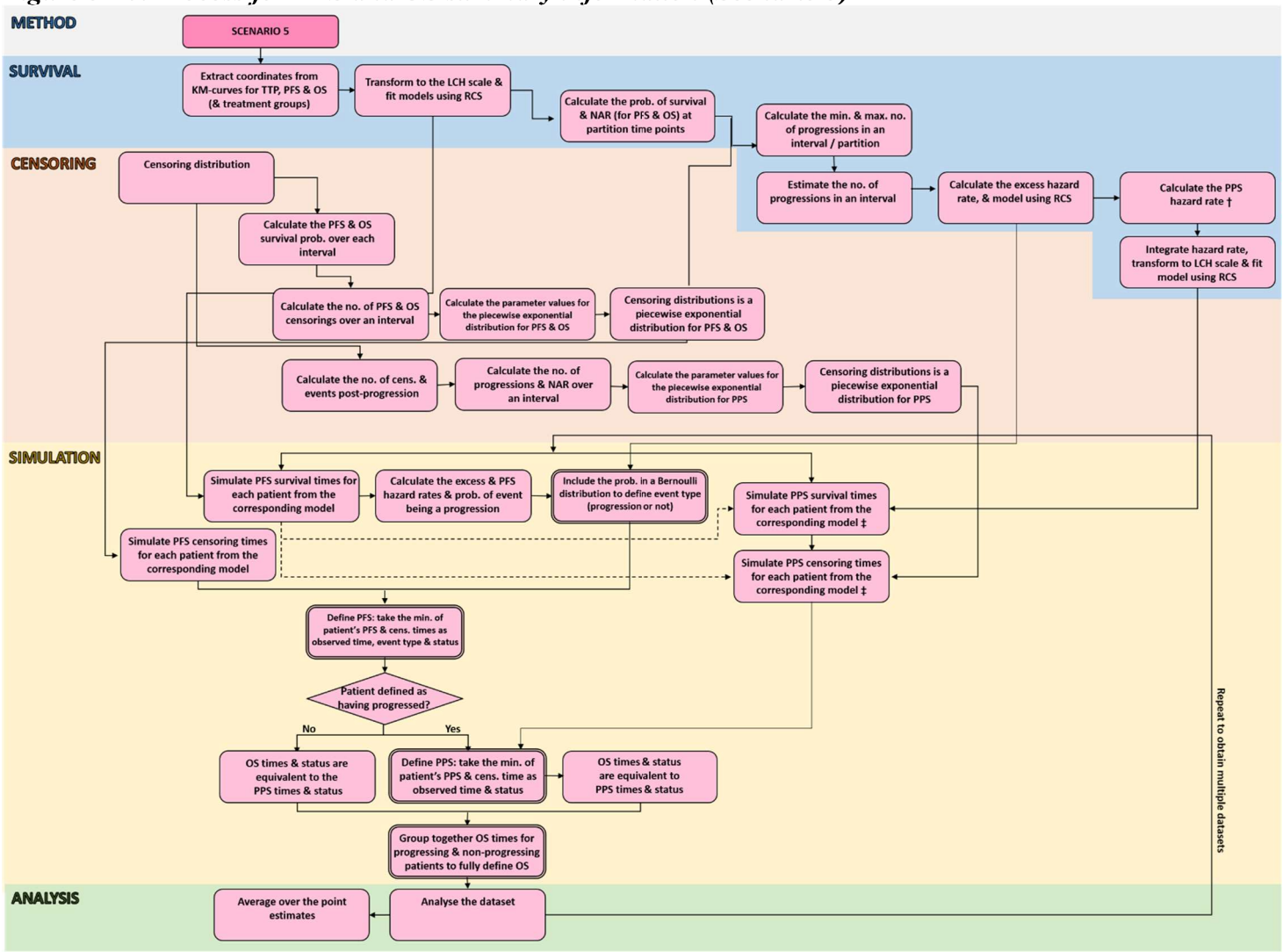
Figure 5-9: Process for TTP, PFS and OS summary information (Scenario 4)



† A 'delayed entry' approach is implicitly used.

‡ The simulated time will be conditional on the patient's progression time.

Figure 5-10: Process for PFS and OS summary information (Scenario 5)



† A 'delayed entry' approach is implicitly used.

‡ The simulated time will be conditional on the patient's progression time.

Table 5-2: Overview and summary of key methods based on available information

Available information	Brief overview of stages	Summary
5.4.2.1 ***All transitions and censoring***	1. Extract coordinates for each outcome and each treatment group 2. Model the survival (or censoring) distribution using RCS splines models fitted on the log-cumulative hazard scale 3. Simulate a time from each outcome model (for that specific simulated treatment group) 4. Combine the information together to define the outcome a. PFS information: take the minimum of the simulated TTP, time to death (before progression) and censoring as the observed survival time and event type (progression, death or censoring) b. OS information: for all patients defined as progression, take the minimum of the simulated PPS and censoring times as the observed PPS and event type. Combine this newly defined PPS information with the PFS censorings and deaths to complete the OS information 5. Analyse the dataset 6. Simulate and analyse additional datasets (using steps 3 - 5) 7. Average over all the datasets to obtain a point estimate	The challenge for this method is in identifying the type of PPS that has been employed. Having done this, the approach is easy to implement in practice.
5.4.2.2 ***TTP, PFS and OS***	1. Extract the coordinates off the scanned Kaplan-Meier curves for TTP, PFS and OS (for each trial arm) using digitizing software. 2. Transform the survival and time coordinates to the LCH and log time scales respectively. 3. Calculate RCS to the data; only outcome and treatment specific information should be used to estimate the knots. 4. Model the survival distribution for a given outcome and treatment, using ordinary least squares regression with the LCH as outcome, and RCS variables as the covariates. 5. Calculate the PPS hazard rate, using the formula given in Section 5.3.1.4.1 evaluated at estimates obtained from the respective survival distribution models. 6. Estimate the censoring distribution, if using the … a. 'Maximum time censoring' approach: determine the maximum study time. b. 'Recruitment times' approach: calculate the minimum and maximum study lengths to incorporate into the uniform distribution, used for the simulation. c. 'Numbers at risk' approach: i. Start by calculating the PFS censoring, which is done exactly as outlined in section 4.4.1.4.3, ultimately resulting in the parameter values for a piecewise exponential distribution. ii. Then for PPS censoring, calculate the number of OS events and censorings iii. Estimate the number of progressions. 7. Calculate the post-progression events, and censorings, and then include within the expression given in Section 5.3.1.3.2 to obtain the parameter values for a piecewise exponential distribution. 8. Simulate a PFS time for each patient from the respective distribution. 9. Include this probability within a Bernoulli distribution, and formally define the event type. 10. Simulate a PPS time for each patient with disease progression, conditional on their PFS (equivalent to their TTP) time. 11. Simulate the censoring times - for the 'Number at risk' approach PFS and PPS must be simulated with PPS censoring being conditional on a patient's progression time. 12. Take the minimum of the PFS and (PFS) censoring time as the patients observed PFS time; the time chosen defines the event status (event or censored). 13. For patients with disease progression, define their PPS by take the minimum of their PPS and (PPS) censoring time as the patients observed PPS time; once again, the time chosen defines the event status (event or censored). Combine the PPS information, with the observed PFS censoring times and PFS event times for patients who died before progression to obtain observed OS time and status information.	As with all the methods, this approach still maintains the broad stages of: extracting the coordinates; modelling the survival and censoring distributions; simulating multiple datasets from the models; analysing datasets separately; and then combining the results. Unlike the methods described in section 5.3.2.1, however, this approach has a much greater dependency on early stages later in the method. For instance, appropriate models for TTP, PFS and OS, are important in ensuring sensible estimates for PPS information, and PFS event type definition. Also, in practice, ensuring harmonious distributions for TTP and PFS can be challenging. By the very definition of PFS, $S_{PFS}(t) \leq S_{TTP}(t)$. However, where almost all of the PFS events are due to progression, extracting coordinates and obtaining models which always comply with this criteria, can be difficult. Nevertheless, it is exceptionally useful to be able to model TTP directly, as this is such an influential factor, particularly with a view to ultimately adjusting for treatment switching; TTP information will define both the eligible population for switching, and the switch time.

Available information	Brief overview of stages	Summary
	14. Simulate multiple datasets, by following stages (7 – 14) multiple times; each dataset should be analysed separately using the methods of choice 15. Results from each separate analysis must be averaged over.	
5.4.2.3 *PFS and OS*	1. Repeat stages 1 – 4 from section 5.4.2.2 this time only for the PFS and OS outcomes (TTP is not available so coordinates cannot be extracted and the distribution modelled directly). 2. Estimate the PFS and OS censoring distribution, if using the … a. 'Maximum time censoring' approach: determine the maximum study time. b. 'Recruitment times' approach: calculate the minimum and maximum study lengths to incorporate into the uniform distribution, used for the simulation. c. 'Numbers at risk' approach: i. Start by calculating the PFS, and then OS censoring; this is done following the stages outlined in section 4.4.1.4.3. Ultimately parameter values for a piecewise exponential distribution should be obtained. 3. Decide upon an appropriate formulation of partitions, calculate the partitions. 4. Estimate the survival and censoring probabilities for PFS and OS for each partition, as explained in section 5.3.1.5.1. 5. Estimate the number of PFS and OS events over a partition. 6. Estimate the number of progressions – TTP events (as outlined in Section 5.3.1.5.1) 7. Calculate and model the excess hazard between the TTP and PFS events. 8. Calculate the PPS hazard rate, using the formula given in Section 5.3.1.5.2 (similar to that Section 5.3.1.4.1, but using the excess hazard rate as opposed to the TTP hazard rate) evaluated at estimates obtained from the respective survival distribution models. 9. If using the 'Numbers at risk' approach, calculate the PPS censoring distribution using steps 6.ii. – 6.iv. of section 5.4.2.2. This could be done either at the partition level or at the interval level. 10. Simulate a PFS time for each patient from the respective distribution. 11. Evaluate the excess and PFS hazard at each of the simulated PFS times, and include in the relevant expression from Section 5.3.1.5.2. The value obtained is still the probability of that specific PFS event being due to disease progression. 12. Follow the stages 9 – 16 given in Section 5.3.2.2.1 to complete the approach.	Once again, the broad stages of the simulation approach carry over to this method. However, this technique consists of more stages, and places a considerable dependency on assumptions and appropriate model specification. It also essentially changes the usual order of approach. In all of the previous variations of the simulation method, typically, all of survival models are obtained before attention is turned to the censoring distribution. Here, though, the censoring for PFS and OS (not PPS), must be calculated early on (stage 2 of 12) in order to allow estimation of TTP information. This approach relies on many assumptions and correct model specifications; this therefore, makes it much more sensitive to changes. Whilst the theory behind this method is robust, the practical application is much more uncertain. To enable estimation, many decisions must be made, all of which may influence the results in some way. These decisions include how to determine the number, and consequently distribution of partitions. Also, in order to numerically integrate to obtain the probability, and hence the number of events / censorings over a partition, the number of points over which to do this, must be chosen. Deciding on the spread of censorings over an interval is choice that must be established. Having to determine these factors is a limitation of this approach, as the choices will largely be arbitrary, and yet could be influential. Nevertheless, it is very necessary to create this method, and reconstruct the data, but as a result it means that sensitivity analysis is very important.

of information is available to the analyst, and how the stages of the approach link together. Thus, this section seeks to address that particular deficiency.

Only the 'key' approaches have been chosen. These are scenario 1, having all transition rates and censoring information; scenario 4, having TTP, PFS and OS; and scenario 5, having PFS and OS. These three scenarios have been classed as 'key', for the following reasons. Scenario 1 demonstrates the underlying process that must be emulated in later

scenarios with less specific information. Scenarios 4 and 5 are important, as these are the most commonly found situations. Scenario 2 has also been briefly mentioned in Figure 5-7 and Figure 5-8, since this only departs slightly from that described in the previous scenario. The process for identifying the appropriate scenario is described in Figure 5-7.

5.4.3 Illustrative example contrasting scenarios 1 (All information) and 4 (Three outcomes only)

5.4.3.1 *Illustrative example*

For ease of explanation and to ensure all the necessary summary information was available, the methods are explained using an example data. The data for this example are simulated based on the results from the TAnDEM trial (Kaufman, 2009) (described in Section 4.7.3.2). Kaplan-Meier curves were produced for each of the transitions, the censoring distribution, PFS and OS. These are shown in Figure 5-6. Figure 5-7 shows the same PFS and OS Kaplan-Meier curves, complete with the respective risk-tables.

5.4.3.2 *Methods*

The methods were carried out as specified in Table 5-2 (Sections 4.4.3.3 and 4.4.3.4). A range of models, with different degrees of freedom and knot locations for the RCS, were tested on the available data, and the most parsimonious ones which also showed the exceptionally good fit graphically, chosen as the final models. For the first scenario (in which all transitions were available), the 'delayed entry' approach was used. For each scenario, two hundred datasets were created for the simulation approach.

The number of events, median survival time and HR obtained from a Cox model are used to compare the representativeness of the IPLD with the original IPD.

5.4.3.3 *Results*

The contrast between the reported summary statistics (number of events, median survival time and HR) for the IPD and the two IPLD scenarios is presented in Table 5-3. In terms of point estimates for PFS, those for the median survival time and HR are very similar across all types of data (IPD and both IPLD). However, for the IPLD there is marginally less variation (e.g. narrow CIs). The number of events is broadly comparable for the IPD and IPLD created using 'All information' (transitions and censoring). For the fourth

Figure 5-11: Kaplan-Meier curves for the simulated example

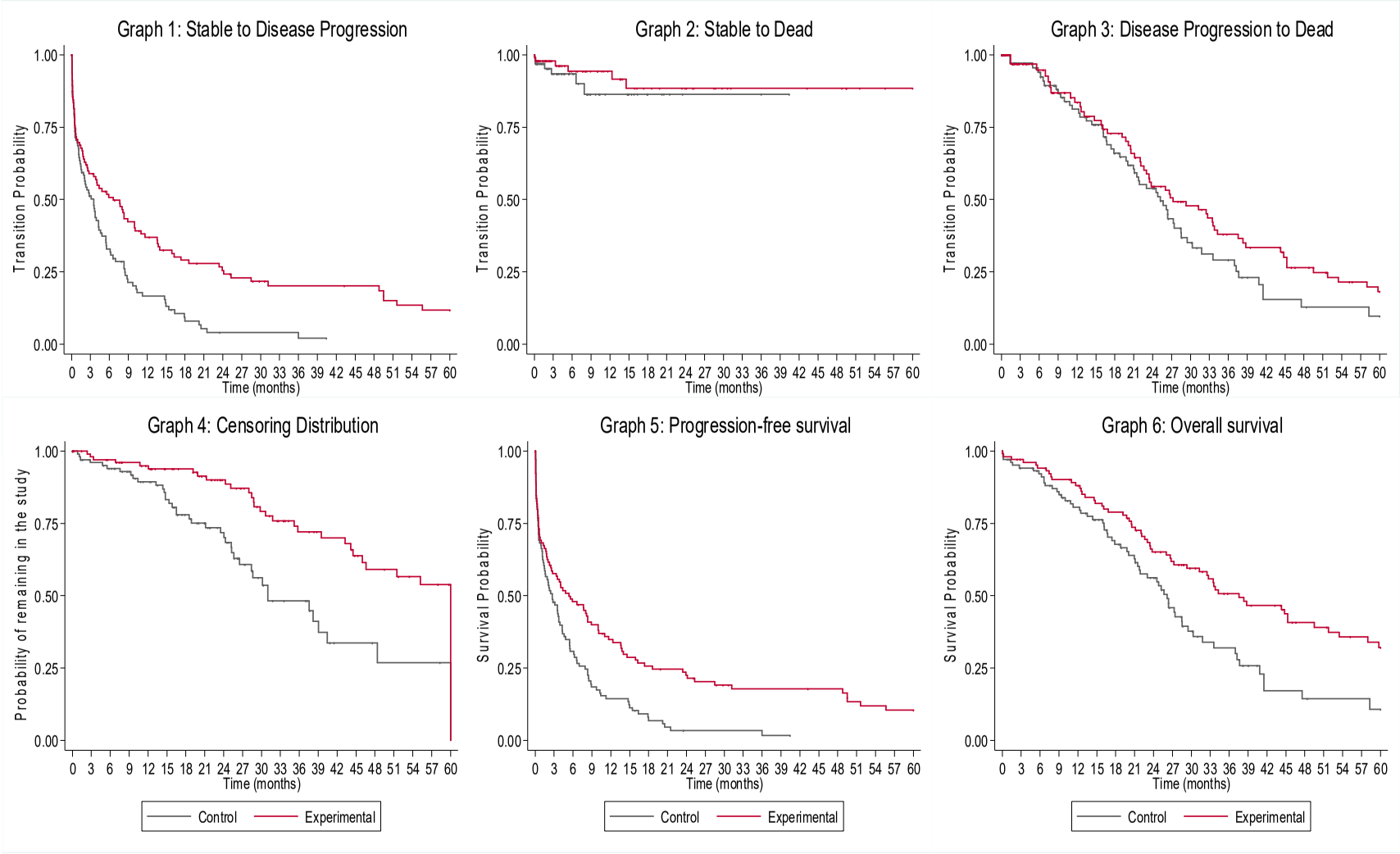
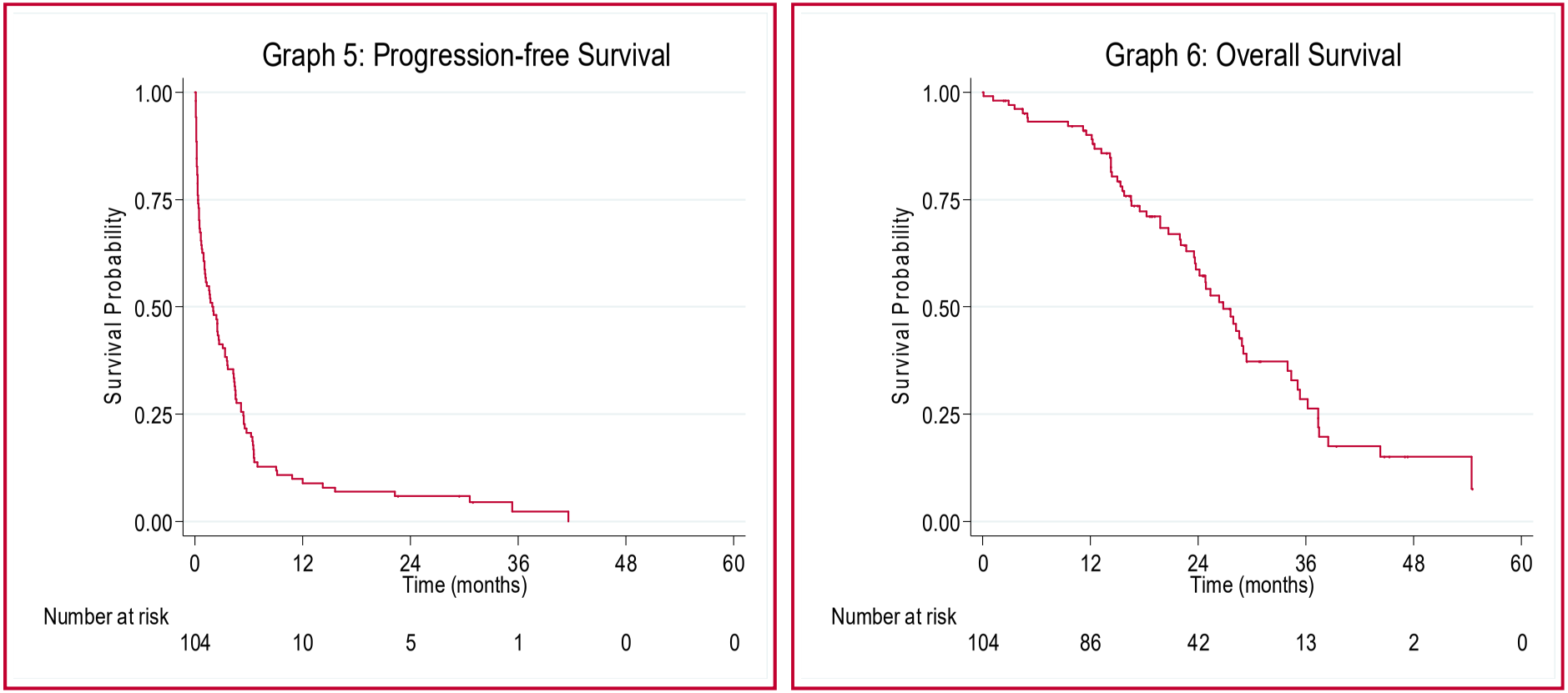


Figure 5-12: PFS and OS Kaplan-Meiers and Risk-tables - Simulated example



scenario, having TTP, PFS and OS, there are more noticeable differences in the number of events, in particular, having four additional events in the control group.

Apart from the number of events, which are almost identical, the IPLD for 'All information' and the IPD point estimates (for median survival time and HR) for OS, although still comparable vary more than for PFS. For these, the uncertainty for the IPLD is slightly increased, compared to that reported from the IPD. Despite the OS median survival time being similar to the IPD for the method relying on TTP, PFS and OS, (with increased variability), the number of events is substantially different. Also, the HR differs considerably.

Figure 5-13 shows the comparison between the extracted coordinates and the survival proportion calculated at every month, and averaged across all datasets for each treatment arm and the two methods (depending on the information available) for PFS and OS.

Table 5-3: ITT analysis results for the IPD and IPLD – Simulated example

				Average over 200 reconstructed datasets	
			Original IPD	**All information**	**TTP, PFS and OS**
PFS	**Number of events**	**Control**	98	97	101
		Experim.	88	87	86
	Median survival time	**Control**	2.7 months (1.4, 3.9)	2.7 months (1.5, 3.9)	2.6 months (1.5, 3.6)
		Experim.	5.4 months (2.5, 8.9)	5.2 months (2.1, 9.1)	5.3 months (2.0, 8.6)
	Hazard ratio		0.582 (0.431, 0.786)	0.585 (0.454, 0.785)	0.573 (0.433, 0.760)
OS	**Number of events**	**Control**	62	62	79
		Experim.	58	58	65
	Median survival time	**Control**	26.2 months (21.5, 29.5)	25.7 months (21.6, 30.4)	25.5 months (21.3, 29.7)
		Experim.	37.6 months (29.2, 49.5)	36.6 months (27.3, 48.1)	38.4 months (30.4, 46.4)
	Hazard ratio		0.556 (0.385, 0.804)	0.563 (0.392, 0.867)	0.499 (0.353, 0.705)

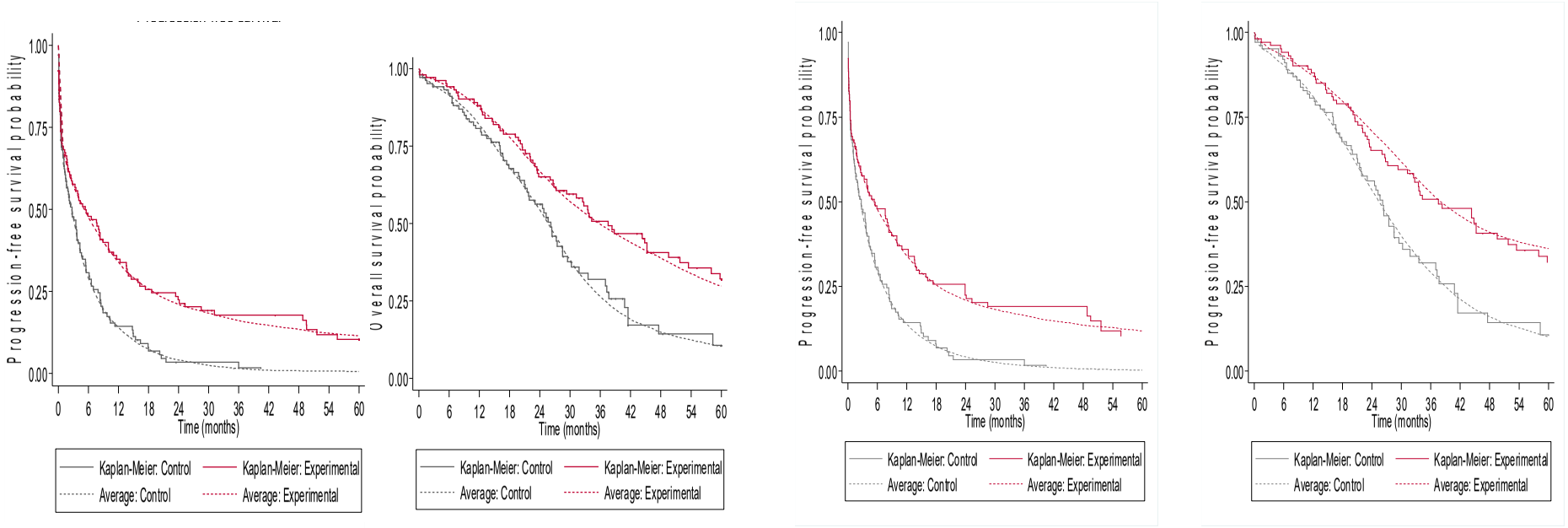
For this example, the average number of events for the IPLD has been rounded to the nearest integer.

5.4.3.4 *Conclusions*

From the results presented here, it is clear that using Scenario 1 (All information) works exceptionally well, leading to results that are highly comparable the IPD. This is very pleasing as it means the underlying theory is working well and therefore, aiming to emulate this approach in the later scenarios is appropriate.

Unfortunately applying the methodology for scenario four (having TTP, PFS and OS) did not work as well as expected, although still sufficiently good enough to be useful. Whilst

Figure 5-13: IPD and IPLD for scenarios 1 and 4 - Simulated example



Comparisons of the Kaplan-Meier curves (based on the extracted coordinates) and the average value (calculated at each month) for the IPLD. (left to right) All information - PFS; All information - OS; TTP, PFS and OS – PFS; TTP, PFS and OS – OS.

the PFS data for the experimental group is relatively comparable, at this stage there were already issues with the number of events in the control group. This issue worsened considerably for the PPS events. This was explored in greater detail subsequently, and the principle cause was found to be the censoring; most attributable to the wide interval length and sudden drop in survival over the interval.

Since the underlying summary data had been simulated, having the risk-table with intervals at every three, and then every one month was investigated. Even using the three-month intervals, the censoring still performed poorly, but once one-month intervals were used, the results were far more reflective of the IPD. This emphasised the importance, particularly for PPS in ensuring that the value of α in the censoring is appropriate.

5.4.4 Illustrative example using the TAnDEM trial

5.4.4.1 *Illustrative example*

Here the TAnDEM trial (Kaufman, 2009) example that was used previously in section 4.7.3.2 has been used. Kaplan-Meier curves, and 'at risk' tables for PFS and OS; and the number of events for TTP, PFS and OS were available. It should be noted that this example, perhaps, has slightly more information than would be typically reported, reporting on events is often variable and in addition, the 'at risk' table uses many intervals (12 intervals).

5.4.4.2 *Methods*

It was decided to only use information for PFS and OS. This choice was made for two reasons;

1) it avoids the practical challenges of extracting the coordinates for TTP and PFS such that $S_{TTP}(t) \leq S_{PFS}(t)$ since there are very few deaths prior to disease progression;
2) it assesses how the method should work if only the two outcomes had been reported.

Figure 5-14: Kaplan-Meier curves for PFS and OS in the TAnDEM trial

Image subject to copyright and so has been removed from the text. Please see the original source (Kaufman, 2009) for image.

5.4.4.2.1 *Overview of reconstructing the IPLD*

This example used the method set out in Section 5.4.2.3. In brief: extracting the coordinates from the Kaplan-Meier curves; fitting models to each arm; calculating the censoring distributions for PFS and OS; estimating the progressions over the interval, excess hazard and TTP hazard rate; estimating and modelling the PPS and PPS censoring distribution; simulating the datasets and analysing the data for common ITT statistics.

5.4.4.2.2 *Specific details*

Table 5-4 shows the respective degrees of freedom for the RCS models and treatment groups (Durrleman, 1989). The partitions were based on the number of coordinates (combined for PFS and OS) in each interval. Tables highlighting the initial calculation PFS censorings in the A group and the process of the scale factor are available in the Appendix G and were achieved because the number of PFS events was also available for each treatment group. The PPS censoring distribution was formed using the partitioned data rather than the original intervals. For this method, two hundred datasets were simulated.

5.4.4.3 *Results*

On average, from Table 5-5, the number of events, median survival times and HR for both PFS and OS are generally comparable (e.g. within acceptable limits) to those obtained from the original IPD. There is slightly better agreement with the control group (A) information compared to the experimental group (T+A).

Table 5-4: Degrees of freedom – TAnDEM
Degrees of freedom for the restricted cubic splines models for each outcome and each treatment group.

Trial arm	Degrees of freedom			
	PFS	OS	Excess hazard	PPS
A	4	9	4	5
T+A	8	6	4	5

5.4.4.4 *Assessing the representativeness of the IPLD*

Table 5-5: ITT results for the IPD and IPLD - TAnDEM

			Original IPD	Average over 200 reconstructed datasets
TTP	**Number of events**	A	92	92.9
		T+A	84	86.6
PFS	**Number of events**	A	99	100.6
		T+A	87	90.6
	Median survival time	A	2.4 months (2.0, 4.8)	2.9 months (1.6, 4.2)
		T+A	4.8 months (3.7, 7.0)	5.1 months (2.9, 7.3)
	Hazard ratio		0.63 (0.47, 0.84)	0.67 (0.52, 0.85)
OS	**Number of events**	A	64	62.2
		T+A	58	55.2
	Median survival time	A	23.9 months (18.2, 37.4)	21.7 months (18.3, 25.2)
		T+A	28.5 months (22.8, 42.4)	26.1 months (22.1, 30.1)
	Hazard ratio		0.84 (0.59, 1.20)	0.86 (0.59, 1.26)

5.4.4.5 *Conclusions*

The IPLD generated was broadly representative of the data, although there were still some reasonably large discrepancies between the IPD and IPLD e.g. approximately four extra PFS events in the T+A group. However, in terms of the OS results, these were far more comparable for the initial analysis than in the previous example using Scenario 4 (Three outcomes) in Section 5.4.3. The main reason for this is probably due to the detailed summary information, in particular the risk-tables. As explained in the Section 0, the simulated example only reported the numbers at risk at five, very widely spaced, intervals. The larger the change in survival, particularly in the early intervals, the less accurate the estimates for the piecewise-exponential models are; this is likely to be that situations where the interval is wide and with a large decrease in survival are indicative of scenarios where assuming a constant hazard rate over the interval is not appropriate. For the example in Section 0, the best approximation to the true underlying censoring function occurred when the numbers at risk were reported every month (e.g. 12-times more frequently than initially reported). In the TAnDEM example, though, the numbers at risk were reported every 5 months, giving a total of 12 intervals. In addition, with the simulated example, the PFS at the end of the first interval for the experimental treatment group was just under 10%, in contrast with slightly less than 50% survival.

Whilst the variation between the IPLD and IPD are greater than would have been hoped, it is still within acceptable limits. This is especially true given how many strong assumptions have to be applied in order to estimate necessary components for reconstructing the IPLD with an 'illness-death' modelling approach and only two outcomes.

5.4.5 Understanding and assessing the underlying driving factors

This process is extremely complex, and relies on many interconnected models and stages, particularly when using more limited and / or general information. Therefore, if the results are less comparable, it is vital to understand, and put into practice how to assess individual parts of the method. Some of this 'testing' will be done through sensitivity analysis, such as trying other models with different degrees of freedom and / or knot locations, using different partitions, etc. However, two parts of the method can be specifically assessed. These are the estimation of the TTP events (using the relative survival framework) and the 'consistent censoring' assumption.

5.4.5.1 *Using the relative survival framework to assess the distribution of TTP events*

The relative survival framework for estimating the TTP distribution is by no means infallible, since it was necessary to adapt this to suit the needs of the 'illness-death' modelling approach. It may still be possible to estimate a negative number of progressions, or alternatively a number exceeding the change in risk-table (e.g. more progressions than the number of progressions and deaths combined). However, in these situations, these issues can actually provide some indication of whereabouts the method is not performing well and why this might be.

One way of checking the model suitability is to plot both the PFS hazard model and estimated excess hazard model against time. Where the excess hazard curve exceeds the PFS hazard, this indicates that too many events have been estimated in that region. If the excess hazard rate falls below zero, this would suggest that not enough events have been estimated in this region. It may help to partition the timescale based on intervals / partitions to identify particular regions. This information could then ultimately be crudely used to perform some type of re-weighting of the events.

5.4.5.2 *Assessing 'consistency' of censoring*

The key assumption for the censoring distribution is that of 'consistency'. This means that once a patient has been censored for PFS, this censoring time and status is also used for OS. As a result, the following statements must hold:

$$w_{PFS} \leq w_{OS} \tag{5-47}$$

Where w_j is the number of censorings for outcome j.

However, in practice, this may not always occur. This could be due to measurement / estimation error, the assumption of the events occurring evenly throughout an interval or because 'consistency' does not hold (e.g. if all the censorings occurred before any event times).

Therefore, this section explores exactly what the 'consistency' assumption means in terms of estimation. To achieve this the starting point must be the life-tables equation for calculating the probability of surviving an interval, i.

$$p_i = 1 - \left(\frac{d_i}{n_i - \frac{1}{2} w_i} \right) \tag{5-48}$$

Here, ½ represents the time which censored patients are assumed to be at risk for. To assume censoring evenly over the interval, the ½ is used. However, essentially any value, α_i, where $0 \leq \alpha_i \leq 1$ could be used and hence, equation (5-48) becomes:

$$p_i = 1 - \left(\frac{d_i}{n_i - \alpha_i w_i} \right) \tag{5-49}$$

Where the values α_i could be different for each interval, and α_i represents the proportion of the interval that censored patients are 'at risk' for.

Letting y_i be the change in risk table over the interval, i, d_i, the number of deaths and w_i, the censorings. For a given outcome, j, and a given interval, i, the estimated censorings $\widehat{w}$ can be defined as:

$$\widehat{w}_{ji} = \frac{y_{ji} - n_{ji}(1 - p_{ji})}{1 - \alpha_{ji}(1 - p_{ji})} \tag{5-50}$$

So, if the consistency conditions hold, then:

$$\frac{y_{PFSi} - n_{PFSi}(1 - p_{PFSi})}{1 - \alpha_{PFSi}(1 - p_{PFSi})} \leq \frac{y_{OSi} - n_{OSi}(1 - p_{OSi})}{1 - \alpha_{OSi}(1 - p_{OSi})} \tag{5-51}$$

Whilst values of y_{ji}, n_{ji}, p_{ji} are fixed, that for α_{ji} is not.

Given that α_{ji} measures the proportion of the interval that censored patients are 'at risk' for, $0 \leq \alpha_{ji} \leq 1$. If consistency for interval *i* holds, then for at least one value of $0 \leq \alpha_{PFSi} \leq 1$, there must exist at least one value of $0 \leq \alpha_{OSi} \leq 1$.

Rearranging this equation (5-51), gives:

$$\alpha_{OSi} \geq \frac{\left(1 - \frac{y_{OSi} - n_{OSi}\, q_{OSi}}{y_{PFSi} - n_{PFSi}\, q_{PFSi}}\right)}{q_{OSi}} + \left(\frac{q_{PFSi}}{q_{OSi}}\right)\left(\frac{y_{OSi} - n_{OSi}\, q_{OSi}}{y_{PFSi} - n_{PFSi}\, q_{PFSi}}\right)\alpha_{PFSi} \tag{5-52}$$

Where, $q_{ji} = 1 - p_{ji}$

This can be written more simply as:

$$\alpha_{OSi} \geq \frac{(1 - \Theta_i)}{q_{OSi}} + \left(\frac{q_{PFSi}}{q_{OSi}}\right)\Theta_i\, \alpha_{PFSi}, \quad where\ \Theta_i = \frac{y_{OSi} - n_{OSi}\, q_{OSi}}{y_{PFSi} - n_{PFSi}\, q_{PFSi}} \tag{5-53}$$

This formulation of the problem essentially takes account of whether the assumption of the events occurring evenly throughout an interval holds (if so, $\alpha_{OSi} = \alpha_{PFSi} = 0.5$ will satisfy the conditions, or if not, what a more reasonable value might be.

However, the question still remains how to use this information? This function could be illustrated graphically, showing the range of all plausible values for $\alpha_{OSi}, \alpha_{PFSi}$, but this would result in many graphs. For example, for a two arm study, with six intervals, twenty-four graphs would be produced. However, since the functions are linear, a simpler strategy is to evaluate the function at α_{PFSi} = 0 and α_{PFSi} = 1. Table 5-6 and Table 5-7 show how these values can be interpreted.

Table 5-6: Values to ensure consistent censoring

	Values of α_{PFSi} for which $0 \leq \alpha_{OSi} \leq 1$	Values at the limits: At $\alpha_{PFSi} = 0$	Values at the limits: At $\alpha_{PFSi} = 1$	Interpretation
1.	$0 \leq \alpha_{PFSi} \leq 1$	$\alpha_{OSi} \leq 0$	$\alpha_{OSi} \leq 0$	This means that any of value of $\alpha_{PFSi}, \alpha_{OSi}$ between 0 and 1 will ensure consistency.
2.	$0 \leq \alpha_{PFSi} \leq \phi_i,$ $0 < \phi_i < 1$	$\alpha_{OSi} \leq 0$	$\alpha_{OSi} \geq 1$	This means that for all values $0 \leq \alpha_{OSi} \leq 1$ but only values α_{PFSi} between 0 and ϕ_i ensure consistency.
3.	No feasible values $\alpha_{PFSi} \leq 0$	$\alpha_{OSi} > 1$	$\alpha_{OSi} > 1$	No values can be found to obtain consistency for this interval.

It may be possible, if, at $\alpha_{PFSi} = 0$, $0 < \alpha_{OSi} < 1$, that α_{OSi} is restricted further (e.g. $\psi_A < \alpha_{OSi} < \psi_B,\ 0 < \psi_A < \psi_B \leq 1$).

Table 5-7: Values to ensure consistency censoring - limits

	Values of α_{PFSi} for which $\psi_A \leq \alpha_{OSi} \leq \psi_B$	Values at the limits: At $\alpha_{PFSi} = 0$	Values at the limits: At $\alpha_{PFSi} = 1$	Interpretation
4.	$0 \leq \alpha_{PFSi} \leq 1$	$\alpha_{OSi} = \psi_A$	$\alpha_{OSi} = \psi_B$	This means that for all values $0 \leq \alpha_{PFSi} \leq 1$ but only values α_{OSi} between ψ_A and ψ_B ensure consistency.

To calculate the value of ϕ_i in scenario 2, the expression on the right-hand side of the inequality needs to be set to one and α_{PFSi} to ϕ_i.

$$\frac{(1-\Theta_i)}{q_{OSi}} + \left(\frac{q_{PFSi}}{q_{OSi}}\right)\Theta_i\,\phi_i = 1, \qquad \Theta_i = \frac{y_{OSi} - n_{OSi}\,q_{OSi}}{y_{PFSi} - n_{PFSi}\,q_{PFSi}} \tag{5-54}$$

Rearranging gives:

$$\phi_i = \frac{\left(1 - \frac{(1-\Theta_i)}{q_{OSi}}\right)}{\left(\frac{q_{PFSi}}{q_{OSi}}\right)\Theta_i} \tag{5-55}$$

To correct for an inaccurate total number of events supposedly caused through estimation / measurement error, the simulation approach, both for a single or multiple outcomes,

applies a scaling factor to the number of censorings in each interval. This scale factor ζ_j for outcome *j* takes the form of

$$\zeta_j = \frac{W_j}{\sum_{i=1}^{I} \widehat{w}_{ji}} \tag{5-56}$$

Where W_j is the total number of censorings reported, and *I* the last interval.

For consistency to hold during and after the application of the scale factor, then

$$\zeta_{PFS}\, \widehat{w}_{PFSi} \leq \zeta_{OS}\, \widehat{w}_{OSi} \tag{5-57}$$

Writing in terms of $\widehat{w}_{ij}$

$$\zeta_{PFS}\left(\frac{y_{PFSi} - n_{PFSi}\, q_{PFSi}}{1 - \alpha_{PFSi}\, q_{PFSi}}\right) \leq \zeta_{OS}\left(\frac{y_{OSi} - n_{OSi}\, q_{OSi}}{1 - \alpha_{OSi}\, q_{OSi}}\right) \tag{5-58}$$

Once again, rearranging to obtain an expression in terms of α_{ji} gives:

$$\begin{aligned} \alpha_{OSi} \geq \frac{1}{q_{OSi}}\Bigg(1 - \frac{\zeta_{OS}}{\zeta_{PFS}}\left(\frac{y_{OSi} - n_{OSi}q_{OSi}}{y_{PFSi} - n_{PFSi}q_{PFSi}}\right)\Bigg) \\ + \frac{\zeta_{OS}}{\zeta_{PFS}}\left(\frac{q_{PFSi}}{q_{OSi}}\right)\left(\frac{y_{OSi} - n_{OSi}q_{OSi}}{y_{PFSi} - n_{PFSi}q_{PFSi}}\right)\alpha_{PFSi} \end{aligned} \tag{5-59}$$

5.4.6 Sensitivity analysis

With this type of simulation approach, sensitivity analysis is strongly advocated. Different models should be investigated to ensure suitability and robustness of the results. This could be done through visual inspection of the curve and comparison of other model fitting statistics such as the AIC and / or BIC. It may also be useful to examine different knot locations as this can also impact on the fit of the model. If possible, it would be desirable to run the whole approach (e.g. simulating and analysing the full datasets) for a variety of different models and for different outcomes. However, this may be infeasible as the process can be very computationally intensive. The whole process can take between two to five hours.

It is important to consider different partition locations and formulations. For instance, these are usually recommended to be proportional to the number of coordinates (or the square root of the number of coordinates – rounded up to the nearest integer). Alternative strategies could also include, exploring the impact of no additional partitions, the same number of partitions, half the number of coordinates etc. This is a vital part of sensitivity analysis for ensuring whether this 'model fitting' stage is influential on the results or not. Ideally, the results should not vary depending on how many partitions there are, or where these are selected.

5.5 Secondary analysis for treatment switching

Simulating IPLD with paired PFS and OS times is not exclusive to examples where treatment switching has occurred, as this stage of the work only aims to reconstruct the survival times. Therefore, in order to complete the objective of this project, information on treatment switching must also be reconstructed.

Continuing with the assumption that patients switch on (or shortly following) disease progression, there are naturally a subset of patients who are eligible to have switched i.e. in every simulated dataset any patient defined as having experienced disease progression. However, the number of eligible patients will vary in each dataset, and it is unlikely that all patients experiencing progression would switch (in general, the switch proportion is between 30% and 80% of all control group patients; whereas PFS / TTP occurs in the majority of patients).

In previous research, (Boucher, 2013b) and Section 4.8, patients were chosen at random from the control group and assigned to switch until the number reported was reached. This meant that all control group patients were at risk of being defined as having switched. However, in this situation, modifying this approach to simply use the eligible patients rather than the whole population may not be appropriate; especially when the number of progressions is similar to the number of switchers. This is because given that the number of progressions will vary from simulation to simulation, there may be circumstances where, for one or more datasets, the number of progressions (i.e. the potential switchers) is less than the reported number of switchers. Therefore, the proportion of switchers to progressions will be fixed, so that for the i^{th} dataset, the number of treatment switchers, x_i, is:

$$x_i = \alpha P_i \tag{5-60}$$

Where P_i is the number of patients who progressed for the i^{th} dataset and α the proportion of switchers to progressions calculated as:

$$\alpha = \frac{X}{P} = \frac{N\rho}{P} \tag{5-61}$$

Where X is the reported number of switchers, and ρ the proportion of patients who have switched (from the whole control group population), and N, P are the reported number of control group patients and number of progressions respectively.

A subset of size x_i is chosen from P_i, where patients for the subset are selected at random (i.e. any patient in P_i has an equal chance of having switched). This defines the treatment switchers, and time to treatment switch is assumed to be equivalent to TTP information.

5.5.1 Exploring other mechanisms

Thus far, the method has been modified, from that described in Section 4.8, to include a selection process; by restricting the population at risk of switching to those control group patients who progressed (Section 5.5) rather than all control patients (Section 4.8). However, the true underlying treatment switching mechanism is likely to be considerably more complex. This current selection process (outlined earlier in Section 5.5) essentially assumes 'exchangeability' of progressors. In other words, any patient who progresses has an equally likely chance of switching. Nevertheless, previous 'simple' simulation study structures have differentiated patients with 'poor' and 'good' prognosis. This is an element that may not be captured if assuming 'exchangeability' of progressors.

This can, though, to some extent be built in quite easily by using different assumptions. The easiest, perhaps, to understand is weighting the treatment switching probability by progression time. In its simplest form, this means defining the treatment switchers as those with:

1) the shortest progression times: it is assumed that patients progressing early are likely to have a poorer prognosis, and could be more likely to switch; or,
2) the longest progression times: this assumes that patients with a 'good prognosis' could be more likely to switch, indicated perhaps by the patients

who progressed later in the time period (the earliest progressors could be already too advanced to be switched).

The easiest way of implementing these additional assumptions is purely to rank the eligible subjects in order (depending on whether the shortest or longest progressors are needed) and then assign the first x% to switch. This would be exceptionally crude in terms of weighting by the progression, and so a refined approach would potentially be more desirable in practice, which could add a stochastic element (e.g. for (2) mostly later progressors would be used, but so would a small proportion of patients with earlier progression times).

Another option would be to select switchers based on the PPS. This is perhaps less intuitive, and possibly harder to justify in some ways. Selecting the patients with the longest PPS times represents the situation whereby the new treatment is considerably more effectively, and patients with longer post-progression times are more likely to have switched (shorter post-progression times are more likely to be the control group patients who don't switch). However, this assumes a potentially drastic effect of the new treatment. More importantly, though, this would completely ignore the effect of TTP.

Whilst the results from these progression / post-progression time dependent mechanisms essentially provide good limits for the true treatment effect, in their simplest form they are completely deterministic and hence, slightly artificially created scenarios. Adding an extra layer of complexity in the selection process could be important. Additionally, exploring other possibilities which aim to select patients with a shorter progression time but longer post-progression time would also be valuable.

Once the selection mechanism has been decided upon, the analyst must then consider the time between progression and treatment switch. In reality, there will be a period of time, possibly quite short, between disease progression occurring and switching to receive the new treatment. For ease, this period of time will be referred to as the *lapse period*. Whilst it is easiest to ignore this lapse, and assume that the progression time is the time of switch, to improve the realism, or perhaps for sensitivity analysis, it might be important to add a *lapse period*.

Ideally, if a Kaplan-Meier curve is available from progression to time of treatment switch this should be used. In practice, though, this is rarely reported. More likely this *lapse period* may need to be informed by clinical opinion (e.g. views on how long the offer to switch treatment is left open for). For example, Latimer (2012) suggests that patients usually switch treatments within two follow-up appointments of documented disease progression if they are planning to. When the length of time has been agreed on, the *lapse period* could potentially be defined using a uniform distribution.

It should be noted that in considering *lapse periods*, this may also inform the selection process. Again, to improve the reality of the mechanism it may be worthwhile to add in more conditions for selecting switchers e.g. set a minimum PPS time. If, for instance, it is known that in practice, it was at least a week before a patient could receive the new treatment after being diagnosed as having progressed, then it is unlikely or even impossible for a patient who died less than seven days after progression to have switched.

5.5.2 Reanalysis of the TAnDEM trial for treatment switching

The TAnDEM trial (Kaufman, 2009) has once again been chosen to illustrate how to reconstruct and reanalyse summary data for treatment switching. Since Section 5.4.3 examined how to reconstruct the underlying survival TTP, PFS and OS data, here only the additional steps needed for the treatment switching reanalysis are discussed.

5.5.2.1 *Treatment switching in the TAnDEM trial*

In terms of treatment switching, 73 of 104 patients receiving anastrazole monotherapy (A), had trastuzumab added to their treatment regimen after disease progression. Based on the information provided, this means that 79.3% (73 / 92) of patients that progressed switched.

5.5.2.2 *Reconstruction of the treatment switching information*

The treatment switching information was reconstructed using the method outlined in the earlier parts of this section (5.4), but under a number of different scenarios. Those considered were treatment switchers selected:

1. At random
2. As those with the shortest progression-times

3. As those with the shortest post-progression times
4. As those with the longest post-progression times

For this example, switchers were assumed to switch at the time of progression (i.e. no lapse period was used). In terms of the RPSFTM used, a log-rank test statistic was chosen and the method conducted with and without re-censoring (Robins and Tsiatis, 1991, White, 1999). Given that the 'end-study' data are unavailable; this was assumed to be the maximum study length (i.e. 60 months) for every patient.

This example was chosen because it reported a RPSFTM analysis and therefore it is possible to make a comparison between an equivalent analysis for treatment switching using IPD and IPLD.

5.5.2.3 *Secondary analysis applied*

Once the final simulated datasets including the treatment switching information have been reconstructed, a RPSFTM can be fitted to the data (Robins and Tsiatis, 1991). In order to calculate an appropriate SE for the RPSFTM HR, this method uses an approach called 'preserving the p-value', whereby the SE is created using the p-value from the ITT analysis. This preserves the extra uncertainty surrounding the adjusted results. However, it also means that only a point estimate is required. Therefore, the model is fitted to the larger dataset containing all the individually simulated datasets; rather than analysing each simulated dataset separately and then averaging over all the results as is necessary for all other analyses.

5.5.2.4 *Results*

Table 5-8: Reanalysis for treatment switching results IPD and IPLD - TAnDEM

	Switching mechanism	Hazard ratio	
		RPSFTM without re-censoring	RPSFTM with re-censoring
	IPD	0.74* (0.39, 1.38)	
1	At random	0.73 (0.38, 1.39)	0.73 (0.39, 1.38)
2	Those with the shortest progression-times	0.72 (0.36, 1.41)	0.72 (0.37, 1.40)
3	Those with the shortest post-progression times	0.77 (0.46, 1.30)	0.78 (0.48, 1.29)
4	Those with the longest post-progression times	0.64 (0.26, 1.58)	0.65 (0.27, 1.56)

* This HR was extracted from the TA, where no details were given as to whether the model re-censored patients or not.

5.5.2.5 *Conclusions*

For this example, results for the re-analysis using the 'at random' switching mechanism were exceptionally close to those of the IPD. The other switching mechanisms performed poorer in comparison; the best of these being the selection of those with the shortest progression time the HR differing slightly, and having increased variability. As can be expected, the results are potentially being biased by selecting patients based on the PPS. By choosing shorter survival times it is assumed the new treatment is less effective leading to an adjusted analysis which only suggests a 23% reduction in the mortality rate. Choosing the longest post-progression times leads to a revised estimate suggesting a reduction of 36% in the mortality rate. Compared with the IPD analysis results, the first (switch mechanism 3) underestimates the treatment effect, whilst (4) overestimates it.
Some of the similarity could be partly explained by the 'end study time' given to the patients during the re-censoring process. In the IPD, this 'end study time' used to re-censor patients will vary from person to person depending on when they were recruited to the trial, i.e. it is the subject-specific maximum length of follow-up. However, this level of information is not accessible, and so typically the 'end study time' is assumed to be the maximum follow-up time and to be the same for everyone.

5.6 Discussion, Strengths and Limitations

The content of this chapter describes and illustrates novel methods for reconstructing IPLD with outcomes paired across patients, provided that the underlying data follows an 'illness-death' modelling structure. The examples presented show that data can be reconstructed given a minimum level of information (namely Kaplan-Meier curves and risk-tables for PFS and OS, and the number of events for TTP, PFS and OS). This 'illness-death' modelling framework is especially useful if the data needs to be reanalysed for treatment switching; as very often analysts may have access to their own RCT but not for other comparators and may rely on published K-Ms

As expected, the quality of the data varies depending on the level and detail of the information reported. The more specific information there is (e.g. on individual transitions and censoring), the more it can be presumed that the IPLD will represent the IPD. The strong assumptions and necessity to estimate several of the crucial parameters /

functions, as summary information becomes more limited, lead to sensitivity and greater variation between the IPLD and IPD summary statistics.

Using the 'illness-death' modelling framework to reconstruct IPLD in order to reanalyse summary data for treatment switching demonstrates great potential. For the TAnDEM trial, the reanalysis using IPD and IPLD were remarkably similar (when a 'random' approach) was used. Whilst the results given in Chapter 4 for the same RPSFTM-analysis on the IPLD where only OS was reconstructed from the summary data, were within acceptable limits; by using this extended methodology the treatment switching mechanism is much better captured. By having both TTP and OS information (assuming patients switched on progression), a selection process can automatically be implemented (e.g. only patients defined as having progressed can switch). Therefore, even if the 'at random' mechanism is used, which may be quite a realistic option rather than alternative more selective mechanisms, there is a form of selection in the background. Also, the variation in the switch time is a much better, more realistic, improvement. In addition, having TTP information could enable other complex methods for addressing treatment switching to be used. Since TTP is often one of the most driving characteristics in patients switching treatments, this could be incorporated as the variable (at secondary baseline) predictive of crossover, in a two-stage analysis (Section 1.1.2.4.3).

Another key advantage, as with the simulation approach in general, is the flexibility. Whilst only a small set of treatment switching mechanisms were explored, the approach has the potential to accommodate many other styles (e.g. "weighted" selection of progressors as switchers; time between diagnosis of progression and switching treatments, etc.). To consolidate this work further, it would benefit from a greater number of examples, and also assessing how the results from a reanalysis fit into a secondary analysis. This has been explored in Chapter 6.

Chapter 6: Addressing treatment switching in assessment for surrogacy

6.1 Chapter overview

This chapter opens by summarising the key assumptions and data requirements for implementing the simulation technique for paired data (e.g. PFS and OS). It then considers another type of secondary analysis, assessment of surrogacy, and the effect treatment switching has on it. A case study has been used to illustrate the effect of reconstructing and re-analysing summary data for treatment switching and its inclusion in this secondary analysis. The case study also serves to highlight the issues, largely due to lack of necessary information and variation of reporting, faced by analysts reconstructing IPLD, and some potential solutions.

6.2 An illustrative case study

6.2.1 Surrogacy in non-small-cell-lung cancer

Whilst this research began by concentrating on the impact unadjusted treatment switching has on IC or MTC of OS, there are other secondary analyses that are affected. The issue of treatment switching in assessing surrogacy was highlighted by Hotta (2013), who described the effect for non-small-cell-lung cancer (NSCLC). The paper examined whether PFS could be used as a surrogate endpoint for OS. The study, using thirty-four trials, found little evidence of correlation between PFS and OS. However, when stratifying studies with less than 1% of crossover (n = 20), and those with more than 1% (n = 15), a strong correlation (R-squared = 0.69) was observed in those which did not permit crossover. Stratifying further based on the proportion of crossover (between 1% and 20%; 20% to 40%; 40% or more), did not show any clear evidence of correlation. Nevertheless, the correlation for studies with less than 1% crossover gave rise to the hypothesis that a strong relationship, in general, would have been observed had all studies prohibited crossover. It should be noted that although the R-squared values were quite high, there would still be a lot of uncertainty in terms of trying to infer differences between them.

Figure 6-1: Association between PFS and OS, depending on crossover (Hotta)

Image subject to copyright and so has been removed from the text. Please see the original source (Hotta, 2013) for image.

The observation and results seem quite intuitive as it is unlikely that any of the studies will have been appropriately adjusted for treatment switching. This will mean that whilst the PFS estimates will be unbiased, those for OS will have been severely underestimated for studies with treatment switching. The proportion of bias will depend on both the true underlying survival effect, proportion of treatment switchers, potentially the length of follow-up, and possibly even the method used if a PP analysis was adopted instead of the ITT. Therefore, just stratifying based on treatment switching proportion alone will not necessarily lead to any clear findings (such as a 'dose response' type of relationship, e.g. weaker correlation as the crossover proportion increases).

To establish clear evidence for whether PFS is a surrogate for OS, in examples such as this, the best solution, is to ensure treatment switching is adequately accounted for, before estimates are included within an assessment for surrogacy. Once again, it is unlikely that access to the IPD for all the studies would be obtained, and hence this is a situation where the methods outlined in Chapter 4 and particularly Chapter 5 would need to be used.

6.2.2 Cochrane review for EGFR positive NSCLC patients

Further examples, published after Hotta (2013), have been identified. One of which was a Cochrane review (Greenhalgh, 2016), for first-line treatment for patients with advanced NSCLC with epidermal growth factor receptor (EGFR) positive disease. This review had included nineteen studies, the majority of which had treatment switching in varying proportions. As there were many examples where treatment switching had not been adequately accounted for, there was much scope for the methodology described in Chapter 5. In addition, this case study had the potential to highlight the extent to which adjusting can affect results.

6.2.2.1 *The trials*

Nineteen trials were included, all of which compared chemotherapy (primarily platinum-based) with a type of targeted therapy (with or without chemotherapy). Some of the key trial characteristics, including the study's patient population, blinding, trial phase, sample size of the EGFR population and endpoints. The exact type of chemotherapy used varied across studies, and in many cases a combination of chemotherapies were given. In a few examples patients could have received one of several possible chemotherapies (e.g. gemcitabine in combination with paclitaxel or docetaxel). Common types included: cisplatin, gemcitabine, paclitaxel, docetaxel, vinorelbine and carboplatin. Four targeted therapies were examined: one monoclonal antibodies, cetuximab; and three Tyrosine-Kinase Inhibitors (TKIs), afatinib, gefitinib and erlotinib. References for the trials can be found in Appendix J.

Some (n=8) of the trials were purely within an EGFR positive population, whilst other used unselected population, and then carried out either a pre-planned or ad hoc subgroup analysis. In addition, the patient population was very varied, with some studies having purely Asian participants; a population where particularly TKIs are thought to perform

well. Trial designs ranged considerably across studies; with some advocating a 'crossover' style design where the second-line treatment was the alternative treatment arm.

Whilst the INTACT 1 and 2 trials were reported separately for the ITT population, the publication reporting EGFR data, chose to pool the information for these two trials. Therefore, no individual estimates for EGFR +ve patients were available at a trial level. Thus, this thesis uses the pooled analysis and so treated INTACT as if it were one study rather than the two trials.

6.2.2.2 *Treatment Switching*

In the standard treatment switching analysis for HTA, adjusting the control arm for treatment switching is usually sufficient. This is since switching within the experimental arm, where switching occurs from a non-reimbursed to an already reimbursed intervention, is representative of current NHS practice, whereas the opposite (switching from an already reimbursed intervention to one that is not reimbursed) is not. This analysis continues under the same conditions, thus ignoring whether treatment switching occurred in the experimental arm.

The reporting of treatment switching information was also very variable, and Table 6-2 describes crossover information for each trial. Only:

- 1 trial specified that treatment crossover was not permitted;
- 2 (combining the INTACT trials as one study) did not mention whether treatment switching was or was not permitted, and so are assumed not to have allowed it;
- 4 specified crossover for second-line treatment;
- 2 specified that treatment switching was permitted in the protocol for at least one arm (and may have had ad hoc crossover in the other);
- 3 allowed treatment switching at the physician's discretion;
- 2 only specified that TKIs had been administered in the post-study treatment regimens with no additional details.

Table 6-1 Details of the trial characteristics included in the Cochrane review

Trial name	Phase	Blindedness	Treatment line	Patient population	Population age restriction	Location of participating centre(s)	Outcomes				EGFR +ve sample size
							PFS-IRC	PFS-Inv.	PFS–unspec.*	OS	
BMS 099	III	Unspecified	First	Unselected	18+	USA	Primary	Secondary		Secondary	17
CHEN	II	Unspecified	Unspecified	Unselected	70+	Taiwan			Secondary	Secondary	24
ENSURE	III	Open-label	Unspecified	EGFR +ve	18+	China, Malaysia, Philippines	Primary	Primary		Secondary	217
EURTAC	III	Open-label	Unspecified	EGFR +ve	18+	Spain, Italy, France			Primary	Secondary	174
FASTACT-2	III	Double blind	First to second-line	Unselected	18+	Asia: China, Hong Kong, Indonesia, S. Korea, Philippines, Taiwan, Thailand			Primary	Secondary	97
GTOWG	II	Unspecified	First to second-line	Unselected	75+	Germany			Primary	Secondary	10
FLEX	III	Open-label	First	EGFR +ve	18+	Multinational			Secondary	Primary	1125
First-SIGNAL	III	Open-label	First	Unselected	18+	Korea			Secondary	Primary	53
INTACT 1	III	Unspecified	First	Unselected	18+	Multi-centre, mostly Europe			Secondary	Primary	32
INTACT 2		Double blind				Multi-centre, mostly USA			Secondary	Primary	
IPASS	III	Open-label	First	Unselected	18+	Asia: Hong Kong, China, Indonesia, Japan, Malaysia, Philippines, Singapore, Taiwan, Thailand			Primary	Secondary	261
LUX-Lung 3	III	Open-label	First	EGFR +ve	None reported	Multi-national			Primary	Secondary	345
LUX-Lung 6	III	Open-label	First	EGFR +ve	None reported	China, Thailand, South Korea	Primary	Secondary		Secondary	364
NEJ002	III	Unspecified	First	EGFR +ve	None reported	Asia			Primary	Secondary	228
OPTIMAL	III	Open-label	First	EGFR +ve	<70	China			Primary	Secondary	155
TOPICAL	III	Double blind	First	Unselected	18+	UK			Primary	Secondary	28
TORCH	III	Open-label	First	Unselected	<70	Italy, Canada			Secondary	Primary	39
WJTOG3405	III	Open-label	Unspecified	EGFR +ve	<75	Japan			Primary	Secondary	118
Yu	II	Open-label	First	Unselected	18+	China			Primary	Secondary	32

Table 6-2: Treatment switching in trials included in the Cochrane review

Trial name	Explicit reporting on crossover	Crossover Direction	How crossover was permitted in the trial	Information on the number / proportion who switched in the control group
BMSO99	Yes	Not permitted	-	
CHEN	Yes	Both	As 'Salvage therapy'	22 patients in the ITT population received TKIs
ENSURE	Yes	Both	Crossover permitted at time of progression	92 (85.6%) of patients randomized to the GP arm received further treatment with EGFR TKIs
EURTAC	Yes	Both	Crossover recommended at time of progression (as part of study design)	65 (74/7%) of patients received erlotinib after discontinuation of randomly assigned treatment
FASTACT 2	Yes	Both	Control: Treatment unblinded & patients offered crossover at time of progression; Post study treatment for TKI arm at physician's discretion	57 (73%) optional cross-over to erlotinib
First-SIGNAL	Yes	Both	Post study treatments at physician's discretion	94 patients received gefitinib
FLEX	Yes	-	Post study treatments	Patients in both groups received subsequent line therapies including TKIs (not cetuximab).
GTOWG	Yes	Both	Crossover recommended at time of progression (as part of study design)	Unspecified amount but as commented recommended at time of progression
INTACT 1 & 2	No	n/a	-	
IPASS	Yes	Both	Subsequent therapy was at the physician's discretion for the control group; crossover recommended for experimental treatment group	83 patients received gefitinib
LUX-Lung 3	Yes	Both	TKIs included in subsequent treatment regimens	78 (65%) patients crossed over to afatinib
LUX-Lung 6	Yes	Both	TKIs included in subsequent treatment regimens	61 patients received afatinib
NEJ002	No	Both	Crossover recommended at time of progression (as part of study design)	112 patients ultimately received gefitinib in subsequent therapy lines
OPTIMAL	Yes	Both	Post study treatments given which included crossover	33.7% were treated with erlobinib, 46 at the initial data cut-off; 50 by the latest date.
TOPICAL	Yes	Both	Post study treatments at physician's discretion	No additional information given.
TORCH	Yes	Both	Crossover recommended at time of progression (as part of study design)	In the ITT population 226 patient switched to erlotinib
WJTOG3405	Yes	Both	Post study treatments at physician's discretion	78 patients switched to gefitinib
Yu 2014	Yes	One	Treatment unblinded & patients offered crossover at time of progression	8 patients received gefitinib

6.2.3 Available information

As shown in Table 6-3, the information available for the trials, which need to be reanalysed, is variable. The GTOWG and the EURTAC trial contained the most detailed information about the survival. With the exception of the TOPICAL and Yu trials, the others all contain, at the very least Kaplan-Meier curves for PFS and OS. Two trials were later excluded from the analysis. The first was the TOPICAL trial, which was removed because it did not report Kaplan-Meier curves for either PFS or OS at subgroup level, making it impossible to reconstruct the data. The second was the trial reported by Yu; whilst a PFS Kaplan-Meier curve was published on the EGFR +ve subgroup, no OS information was given on the subgroup, and thus, the OS data could not be reconstructed or reanalysed for the EGFR +ve population.

6.2.4 Methodology: overview based on the initial intentions of data reconstruction

For 11 (CHEN, ENSURE, FASTACT-2, First-SIGNAL, IPASS, LUX-Lung 3, LUX-Lung 6, NEJ002, OPTIMAL, TORCH, WJTOG 3405) of the 13 trials with treatment switching that needed reconstructing, PFS and OS Kaplan-Meier curves for the EGFR +ve patients were extracted. The information for the GTOWG study was reported in the form of a poster presentation which contained a table including the survival times (both for PFS and OS) for the ten patients with EGFR activating genes. Therefore, for this study no data reconstruction took place. For the last trial, the EURTAC trial, detailed summary information was provided specifically for this analysis by the manufacturer. This included life tables, Kaplan-Meier curves, risk-tables, and the number of events for TTP, PFS, death before progression, PPS, OS, and PFS and OS censoring.

For the EURTAC study, the survival proportion and times from the life-tables for the individual transitions and OS censoring were used to model these distributions, using the methodology outlined in Section 5.4.2.1. For the remaining studies (where the PFS and OS graphs had been extracted), the methods, described in Section 5.4.2.3, were used. For all studies, where data were reconstructed, 1000 datasets (per example) were created. Once the PFS and OS data had been reconstructed, the treatment switching information was then reconstructed using the ‘at random’ switching mechanism, and the RPSFTM applied to the data (detailed in Section 5.5). Having obtained all adjusted estimates, a meta-regression model was fitted, for both the adjusted and unadjusted OS

Table 6-3: Information available for the Cochrane review trials

Trial name	Patient population	EGFR +ve survival IPD	EGFR +ve crossover IPD	Life tables & KM curves for transitions & cens.	Life tables for PFS & OS	No. of PFS events	No. of OS events	No. of TTP events	No. of switchers	KM curves for PFS & OS	Risk-tables for PFS & OS
CHEN	Unselected	No	No	No	No	Yes	Yes	No	ITT level only	Yes	No
ENSURE	EGFR +ve	No	No	No	No	No	No	No	Yes	Yes	Yes
EURTAC	EGFR +ve	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
FASTACT-2	Unselected	No	No	No	No	Yes	Yes	No	ITT level only	Yes	Yes
GTOWG	Unselected	Yes	No	No	No	Yes	Yes	No	No	No	No
First-SIGNAL	Unselected	No	No	No	No	No	No	No	ITT level only	Yes	No
IPASS	Unselected	No	No	No	No	Yes	Yes	No	Yes	Yes	Yes
LUX-Lung 3	EGFR +ve	No	No	No	No	Yes	Yes	No	Yes	Yes	Yes
LUX-Lung 6	EGFR +ve	No	No	No	No	No	Yes	No	Yes	Yes	Yes
NEJ002	EGFR +ve	No	No	No	No	Yes	Yes	No	Yes	Yes	Yes
OPTIMAL	EGFR +ve	No	No	No	No	Yes	Yes	No	Yes	Yes	Yes
TOPICAL	Unselected	No	No	No	No	No	No	No	No	No	Yes
TORCH	Unselected	No	No	No	No	Yes	Yes	ITT level only	ITT level only	Yes	No
WJTOG3405	EGFR +ve	No	No	No	No	Yes	Yes	No	Yes	Yes	Early FU only
Yu	Unselected	No	No	No	No	No	No	No	No	PFS only	PFS only

results, where the outcome was the log-HR for OS and with the log-HR for PFS as the covariate of interest. In order to compare the findings to those presented in Hotta (2013), a linear regression model with the OS HR as outcome, and PFS HR as covariate was also fitted, where observations were weighted by sample size. The regression models were plotted graphically, and the R^2 value recorded. Results for the overall population, and stratifying by treatment switching were collected.

6.2.5 Issues encountered and possible solutions

To develop this method, a 'standard' set of information was assumed to be available; information which allows 'consistency' between outcomes. In other words, follow-up, time of reporting and intervals for the 'numbers at risk' table are the same for both outcomes. Nevertheless, the variability in the design and reporting of survival analysis trials is great in practice. This meant that, in order to fulfil the aim of reconstructing these studies, amendments had to be made to the methodology to accommodate the differences.

The key issues included:

- Differential reporting times for the 'numbers at risk' table
- Absence of the 'numbers at risk' table – but inclusion of some information on censoring
- Differential reporting times for outcomes
- Extracting subgroup specific information
- Little or no information on treatment switching
- Estimating the number of events
- Differential censoring

The complications caused by these issues, potential solutions and the approaches taken here are explained in detail in the following sections and were instrumental in forming the suggested guidelines for reporting given in Chapter 7:.

6.2.5.1 *Differential reporting times for the 'numbers at risk' table*

There can often be a great difference in the shape of the survival curve for PFS and OS; this may mean that it is more reasonable to use different scales for different outcomes when producing Kaplan-Meier curves and risk tables. For example, at eighteen months,

the survival proportion may be less than ten percent for PFS, and more than sixty percent for OS. Therefore, there is little benefit in reporting PFS past this point, and it may be useful to report 'numbers at risk' every three months say (six monthly intervals may not fully represent the data). However, if follow-up data are available for up to sixty months, by which time the OS proportion has reached approximately twenty percent, it would be useful to report this. Using three monthly intervals on the graph for OS, might, nevertheless, appear cluttered. Therefore, a scale of either six or twelve months may be chosen instead. This may be particularly common as well where outcomes are reported separately.

Figure 6-2: Differences in the intervals for PFS and OS in ENSURE

Image subject to copyright and so has been removed from the text. Please see the original source (Wu, 2015) for images.

Reported Kaplan-Meier curves for (top) PFS and (bottom) OS, complete with risk-tables. Follow-up time and number at risk intervals vary by outcome.

An example of this can be seen with the ENSURE trial (Figure 6-2), where the authors show PFS data up to fourteen months, reporting the 'number at risk' every two months, whilst OS follow-up lasts up to forty-two months, with the 'number at risk' given every six months.

The impact this issue has on the method is marginal. This means that the same level of agreement for the two endpoints cannot be guaranteed, when scaling is used to re-calculate the number at risk, events and censorings for a given partition.

6.2.5.2 *Absence of the 'Numbers at risk' table*

Whilst the majority of Kaplan-Meier curves were accompanied by risk tables there were several exceptions: CHEN, First-SIGNAL, TORCH and the long-term follow-up for WJTOG 3405. Nevertheless, for the CHEN, the number of censorings were reported. Therefore, the study length (approximately 36 months), was treated as a single interval, and so the censoring distribution was calculated as:

$$S_{cens_j}(t) = \exp(-\lambda_j t), \qquad \lambda_j = -\frac{\ln\left[1 - \left(\frac{c_j}{n_j - \frac{1}{2}(n_j - c_j)}\right)\right]}{36} \tag{6-1}$$

Where j indicates the treatment group, and c_j, n_j the number of censorings and total number of patients in the j^{th} group. An exponential distribution, and thus a constant hazard rate, is assumed.

For the other three (First-SIGNAL, TORCH and WJTOG 3405), the censorings were indicated on the Kaplan-Meier curve. In each case, however, the locations of the censoring provided evidence against using either the recruitment times or maximum censoring time methods. Both the recruitment times or maximum censoring time methods assume purely administrative censoring, but since all trials indicate censoring within the minimum follow-up, this assumption is violated, invalidating the use of these approaches. Thus, for WJTOG 3405, it was decided to use the data reported at the initial data cut-off, which contained a risk table, and effectively ignore this later follow-up. Both the First-SIGNAL and TORCH trials were comparatively small trials (n=39 and n=53 respectively), and thus the coordinates were expected to capture the majority of event times. The 'tick' marks for specifying the censorings were then digitized; here the time

coordinate was of the utmost importance whilst the survival proportion was largely irrelevant. The survival proportion was ignored because this information primarily related to events, rather than censorings. The timescale was then divided into intervals, based on the scale for the time on the PFS Kaplan-Meier curve (e.g. every 10 months for First-SIGNAL). The number of coordinates for 'events' were then calculated for each interval, and the same was done for those for censorings. These were then treated as proxies for the number of events and censorings for an interval and as such are scaled to ensure that the total 'number of events' matched that which was reported. The 'number at risk' at the start of an interval is then estimated as:

$$\hat{n}_i = \hat{n}_{i-1} - \left(\hat{d}_{i-1} + \hat{c}_{i-1}\right) \tag{6-2}$$

This achieved, the censoring distribution for interval *i* and treatment group *j*, is calculated as:

$$S_{cens_{ji}}(t) = \exp\left(-\lambda_{ji} t\right), \qquad \lambda_{ji} = -\frac{\ln\left[1 - \left(\frac{\hat{c}_{ji}}{\hat{n}_{ji} - \frac{1}{2}\left(\hat{d}_{ji}\right)}\right)\right]}{l_i} \tag{6-3}$$

6.2.5.3 *Differential reporting times for outcomes*

'Differential reporting times' for outcomes essentially occurs when there are different lengths of follow-up for PFS and OS. This is a very key issue and proved to be a common one, occurring in at least eight of the studies which were to be reconstructed. Primarily, it is due to PFS data reaching maturity much earlier than the OS data. Often once PFS data are considered mature, there is less reason to record PFS data for any events beyond this cut-off point. For OS, though, the patients must be followed-up until the data for that outcome are mature, and so it is possible that the length of follow-up for OS is longer than for PFS. This potentially violates one of the most vital assumptions of the simulation methods; that which assumes that patients who are censored for PFS, are also censored at the same time for OS. In this situation, however, even assuming purely administrative censoring, it is perfectly possible that a patient might be administratively censored for PFS, but then followed up and either be administratively censored at a different time OR die during the subsequent follow-up.

Assuming that in the context of a particular application, only administrative censoring occurs, then the problem can be considered as essentially having two time scales: calendar

time and study time. This issue will be illustrated further, using the OPTIMAL trial as the example, and at present assuming that only administrative censoring occurred (this may or may not be a valid assumption). In this example, PFS and OS were reported twice, however, none of these analyses used the same data cut-off date (PFS: 16th August 2010, 7th January 2011; OS: 31st December 2011; 21st December 2012). Taking the updated analysis for both PFS and OS, this means that there is a difference of 23.5 months follow-up.

Adopting a similar approach to that used during the 'recruitment time' censoring distribution (Section 4.4.1.4.2) can help to understand more about the issue. Given that no patients could leave the study for any other reason that experiencing an event or being administratively censored, any patient not experiencing a PFS event, must be in the study for at least 17.6 months (between 17th July 2009 and 7th January 2011). The maximum length of time a person could have participated in the study would be 28.4 months (this would mean that that person had been enrolled / randomised on the 24th August 2008 – the start of recruitment). Thus, the observed time of any censored patient would lie between 17.6 months and 28.4 months for PFS. Now consider OS; the follow-up for this continues for a further 23.5 months. However, assuming these people attend follow-up, there is the possibility that within that time they could either die (before or following progression) or be censored administratively, at which point their OS time of censoring would be their PFS time plus 23.5 months (hence between 41.1 and 51.9 months). If they experienced an OS event, then this must occur after the PFS censoring time.

It may be possible to decipher some of this information (e.g. difference in follow-up, minimum and maximum length of PFS follow-up, etc.) from the trial publications. However, it is still reliant on the assumption that patients are only censored administratively, which may or may not be valid; and additionally, does not aid the calculations needed to define distributions for post-progression events. This is because it is impossible to distinguish between post-progression events / censorings for known progressors and those for patients who were administratively censored for PFS.

This situation is potentially most problematic when the data cut-off date, at least for PFS, occurs soon after the end of recruitment, as it is more likely that there will be a proportion of patients who will not have been followed up for sufficient time to enable disease

progression to be recorded. Ignoring this differential follow-up would potentially seriously affect the simulation technique; since it is unlikely the censoring distribution will be correct (e.g. censoring for OS occurs too early; not enough censoring for OS patients towards the end of the timescale). Also, if administratively censored patients experience events, these events will be attributed to 'post-progression', and thus the PPS distribution will likely underestimate the survival proportion. This in turn could bias the treatment switching reanalysis, if the PPS is shortened because the administrative censoring events have not been taken into account. Differential reporting of outcomes definitely affected the following trials: ENSURE, First-SIGNAL, IPASS, LUX-Lung 3, LUX-Lung 6, NEJ002 and OPTIMAL.

6.2.5.3.1 Considering solutions

This problem is very complex, and whilst, at this stage, the effects ignoring this could have in practice cannot be predicted; theoretically they could be quite serious. Therefore, it is important to consider how adjustments could be made to account for the differential follow-up and still allow the method to work. It should be noted though, that whatever, amendments are made, these will no doubt rely on some strong assumptions.

The possible options are to:

1. Assume administrative censoring for both PFS and OS, where patients who are administratively censored at PFS have to also be censored for OS (prohibiting events from PFS censored patients)
2. Restrict the OS timescale, by using information on recruitment times to re-censor OS
3. Consider the problem in terms of multi-state model diagram; essentially considering patient's moving into an 'administrative censoring' state before moving into one of the other health states (e.g. death).

6.2.5.3.2 Administrative censoring

Option 1 could be very easy to implement in practice, and modifications could be made to allow some patients to have been censored for other reasons. Below gives an example

Figure 6-3: Example of censoring due to differential reporting times

Colour coding: Light grey indicates periods in which any censorings cannot be attributed to administrative censoring, blue indicates periods in which censorings would be attributable to administrative censoring, dark grey indicates periods in follow-up was not recorded for that outcome

Time, months	Study time		Intervals	Total censorings		Pre-progression censorings		Post-progression censorings	
				PFS	OS	Due to drop-out	Administrative	Due to drop-out	Administrative
1	Minimum study time for PFS	Minimum study time for OS	1	2	2	2	0	0	0
2									
3									
4									
5			2	3	5	3	0	2	0
6									
7									
8									
9			3	0	9	0	0	9	0
10									
11									
12									
13			4	5	2	0	5	2	0
14									
15									
16	Admin. cens.								
17			5	6	2	0	6	2	0
18									
19									
20									
21			6	0	21	0	0	0	21
22		Admin. Cens							
23									
24									

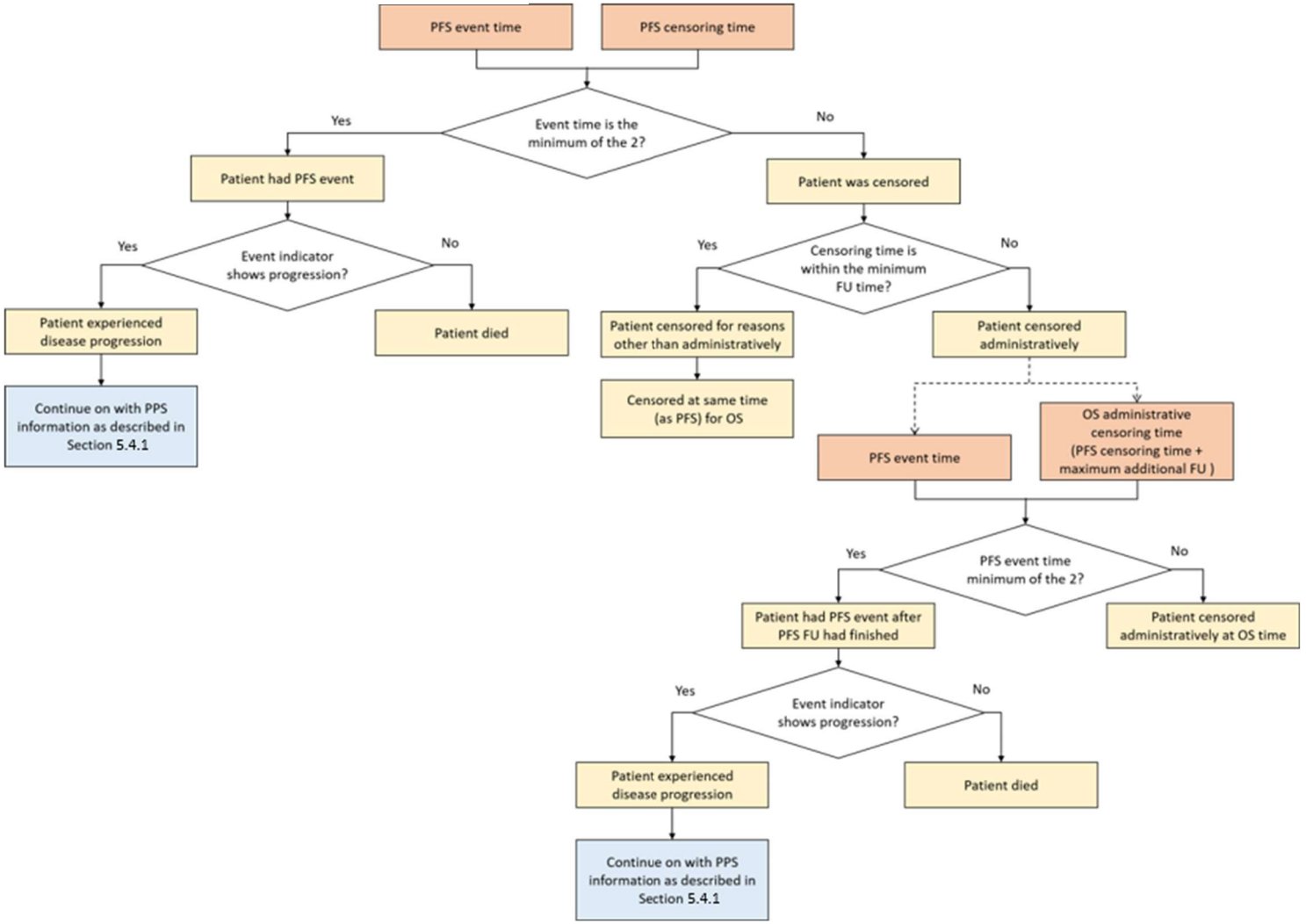
The first stage of the method would be to calculate the PFS and OS events as usual. Then, the minimum and maximum length of study time should be computed. Any PFS censorings estimated as occurring before the minimum length, must be attributed to drop-out for reasons other than administrative, and assumed to also occur for OS. For intervals

which occur after the minimum study time for the outcome, these censorings are all assumed to be administrative. If the minimum study time occurs during an interval (rather than between them), it could either be decided to treat all the censorings during that interval as administrative or only consider a proportion as administrative.

This is demonstrated in the above example, where all censorings in the interval containing the minimum time, have been attributed to administrative censoring. This approach means than of the twenty-one estimated administrative OS censorings, eleven (five plus six) censorings would be the administrative censorings from PFS, re-censored after the end of follow-up. Using this framework and assuming the original calculations were 'consistent' (i.e. the number of OS censoring is greater than or equal to the numbers of censorings for PFS) for all intervals before the minimum time, this framework allows the censoring distributions for pre- and post-progression to be calculated easily.

PFS information is simulated as usual (Chapter 5), however for patients whose censoring time falls into the 'administratively censored' time period, the data must be treated differently. For these patients, a new administrative censoring time (for OS) is constructed by adding the additional follow-up length to their PFS censoring time. In contrast with previous descriptions of this technique for 'PFS censored' patients, this revised administrative time is then compared with the patient's PPS time. If the PPS time is the minimum of the two values, then the patient is defined as having had a PPS event. If, instead the censoring time is the minimum, the patient is 'censored'. This approach, therefore, takes account of the situation where insufficient follow-up has been allowed for all patients to experience progression, and thus it is likely that some patients would have experienced progression and death post-progression subsequent to the publication of the PFS outcome. This is demonstrated in the flowchart in Figure 6-4.

Figure 6-4: Flowchart for the process



6.2.5.3.3 Restricting the timescale

This method is perhaps the most straight-forward approach of the three. Initially, the issue of differential reporting is essentially ignored: PFS events and censoring distributions are calculated as usual as is the PPS event model. The only minor deviation for the PPS censoring is that where the number of PFS censorings is greater than OS censorings there is assumed to be no difference (i.e. Equation 6-4).

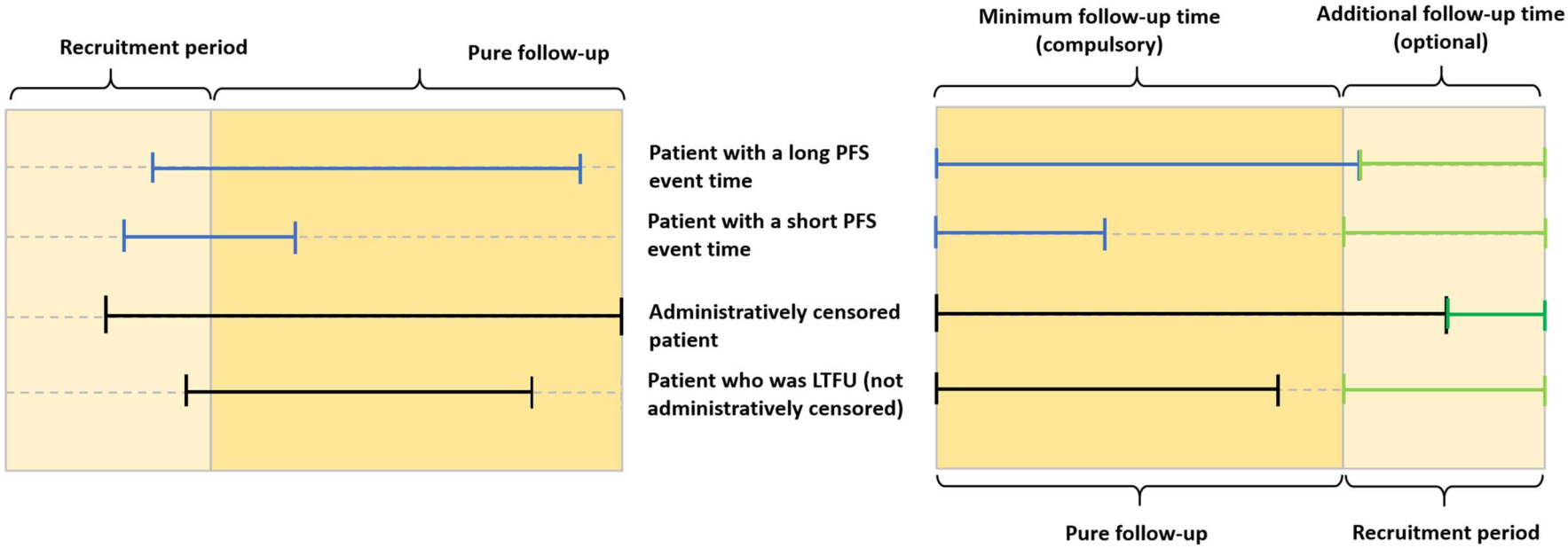
$$c_{PPS_i} = \max(0, c_{OS_i} - c_{PFS_i}) \quad (6\text{-}4)$$

Once the entire dataset (including survival times for PFS and OS has been created), recruitment time information is added. In earlier chapters (Chapter 4 and Chapter 5) of the thesis, using recruitment times censoring only involved simulating a maximum study length from a uniform distribution (independent of any other information). However, here the PFS data includes vital information on censoring which must be taken account of. In other words, if a PFS event did occur then the recruitment time must be such that it allows sufficient time for that event to have occurred in the study's duration. Figure 6-5 explains the four possibilities that may occur.

First, it is vital to understand what is meant by the 'compulsory' and 'optional' parts of follow-up (as shown on right side of Figure 6-5). Unless a patient has left the study before the end of the study (e.g. through withdrawal of consent, LTFU etc.) or experienced an event, they must have remained in the study at least, from the end of recruitment to the data cut-off time. This length of time has been termed 'minimum follow-up time' and is 'compulsory'.

The length of recruitment period, has been termed as 'additional' and 'optional' follow-up', as whilst all patients should experience some proportion of this, the amount will vary from person to person, i.e. the earlier the patient was recruited (e.g. at the start of recruitment), the more 'additional follow-up' they will have. 'Maximum follow-up' refers to the sum of the minimum (compulsory) and additional (optional) follow-up times, and is the maximum length of time a patient could have been in the study.

Figure 6-5: Recruitment period and follow-up



In Figure 6-5, the panel denoted (A) shows what the situation might look like in calendar time; patients enter the study at some point throughout the recruitment period and are followed up until they

1. have an event,
2. leave the study for some reason, or
3. are administratively censored at the data cut-off time.

Panel (B) translates this over to study time, where all patients enter the study at time = 0.

For the first patient (with long PFS event time), their end study time must occur between their event time and the maximum length of follow-up. With the second patient, their event time is such that any of the simulated time from the 'additional' follow-up period would be satisfactory. Thus, for patients who are defined as experiencing an event, their t'_{max} ('end study time' for the patient).

$$t'_{max} = \begin{cases} t_{PFS} + (t_{max} - t_{PFS})R & \text{if } t_{PFS} \geq t_{min} \\ t_{min} + (t_{max} - t_{min})R & \text{if } t_{PFS} < t_{min} \end{cases} \quad (6\text{-}5)$$

Where $R \sim \text{Uniform}(0,1)$

For patients defined as being censored, if their censoring time falls within the additional follow-up time, then these are assumed to be administratively censored patients. Therefore, their recruitment time is exactly the difference between the defined censoring time and the maximum follow-up time (no proportion unlike the other situations mentioned). The final individual is illustrative of a patient who must have been censored for reasons other than administrative, as the simulated censoring time is within the minimum follow-up time. Hence the recruitment time, for this patient, is simulated from the uniform distribution between zero and length of the additional follow-up time. In other words, for censored patients:

$$t'_{max} = \begin{cases} t_{max} & \text{if } t_{cens} \geq t_{min} \\ t_{min} + (t_{max} - t_{min})\,R & \text{if } t_{cens} < t_{min} \end{cases} \quad (6\text{-}6)$$

Where $R \sim \text{Uniform}(0,1)$

Once an end study time has been generated for each patient, the OS data are re-censored based on these study times. This re-censored analysis should then be representative of the OS data at the PFS data cut-off.

6.2.5.3.4 Multistate model framework

The third approach is to consider administrative censoring as if it were one of the health states, by essentially formulating a multi-state model framework. This framework could be particularly valuable where the PFS data cut-off occurs soon after the end of recruitment. This is because the follow-up period may be insufficient to capture progression of patients only recently recruited, but the longer OS follow-up might capture their death (possibly post-progression). This approach explicitly considers this happening.

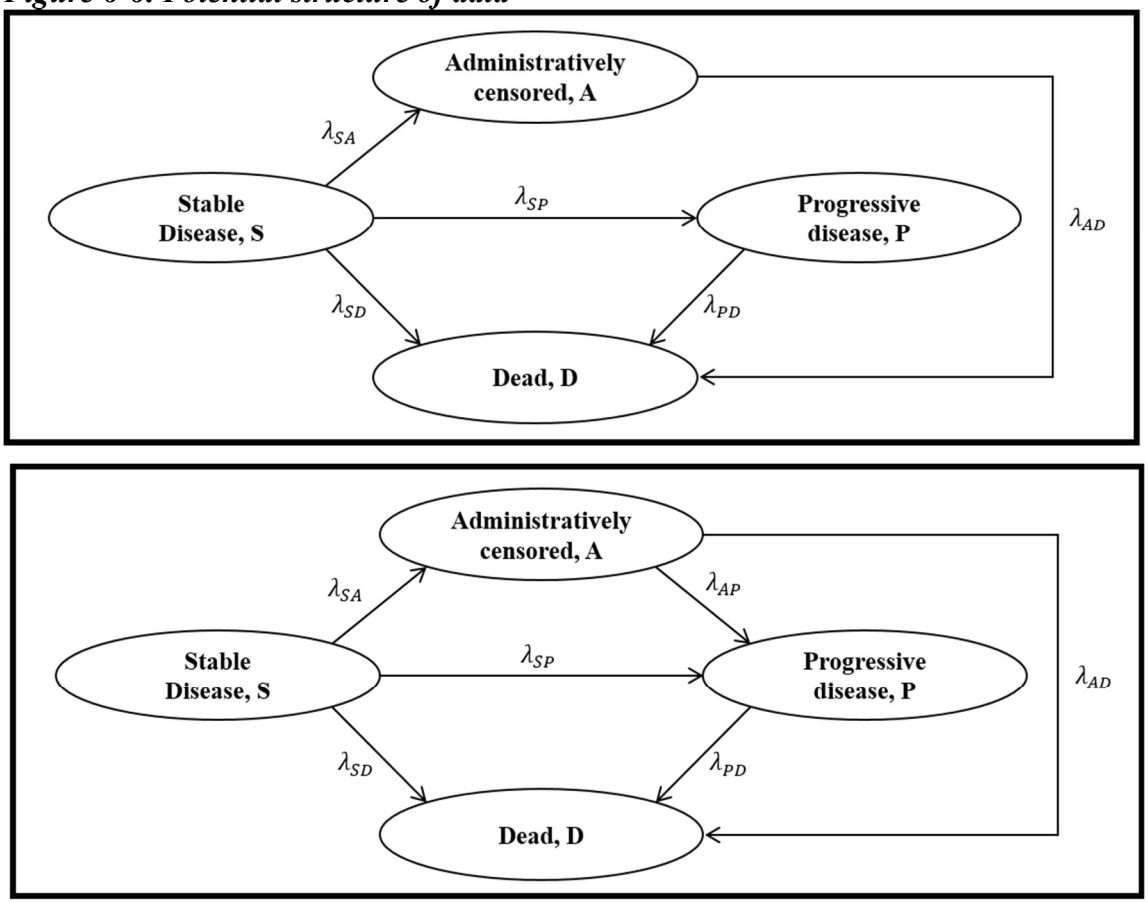
However, within this structure there are essentially two possibilities for models that could be used. The first model assumes that once in the 'administratively censored' state, the only way of leaving it, is by dying (i.e. transitioning to progressive disease cannot occur). The second model is potentially more realistic, but relies on two exceptionally strong assumptions which are:

1. The same care pathways can occur after being administratively censored e.g. patient can either progress or die before progression.
2. Once a patient who was administratively censored (for PFS) has progressed, they transition from progressive disease to dead at the same rate as those who progressed from stable disease.

These models are very useful for understanding the problem and the mechanism that needs to be represented, but this adds in several additional functions to be estimated (namely $\lambda_{SA}, \lambda_{AD}$ and for model 2, also λ_{AP}). It may be possible to estimate λ_{SA} , at the very least using information from the recruitment and follow-up times, and possibly in conjunction with the 'numbers at risk' table. The other functions would be exceptionally hard to estimate as there is almost no distinct information on these. It may perhaps be possible to continue with the method under the following assumptions: for model 1, $\lambda_{PD} = \lambda_{AD}$ (ideally it should be better if $\lambda_{PD} = \beta\, \lambda_{AD}$, however the limited data available will not be sufficient to estimate β); for model 2, it would need to assume that the total hazard rate from stable to progression via administrative censoring, is the same as that from transitioning directly from stable to progression. Similarly, it requires that the total hazard rate from stable to death via administrative censoring, is the same as that from transitioning directly from stable to death. Whilst for model 2, this set-up should seem

trivial and collapsible to the three state model, for the estimation of PPS and simulation of this data this underlying structure would be very valuable.

Figure 6-6: Potential structure of data



6.2.5.3.5 Solution used in this example

In order to address the issue within this example, it was decided to take the approach described in Section 6.3.5.3.3, restricting the timescale, as this was relatively easy to implement, very understandable and perhaps the most intuitive. It is also less likely to be biased, but does result in a loss of precision. In addition, of the three potential solutions it possibly relied on the least assumptions.

6.2.5.4 *Extracting subgroup specific information from an ITT population*

As explained in section 6.2.2.1, whilst the review concentrated on the EGFR positive mutation patients, several of the trials used an unselected population, some of which allowed treatment switching. Although all trials report some subgroup-specific information, which statistics are reported vary across papers. For very few papers, all

information (Kaplan-Meier curves for both PFS and OS, number of events, 'at risk' tables and treatment switching proportion) is available at subgroup level.

There are several options open to do this:

1) Reconstruct the entire unselected population:
    a. If data for every single subgroup are available, but no subgroup-specific information on treatment switching is:
        i. Reconstruct data for each subgroup separately and combine, making sure to have a subgroup specific indicator variable.
        ii. Apply the treatment switching proportion as if conducting the analysis on the unselected population.
        iii. Define the final dataset for the unselected population.
        iv. Analyse the EGFR positive mutation data (based on the subgroup indicator).
    b. If data for not every subgroup (only that of interest) are provided:
        i. Simulate data from the unselected population.
        ii. Apply the treatment switching information, ignoring the interest in any specific subgroup.
        iii. Model the subgroup specific hazard function for one outcome.
        iv. Define the EGFR status using a Bernoulli distribution with probability $(EGFR\ status = +ve) = \frac{\lambda_{EGFR+ve}}{\lambda_{Unselected}}$. Where λ_i is the hazard function for either PFS or OS.
        v. Analyse the data using this EGFR status to stratify mutation positive patients
2) Reconstruct data for the EGFR mutation positive patients only
    a. Assume the treatment switching proportion is the same for subgroup as it is for the unselected population
    b. Assume the treatment switching proportion for the subgroup is proportional to that of the unselected population and the progression events for the subgroup and unselected population.

For this case study, approach (2) was implemented.

6.2.5.5 *Limited or no information on treatment switching*

For those trials which did not report that treatment switching had or had not occurred, it was assumed that no treatment switching had occurred. However, a smaller group of studies specified that treatment switching had been permitted at the physician's discretion but did not quantify how often this had happened. Therefore, for these studies, it was decided to assume that all patients who could have switched (i.e. all patients experiencing progressive disease) did switch.

6.2.5.5.1 Estimating treatment switching information for the subgroup

Five of the trials used an unselected population, and thus for four of these, no subgroup-specific information was reported on treatment switching. Information was available, however, on the full ITT population. In an attempt to capture both the association between TTP and switching, for these trials, the proportion of switchers in the estimated TTP unselected population was calculated, and then this proportion was applied to the estimated TTP population for the subgroup. In the practical application of assigning treatment switchers in the reconstructed data, if the estimated / reported number of treatment switchers exceeded the estimated / reported number of progressors, all patients who progressed were assumed to have switched.

6.2.5.6 *Estimating the number of events*

6.2.5.6.1 Estimating PFS events

Looking at the papers there is a variety of reporting with regard to either specific subgroup, and / or ITT information on events. These quantities are important for scaling censoring and in comparing the reconstructed data with the IPD. However, several papers did not report these. Originally it had been hoped that modelling objective response rate (ORR) and the number of PFS events using ordinary least squares regression, at subgroup level, where both had been reported, could provide useful predictions for an approximate number. However, the predictions were considered very unreliable when compared with the total number of events across both groups (for the LUX-Lung 6 trial), and also for the approximate number of censorings (calculated at a later date, by counting the censoring tick marks on the Kaplan-Meier). Therefore, these predictions were scaled by the reported total number of events for the LUX-Lung 6 trial, and by using the estimated number of censorings (from the tick marks) for the others.

6.2.5.6.2 Estimating TTP events

None of the studies provided information at subgroup level for the number of progressions, and only one trial reported it at the unselected population level (TORCH). It seemed rare that in any of the studies, some patients had died before disease progression. Therefore, it was decided to use the TORCH trial information to determine the average proportion for the number of patients expected to die before progression. This meant for the control group that, 94.3% (316/335) of PFS patients were expected to experience progression and 93.5% (333/356) for the experimental group. These proportions were applied to the reported number of PFS events (and rounded to the nearest integer) to give an estimate of the number of progressions. Where the PFS event total had not been actually published, that obtained when estimating the censoring distribution was used. The only exception was the control arm of the NEJ002 trial, in which all included patients received subsequent treatment (secondline or later) with an TKI, and therefore, it had to be concluded that no patients died before progression.

6.2.5.7 *Differential censoring*

This issue was highlighted by the EURTAC trial where a greater amount of summary data was available. Having the life tables illustrated that whilst there were a greater number of censorings early on for PFS, these did not all appear in the OS life table. This would affect the PPS censoring and the assumption of 'consistent censoring'. The reason for this seemed largely unclear, but it was ultimately suggested that it could be that although a patient may withdraw from or be LTFU for PFS, their death may be flagged up and thus recorded. For this analysis, this issue has been ignored.

6.2.5.8 *Practical issues in model fitting*

Choosing appropriate models is an ongoing challenge for this method, and was certainly in this example. Some of the most crucial models were for OS and PPS, and whilst the RCS are flexible, at times it was a struggle to find distributions that were flexible enough to cope with the large decrease in survival that occurred soon after randomisation, such as when there were many tied events very early in the trial. Also, in some cases the data negatively affected the choice of model; towards the tail of the distribution large steps influenced the curve into a rapidly decreasing survival function, when a plateau was more

reasonable. This was ultimately achieved by restricting the timescale, but did require additional and in some cases time consuming input to find realistically suitable models.

6.2.5.9 *Using earlier follow-up*

For two studies, IPASS and WJTOG 3405, the data reconstructed was all for the earliest follow-up time. In the case of IPASS, this was chosen as it avoided contending with the differential reporting of outcomes and the additional assumptions that would otherwise have needed to be used. For WJTOG 3405, this earlier time point was chosen because additional information ('numbers at risk tables') was available. As explained previously, unlike the initial publication, the poster showing subsequent follow-up did not include risk-tables, and thus using this follow-up could have led to issues with the censoring.

Table 6-4: Comparison of reported and reconstructed HRs

Name	Reported information						Average of Reconstructed data			
	PFS		OS (early FU)		Extended FU OS		PFS		OS	
	HR	SE	HR	SE	HR	SE	HR	SE	HR	SE
CHEN	0.544	0.491	n/a		2.355	0.608	0.620	0.531	0.753	0.570
ENSURE	0.340	0.222	n/a		0.910	0.188	0.356	0.189	0.863	0.343
EURTAC	0.370	0.200	n/a		1.040	0.248	0.366	0.218	1.043	0.219
FASTACT-2	0.250	0.228	n/a		0.480	0.294	0.207	0.303	0.464	0.290
First-SIGNAL	0.544	0.359	n/a		1.043	0.377	0.556	0.333	0.938	0.442
IPASS	0.480	0.147	0.780	0.227	1.000	0.140	0.470	0.160	0.831	0.227
LUX-Lung 3	0.580	0.153	n/a		0.880	0.147	0.551	0.143	0.909	0.223
LUX-Lung 6	0.280	0.172	n/a		0.930	0.131	0.251	0.159	0.921	0.219
NEJ002	0.322	0.159	0.798	0.171	0.887	0.171	0.297	0.195	0.814	0.243
OPTIMAL	0.160	0.240	1.040	0.406	1.190	0.184	0.173	0.226	0.910	0.318
TORCH	0.600	0.354	n/a		1.580	0.415	0.549	0.433	1.701	0.436
WJTOG3405	0.520	0.163	1.638	0.399	1.185	0.222	0.504	0.187	2.237	0.475

Table 6-5: Comparison of reported and reconstructed events

Name	Reported information						Average of Reconstructed data			
	PFS		OS (early FU)		Extended F OS		PFS Events		OS Events	
	Chemo.	Target.	Chemo.	Target.	Chemo.	Target.	Chemo.	Target.	Chemo.	Target.
CHEN	13	8	n/a		6	6	13.4	8.3	11.0	7.3
ENSURE	Not reported		n/a		Unclear		75.6	45.8	29.1	39.2
EURTAC	59	52	n/a		31	38	67.9	57.0	38.0	46.3
FASTACT-2	Not reported		n/a		Not reported		47.7	37.1	15.7	20.8
First-SIGNAL	Not reported		n/a		Not reported		15.0	24.5	9.6	16.4
IPASS	111	97	43	38	95	104	115.8	101.3	43.3	37.7
LUX-Lung 3	83	155	n/a		73	140	0.551	79.4	154.2	46.7
LUX-Lung 6	221 (overall)		n/a		84	162	0.251	89.8	187.1	38.3
NEJ002	100	87	43	39	69	69	105.6	89.8	35.0	36.6
OPTIMAL	64	58	42	50	52	68	64.4	54.6	19.2	24.0
TORCH	82	77	n/a		10	14	18.3	18.2	9.8	14.6
WJTOG3405	111	97	10	17	39	43	67.8	53.1	8.1	17.0

6.2.6 Results

6.2.6.1 *Reconstructed data*

Table 6-4 illustrates the reported HRs for PFS and OS for the reconstructed data and the average over the 1000 datasets, whilst Table 6-5 gives the comparison for the number of events. Considering PFS HR, the point estimates are largely comparable, indeed in most cases there is very good agreement. CHEN and TORCH are in general the exceptions, with there being a difference of more than 0.05 between the reported and IPLD HR. However, it should be noted these were very small studies (n<40). There is also some difference in the log-SE for the PFS HR, but this is, generally comparable. Differences are more apparent in OS, although for some studies, this may be due to the restricted length of follow up. The least similar estimates are for the CHEN trial. The WJTOG 3405 trial also has poorer agreement, as do First-SIGNAL and TORCH to a slightly less extent. For some trials, the log-SE is increased, but this could be attributable to the restricted timescale, which would naturally lead to increased variability.

6.2.6.2 *Impact of treatment switching on surrogacy*

Table 6-6: Revised estimates accounting for treatment switching

Name	N	Extended OS		Subject to 'crossover'	IPLD OS		Adjusted OS	
		HR	SE		HR	SE	HR	SE
BMS 099	17	1.620	0.561	No	n/a			
CHEN	24	2.355	0.608	Yes	0.753	0.570	0.653	0.857
ENSURE	217	0.910	0.188	Yes	0.863	0.343	0.594	1.209
EURTAC	174	1.040	0.248	Yes	1.043	0.219	1.089	0.439
FASTACT-2	97	0.480	0.294	Yes	0.464	0.290	0.370	0.376
GTOWG	10	0.781	0.837	Yes	n/a		0.451	2.700
FLEX	1125	0.871	0.068	No?	n/a			
First-SIGNAL	53	1.043	0.377	Yes	0.938	0.442	0.948	0.370
INTACT 1 & 2	32	1.770	0.645	No?	n/a			
IPASS	261	1.000	0.140	Yes	0.831	0.227	0.788	0.291
LUX-Lung 3	345	0.880	0.147	Yes	0.909	0.223	0.807	0.505
LUX-Lung 6	364	0.930	0.131	Yes	0.921	0.219	0.903	0.271
NEJ002 (NEJSG)	228	0.887	0.171	Yes	0.814	0.243	0.736	0.362
OPTIMAL	155	1.190	0.184	Yes	0.910	0.318	0.841	0.584
TORCH	39	1.580	0.415	Yes	1.701	0.436	2.090	0.605
WJTOG3405	118	1.185	0.222	Yes	2.237	0.475	4.412	0.876

There are differences in the number of events, both for PFS and OS between the average IPLD and the IPD. Where the timescale has been restricted it is pleasing to note that the average IPLD at the early timescale is, in all cases, less than was reported at the later date.

Table 6-6 shows the reported OS HR and SE for the latest available follow-up for 17 (16 treating INTACT 1 & 2 as one study), the IPLD HR for OS for the reconstructed data and adjusted OS estimates. In all cases the point estimate is affected by the treatment switching reanalysis, for some studies the change is fairly minor, but for others it is quite significant; FASTACT-2 for instance, the point estimate had an absolute decrease by 0.11. In all except one study, the SE for the adjusted estimate is substantially increased as is usual with an RPSFTM analysis, and SE calculated by preserving the p-value.

For the subsequent analyses, because of the change in follow-up, several different scenarios were considered:

1) using the reported PFS and latest OS follow-up (estimates typically used in practice);
2) using the reported PFS and earlier OS follow-up, where appropriate (more representative of the IPLD results, unadjusted for treatment switching);
3) using the reconstructed PFS and unadjusted OS estimates;
4) using the reconstructed PFS and adjusted OS estimates.

For the INTACT studies, FLEX and BMS 099 the estimates remain the same across all four scenarios. The estimates for the GTOWG study are the same for the first three scenarios; but the adjusted estimate for OS is used in the last scenario.

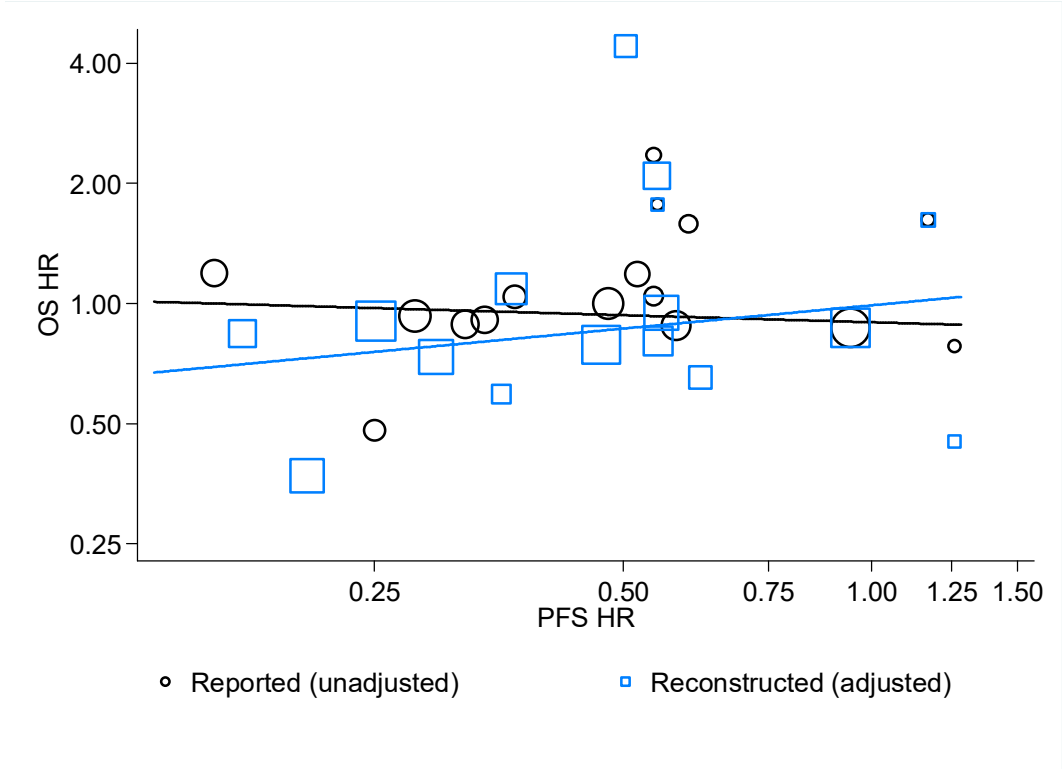
6.2.6.2.1 Meta-regression

To explore the relationship between PFS and OS using these studies, a meta-regression was conducted, the findings of which are given in Table 6-7 and Figures Figure 6-7 and Figure 6-8.

Table 6-7: Coefficients for the PFS log-HR from the meta-regressions

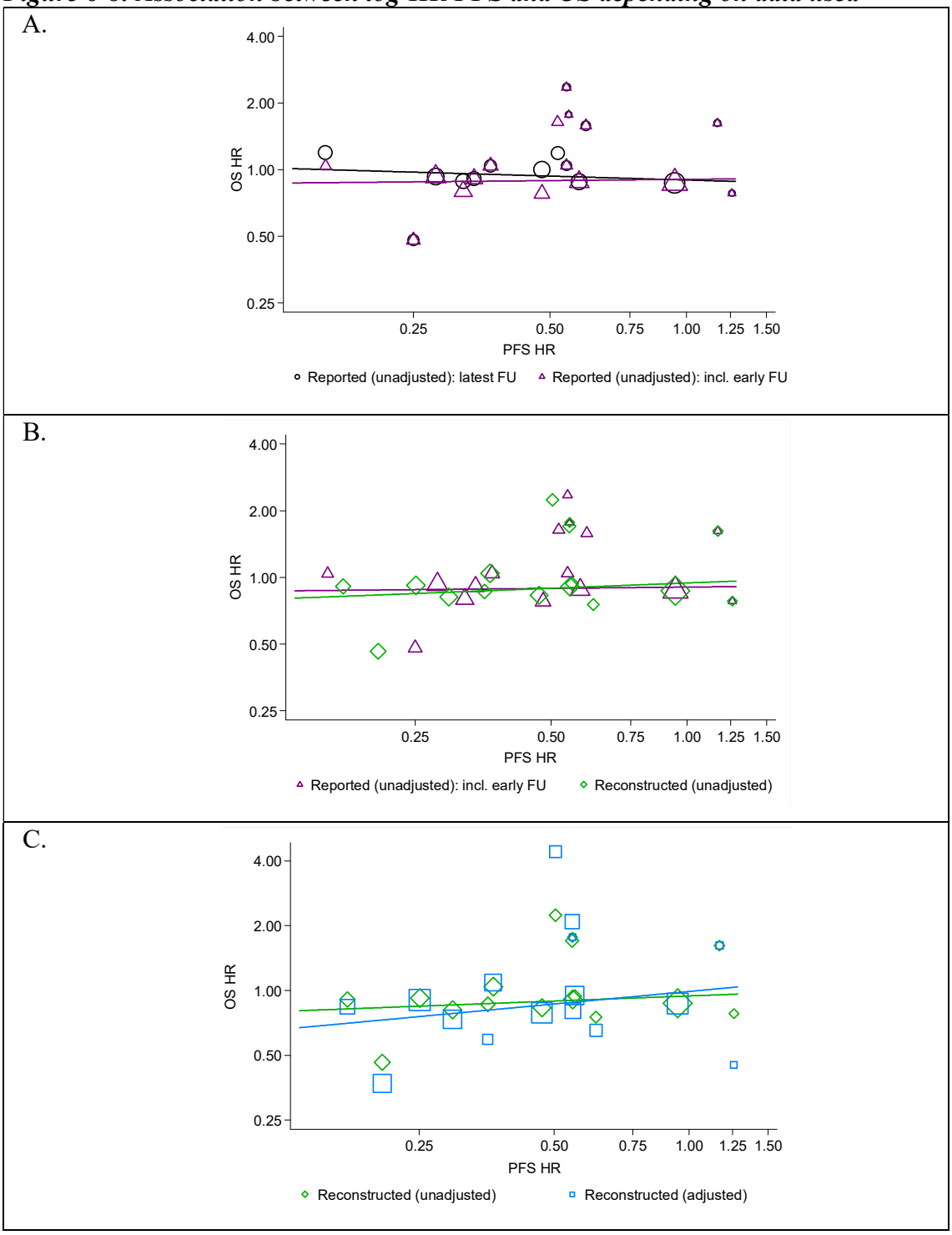
Scenario	Coefficient for PFS log-HR	(95% CI; p-value)
1. Reported PFS and latest OS estimates	-0.059	(-0.233, 0.115; p = 0.479)
2. Reported PFS and OS, with estimates for earlier OS FU used where appropriate	0.019	(-0.179, 0.217; p = 0.839)
3. Reconstructed PFS and unadjusted OS estimates	0.079	(-0.166, 0.324; p = 0.501)
4. Reconstructed PFS and adjusted OS estimates	0.194	(-0.172, 0.560; p = 0.274)

Figure 6-7: Association between log-HRs for PFS and OS depending adjustment



Contrasting scenarios 1 and 4, it is clear to see that there is a difference in the relationship between PFS and OS, having taken treatment switching into account. Whilst the graded approach (looking from scenario 1 to 4), highlights some of this could be due to restricting the timescale and looking at earlier follow-up, the difference between scenario 3 and 4 suggests that treatment switching does have an effect. A potential issue with these plots is that they assume PH for both PFS and OS. If the PH assumption is justified then length of follow-up is not important. However, with differential follow-up and different HRs, the agreement between HRs will be reduced.

Figure 6-8: Association between log-HR PFS and OS depending on data used



The association between PFS and OS, (A.) comparing the reported data with different follow-up lengths; (B.) comparing the reported data with shorter follow-up with the reconstructed (unadjusted) data; (C.) comparing the reconstructed data with and without adjustment.

6.2.6.2.2 Linear regression model, weighted by sample size

In contrast to the analysis given in 6.2.6.2.1, these models were fitted on the HR scale (as opposed to the log scale). Whilst, the log scale may seem a more reasonable choice, the analysis presented here follows that of Hotta (2013).

Scenario	R^2		
	Overall population	Crossover prohibited	Crossover permitted
1. Reported PFS and latest OS estimates	0.0196	0.3367	0.0185
2. Reported PFS and OS, with estimates for earlier OS FU used where appropriate	0.016		0.0468
3. Reconstructed PFS and unadjusted OS estimates	0.001		0.0787
4. Reconstructed PFS and adjusted OS estimates	0.000		0.0546

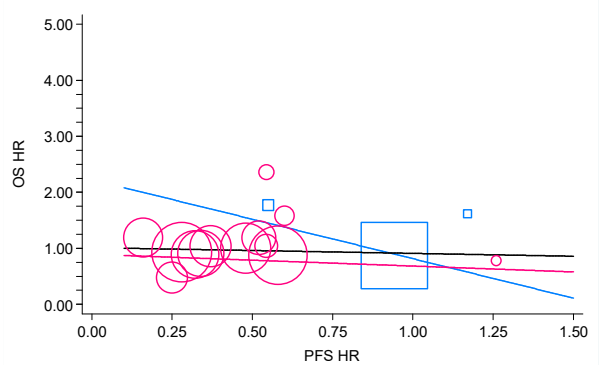
6.2.7 Implications from this case study

6.2.7.1 *Reconstructing data*

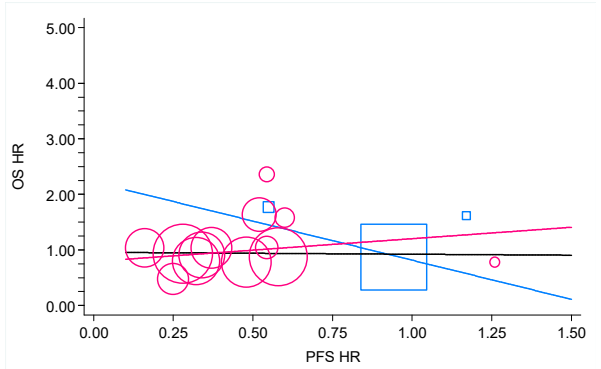
This analysis highlighted complications that the poor and inconsistent reporting of trials, investigating time-to-event endpoints, can cause for the method. This will also have had implications on the quality of the reconstructed data. Particularly for the CHEN study, which had a small sample size, the IPLD performed poorly. In the case of CHEN, some of this could be attributable to the lack of information on the EGFR +ve population, to calculate the HRs initially, the Guyot method was used. Based on the previous contrast between Guyot and the simulation approach, differences could, perhaps have been, expected. There were also other features that could have impacted on the quality; for example the fact that there was no 'risk table'. As described above, this resulted in the censorings being evenly distributed throughout the timescale, something which may not have been appropriate. Using a model for the simulation process may have contributed as well. In larger studies, the steps are smaller, leading to a smoother curve, which means that the model typically captures the shape well, and generates representative data. For this example, since the steps were so steep, the model has difficulty fitting well, and the data will no doubt vary quite considerably across datasets, and if compared to the original IPD.

Figure 6-9: Association between PFS and OS HR, overall and stratified by crossover

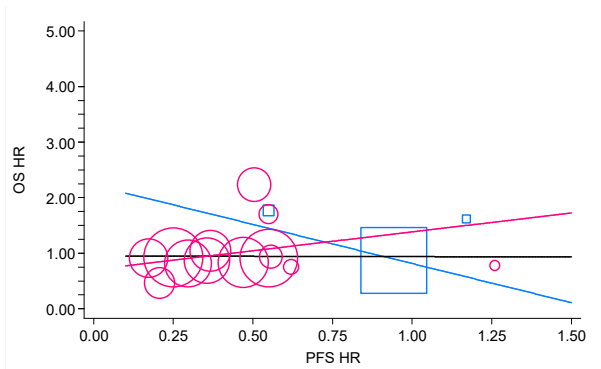
A.



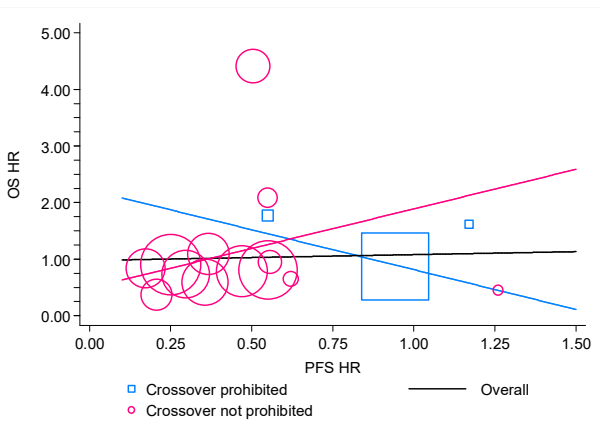
B.



C.



D.



The association between PFS and OS, both when all studies are included (overall) and when stratified by crossover for the 4 scenarios. (A.) 1. Reported PFS and latest OS estimates; (B.) 2. Reported PFS and OS, with estimates for earlier OS FU used where appropriate; (C.) Reconstructed PFS and unadjusted OS estimates (D.) 4. Reconstructed PFS and adjusted OS estimates

Restricting the timescale will also have had implications. The OS HR obtained at the later date is only truly comparable with the restricted analysis if PH are assumed; something which may or may not be valid. The findings from the IPLD would suggest that this could be the case, but trials which did report an earlier OS HR challenged this with estimates at earlier time points reporting a more effective / less harmful effect than the final one.

One study for which some concern with the reanalysis remains is the LUX-Lung 6; this is primarily given its similarity to the LUX-Lung 3 trial. A principle difference between the two LUX-Lung trials is the patient population; LUX-Lung 3 was a multinational trial whilst LUX-Lung 6 was conducted purely in an Asian population. Since the targeted therapies have often performed better in Asian populations, this could explain some of the variation, at least at PFS level, as the LUX-Lung 3 trial showed considerably less benefit for PFS (0.58 compared to 0.28, SE was similar in both groups). However, this difference was not reflected in the OS, and certainly not when treatment switching was taken account of (LUX-Lung 3 – 0.807; LUX-Lung 6 – 0.903).

The First-SIGNAL and OPTIMAL studies should perhaps be treated with some caution since both reported OS HRs slightly above one, and yet the reconstructed data estimated one below one. Since the direction of the adjustment does appear to depend on the ITT HR, (adjustment increasing numerically is OS HR greater than 1, and OS HR decreasing numerically if ITT HR is less than 1), this may have some impact on the validity of the findings. With OPTIMAL, however, it is noticeable that the HR has increased over time, and since at the early time point it is only just exceeding one, it is conceivable that since the timescale has been restricted even further this value could be appropriate. With the exception of LUX-Lung 6 and OPTIMAL, the other trials where concern has been expressed are all small studies, and as the subsequent analyses (meta-regressions) are weighted inversely based on sample size, these should not have too large an impact.

6.2.7.2 *Relationship with PFS and OS in a EGFR +ve NSCLC population*

Comparing the outcomes from the meta-regression using the reported PFS and latest possible OS results (scenario 1), and those obtained with the reconstructed data and adjusted OS (scenario 4) results shows a key difference in the direction of that relationship. Whilst in neither of the analyses, the findings were statistically significant,

in the first (scenario 1), there is a negative correlation between the two, with the OS log-HR decreasing as the PFS log-HR increases. In contrast, the final scenario (scenario 4) shows a stronger positive relationship, with the OS log-HR increasing as the PFS log-HR increases.

Looking at the two intermediate stages (scenarios 2 and 3) demonstrated that there may well be some effect due to the timescale, since, as the OS follow-up is restricted, so the positive relationship between PFS and OS log-HRs increased. This could be something to investigate further. It could, however, perhaps be argued that some of this could still be attributable to the impact of treatment switching, as the earlier the timescale is, the less time the effect may have to manifest itself; i.e. because patients tend to switch treatments earlier, this may have implications in different type of cancers, e.g. depending on the length of PPS.

Examining the results described in Section 6.2.6.2.2, it is pleasing to see some agreement with those from Hotta (2013). There is a similar trend when comparing the reported estimates (scenario 1), despite the R^2 values being considerably lower. In addition, the crossover prohibited studies show a much stronger association, observed overall, and the association between studies where treatment switching had not been prohibited was weaker still. Whilst the association between the studies which did not prohibit crossover improves when treatment switching is accounted for, the direction of the association directly contrasts with that for the studies which prohibited treatment switching. This, consequently, leads to the final results demonstrating an even weaker association between PFS and OS HRs.

Of the two analyses, the meta-regression provides greater and more robust evidence of the association between the two outcomes. A key reason for this is because of the weights used in the regression models. The choice of the SE is a better measure of a study's reliability as it takes into account the number of events in addition to the study's size, as it is important to include the number of events (a large study with few events will still contain a large amount of uncertainty which needs to be captured). In addition, using the SE means that the greater uncertainty surrounding the adjusted estimates is also taken account of, once treatment switching is adjusted for. This is not the case if the sample size is used, as the sample size remains unaffected by treatment switching.

6.3 Conclusions

This case study proved exceptionally enlightening. Up until now, the example (TAnDEM trial) had reported a good level of summary data (e.g. numbers of events, detailed risk table, same data cut-off date, etc.) making it relatively easy to reproduce the data. In these studies such a level of information was rare, and the reporting of basic statistics variable. Finding sufficient information to effectively reconstruct the data did prove challenging. Where possible, other information was used to improve quality, e.g. estimating censorings by counting the tick marks. These may have partially improved particular examples, but the approaches would not necessarily be feasible with larger studies. Consideration of whether this information could be used to further effect in a more systematic way would be an asset.

It was reassuring to see how effectively the amendments did work, and the similarity between the IPLD and IPD. Nevertheless, given that unlike other studies, hitherto used where additional IPD analyses (e.g. treatment switching analysis) were available for comparison, more exploration perhaps should be given to investigating the impact; either by gaining further examples where IPD are available, or by using simulated data. Nevertheless, the amendments will have had some effect. These analyses demonstrated areas for further development. Having the indications for censoring did indicate that in some situations the assumption of even censoring, even throughout an interval, might not be particularly appropriate. Developing an approach to use in these circumstances could prove useful. How to account for treatment switching when the number of treatment switchers exceeds the number of reported progressions, or if crossover occurs before documented progression is also an issue. This is essentially developing strategies for addressing treatment switching when it is not related to progression. In order to do this, additional information would no doubt be needed. Further steps would be to investigate how treatment switching would be introduced in these circumstances (e.g. for what reasons treatment switching would be permitted) and when in the timescale this could occur (e.g. anytime, after a particular date etc.), and consequently, if information was to become available, what would be useful. Once identified, this could then be developed further. Finally, for studies such as IPASS as well as the ‘immature’ OS, an updated analysis was available. It would be incredibly useful if essentially the ‘immature’ OS IPLD could be updated using this additional information. This may well be possible, but

would rely on an improved understanding of how all the different outcomes interact with each other and who would be ‘at risk’ of events. It could also be of interest to explore the alternative solutions for resolving the issues outlined in section 6.3.5.

Access to more detailed summary information proved useful. Having detailed lifetables permitted additional information to be gathered and highlighted potential issues with the censoring distribution, even though these did not appear to have an effect. One surprising finding was how difficult the CHEN study was to reproduce, largely due to the small sample size. Even from the start of the process, model fitting proved challenging, given the large steps in the Kaplan-Meier curve. The other issue was in replicating the ITT statistics. With such a small sample size, any results are heavily influenced by chance, and thus the estimates obtained from the reported statistics are exceptionally wide ranging. It might be worth therefore, determining whether for smaller sample sizes (e.g. less than 40 patients) there is a preferable way of reconstructing the data; for example, using an alternative method to the simulation approach, simulating from the survival times rather than a model etc. (similar to bootstrapping).

It is interesting to see how comparable the findings were with those from Hotta (2013). It further confirms that treatment switching can play a vital role in assessing surrogacy within advanced / metastatic cancer trials. In the wider context, it demonstrates how important it is to consider the impact of treatment switching in any secondary analysis involving OS. Chapter 3 investigated the impact of treatment switching in secondary analysis, where primarily only two studies are involved. Here, where there are many trials, and with the quantity of treatment switching involved, it is hardly surprising that there should be a noticeable difference when crossover is taken into account.

In summary, the two most key findings were:

1. the poor quality of reporting of trials;
2. that, as previously shown with Hotta, treatment switching can impact on the proving of surrogacy, but this can be partly addressed by reconstructing and adjusting summary data.

Chapter 7: Suggested reporting guidelines, summary and discussion

7.1 Chapter overview

Chapter 7 draws on the knowledge learnt through the review in Chapter 2 and the example in Chapter 6 to produce guidelines for reporting studies in treatment switching. In addition, it provides processes for analysts to follow to evaluate studies for treatment switching prior to performing secondary analyses. This chapter continues by summarising the entire project, highlighting the key findings and methodological developments. In particular it also discusses how these could be translated into practice, or affect current practice

7.2 Specific information needed to assess the impact of treatment switching on summary data and to adjust accordingly

7.2.1 Understanding the impact of treatment switching on secondary analysis

The findings from this thesis highlight how not accounting for treatment switching can have a key impact on the results of secondary analysis, such as ICs (sections 3.3.3, 3.3.4, 3.3.5, 3.4, 3.5) and in investigating surrogate endpoints (6.3.1, 6.3.6.2). Since the impact is not negligible, it is vital that reviewers, especially decision-makers are aware of the potential bias caused by treatment switching in secondary analysis for a particular example. This emphasises the importance of clear reporting about the inclusion of trials with treatment switching in any secondary analysis involving OS (or possibly even PFS). As found in Section 2.3, historically, the description of the studies included in the secondary analysis is varied, and does not necessarily state whether treatment switching was permitted. It is strongly advocated that, at the very least, the following details are reported:

- whether treatment switching occurred (this should include any indication of whether it is likely or unlikely if the publication / information used does not explicitly state if it was prohibited or permitted);
- the proportion of treatment switchers;
- whether any adjustment has been made to account for treatment switching.

This would ensure that the impact of treatment switching has been fully assessed, and that even if no reanalysis is undertaken to adequately adjust data with crossover, more assessment could be made of the impact of treatment switching, and the results approached with caution, if necessary.

Should the analyst choose to reconstruct and reanalyse the studies with unadjusted treatment switching (using methodology from Sections 4.3, 4.4 and 5.4.2), then a certain level of information is required (Section 5.3.2). To start with the analyst needs conviction that treatment switching occurred at, or soon after, disease progression, since the methods were only designed to address this reason for treatment switching. Given the development in Section 6.3, it may also be possible to address switching if this occurs after a secondary timepoint (e.g. interim analysis or unblinding). This is discussed further in Section 7.3.2.2

In addition to the treatment switching information, at minimum, the following must be available to enable the survival data to be reconstructed:

(1) either a K-M curve for OS and median PFS and OS times; or
(2) both PFS and OS K-M curves must be available.

Without these, a different approach would be required.

Of the two approaches, the second will produce more realistic data, but to improve the quality of the data, requires:

- PFS and OS data having been reported using the same data cut-off date,
- the number of TTP, PFS and OS events

Further improvement would be seen, if the analyst could access K-M curves for TTP, PPS, death before progression and censoring.

7.3 Proposed guidance

The development of effective guidance continues to be a priority, and remains on-going (Latimer, 2018b). There are several areas in which work has already commenced. The initial recommendations revolved around which methods should be used in practice (Latimer, 2013, Latimer, 2014, Latimer, 2018c). Further work has focused on how to check the appropriateness of the approach, and provide valid justification (Bell, 2014,

Bell, 2015. Watkins, 2016). In 2016, The Center for Medical Technology Policy (CTMP) published best practice guidance on conducting, analysing and reporting trials with treatment switching, which looked at the problem in a considerably broader context (Conley, 2016, The Green Park Collaborative, 2016). This guidance document essentially spans the whole life of a trial, from conception to final report, and features details such as how to assess necessity of including treatment switching when designing the trial to the choice of the final analysis method, and its reporting.

Sufficient planning and thought for treatment switching throughout the whole duration of the trial should improve outcomes and the choice / implementation of analysis methods. For example, it will ensure that relevant information is collected for the appropriate methodology or methodologies. Another key impact is likely to be that decision-makers will have more consistent access to detailed treatment switching data. However, it should be noted that the CTMP guidance has a strong emphasis on the reporting for documents intended for decision-makers, rather than general publications (The Green Park Collaborative, 2016).

Identifying original manufacturer's submissions, or evidence review group reports for Chapter 2 proved challenging, and so the majority of the detailed information was obtained came from publications, and similar challenges are likely to face manufacturers when identifying information for a competitor's product. The CTMP guidelines (The Green Park Collaborative, 2016) do not currently reiterate the need to include this high level of information in the journal article, or other publicly available document. Therefore, whilst this might be feasible for manufacturers to report, it seems unlikely that this information would be readily available, such as in publications, if practice continues as it is at present.

In addition, the guidance documents described above do not account for the variable reporting in survival analysis trials in general (Pelissier, 2008, Batson, 2016), nor do they provide any guidance on how to successfully review trials with treatment switching to determine whether sufficient evidence has been documented in advance of their inclusion in secondary analyses.

The the primary aim of the guidance presented in this thesis was to ensure that sufficient information would be available, so that IPLD could be reconstructed more easily and accurately. However, these suggestions have the capacity to also address some of the deficiencies identified in the reporting of studies with a 'time-to-event' outcome by Pelissier (2008) and build on the finding of Batson (2016). For example the guidance recommends detailed 'at risk' tables and the reporting of the number of events, both of which were found to have been reported inconsistently (Pelissier, 2008, Batson 2016).

7.3.1 Reporting of studies

This section divides into three parts:

(1) comments about the reporting of survival analysis studies in general;
(2) the reporting of treatment switching studies in general; and,
(3) the reporting of secondary analyses where studies with treatment switching are likely to occur.

The suggestions are separated into those which should always be reported, referred to as 'Good practice', and other extra information ('Additional Features'); in some cases this relates to more complex analysis, which although not necessarily compulsory, strengthens the evidence base and the reader / analysts understanding (and would improve the quality if reanalysis was needed). These suggestions primarily use the findings from Section 2.3 on the current state of reporting; and the information needed for effective data reconstruction and treatment switching reanalysis (sections 4.4, 4.5, 5.3.2 and 6.3.5). These recommendations are intended to be used alongside other reporting guidance such as the Consolidated Standards of Reporting Trials (CONSORT) checklist and the CTMP guidelines, not to replace it.

7.3.1.1 *Suggestions for improving the reporting of survival analysis studies*

Good practice

- Clearly specify the data cut-off date for PFS and OS, particularly if different within the same paper
- State the recruitment dates
- Include K-M curves for PFS and OS. If the data is immature for OS, this could be included in the supplementary appendix rather than the main text, so long as it is still publicly reported
- Report the number of TTP, PFS and OS events for each treatment group, alongside any important subgroup populations *(e.g. those where a K-M curve has been presented)*
- Produce a detailed 'at risk table', with at least five intervals where the 'number at risk' is greater than 1 and where the survival proportion in the first interval does not decrease by more than 30%
- In studies where crossover might be possible, clearly state whether crossover was permitted or not in the protocol and if it may have occurred during the post-protocol follow-up

Additional features (cont.)

- Alongside the 'at risk table' report the number of events that have occurred during a particular interval
- Produce K-M curves for any or all of the following outcomes: TTP, PPS, time to death before progression; time to censoring; for inclusion in the supplementary appendix
- Detail why patients were censored and whether any patients censored for PFS were subsequently followed up (and received an event) for OS

7.3.1.2 *Suggestions for improving the reporting of treatment switching studies*

Good practice

- Specify how treatment crossover occurred, clearly giving all reasons *(e.g. allowed at or after disease progression, permitted to all patients after interim analysis (or after a given date) etc.)*
- Report the number and / or proportion of patients who crossed over for each treatment arm, and include estimates for any key subgroups of patient populations *(e.g. those for which specific additional analyses have been given)*
- Document the number and / or proportions of patients who crossed over for each reason, if crossover could have occurred in more than one way *(e.g. 'x' patients switched after disease progression until interim analysis and 'y' patients switched before progression after this was allowed following interim analysis)*
- Specify whether any methods have been used to adjust for treatment switching, and which these were
- Include ITT K-M curves for PFS and OS

Additional features

- Include a K-M curve showing time from either start of study or time of progression, depending on which is more appropriate, until time of treatment switch
- Report median time (and / or other centiles) and range until treatment switch, either from the start of study or progression time, depending on which is most appropriate

7.3.1.3 *Recommendations for the reporting of secondary analyses involving OS in disease areas which may involve data with treatment switching*

Good practice	Review all included studies for treatment switching Report how many studies included treatment crossover State the proportion of patients crossing over between treatment arms Describe the methods used to address the treatment switching
Additional features	Report the secondary analysis using (1) ITT estimates and (2) crossover-adjusted* estimates **Crossover-adjusted: this either means the use of an appropriately justified recommended method on the IPD; or, where possible using the data reconstruction and reanalysis process (Detailed in Section 7.3.1.4)*

7.3.1.4 *Procedures to be followed when conducting secondary analyses involving OS in disease areas which may involve data with treatment switching*

7.3.1.4.1 Screening process and determining whether data can be reconstructed

As part of the usual screening processes (e.g. relevance, matching the inclusion criteria, bias assessment etc.) for trials, before their inclusion into a secondary analysis, treatment switching should also be taken into account. In terms of treatment switching, it is important to ascertain whether it occurred; if so, for what reasons it was permitted, which treatment groups were affected, how many patients switched; and what methodology, if any, was employed. A suggestion for an assessment tool is given in Table 7-1. This tool comprises of eight questions divided over four key areas (denoted Parts A – D). The questions and possible answers have been formed based on the findings in Section 2.3, and from the quality of reporting for the example used in Section 6.3. Part A focuses on determining whether treatment switching occurred. If treatment switching did not occur, the analyst does not need to proceed with any further questions. If crossover did or was likely to have occurred, then the subsequent sections should be completed. Part B aims to identify the reasons for why treatment switching occurred; this is a key element, should the data need to be reconstructed and

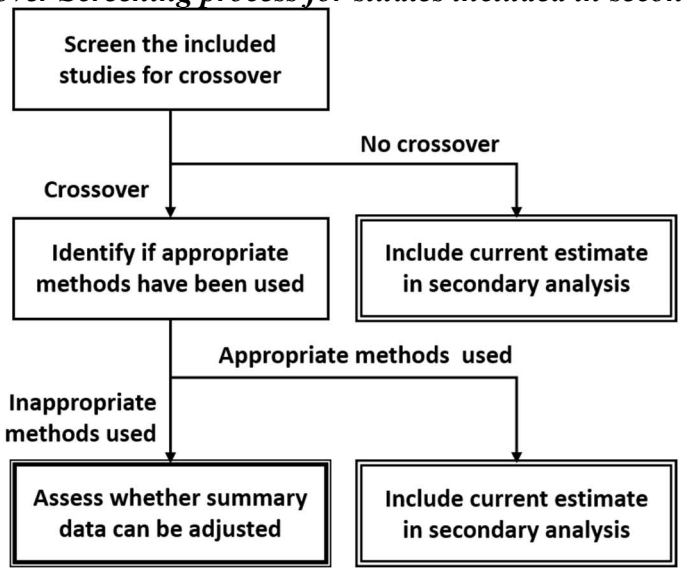
Table 7-1: 'Crossover' assessment tool

PART A: Likelihood of treatment 'crossover'	
1. The publication:	☐ States that 'crossover' was not permitted *(no further questions required)* ☐ Reports no information about 'crossover', making it impossible to determine if it was likely to occur *(no further questions required)* ☐ Reports no specific information about 'crossover', but trial design / information implies that 'crossover' was likely to have occurred ☐ Indicates that the post treatment phase contained trial interventions ☐ Specifies that 'crossover' did occur
PART B: Reasons for 'crossover'	
2. The reasons for crossover are: (tick all that apply)	☐ Not reported ☐ As 'protocol-specified' first choice for second- (or subsequent) line therapy ☐ As a possible option for therapy after disease progression [may be subject to other conditions, such as patient health] ☐ As 'superior' treatment on un-blinding ☐ For other reasons, *state these:*
PART C: Details of 'crossover'	
3. Due to 'crossover', the following treatment groups were affected:	☐ It is unclear how 'crossover' affected the treatment groups ☐ Control patients 'crossed over' to the experimental treatment; Experimental group patients did not switch ☐ Experimental group patients 'crossed over' to the control treatment; Control group patients did not switch ☐ Control patients 'crossed over' to the experimental treatment; Experimental group patients 'crossed over' to the control treatment;
4. The proportion of 'crossover':	☐ Is not reported ☐ Is not clearly reported (some details may be given, but the overall number or proportion is unclear) ☐ Is clearly reported
5. The following outcomes were affected by crossover:	☐ The impact of crossover on the outcomes was unclear ☐ OS only ☐ PFS and OS
PART D: Methodology employed to account for 'crossover'	
6. Was 'crossover' specifically addressed in the analysis?	☐ It is unclear how 'crossover' was addressed *(no further questions required)* ☐ No; no additional methods were used (underlying analysis was ITT) ☐ Yes; but the methods used were not those recommended in NICE TSD 16 (e.g. PP excluding or censoring patients) ☐ Yes; using methods recommended in NICE TSD 16 (e.g. RPSFTM, IPCW, two-stage)
7. Has sensitivity analysis been performed? E.g. several recommended methods applied	☐ No ☐ Unclear ☐ Yes
8. Justification for the methodology	☐ Was not given ☐ Was given, but was not valid (e.g. not based on reasonable arguments relating to data requirements and / or assumptions) ☐ Was given, and considered appropriate (e.g. based on reasonable arguments relating to data requirements and / or assumptions)

reanalysed, as current methods only cover certain situations. The objective of Part C, is to assess the magnitude of treatment switching, as well as obtaining estimates for any further reanalysis. Knowing how many patients switched can provide some measure of the enormity of the issue; a smaller proportion of treatment switcher is likely to have a more marginal effect, than if practically the entire control group had switched. It also aims to determine which outcomes and treatment groups are affected. The final part (Part D) investigates how treatment switching has been handled. After the identification of studies with treatment switching, the next stages are largely governed by the methodology used, and the analyst's propensity to perform additional analyses. To conclude whether the adjustment for crossover is really appropriate, even if recommended methods have been used the justification for the choice of approach is key. Where this is presented, it should be scrutinized. Nevertheless, it may be difficult to determine whether the justification is appropriate in practice, and so some level of trust is required. However, reasons such as the method being 'acceptable' in other examples should not be classed as sufficient and thus disregarded. Caution should also be given to those that say a method was chosen because it 'performed well'. Additionally, sensitivity analysis could prove useful in terms of checking the variability of the results due to method, and provide alternative estimates if data requirements / assumptions are deemed more appropriate by the reviewer or analyst. For examples, where there is considerable doubt as to a recommended methods suitability, this should be clearly documented.

Once the 'crossover' information has been collected, the flow-chart in Figure 7-1 shows the preliminary steps to be taken, depending on the amount of information available. As shown in the figure, if no treatment switching is identified, the estimates can be immediately included in the secondary analysis (subject to the other suitability checks, e.g. bias, appropriateness of trial population and methodology etc.). Similarly, if treatment switching did occur but has been adequately accounted for, then the 'adjusted' estimates can be included in the secondary analysis without any further work. The key issues arise if crossover is identified and the analyst considered the methods used to have been inadequate for addressing treatment switching.

Figure 7-1: Crossover Screening process for studies included in secondary analysis



At this stage, there are essentially two choices:

(1) to endeavour to conduct a reanalysis of the data, appropriately addressing treatment switching; or,

(2) to continue with the inappropriate estimates.

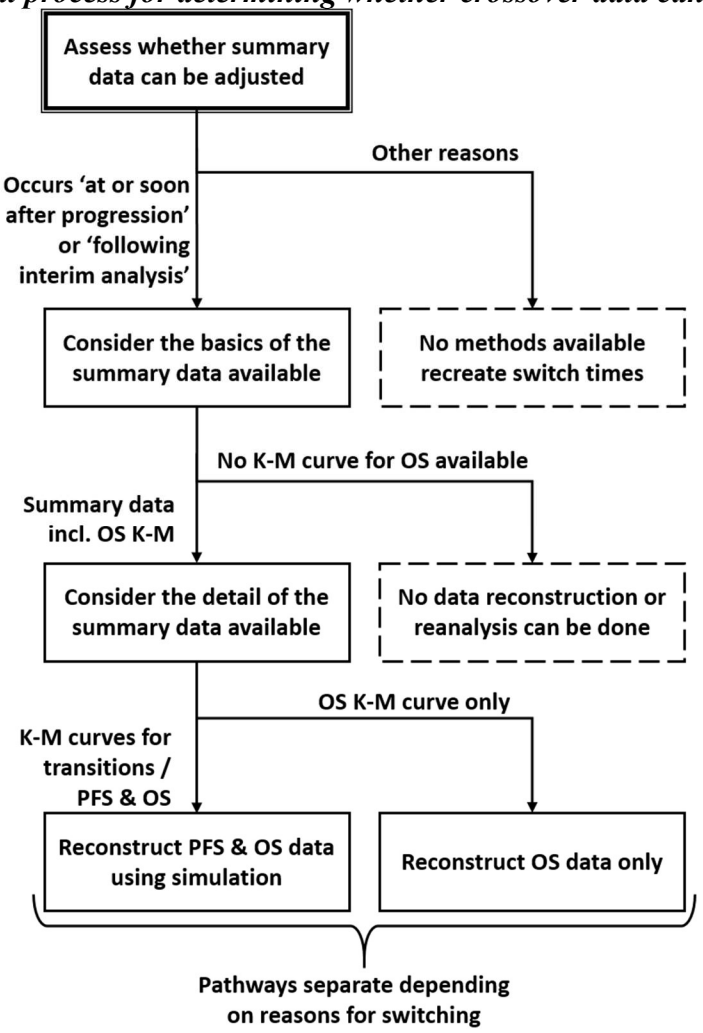
The first is advocated of these two options; however, if the analyst chooses to adopt the second option, they must, at the very least, report which studies crossover occurred in, and where possible the outcomes affected and the proportion of treatment switching. It is highly recommended conducting sensitivity analysis, in terms of including and excluding studies permitting treatment switching, also possibly a stratified analysis based on crossover, if using meta-analysis or a regression approach to test for surrogacy.

7.3.2 Procedures to be followed if data are to be reconstructed and reanalysed

Should the analyst be prepared to attempt the reanalysis, and thus reconstruction of the data, the next stages of the process are described in Figure 7-2. These stages assess whether approaches exist to reconstruct the treatment switching data for the particular switching mechanism used; and secondly, whether there is sufficient evidence to be able to reconstruct the data.

To-date the methods have been specifically developed to adjust for treatment switching only when it occurs around the time of disease progression. Based on some of the theory developed around recruitment times, designed to account for differential times

Figure 7-2: Initial process for determining whether crossover data can be reanalysed



for reporting outcomes, it would be possible to also adjust for patients switching at a particular time point, e.g. at or just after the date of interim analysis, or un-blinding of treatment regimen. To adjust for treatment switching, where the time of switch was not:

(1) at or soon after TTP (TTP as a proxy for switch time); or

(2) at or soon after a specific time that can calculated from recruitment information and a specific date (e.g. interim analysis, un-blinding),

further methods development would be required.

To produce the OS data, either of the approaches outlined in Chapter 4 and 5 can be used depending on the amount of information available. If K-M curves only exist for OS, then whilst this may limit and consequently impact on the reconstruction of the treatment switching information (imposing additional strong assumptions), OS data can be reconstructed. Whilst it is advocated to use the methodology in Chapter 4 to incorporate the additional uncertainty, the Guyot (2012) or Hoyle and Henley (2011) methods are alternatives. Wherever possible though, if enough information (e.g. K-M curves) matching any of the scenarios described in Chapter 5 is available, the corresponding approach should be used. Modifications described in Chapter 6 (Section 6.2.5) may also be necessary.

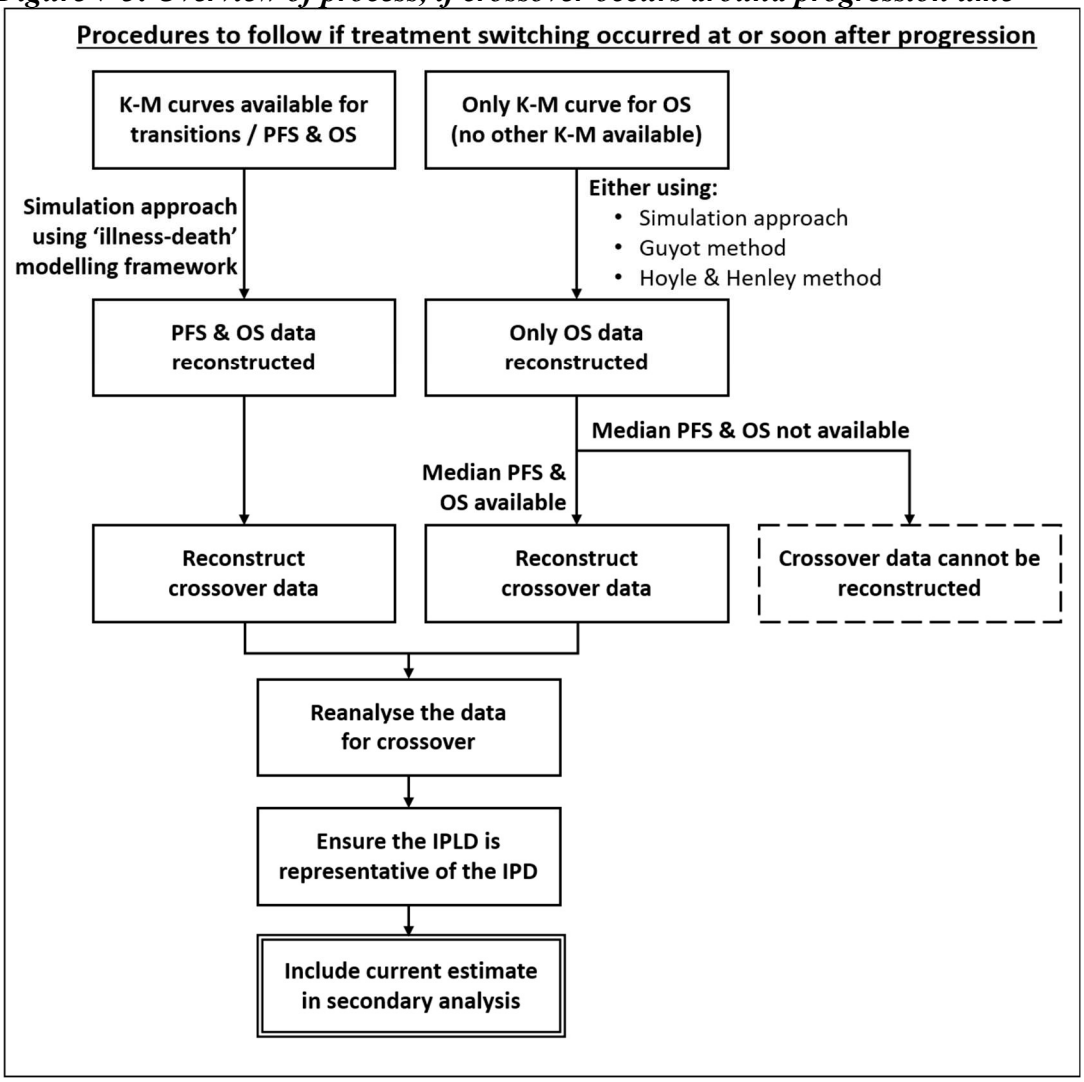
Once the data for OS (and PFS, if feasible) has been constructed, the information for treatment switching must be reconstructed. These differ slightly depending on the reasons for treatment switching and the information available, and the basic processes are described in Figure 7-3 and Figure 7-4.

Sensitivity analysis remains integral both during and after the data reconstruction. The models from which any data are simulated should be investigated, and the impact both of the number and location of knots examined. Where paired data have been reconstructed, understanding the effect of the individual parts of the process is key. It is strongly recommended that the investigations described in Section 5.4.5 are conducted, where appropriate, and especially if there are noticeable differences between the IPLD and IPD.

7.3.2.1 *Using TTP as a proxy for switch time*

Figure 7-3 shows an overview of the process, when the progression time can essentially be used as a proxy for switch time. Where the data has been reconstructed using an 'illness-death' modelling framework, reconstructing treatment switching follows the procedure outlined in Chapter 5. If the PFS K-M curve was not available, and only OS data were reconstructed, then to form the treatment switching information, it is essential that median survival times for PFS and OS are reported. Where these are available, then the treatment switching information is constructed as described in Sections 4.8.2.1 and 4.8.2.3.

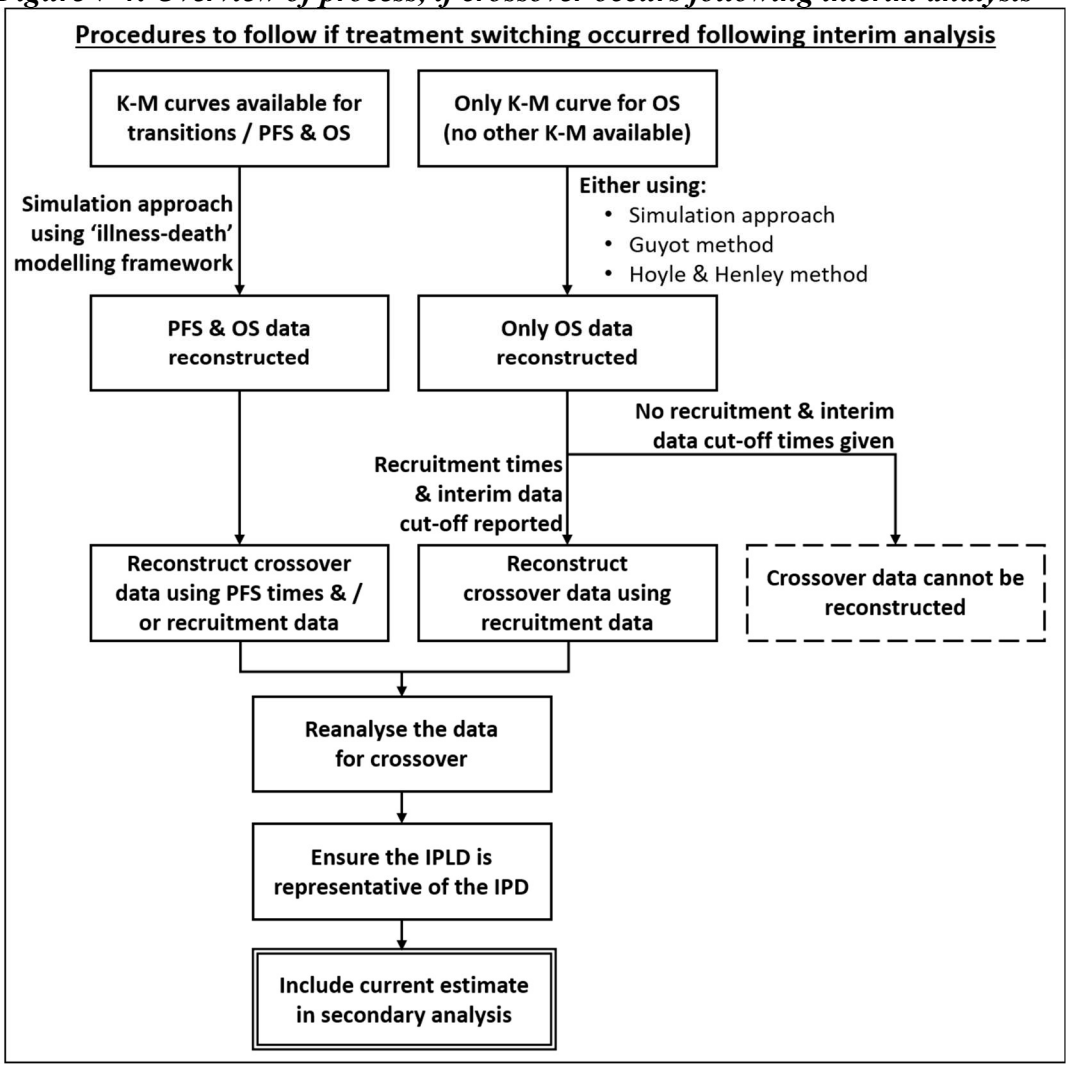
Figure 7-3: Overview of process, if crossover occurs around progression time



7.3.2.2 *Using a particular time point to calculate switch times*

If instead the treatment switching occurs at or soon after a particular time point (e.g. interim analysis as described in Figure 7-4 or the date of un-blinding, etc.), then it may still be possible to reconstruct the treatment switching information; provided that at least the start date of the trial and the date of this time point have been reported. Preferably the date of the end of recruitment would also be available. Where only OS data has been reconstructed, a similar approach to that for creating censoring times using the 'recruitment times' method (Section 4.4.1.4.2) would be adopted, but instead of using the final follow-up date, the time point (e.g. interim analysis data cut-off; date of un-blinding) would be used. Instead of producing a 'length of follow-up', these values would represent 'time to switch'; should this occur within a patient's survival time (e.g. this value occurs before their observed OS time (be it event or censoring)), then the patient could have

Figure 7-4: Overview of process, if crossover occurs following interim analysis



switched. In terms of assigning switchers, these would then be chosen randomly from those who 'could' have switched, based on the switch times.

If both PFS and OS data have been generated, then as specified in Section 6.2.5.3.5, the recruitment times can be constructed. Each patient will then have a suitable switch time calculated by deducting their individual simulated recruitment time, the time between the start of the study and permitting 'crossover' to take place. As before, patients where this switch time occurs before their final OS survival time, are eligible to switch. As usual the appropriate proportion can then be selected at random from the eligible population.

When all the data has been reconstructed, it can then be analysed, first for the ITT statistics to check it is reasonably representative of the IPD, and secondly for the treatment switching reanalysis. If using a simulation approach (either for a single outcome or for

paired data), this process will involve creating, analysing and averaging over many datasets (as described in Sections 4.4 and 5.4).

7.4 Summary

This work started by investigating the methodology used, in practice, to analyse data with treatment switching for NICE submissions. Findings showed that there has been a change in the approaches taken, which suggest that the promotion of appropriate methods is succeeding. Whilst there continues to be a proportion of TAs which choose to ignore the presence of treatment switching, those which do decide to acknowledge and address the problem, implement the recommended methods. In addition, there now appears a greater willingness to consider the use of and / or test all three approaches. Also, with regard to the final choice of method, sound reasons (such as choices based on data requirements or method assumptions) are now being used as justification. There is clear evidence that ICs and NMAs are becoming more prevalent in NICE TAs, which gives greater scope for the inclusion of biased estimates, caused by inappropriate analysis of treatment switching, in otherwise suitable TAs. The example, TA417 highlighted some of the difficulties in obtaining crossover-adjusted ICs and NMAs as the manufacturers are limited by previously conducted, publicly available analyses. No conclusive evidence was found that treatment switching or ICs / NMAs have a specific impact on the HTA recommendation.

The impact of using inappropriate (e.g. ITT) or appropriate (e.g. RPSFTM) analyses, when conducting a simple IC, was then assessed. Initial examples demonstrated that not only could the point estimate of the IC HR drastically change (depending on whether adjusted or unadjusted estimates were included), but potentially the statistical significance might too. Exploring this further through simulation ascertained that, in the majority of cases using an adjusted analysis for both studies with treatment switching showed a marked improvement in terms of the point estimate. The coverage, however, suggested that the SE was too great. There were a few cases where using both ITT estimates, or both RPSFTM estimates, performed equally well. These were when the crossover proportion for both studies were the same, as was the underlying treatment effect. Identifying this situation in practice would be very difficult as the true underlying treatment effect is unknown.

This early work highlighted two key findings:

(1) a large base of evidence exists in which treatment switching has not been accounted for appropriately;
(2) using adjusted estimates is the most reliable way of obtaining reasonable estimates from an IC when treatment switching data are involved.

This would mean that before inclusion into an IC, unadjusted estimates would have to be reanalysed using appropriate methods; a step which would require IPD in order to apply any of the recommended adjustment approaches. Given that the analysts rarely have access to IPD for the comparator interventions, the next stage was to develop methods that allowed treatment switching to be addressed using summary data. Previous research had highlighted the potential for reanalysing 'reconstructed data', but described limitations with the reconstruction methods. This led to the huge project of creating a reconstruction approach for time-to-event data which utilised simulation techniques, thus incorporating a sufficient amount of the uncertainty around the process. Initially the aim was to reconstruct a single outcome. The novel method relies on modelling the survival and censoring distributions, (primarily using extracted coordinates from the Kaplan-Meier curve, information on the numbers at risk or follow-up length), and produces multiple datasets which can be averaged over. Preliminary examples show that in terms of replicating reported statistics, conducting a simple reanalysis for non-proportional hazards and reproducibility, the proposed simulation technique is an excellent, possibly superior, alternative to other reconstruction methods (e.g. Guyot, 2012).

However, in order to account for treatment switching, a number of strong assumptions are necessary to reconstruct the crossover information. This, in turn, means that, even if the results seem plausible, the underlying switching mechanism is not appropriately modelled. To enable some of these strong assumptions to be relaxed or modified to become more realistic, PFS needed to be reconstructed alongside OS, and outcomes matched across patients.

To reconstruct paired data, such as PFS and OS, an 'illness-death' modelling structure should be adopted. To implement this effectively, summary data, such as Kaplan-Meier curves on the individual transitions and censoring are required. Having this means that the models for each transition and censoring can be fitted, and simulated from, to create the final dataset. As with the single outcome technique, multiple datasets are simulated

and the results averaged over. This high level of information is almost always unavailable, and consequently this original method must be replicated by estimating the relevant functions from available data on other outcomes (e.g. PFS, OS). The less specific information there is, the greater the margin for error, and the more variation between the reported results and any from the reconstructed data. The treatment switching mechanism which accompanies this more advanced simulation technique is, in its simplest form, easy to implement, but captures the underlying mechanism much more effectively.

The inclusion of data with treatment switching not only affects ICs, but also impacts on other secondary analysis. Hotta (2013) highlighted the issue of treatment switching affecting surrogacy in NSCLC. A Cochrane review (Greenhalgh, 2016) was used to illustrate the impact of adjusting adequately for treatment switching, before assessing the relationship between PFS and OS. This review (Greenhalgh, 2016) comprised of nineteen studies, of which fifteen studies, permitted treatment switching, with at least 30%, and more often 65% or more of patients switching from the control intervention (of Chemotherapy) to an experimental treatment (Targeted therapy). The adjustment showed considerable impact on the point estimate for the meta-regression coefficient for the log-HR for PFS. This highlighted that adjusting the summary data can have a substantial difference on a secondary analysis. However, the example also demonstrated a number of difficulties in implementing the method because of the nature of the reported information. In many cases, suitable solutions were developed to address these problems, although these were often bespoke modifications, designed to contend with the available information and not always generalisable.

7.5 Limitations, Discussion and Context

The findings from the review proved very interesting, since in terms of the final recommendation, no noticeable difference was observed between studies which permitted treatment switching and those that did not. One possible explanation for this is that, perhaps, although the final outcome does not differ majorly, the length of the appraisal might. It is conceivable that, in order to reach a conclusion, the HTA bodies request additional information from studies which would not be necessary for studies without treatment switching. This could, thus, lengthen the appraisal process but not affect the recommendation. Therefore, in terms of updating or conducting further reviews, one

additional part could be to investigate the length of the decision process (from submission to recommendation).

The change over time in the methods used to analyse data with treatment switching is especially pleasing, as there is clearly a trend with more TAs exploring the use of different recommended methods, and far fewer using the most inappropriate or experimental 'ad hoc' approaches. This demonstrates the success, that particularly Latimer (2014) has had, in promoting the most appropriate methodology. In addition, it highlights a 'learning curve effect'; initially as the number of methods reduced there was a compulsion to use the RPSFTM, as this had 'been previously accepted as appropriate' in crossover TAs, and perhaps as this was a longer-standing method specifically designed for treatment switching. This showed an understanding of which approaches should not be used, and that the RPSFTM was a more appropriate model, but demonstrated a key lack of knowledge of the modelling assumptions, and thus appropriateness of the various recommended methods. Based on the increase in the use of the other methods, the variety of methodology being documented per appraisal (where the analysts have identified and aimed to address treatment switching), and the type of justification being given, it is clear that this deficiency is being dealt with.

One area which the review was not fully able to determine was whether the application of appropriate methodology affects the recommendation. These findings appear to be in accordance with those of Gurskyte (2018). The principle difficulty faced is likely to be that of insufficient evidence, since there are still relatively few appraisals which have used recommended methods. This is most likely due to the research on suitable methods still being quite recent; given that RCTs are usually three to five years duration, very few trials would have been able to specify the recommended methods at the protocol stage, those that may have, may be now coming to completion. Views about using methods that were not specified in the protocol can be quite divergent; some maintaining that the protocol must be followed exactly with no deviation; others advocating the use of different methods if it means conducting a more appropriate analysis. Should the review be updated in five or ten years' time, there would no doubt be a greater number of examples using the recommended methods, and the results less likely to be affected by chance. However, that analysis may face similar difficulties for the comparator. To maintain equipoise in decision making, it is vital that the timescale is restricted to the period in which clear

findings had been given on appropriate methodology. If, as is indicated, the situation continues to improve, then more and more trials with treatment switching should be relying on recommended methods rather than an ITT (or even less appropriate e.g. PP) approach. This could mean, however, that there is an insufficient number of TAs using an ITT analysis to compare decisions across; which though positive in many ways, would hamper this assessment.

A vital aspect of this research was the simulation study conducted in Chapter 3. This assessment of the effect of incorporating unreliable summary data within a quite simple form of secondary analysis is most enlightening. It confirms the existing supposition that, in general, using appropriately adjusted estimates is key to obtaining most appropriate results for an IC. As always with a simulation study, its main advantage is that of having the true underlying values, which means that the appropriateness of the treatment switching analysis could also be assessed alongside that of the IC; something not possible with a case-study.

Whilst many different scenarios were covered, this was by no means exhaustive. This study concentrated on the proportion of switchers and the treatment effect. However, study sample size could also have a bearing on the findings; in addition, there could potentially be more than one trial per comparator (e.g. two trials for treatment 'X' compared with 'A', one trial for 'X' compared with 'B'). Using more complex methods of simulating the underlying data might also have been beneficial, as would simulating data which may suit other methods better and applying a wider variety of methods. Understanding the implications of these could be very important, both for analysts and for decision-makers. Gaining a better understanding of the underlying situation could also be useful, alongside guidance on suitable treatment switching approaches, in allaying uncertainty regarding the appropriateness of adjustment methods in ICs (Ishak, 2018).

Here, only simple ICs have been considered, but in practice, and based on the review it is clear to see that NMA / MTCs are gaining popularity. The data structure for these would be much more complex. NMA has sometimes been suggested as a solution for studies with treatment switching (Thorland, 2013), as strength could be borrowed from other studies which prohibited treatment switching, and used to reduce the bias caused by treatment switching. However, this will introduce more variability and potentially

inconsistency. This, in turn, could lead to additional complexity, as to account for this, methods which address heterogeneity and inconsistency in NMA (Jansen and Cope, 2012) would be required. Whilst this might provide a solution where there is sufficient evidence to conduct a NMA, this approach would not necessarily be generalisable to other secondary analysis e.g. ICs with two studies or meta-analysis. Therefore, as adjusting for the treatment switching, by reconstructing IPLD, would rectify the unreliability of the estimates, this may still be preferable.

Prior to this work, Hotta (2013) had already demonstrated, through a case-study, that treatment switching could have a bearing on the proving of surrogacy. These conclusions were reiterated in the work of Hernadez-Villafuerte (2018), where proof of surrogacy in other cancers such as melanoma and RCC was deemed to have been affected by crossover. These papers concentrated on demonstrating that treatment switching causes issues. They did not suggest any solutions to rectify the issue, except to potentially stratify the analysis by whether treatment switching occurred or not. However, this suggestion means that the wealth of information is not being retained. Moreover, there could be the potential to introduce bias if the studies which permit treatment switching are fundamentally different to the others. For example, if only those studies which saw a considerable PFS benefit permitted treatment crossover. Therefore, analyses which included estimates unadjusted for treatment switching, and those which stratified by crossover, could provide some evidence, but would not be firm conclusions on which to prove PFS as a surrogate outcome for OS.

Treatment switching was a key issue for the Hotta study (Hotta, 2013) as there was a sizeable proportion (15 / 35) of studies with crossover. Nonetheless, supposing that, in fact, the EGFR +ve population, as described in the Cochrane review (Greenhalgh, 2016), had been the population of interest, the problem would have been fundamental. Excluding the studies with treatment switching would be an incredibly severe approach, as this would have just left three / four studies, of which only one was of any size ($n > 40$).

The outcome of readjusting the simulation approach is a very crucial finding. Whilst this is not conclusive, the surrogacy analysis is rather naïve, and a more robust approach is required to validate PFS as a surrogate in the EGFR +ve NSCLC population, it provides valuable evidence that, readjusting summary data for treatment switching could address

the crossover issue sufficiently to give appropriate and valid results. This could fulfil a fundamental deficiency in the current methodology and evidence base for determining surrogacy, as there will be further examples, in other types of cancer, where the proof of surrogacy will be affected by treatment switching. This is particularly likely in cancers where patients typically have short progression times, and where a number of studies have allowed treatment switching on progression. Furthermore, since it worked so effectively for the surrogacy example, it provides support that the reanalysis of IPLD could also extend over to address other issues, such as meta-analysis of OS.

The reconstruction methodology is designed to be used when access to the IPD are unavailable. Whilst it often gives reasonable agreement to the IPD, it will always be an inferior substitute. Wherever possible, IPD should be obtained, and this methodology only used as a secondary alternative. However, Simmonds (2005) found that, for systematic reviews and meta-analysis, it is usually only possible to obtain IPD for 50% of the studies. Given the drive towards data sharing, and thus permitting access to IPD, this percentage should increase, which theoretically could minimise the use of the reconstruction methodology in its original context. Nevertheless, the time from the application for IPD access to approval and the release of the IPD (if approved) is often lengthy (e.g. in ensuring all legal requirements are met with relation to the General Data Protection Regulation (2018) and Data Protection Act, (2018)) and the whole process time consuming. Therefore, these reconstruction techniques could be applied in the meantime to provide a preliminary estimate, which would then be updated once the analyst has access to the IPD. This means that although the methods may not be required for the primary analysis, they could still prove useful. In addition, there will always be an abundance of historical trials for which these methods may be needed, as access to the IPD will still be prohibited, or where the IPD may no longer actually exist. In the meantime, at least, the methods developed and described in Chapters 4 and 5 may allow for a level of compromise. Whilst manufacturers may not be prepared to disclose the IPD, or rerun new additional analysis, they may be prepared to release additional basic summary information, if necessary, such as the number of events or K-M curves for a particular outcome, or even life-tables. These are very simple analysis, that could easily be produced. Taking previously published and / or additional information, and using this within the methods could solve the issue; the manufacturer conducting the secondary

analysis would have more appropriate valid results, and its competitor will have maintained confidentiality.

The simulation method has shown itself to be an excellent and exceptionally flexible alternative to other data reconstruction processes, but it should be noted that no method of data reconstruction is ever going to be perfect. Whilst the simulation approaches do account for some of the uncertainty surrounding the data reconstruction process (unlike other existing methods), they still do not fully capture all the uncertainty. Therefore, sufficient sensitivity analysis is required to assess the robustness of the models and the findings (as detailed in Sections 4.7 and 5.4.6). For example, the uncertainty around the model parameters is completely ignored. This is perhaps a deficiency that should be understood, but it should not impede or violate the findings from it in any way (Bennett, 2018). The fact that it does account for at least some of the uncertainty still distinguishes it from its alternatives. One criticism that the simulation process has received was that all simulated datasets are included in the final averaging. Naturally as with simulation some datasets may be closer to the original dataset than others (findings that are perhaps closer to the extreme limits of the distribution). Therefore, restricting to a subset of datasets which most closely resemble the original was recommended, to improve comparability of results. This would lead to an analyst determining suitable criteria for the restriction, which could be very subjective. It would also impact on the calculation of the SE. Alternatively, the datasets could be weighted by their comparability, but similarly this would impact on the SE, and require the weights to be defined. This has meant that at present, no action has been taken to implement this suggestion, however, it could be something that is examined long term.

The key limitation, relating to regular implementation of the data reconstruction methodology, is its complexity. The simulation approach for the single outcome (Chapter 4) remains relatively straightforward to understand; whilst that for the paired data is exceptionally intricate and complicated. It draws on methodology from numerous areas of survival analysis and manipulates them, often using them in unusual contexts (e.g. allowing censoring to be a competing risk). Programming into frequently used software, such as creating a Stata or R package, would no doubt remove some of the difficulties in using the method. However, in conjunction with developing this, some possible issues may still need to be resolved, mostly relating to the model fit. As has been explained

throughout the thesis, whilst in theory, the coordinates should be such that turning points in the survival (not the hazard) distribution are prohibited, occasionally these occur. These can often be overcome by manipulating the data (e.g. extracting additional coordinates in particular places), but this type of approach would not be suitable for a novice using the software. Therefore, investigating the use of further splines functions, which are restricted to producing monotonic models, would be beneficial.

There can be no doubt that the data are affected by the type and quality of the summary information. Clearly, the paired data approach using the summary data on the transitions and censoring provides a better level of agreement and realism. Whilst these are not necessarily outcomes that are considered as meaningful, in these days of online appendices, they could be exceptionally useful to report; especially those relating to TTP and PPS when treatment switching has occurred. As already mentioned, the quality of the reporting also impacts substantially on the work. One particular area of concern is the differential timing in reporting outcomes. This was exceedingly common, and although resolved here by restricting the timescale, could lead to criticism such as was discussed in the example given in Chapter 2. By restricting the timescale, reliance is being put on immature data, which may be seen by decision-makers as being inappropriate and / or unreliable, due to it potentially being highly influenced by chance. Whilst the guidance given at the start of the chapter should improve reporting in the future, it cannot undo previous poor reporting. In order to tackle this issue further, additional examples could prove beneficial, in terms of revealing any other potential problem. These could also provide greater validation of the approach.

At its current stage, unfortunately this methodology cannot encompass all the potential treatment switching scenarios or methodology. To date, this approach does not consider situations where patients do not switch at progression (e.g. at interim analysis). So far this work has not applied other approaches recommended for treatment switching. At present, although it does not permit a method such as the IPCW to be applied, since this would require covariates, the two-stage model may be a possibility. Given the necessity that the chosen approach does actually fulfil the method's assumptions and data requirements, improving the range of possible models would be key. Whilst an application of the two-stage model with PFS time as the covariate is a feasible option, it is unlikely that the

covariates needed to fit the IPCW could ever be reconstructed; and so the implications of this should be investigated further.

Despite its many limitations, these simulation approaches facilitate reanalysis for treatment switching where IPD are unavailable, fulfilling a fundamental gap in the treatment switching research. A few years ago, there were no possibilities for addressing treatment switching when the IPD were unavailable. This would have led to the exclusion of valuable evidence or the use of flawed findings. Whilst this work cannot fully address every kind of treatment switching, and is wholly reliant on 'good' reporting of trials, it provides a solid foundation for the reanalysis of summary data with crossover and an indispensable platform for further research. Nevertheless, the use of sensitivity analysis subsequent to using these approaches is crucial in determining the appropriateness and robustness of the findings.

7.6 Further work

The potential extensions of this project fall into three main themes, that have been covered throughout this thesis. These are:

- Understanding the impact of treatment switching in secondary analysis
- The simulation technique
- Addressing the impact of treatment switching

7.6.1 Understanding the impact of treatment switching in secondary analysis

The simulation studies in Chapter 3 provided a rationale for ensuring that estimates for studies with treatment switching are appropriately analysed before their inclusion in an IC. As discussed in Section 7.5, there are a variety of extensions that could be done to improve understanding even further. However, as that section also suggested, a more important area of continued research would be to explore the impact on NMAs, given their increasing use. Developing a good knowledge of the effect of the estimates could be especially important, given how NMAs include direct and indirect evidence. Treatment switching has the potential to create a vast amount of sensitivity in NMA, as it may lead to direct and indirect evidence being inconsistent, depending on where crossover has been permitted. In addition, there would be far more room for error, since more studies would be involved, and possibly lead to a high number containing treatment switching.

More importantly it might be even more necessary with NMAs to carefully consider how to address crossover, for example if two-way switching has been allowed. This work has approached the treatment switching problem in the same way as other similar research projects; by only considering and adjusting for treatment switching in the control arm, regardless of whether crossover also occurred in the experimental arm. This decision has been fundamentally decided upon as typically switching from the experimental to control arm does not affect the decision-making problem. However, if treatment switching has occurred in the experimental arm, there may be some additional bias still within the estimate, even if the control arm has been adjusted. One concern would be that this may affect the consistency of the direct and indirect evidence, which could also be particularly problematic if looking at a trial where potentially two quite novel and possibly very effective treatments have been compared. In this case, it could be difficult to ascertain which of the treatments would essentially be the 'control' treatment.

7.6.2 The simulation process

The simulation process has been developed to such an extent that it is a usable and relatively reliable method, however, there are some ways in which it could be improved further. As previously described in Section 7.5, to facilitate the uptake of this methodology, producing programs / packages for commonly used computer software (e.g. Stata, R etc.) would be key. Equally, as imparted before, this would potentially involve investigating and ultimately, using splines functions that ensure monotonic functions. Given that these splines functions are already available in an R package, this may be possible to achieve in R.

There are other important areas of development which relate to model fit, for example, some of the models require even more flexibility than is possible, using RCS models. These are particularly those that demonstrate a 'cure' shape (Othus, 2012), such as in example 1 (Figure 4-4). Building this into the method would improve model fit and generalisability. Another potential change would be to model outcomes such as TTP, PFS and OS altogether using multivariate survival models. Since these outcomes are interrelated, it would be excellent if this was continued into the modelling. It would also, hopefully, ensure that all quantities were consistent with each other, and avoid any errors caused by the modelling process (e.g. models for PFS survival proportion which exceed the TTP or OS survival proportion).

Developing the censoring distribution more is an area which could be exploited further. The simulation framework allows for a greater flexibility in the creation of the censoring distribution compared with other methods. However, this flexibility may not be used to its maximum potential. The case study demonstrated that often the location of censorings is indicated on the K-M curve. It would be useful if this could be utilised further to help construct more realistic censoring distributions. This may be improved by better reporting, (e.g. detailing the number of events per interval). In addition, given that interval censoring is exceptionally common, particularly for PFS, it would be beneficial to adapt the methods to include this.

The final area where more research is needed is in addressing the issues discussed in Chapter 6; in particular, how to address modelling of small studies, lack of information in terms of events or numbers at risk, and differential timing in the reporting of outcomes. Here, primarily the simplest technique was employed, but that is not necessarily to say that the approach was the best. This is particularly with regard to

(1) using the shorter term follow-up (for at least two studies); and,
(2) restricting the time period to account for differential timing of reporting outcomes.

In the first case, the aim would to be to update the simulated OS information, based on the later term follow-up. For the second, this would involve investigating alternative approaches to restricting the timescale. In these cases, the current approach taken means that important evidence is being disregarded. In addition, given the HTA panel's reluctance and criticism of using the short-term follow-up in the example highlighted in Section 2.2.1.3.4, the concern would be that these results would be treated with the same response if simulated data and reanalysis used a restricted timescale.

7.6.3 Addressing treatment switching

There are two key ways in which the simulation technique could be used to address treatment switching. The first is the same as it has been used to address treatment switching in this thesis; by reconstructing data for crossover reanalysis when only summary data are available. To improve this further, additional work must concentrate on widening the types of treatment switching that can be adjusted for. If treatment switching for all control group patients has been permitted following interim analysis, providing the date is given, then combining the usual approach, with that used to restrict

the timescale could be developed to produce the necessary information. However, there may be other situations, which could be harder to reconstruct. As suggested in the discussion in Chapter 6, reasons for treatment switching, and the relevant information could be explored further. Even developing the current methodology for reconstructing the treatment switching information could be an area for improvement.

In improving the use of reconstructing IPLD for the accounting of treatment switching, the implementation of the alternative recommended approaches to the RPSFTM should be examined. This is of key importance, given the strong assumptions that the RPSFTM rely on, and that these may not be reasonable in the context in which they are being used. The two-stage method would be most easily transferable into this setting, as initially it could be fitted just using the PFS time as the only covariate. This could be justified as it is one of the most, if not the most informative variable as to why a patient switched, if patients switch on progression. The IPCW would be considerably harder to implement, as it requires covariate data, which at present are not reconstructed during the process described in this thesis. The feasibility of reconstructing this data would be integral to allowing this method to be explored further in this setting, and the inability to do this would prohibit the use of the IPCW on reconstructed data.

A completely different application for this simulation method would be in trying to adjust for treatment switching at the IPD level. The use of external data to account for treatment switching has been suggested; this would require the identification of a study (external data) which prohibited treatment switching, and was sufficiently similar to the trial which allowed crossover. This external data could then be used to predict counterfactual times for the patients who switched treatments. An example of this approach was implemented, using a relatively naive framework, in TA171 (NICE, 2009b). For this example, the PPS model was calibrated such that the median OS in the control group after progression was equal to that which had been observed in the UK Medical Research Myeloma trials (NICE, 2009b). Whilst a more formal approach would need to be developed than that used in TA171, the simulation method could provide a useful tool during this approach. The external data to be used has always been required at IPD level. Having the simulation approach would mean that the analyst would not necessarily require IPD level information. They would just need to identify a similar trial, and either fully or possibly partially reconstruct the data.

7.6.4 Additional areas

A more general area of further work could be to investigate the use of external information for strengthening results. Section 7.6.3 briefly discussed how the simulation method could be used to enable external information to be combined with trial data, where treatment switching occurred. Whilst developing a robust framework for using external data could be beneficial for adjusting for treatment switching, there are other contexts in which external data could be useful, e.g. in validating extrapolation results. In addition, to the other contexts, there are also different ways in which a framework could be developed. For example, by using a 'multiple imputation' (Carpenter and Kenward, 2014) style approach or a Bayesian predictive model.

Another area of interest could be in examining how increasing levels of summary data may affect the methods. This superior information, for example, could involve having a life-table at all distinct event / censoring times for each of the transitions. With more detailed summary information, it could be worthwhile investigating if there were any simpler methods, which would be as efficacious as the paired simulation approach at producing results comparable to the IPD.

7.7 Conclusions

Previous research into appropriate methodology for data with treatment switching is being promoted successfully, and a change in the type of methods used in practice was observed. However, a gap in the research was detected; this was, given the wealth of studies with treatment switching that have inappropriately analysed, the inclusion of biased estimates within secondary analysis.

Preliminary exploration, through simulation studies, demonstrated that even with a simple secondary analysis such as an IC, using ITT estimates, and particularly using a mixture of ITT and adjusted estimates leads to exceptionally biased and unreliable findings. The only really reliable way of accounting for treatment switching, was to have an 'adjusted' estimate for each study, which was then used in the IC. This, nevertheless, meant that summary data must be reanalysed for treatment switching.

Complex data reconstruction methods were developed, which used a simulation approach. Where PFS and OS K-M curves were both reported, paired data could be reconstructed, allowing a more realistic treatment switching mechanism to be implemented. These methods were influenced by the amount of information available and the quality of the reporting. Once the data had been reconstructed, existing treatment switching techniques, namely the RPSFTM could be applied.
The methods were applied to studies used in a Cochrane review for EGFR +ve NSCLC treatments (Greenhalgh, 2016). The aim, however, was to investigate PFS as a surrogate endpoint for OS. This case study highlighted many issues caused by poor reporting, and developed solutions for them. Comparing a meta-regression for the log-HR for PFS against log-HR for OS, showed a difference in the gradient depending on whether the treatment switching was adjusted (in the reconstructed IPLD) or not (using the reported estimates).

Whilst there remain a variety of areas for further research with this methodology, the approaches can be used to account for treatment switching in summary data, fulfilling the original void in research. The results produced, have typically demonstrated good agreement with the ITT summary statistics, and in one example where a RPSFTM analysis was available. This gives more confidence to the results from any secondary analysis (e.g. treatment switching reanalysis).

Appendices

Appendix A: Technology Appraisals included in the review

A-1: List of Technology Appraisals included in the review

List of the Technology Appraisals included in the review alongside the year of publications and detailing the type of cancer.

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA3	Ovarian cancer - taxanes (TA3) (replaced by TA55) (withdrawn)	2000	Ovarian
TA6	Breast cancer - taxanes (TA6) (replaced by TA30) (withdrawn)	2000	Breast
TA17	Colorectal cancer - laparoscopic surgery (TA17) (replaced by TA105)	2000	Colorectal
TA23	Brain cancer - temozolomide (TA23)	2001	Brain
TA25	Pancreatic cancer - gemcitabine (TA25)	2001	Pancreatic
TA26	Lung cancer - docetaxel, paclitaxel, gemcitabine and vinorelbine (TA26) (replaced by CG24)	2001	Lung
TA28	Ovarian cancer - topotecan (TA28) (replaced by TA91) (withdrawn)	2001	Ovarian
TA29	Leukaemia (lymphocytic) - fludarabine (TA29)	2001	Leukaemia
TA30	Breast cancer - taxanes (review) (TA30) (replaced by CG81)	2001	Breast
TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (TA33) (replaced by TA93) (withdrawn)	2002	Colorectal
TA34	Breast cancer - trastuzumab (TA34)	2002	Breast
TA37	Lymphoma (follicular non-Hodgkin's) - rituximab (replaced by TA137) (TA37) (replaced by TA137) (withdrawn)	2002	Lymphoma
TA45	Ovarian cancer (advanced) - pegylated liposomal doxorubicin hydrochloride (TA45) (replaced by TA91) (withdrawn)	2002	Ovarian
TA50	Leukaemia (chronic myeloid) - imatinib (TA50) (replaced by TA70) (withdrawn)	2002	Leukaemia
TA54	Breast cancer - vinorelbine (TA54) (replaced by CG81)	2002	Breast
TA55	Ovarian cancer - paclitaxel (review) (TA55)	2003	Ovarian
TA61	Colorectal cancer - capecitabine and tegafur uracil (TA61)	2003	Colorectal
TA62	Breast cancer - capecitabine (TA62) (replaced by CG81)	2003	Breast

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA65	Non-Hodgkin's lymphoma - rituximab (TA65)	2003	Lymphoma
TA70	Leukaemia (chronic myeloid) - imatinib (TA70) (partially updated by TA241 and TA251)	2003	Leukaemia
TA86	Gastrointestinal stromal tumours - imatinib (TA86)	2004	Gastrointestinal
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review) (TA91)	2005	Ovarian
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (TA93) (replaced by CG131)	2005	Colorectal
TA101	Prostate cancer (hormone-refractory) - docetaxel (TA101)	2006	Prostate
TA116	Breast cancer - gemcitabine (TA116)	2007	Breast
TA118	Colorectal cancer (metastatic) - bevacizumab and cetuximab (TA118) (partially updated by TA242)	2007	Colorectal
TA119	Leukaemia (lymphocytic) - fludarabine (TA119)	2007	Leukaemia
TA121	Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide (TA121)	2007	Glioma
TA124	Lung cancer (non-small-cell) - pemetrexed (TA124)	2007	Lung
TA129	Multiple myeloma - bortezomib (TA129)	2007	Myeloma
TA135	Mesothelioma - pemetrexed disodium (TA135)	2008	Lung
TA137	Lymphoma (follicular non-Hodgkin's) - rituximab (TA137)	2008	Lymphoma
TA145	Head and neck cancer - cetuximab (TA145)	2008	Head & Neck
TA147	Breast cancer (advanced & metastatic) - bevacizumab (withdrawn) (TA147)	2008	Breast
TA162	Lung cancer (non-small-cell) - erlotinib (TA162)	2008	Lung
TA169	Renal cell carcinoma - sunitinib (TA169)	2009	Renal cell
TA171	Multiple myeloma - lenalidomide (TA171)	2009	Myeloma
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab (TA172)	2009	Head & Neck
TA174	Leukaemia (chronic lymphocytic, first line) - rituximab (TA174)	2009	Leukaemia

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA176	Colorectal cancer (first line) - cetuximab (TA176)	2009	Colorectal
TA178	Renal cell carcinoma (TA178)	2009	Renal Cell
TA179	Gastrointestinal stromal tumours - sunitinib (TA179)	2009	Gastrointestinal
TA181	Lung cancer (non-small-cell, first line treatment) - pemetrexed (TA181)	2009	Lung
TA183	Cervical cancer (recurrent) - topotecan (TA183)	2009	Cervical
TA184	Lung cancer (small-cell) - topotecan (TA184)	2009	Lung
TA185	Soft tissue sarcoma - trabectedin (TA185)	2010	Soft tissue sarcoma
TA189	Hepatocellular carcinoma (advanced and metastatic) - sorafenib (first line) (TA189)	2010	Liver
TA191	Gastric cancer (advanced) - capecitabine (TA191)	2010	Gastric
TA192	Lung cancer (non-small-cell, first line) - gefitinib (TA192)	2010	Lung
TA193	Leukaemia (chronic lymphocytic, relapsed) - rituximab (TA193)	2010	Leukaemia
TA202	Chronic lymphocytic leukaemia - ofatumumab (TA202)	2010	Leukaemia
TA208	Gastric cancer (HER2-positive metastatic) - trastuzumab (TA208)	2010	Gastric
TA209	Gastrointestinal stromal tumours (unresectable/metastatic) - imatinib (TA209)	2010	Gastrointestinal
TA212	Colorectal cancer (metastatic) - bevacizumab (TA212)	2010	Colorectal
TA214	Breast cancer - bevacizumab (in combination with a taxane) (TA214)	2011	Breast
TA215	Renal cell carcinoma (first line metastatic) - pazopanib (TA215)	2011	Renal cell
TA216	Leukaemia (lymphocytic) - bendamustine (TA216)	2011	Leukaemia
TA219	Everolimus for the second-line treatment of advanced renal cell carcinoma (TA219)	2011	Renal cell
TA222	Ovarian cancer (relapsed) - trabectedin (TA222)	2011	Ovarian
TA226	Lymphoma (follicular non-Hodgkin's) - rituximab (TA226)	2011	Lymphoma
TA227	Lung cancer (non-small-cell, advanced or metastatic maintenance treatment) - erlotinib (monotherapy) (TA227)	2011	Lung

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA228	Multiple myeloma (first line) - bortezomib and thalidomide (TA228)	2011	Myeloma
TA235	Osteosarcoma - mifamurtide (TA235)	2011	Osteosarcoma
TA239	Breast cancer (metastatic) - fulvestrant (TA239)	2011	Breast
TA241	Leukaemia (chronic myeloid) - dasatinib, nilotinib, imatinib (intolerant, resistant) (TA241)	2012	Leukaemia
TA242	Colorectal cancer (metastatic) 2nd line - cetuximab, bevacizumab and panitumumab (review) (TA242)	2012	Colorectal
TA243	Follicular lymphoma - rituximab (review) (TA243)	2012	Lymphoma
TA250	Breast cancer (advanced) - eribulin (TA250)	2012	Breast
TA251	Leukaemia (chronic myeloid, first line) - dasatinib, nilotinib and standard-dose imatinib (TA251)	2012	Leukaemia
TA255	Prostate cancer - cabazitaxel (TA255)	2012	Prostate
TA257	Breast cancer (metastatic hormone-receptor) - lapatinib and trastuzumab (with aromatase inhibitor) (TA257)	2012	Breast
TA258	Lung cancer (non small cell, EGFR-TK mutation positive) - erlotinib (1st line) (TA258)	2012	Lung
TA259	Prostate cancer (metastatic, castration resistant) - abiraterone (following cytoxic therapy) (TA259)	2012	Prostate
TA263	Bevacizumab in combination with capecitabine for the first-line treatment of metastatic breast cancer (TA263)	2012	Breast
TA265	Bone metastases from solid tumours - denosumab (TA265)	2012	Bone metastases
TA268	Melanoma (stage III or IV) - ipilimumab (TA268)	2012	Melanoma
TA269	Melanoma (BRAF V600 mutation positive, unresectable metastatic) - vemurafenib (TA269)	2012	Melanoma
TA272	Urothelial tract carcinoma (transitional cell, advanced, metastatic) - vinflunine (TA272)	2013	Urothelial tract
TA284	Bevacizumab in combination with paclitaxel and carboplatin for first-line treatment of advanced ovarian cancer (TA284)	2013	Ovarian

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA285	Ovarian, fallopian tube and primary peritoneal cancer (recurrent advanced, platinum-sensitive or partially platinum-sensitive) - bevacizumab (TA285)	2013	Ovarian, Fallopian tube and peritoneal
TA295	Breast cancer (HER2 negative, oestrogen receptor positive, locally advanced or metastatic) - everolimus (with an aromatase inhibitor) (TA295)	2013	Breast
TA296	Lung cancer (non-small-cell, anaplastic lymphoma kinase fusion gene, previously treated) - crizotinib (TA296)	2013	Lung
TA299	Leukaemia (chronic myeloid) - bosutinib (TA299)	2013	Leukaemia
TA306	Lymphoma (non Hodgkin's, relapsed, refractory) - pixantrone monotherapy (TA306)	2014	Lymphoma
TA307	Colorectal cancer (metastatic) - aflibercept (TA307)	2014	Colorectal
TA309	Lung cancer (non small cell, non squamous) - pemetrexed (TA309)	2014	Lung
TA310	Lung cancer (non small cell, EGFR mutation positive) - afatinib (TA310)	2014	Lung
TA311	Multiple myeloma - bortezomib (induction therapy) (TA311)	2014	Myeloma
TA316	Enzalutamide for metastatic hormone-relapsed prostate cancer previously treated with a docetaxel-containing regimen	2014	Prostate
TA319	Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma	2014	Melanoma
TA321	Dabrafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma	2014	Melanoma
TA326	Imatinib for the adjuvant treatment of gastrointestinal stromal tumours	2014	Gastrointestinal
TA333	Axitinib for treating advanced renal cell carcinoma after failure of prior systemic treatment	2015	Renal cell
TA338	Pomalidomide for relapsed and refractory multiple myeloma previously treated with lenalidomide and bortezomib	2015	Myeloma
TA343	Obinutuzumab in combination with chlorambucil for untreated chronic lymphocytic leukaemia	2015	Leukaemia

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA344	Ofatumumab in combination with chlorambucil or bendamustine for untreated chronic lymphocytic leukaemia	2015	Leukaemia
TA347	Nintedanib for previously treated locally advanced, metastatic, or locally recurrent non-small-cell lung cancer	2015	Lung
TA357	Pembrolizumab for treating advanced melanoma after disease progression with ipilimumab	2015	Melanoma
TA359	Idelalisib for treating chronic lymphocytic leukaemia	2015	Leukaemia
TA360	Paclitaxel as albumin-bound nanoparticles in combination with gemcitabine for previously untreated metastatic pancreatic cancer	2015	Pancreatic
TA366	Pembrolizumab for advanced melanoma not previously treated with ipilimumab	2015	Melanoma
TA370	Bortezomib for previously untreated mantle cell lymphoma	2015	Lymphoma
TA371	Trastuzumab emtansine for treating HER2-positive, unresectable locally advanced or metastatic breast cancer after treatment with trastuzumab and a taxane	2015	Breast
TA374	Erlotinib and gefitinib for treating non-small-cell lung cancer that has progressed after prior chemotherapy	2015	Lung
TA376	Radium-223 dichloride for treating hormone-relapsed prostate cancer with bone metastases	2016	Prostate
TA377	Enzalutamide for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated	2016	Prostate
TA378	Ramucirumab for treating advanced gastric cancer or gastro–oesophageal junction adenocarcinoma previously treated with chemotherapy	2016	Gastric or gastro-oesophageal
TA380	Panobinostat for treating multiple myeloma after at least 2 previous treatments	2016	Myeloma
TA381	Olaparib for maintenance treatment of relapsed, platinum-sensitive, BRCA mutation-positive ovarian, fallopian tube and peritoneal cancer after response to second-line or subsequent platinum-based chemotherapy	2016	Overian, fallopian tube and peritoneal
TA384	Nivolumab for treating advanced (unresectable or metastatic) melanoma	2016	Melanoma

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA389	Topotecan, pegylated liposomal doxorubicin hydrochloride, paclitaxel, trabectedin and gemcitabine for treating recurrent ovarian cancer	2016	Overian
TA395	Ceritinib for previously treated anaplastic lymphoma kinase positive non-small-cell lung cancer	2016	Lung
TA396	Trametinib in combination with dabrafenib for treating unresectable or metastatic melanoma	2016	Melanoma
TA399	Azacitidine for treating acute myeloid leukaemia with more than 30% bone marrow blasts	2016	Leukaemia
TA400	Nivolumab in combination with ipilimumab for treating advanced melanoma	2016	Melanoma
TA401	Bosutinib for previously treated chronic myeloid leukaemia	2016	Leukaemia
TA402	Pemetrexed maintenance treatment for non-squamous non-small-cell lung cancer after pemetrexed and cisplatin	2016	Lung
TA403	Ramucirumab for previously treated locally advanced or metastatic non-small-cell lung cancer	2016	Lung
TA404	Degarelix for treating advanced hormone-dependent prostate cancer	2016	Prostate
TA405	Trifluridine–tipiracil for previously treated metastatic colorectal cancer	2016	Colorectal
TA391	Cabazitaxel for hormone-relapsed metastatic prostate cancer treated with docetaxel	2016	Prostate
TA408	Pegaspargase for treating acute lymphoblastic leukaemia	2016	Leukaemia
TA410	Talimogene laherparepvec for treating unresectable metastatic melanoma	2016	Melanoma
TA411	Necitumumab for untreated advanced or metastatic squamous non-small-cell lung cancer	2016	Lung
TA412	Radium-223 dichloride for treating hormone-relapsed prostate cancer with bone metastases	2016	Prostate
TA414	Cobimetinib in combination with vemurafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma	2016	Melanoma
TA416	Osimertinib for treating locally advanced or metastatic EGFR T790M mutation-positive non-small-cell lung cancer	2016	Lung

Reference	Technology Appraisal Title	Year of publication	Type of cancer
TA417	Nivolumab for previously treated advanced renal cell carcinoma	2016	Renal cell
TA421	Everolimus with exemestane for treating advanced breast cancer after endocrine therapy	2016	Breast
TA422	Crizotinib for previously treated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer	2016	Lung
TA423	Eribulin for treating locally advanced or metastatic breast cancer after 2 or more chemotherapy regimens	2016	Breast
TA424	Pertuzumab for the neoadjuvant treatment of HER2-positive breast cancer	2016	Breast
TA425	Dasatinib, nilotinib and high-dose imatinib for treating imatinib-resistant or intolerant chronic myeloid leukaemia	2016	Leukaemia
TA426	Dasatinib, nilotinib and imatinib for untreated chronic myeloid leukaemia	2016	Leukaemia

A-2: List of TAs with treatment switching included in the review

A list of the TAs included in the review, in which treatment switching was clearly identified as having occurred in the pivotal evidence.

Reference	Technology Appraisal Title
TA3	Ovarian cancer - taxanes (TA3) (replaced by TA55) (withdrawn)
TA6	Breast cancer - taxanes (TA6) (replaced by TA30) (withdrawn)
TA28	Ovarian cancer - topotecan (TA28) (replaced by TA91) (withdrawn)
TA30	Breast cancer - taxanes (review) (TA30) (replaced by CG81)
TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (TA33) (replaced by TA93) (withdrawn)
TA34	Breast cancer - trastuzumab (TA34)
TA55	Ovarian cancer - paclitaxel (review) (TA55)
TA70	Leukaemia (chronic myeloid) - imatinib (TA70) (partially updated by TA241 and TA251)
TA86	Gastrointestinal stromal tumours - imatinib (TA86)
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review) (TA91)
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (TA93) (replaced by CG131)
TA101	Prostate cancer (hormone-refractory) - docetaxel (TA101)
TA116	Breast cancer - gemcitabine (TA116)

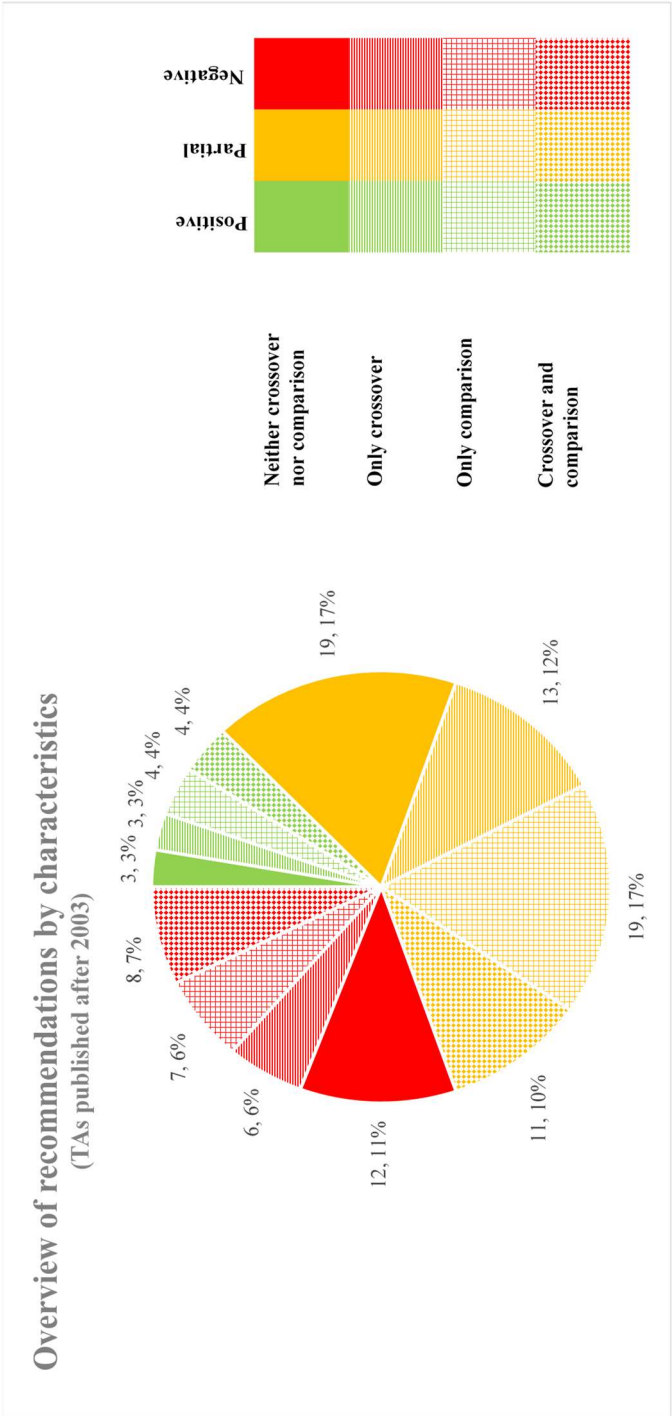
Reference	Technology Appraisal Title
TA118	Colorectal cancer (metastatic) - bevacizumab and cetuximab (TA118) (partially updated by TA242)
TA119	Leukaemia (lymphocytic) - fludarabine (TA119)
TA121	Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide (TA121)
TA124	Lung cancer (non-small-cell) - pemetrexed (TA124)
TA129	Multiple myeloma - bortezomib (TA129)
TA162	Lung cancer (non-small-cell) - erlotinib (TA162)
TA169	Renal cell carcinoma - sunitinib (TA169)
TA171	Multiple myeloma - lenalidomide (TA171)
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab (TA172)
TA176	Colorectal cancer (first line) - cetuximab (TA176)
TA178	Renal cell carcinoma (TA178)
TA179	Gastrointestinal stromal tumours - sunitinib (TA179)
TA192	Lung cancer (non-small-cell, first line) - gefitinib (TA192)
TA214	Breast cancer - bevacizumab (in combination with a taxane) (TA214)
TA215	Renal cell carcinoma (first line metastatic) - pazopanib (TA215)
TA219	Everolimus for the second-line treatment of advanced renal cell carcinoma (TA219)
TA257	Breast cancer (metastatic hormone-receptor) - lapatinib and trastuzumab (with aromatase inhibitor) (TA257)
TA258	Lung cancer (non small cell, EGFR-TK mutation positive) - erlotinib (1st line) (TA258)
TA263	Bevacizumab in combination with capecitabine for the first-line treatment of metastatic breast cancer (TA263)
TA268	Melanoma (stage III or IV) - ipilimumab (TA268)
TA269	Melanoma (BRAF V600 mutation positive, unresectable metastatic) - vemurafenib (TA269)
TA284	Bevacizumab in combination with paclitaxel and carboplatin for first-line treatment of advanced ovarian cancer (TA284)
TA285	Ovarian, fallopian tube and primary peritoneal cancer (recurrent advanced, platinum-sensitive or partially platinum-sensitive) - bevacizumab (TA285)
TA319	Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma
TA321	Dabrafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma
TA326	Imatinib for the adjuvant treatment of gastrointestinal stromal tumours
TA338	Pomalidomide for relapsed and refractory multiple myeloma previously treated with lenalidomide and bortezomib
TA357	Pembrolizumab for treating advanced melanoma after disease progression with ipilimumab
TA359	Idelalisib for treating chronic lymphocytic leukaemia

Reference	Technology Appraisal Title
TA371	Trastuzumab emtansine for treating HER2-positive, unresectable locally advanced or metastatic breast cancer after treatment with trastuzumab and a taxane
TA377	Enzalutamide for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated
TA389	Topotecan, pegylated liposomal doxorubicin hydrochloride, paclitaxel, trabectedin and gemcitabine for treating recurrent ovarian cancer
TA396	Trametinib in combination with dabrafenib for treating unresectable or metastatic melanoma
TA404	Degarelix for treating advanced hormone-dependent prostate cancer
TA422	Crizotinib for previously treated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer
TA425	Dasatinib, nilotinib and high-dose imatinib for treating imatinib-resistant or intolerant chronic myeloid leukaemia
TA426	Dasatinib, nilotinib and imatinib for untreated chronic myeloid leukaemia

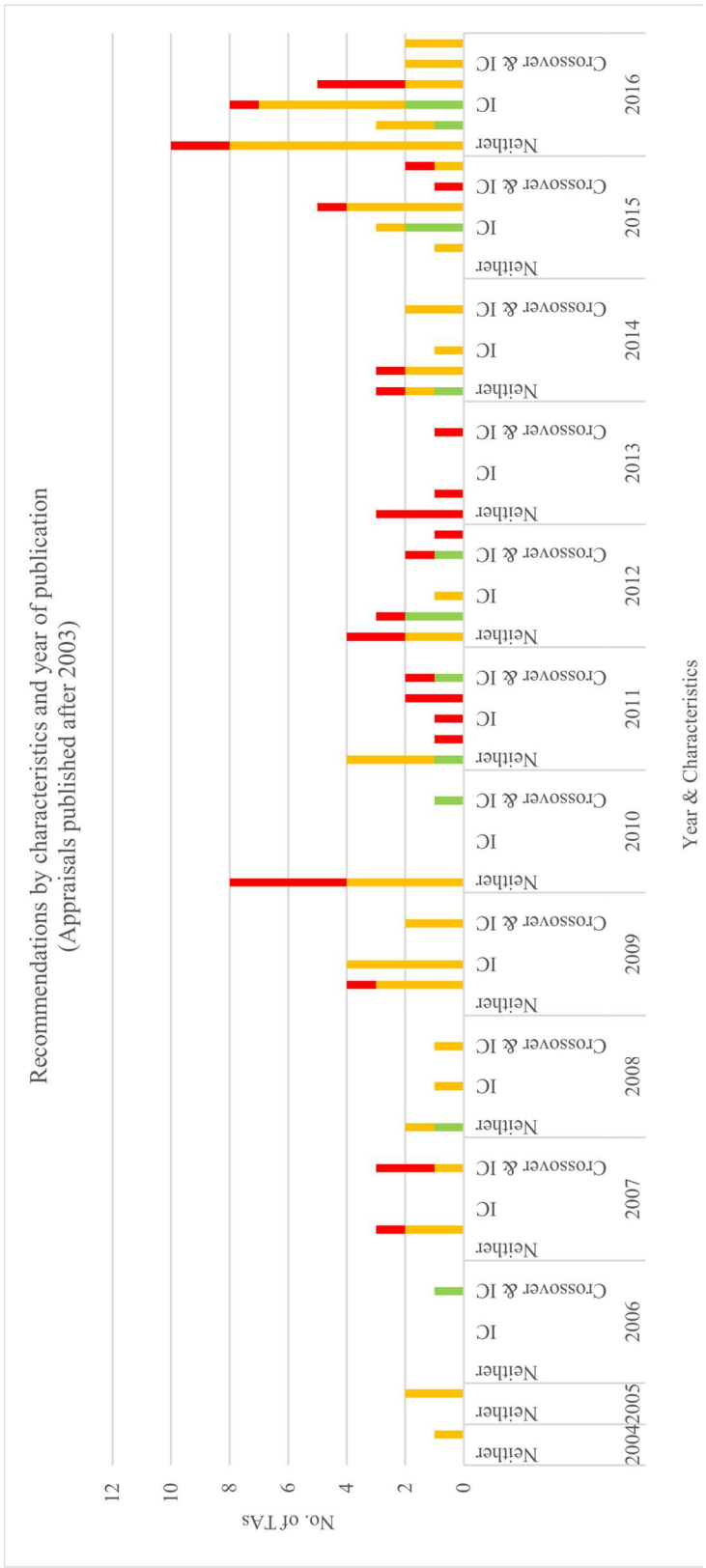
Appendix B: Replication of analysis, restricting the time scale to TAs published after 2003

In the following pie charts, the value next to each segment show the number of TAs, followed by the percentage.

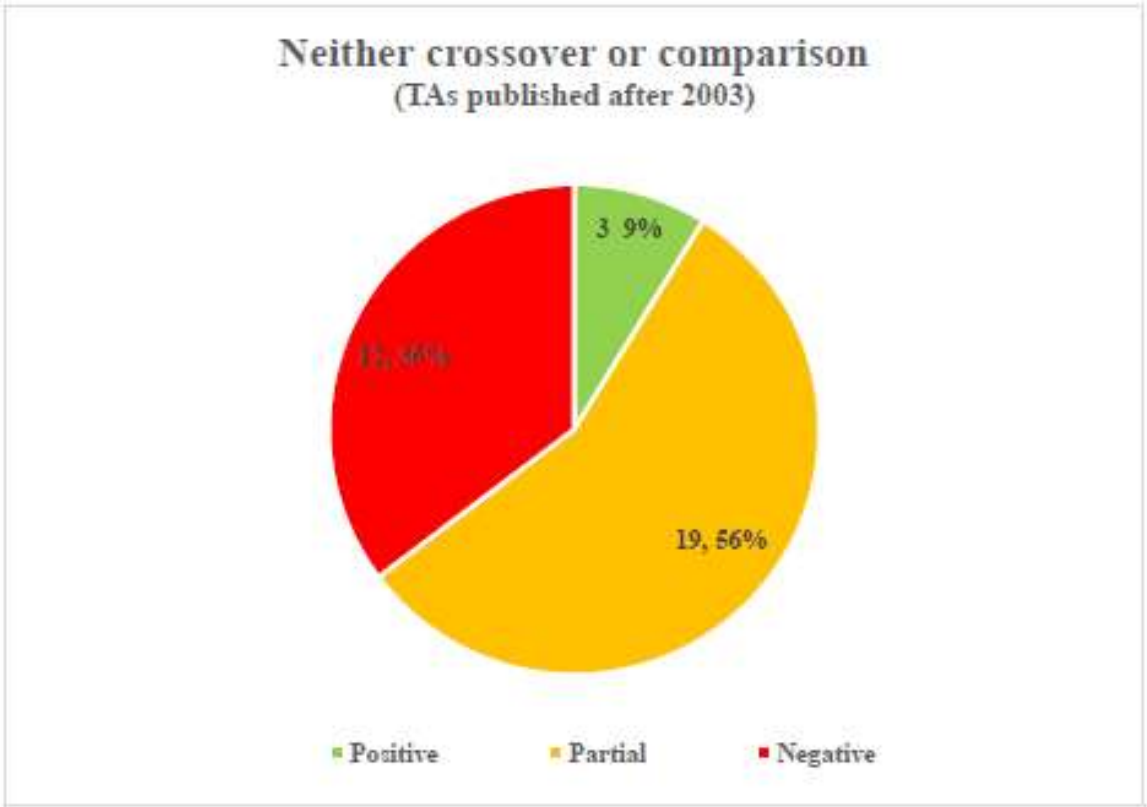
B-1: Recommendations by the characteristics of TAs

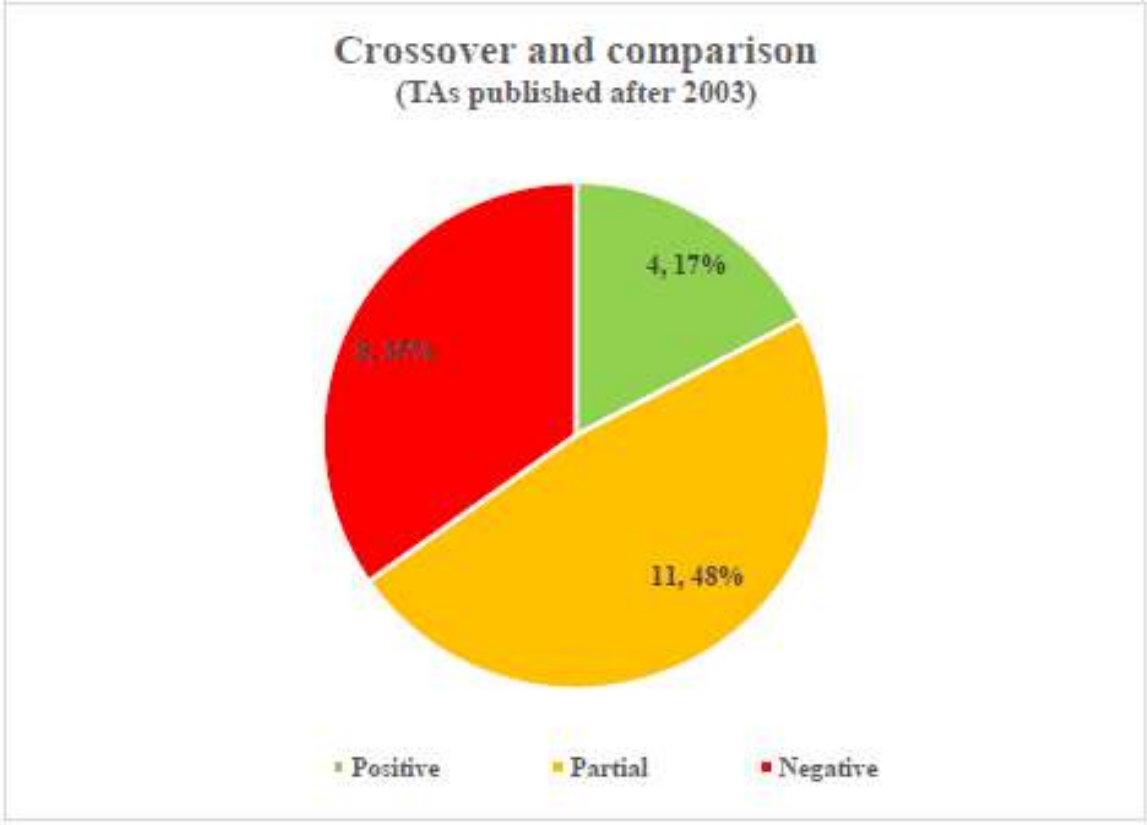


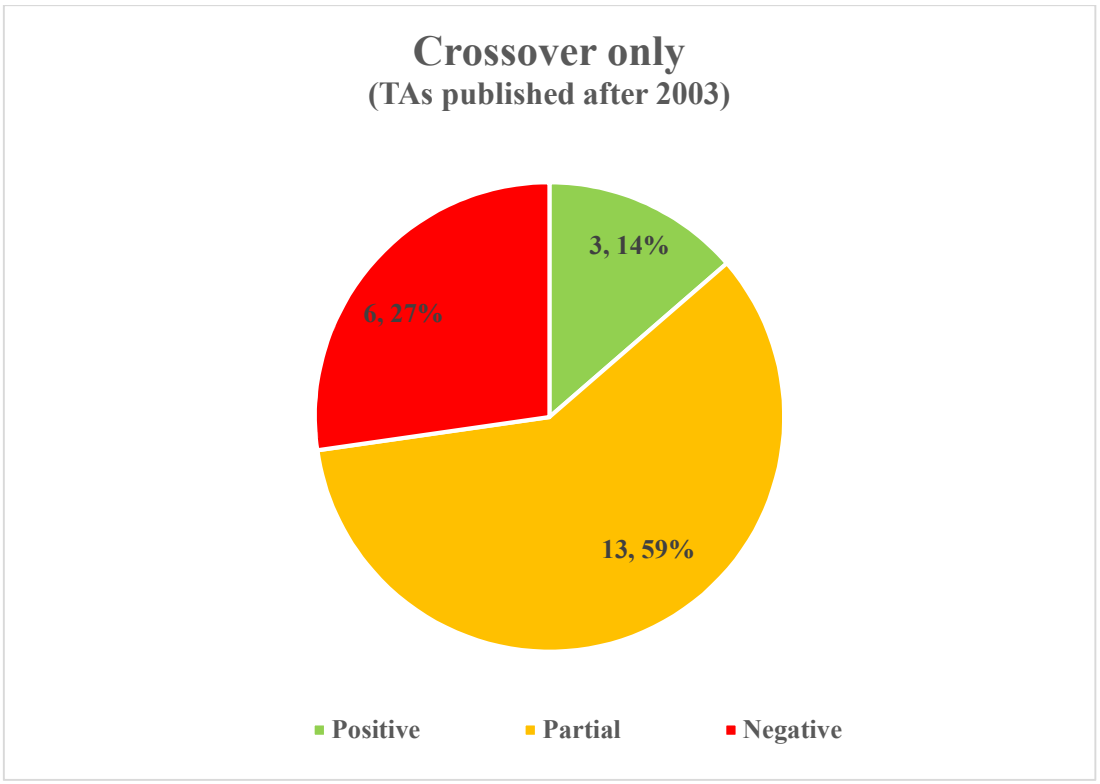
B-2: Recommendations by the characteristic and year of publication

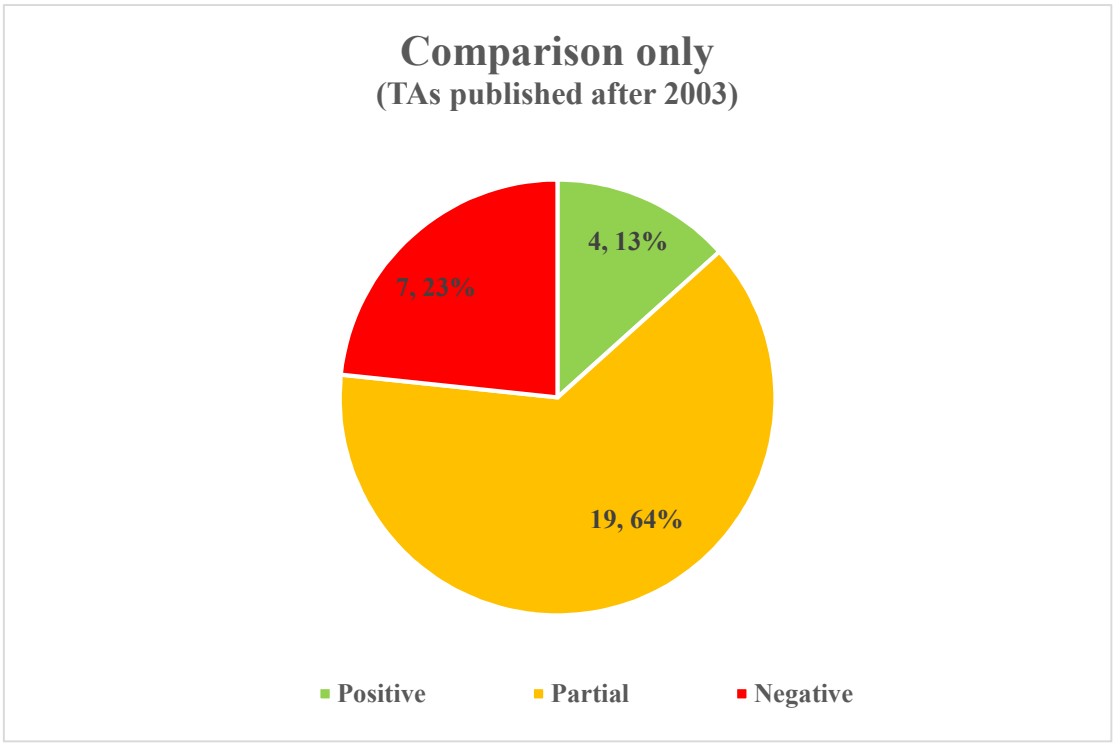


B-3: Summary of results by crossover adjustment and comparison






Crossover only
(TAs published after 2003)
3, 14%
6, 27%
13, 59%
Positive
Partial
Negative


Comparison only
(TAs published after 2003)
4, 13%
7, 23%
19, 64%
Positive
Partial
Negative

Appendix C: List of evidence reviewed in Section 2.3

C-1: Reviewed as 'Manufacturers Submission' Evidence

The following sources were reviewed as indicative of evidence of contained within a Manufacturers Submission to NICE.

TA Ref.	Reference
TA116	Eli Lilly and Company. Gemcitabine for the treatment of metastatic breast cancer. Single Technology Appraisal (STA) submission to the National Institute for Health and Clinical Excellence. 18th May 2006.
TA214	Roche. Single Technology Appraisal (STA) Bevicizumab in combination with taxanes for the treatment of HER2-negative 1st line metastatic breast cancer. 8th March 2010
TA215	Heron Evidence Development Systematic Review. Clinical and Economic Systematic Reviews in Treatment Naïve Advanced/Metastatic Renal Cell Carcinoma. Version 3. December 2010* * *The evidence presented in this review was deemed to be representative of the Manufacturers Submission* GlaxoSmithKline. Pazopanib for the first-line treatment of patients with advanced renal cell carcinoma (RCC): Addendum to GSK's submission to NICE. July 2010

C-2: Reviewed as 'Evidence Review Group' Evidence

The following sources were reviewed as indicative of evidence of contained within a Evidence Review Group's report.

TA Ref.	Reference
TA101	Collins R, Fenwick E, Trowman R, Perard R, Norman G, Light K et al. A systematic review and economic model of the clinical effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer. Health Technol. Assess 2007; 11(2)* * *The evidence presented in this review was deemed to be representative of the Evidence Review Group's Report*
TA214	Rodgers M, Soares M, Epstein D, Yang H, Fox D, Eastwood A. Bevacizumab in combination with a taxane for the first-line treatment of HER2-negative metastatic breast cancer. Evidence Review Group's Report. 17th May 2010
TA215	Kilonzo M, Hislop J, Elders A, Fraser C, Bissett D, McClinton S, Mowatt G, Vale L. Pazopanib for the first line treatment of patients with advanced and/or metastatic renal cell carcinoma: A Single Technology Appraisal. 15 September 2010.

C-3: Trial publications reviewed

Note: Jones 2005 was used as evidence in TA116 in addition to TA214 and the publications for BRIM-3 (Chapman, 2011, McArthur, 2014) were used as evidence in TA321 in addition to TA319.

TA Ref.	Trial Name (if any)	Type of publication	Reference
TA101	TAX327	Original	Tannock IF, de Wit R, Berry WR, Horti J, Pluzanska A, Chi KN, Oudard S, Theodore C, James ND, Turesson I, Rosenthal MA, Eisenberger MA, and TAX 327 Investigators. Docetaxel plus Prednisone or Mitoxantrone plus Prednisone for Advanced Prostate Cancer. New England Journal of Medicine 2004;351:1502-12.
TA171	MM-009	Original	Weber DM, Chen C, Niesvizky R, Wang M, Belch A, Stadtmauer EA, Siegel D, Borrello I, Rajkumar V, Chanan-Khan AA, Lonial S, Yu Z, Patin J, Olenyckyj M, Zeldis JB, Knight RD and Multiple Myeloma (009) Study Investigators. Lenalidomide plus Dexamethasone for Relapsed Multiple Myeloma in North America. New England Journal of Medicine 2007;357:2133-42.
	MM-010	Original	Dimopoulos MA, Spencer A, Attal M, Prince HM, Harousseau JC, Dmoszynska A et al. Multiple Myeloma (010) study investigators. Lenalidomide plus dexamethasone for relapsed or refractory multiple myeloma. N Engl J Med 2007; 357: 2123–2132.
	MM-009 & MM-010	Follow-up	Dimopoulos MA, Chen C, Spencer A, Niesvizky R, Attal M, Stadtmauer EA, Petrucci MT, Yu Z, Olesnyckyi M, Zeldis JB, Knight RD, Weber DM. Long-term follow-up on overall survival from the MM-009 and MM-010 phase III trials of lenalidomide plus dexamethasone in patients with relapsed or refractory multiple myeloma. Leukemia (2009) 23, 2147–2152;
	APEX	Original	Richardson PG, Sonneveld P, Schuster MW, Irwin D, Stadtmauer EA, Facon T, Harousseau JL, Ben-Yuhuda D, Lonial S et al. Bortezomib or High-Dose Dexamethasone for Relapsed Multiple Myeloma. N Engl J Med 2005;352:2487-98.
		Follow-up	Richardson PG, Sonneveld P, Schuster M, Irwin D, Stadtmauer E, Facon E, Harousseau JL et al. Extended follow-up of a phase 3 trial in relapsed multiple myeloma: final time-to-event results of the APEX trial. Blood. 2007. 110(10)
TA214	E2100	Original	Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, Shenkier T, Cella D, Davidson NE. Paclitaxel plus Bevacizumab versus Paclitaxel Alone for Metastatic Breast Cancer. New England Journal of Medicine 2007. 357(26) pp 2666 - 2676
	Albain, 2008	Original	Albain KS, Nag SM, Calderillo-Ruiz G, Jordaan JP, Llombart AC, Pluzanska A, Rolski J, Melemed AS, Reyes-Vidal JM, Sekhon JS, Simms L, O'Shaughnessy J. Gemcitabine Plus Paclitaxel Versus Paclitaxel Monotherapy in Patients With Metastatic Breast Cancer and Prior Anthracycline Treatment. Journal of Clinical Oncology 2008. 26(24) pp 3950 - 3957

TA Ref.	Trial Name (if any)	Type of publication	Reference
TA214	Seidman (CALBG)	Original	Seidman AD, Berry D, Cirrincione C, Harris L, Muss H, Marcom PK, Gipson G, Burstein H, Lake D, Shapiro CL, Ungaro P, Norton L, Winer E, Hudis C. Randomized Phase III Trial of Weekly Compared With Every-3-Weeks Paclitaxel for Metastatic Breast Cancer, With Trastuzumab for all HER-2 Overexpressors and Random Assignment to Trastuzumab or Not in HER-2 Nonoverexpressors: Final Results of Cancer and Leukemia Group B Protocol 9840. Journal of Clinical Oncology 2008. 26(10) pp 1642-1649
	Jones 2005	Original	Jones SE, Erban J, Overmoyer B, Budd GT, Hutchins I, Lower E, Laufman S et al. Randomized Phase III Study of Docetaxel Compared with Paclitaxel in Metastatic Breast Cancer. 2005. 23(24)pp5542-5551
	RIBBON-1	Original	Robert NJ, Dieras V, Glaspy J, Brufsky AM, Bondarenko I, Lipatov ON, Perez EA et al. RIBBON-1: Randomized, Double-Blind, Placebo-Controlled, Phase III Trial of Chemotherapy With or Without Bevacizumab for First-Line Treatment of Human Epidermal Growth Factor Receptor 2–Negative, Locally Recurrent or Metastatic Breast Cancer. Journal of Clinical Oncology 2011. 29(10) pp1252-1260
TA215	VEG105192	Original	Sternberg CN, Davis ID, Mardiak J, Szczylik C, Lee E, Wagstaff J, Barrios CH, Salman P, Gladkov OA, Kavina A, Zarba JJ, Chen M, McCann L, Pandite L, Roychowdhury DF, Hawkins RE. Pazopanib in Locally Advanced or Metastatic Renal Cell Carcinoma: Results of a Randomized Phase III Trial. Journal of Clinical Oncology 2010. 28(6) pp 1061-1068
	Steineck 1990	Original	Steineck G, Strander H, Carbin BE, Borgstrom E, Wallin L, et al. (1990) Recombinant leukocyte interferon alpha-2a and medroxyprogesterone in advanced renal cell carcinoma. A randomized trial. Acta Oncol. 29(2): 155-162.
	Kriegmair	Original	Kriegmair M, Oberneder R, Hofstetter A. (1995) Interferon alfa and vinblastine versus medroxyprogesterone acetate in the treatment of metastatic renal cell carcinoma. Urology. 45(5): 758-762.
	Pyrhonen 1999	Original	Pyrhonen S, Salminen E, Ruutu M, Lehtonen T, Nurmi M, et al. (1999) Prospective randomized trial of interferon alfa-2a plus vinblastine versus vinblastine alone in patients with advanced renal cell cancer. J Clin Oncol. 17(9): 2859-2867.
	Motzer 2009	Original	Motzer RJ, Hutson TE, Tomczak P, Dror Michaelson M, Bukowski RM, Rixe O, Oudard S, Negrier S, Szczylik C, Kim ST, Chen I, Bycott PW, Baum CM, Figlin RA. Sunitinib versus Interferon Alfa in Metastatic Renal-Cell Carcinoma. New England Journal of Medicine 2007. 356(2) pp115-124
		Follow-up	Motzer RJ, Hutson TE, Tomczak P, Michaelson MD, Bukowski RM, et al. (2009) Overall survival and updated results for sunitinib compared with interferon alfa in patients with metastatic renal cell carcinoma. J Clin Oncol. 27(22): 3584-3590
	CRECY	Original	Negrier S, Escudier B, Lasset C, Douillard JY, Savary J, et al. (1998) Recombinant human interleukin-2, recombinant human interferon alfa-2a, or both in

TA Ref.	Trial Name (if any)	Type of publication	Reference
			metastatic renal-cell carcinoma. Groupe Francais d'Immunotherapie. N Engl J Med. 338(18): 1272-1278.
	Negrier 2007	Original	Negrier S, Perol D, Ravaud A, Chevreau C, Bay JO, et al. (2007) Medroxyprogesterone, interferon alfa-2a, interleukin 2, or combination of both cytokines in patients with metastatic renal carcinoma of intermediate prognosis: results of a randomized controlled trial. Cancer. 110(11): 2468-2477
TA215	AVOREN	Original	Melichar B, Koralewski P, Ravaud A, Pluzanska A, Bracarda S, et al. (2008) First-line bevacizumab combined with reduced dose interferon-alpha2a is active in patients with metastatic renal cell carcinoma. Ann Oncol. 19(8): 1470-1476.
	TARGET	Original	Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, et al. Sorafenib for treatment of renal cell carcinoma: Final efficacy and safety results of the phase III treatment approaches in renal cancer global evaluation trial. J Clin Oncol. 2009b 27(20): 3312-3318.
	CALGB 90206	Original	Rini BI, Halabi S, Rosenberg JE, Stadler WM, Vaena DA, et al. (2008a) Bevacizumab plus interferon alfa compared with interferon alfa monotherapy in patients with metastatic renal cell carcinoma: CALGB 90206. J Clin Oncol. 26(33): 5422
	Global ARCC	Original	Hudes G, Carducci M, Tomczak P, Dutcher J, Figlin R, et al. (2007) Temsirolimus, interferon alfa, or both for advanced renal-cell carcinoma. N Engl J Med. 356(22): 2271-2281.
	MRC RE01	Original	Ritchie AWW, Griffiths G, Parmar M. (1999) Interferon-alpha and survival in metastatic renal carcinoma: early results of a randomised controlled trial. Medical Research Council Renal Cancer Collaborators. Lancet. 353(9146): 14
TA258	EURTAC	Original	De Marinis F, Rosell R, Vergnenegre A, Massuti B, Felip E, Gervais R, et al. Erlotinib vs chemotherapy (CT) in advanced non-small cell lung cancer (NSCLC) patients with epidermal growth factor receptor (EGFR) activating mutations − the EURTAC Phase II randomized trial interim results. European Journal of Cancer 2011;47: S597.
	IPASS	Original	Mok TS, Wu, YL. Thongprasert S, Yang, CH, Chu DT, Saijo N.et al. Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma N Engl J Med 2009; 361:947
		Follow-up	Fukuoka M, Wu YL, Thongprasert S, Sunpaweravong P, Leong SS, Sriuranpong V, et al. Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/ paclitaxel in clinically selected patients with advanced nonsmall cell lung cancer in Asia (IPASS). Journal of Clinical Oncology 2011;29(21):2866–74
	FIRST-SIGNAL	Original	Han JY, Park K, Kim SW, Lee DH, Kim HY, Kim HT, et al. First-SIGNAL: first-line single-agent Iressa versus gemcitabine and cisplatin trial in never-smokers with adenocarcinoma of the lung. Journal of Clinical Oncology 2012;30(10):1122–8
	NEJ002	Original	Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for

TA Ref.	Trial Name (if any)	Type of publication	Reference
			non-small cell lung cancer with mutated EGFR. The New England Journal of Medicine 2010;362(25):2380–8.
		Follow-up	Inoue A, Kobayashi K, Maemondo M, Sugawara S, Oizumi S, Isobe H, et al. Updated overall survival results from a randomized phase III trial comparing gefitinib with carboplatin-paclitaxel for chemo-naïve non-small cell lung cancer with sensitive EGFR gene mutations (NEJ002). Annals of Oncology 2012;24(1):54–9.
TA258	FIRST-SIGNAL	Original	Han JY, Park K, Kim SW, Lee DH, Kim HY, Kim HT, et al. First-SIGNAL: first-line single-agent Iressa versus gemcitabine and cisplatin trial in never-smokers with adenocarcinoma of the lung. Journal of Clinical Oncology 2012;30(10):1122–8
	WJTOG3405	Original	Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Tsurutani J, et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. The Lancet Oncology 2009;11(2):121–8.
	OPTIMAL	Original	Zhou C, Wu YL, Chen G, Feng J, Liu XQ, Wang C, et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation positive nonsmall cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. The Lancet Oncology 2011;12(8):735–42.
		Follow-up	Zhou C, Wu YL, Chen G, Feng J, Liu X, Wang C, et al. Final overall survival results from a randomised, phase III study of erlotinib versus chemotherapy as first-line treatment of EGFR mutation-positive advanced non smallcell lung cancer (OPTIMAL, CTONG-0802). Annals of Oncology 2015;26:1877–83.
TA319	BRIM-3	Original	Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. N Engl J Med 2011;364:2507-16.
		Follow-up	McArthur, GA. Chapman, PB., Robert C, Larkin J., Haanen, JB., Drummer, R. Ribas, A et al. Safety and effcacy of vemurafenib in BRAFV600E and BRAFV600K mutation-positive melanoma (BRIM-3): extended follow-up of a phase 3, randomised, open-label study. Lancet Oncol. 2014; 15: 323–32.
TA377	PREVAIL	Original	Beer TM, Armstrong AJ, Rathkopf DE, Loriot Y, Sternberg CN, Higano CS, Iversen P, Bhattacharya S, Carles J, Chowdhury S, Davis ID, de Bono JS, Evans CP, Fizazi K, Joshua AM, Kim CS, Kimura G, Mainwaring P, Mansbach H, Miller K, Noonberg SB, Perabo F, Phung D, Saad F, Scher HI, Taplin ME, Venner PM, Tombal B and the PREVAIL Investigators. Enzalutamide in Metastatic Prostate Cancer before Chemotherapy. N Engl J Med 2014;371:424
		Follow-up	Skaltsa K, Ivanescu C, Naidoo S, Phung D, Holmstrom S, Latimer N. Adjusting Overall Survival Estimates after Treatment Switching: a Case Study in Metastatic Castration-Resistant Prostate Cancer. Targeted Oncology. 2017; 12: 111 - 121. DOI: 10.1007/s11523-016-0472-3
	COU-AA-302	Original	Ryan CJ, Smith MR, de Bono JS, Molina A, Logothetis CJ, de Souza P, Fizazi K, Mainwaring P et al. Abiraterone

TA Ref.	Trial Name (if any)	Type of publication	Reference
			in Metastatic Prostate Cancer without Previous Chemotherapy. N Engl J Med 2013;368:138-48.
		Follow-up	Ryan CJ, Smith MR, Fizazi K, Saad F, Mulders PFA, Sternberg CN, Miller K, Logothetis CJ et al. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. Lancet Oncol 2015; 16: 152–60

Appendix D: Results of the Systematic Simulation Study

D-1: ITT Results for Study A (for Group 1)

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over all datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
1	0.31	25%	5%	47.62	0.38	-0.01	-4%	1.74	0.50
17	0.31	50%	25%	105.88	0.46	0.02	6%	1.27	0.04
32	0.31	70%	25%	141.77	0.51	0.01	4%	1.06	0.00
46	0.31	90%	40%	187.64	0.58	0.05	14%	0.83	0.00
59	0.50	25%	5%	47.78	0.55	0.12	27%	1.29	0.84
71	0.50	50%	25%	105.81	0.61	0.06	17%	1.08	0.48
82	0.50	70%	25%	140.89	0.64	0.06	17%	0.98	0.25
92	0.50	90%	40%	187.57	0.69	0.06	15%	0.85	0.09
101	0.70	25%	5%	47.69	0.73	-0.01	-3%	1.13	0.92
109	0.70	50%	25%	106.68	0.76	0.01	4%	1.04	0.86
116	0.70	70%	25%	140.93	0.77	0.05	14%	1.01	0.80
122	0.70	90%	40%	187.30	0.80	0.08	20%	0.94	0.70
127	0.95	25%	5%	47.02	0.95	0.07	19%	1.09	0.96
131	0.95	50%	25%	106.37	0.95	0.05	13%	1.08	0.95
134	0.95	70%	25%	141.42	0.96	0.13	29%	1.07	0.94
136	0.95	90%	40%	187.49	0.96	0.03	9%	1.07	0.95

D-2: ITT Results for Study A (for Group 2)

This table continues on the next page.

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
20	0.31	50%	25%	106.17	0.47	0.16	33%	0.21	4%
24	0.31	50%	25%	106.45	0.47	0.16	33%	0.21	4%
28	0.31	50%	25%	105.84	0.47	0.16	33%	0.20	4%
34	0.31	70%	25%	140.89	0.52	0.21	39%	0.26	0%
35	0.31	70%	25%	141.30	0.52	0.21	39%	0.26	0%
38	0.31	70%	25%	141.32	0.52	0.21	39%	0.26	0%
39	0.31	70%	25%	141.10	0.52	0.21	39%	0.26	0%
42	0.31	70%	25%	141.23	0.52	0.21	39%	0.26	1%
43	0.31	70%	25%	141.74	0.52	0.21	39%	0.26	0%
47	0.31	90%	40%	187.45	0.59	0.28	47%	0.34	0%
48	0.31	90%	40%	187.39	0.59	0.28	47%	0.34	0%
49	0.31	90%	40%	187.31	0.59	0.28	47%	0.34	0%
51	0.31	90%	40%	187.38	0.59	0.28	47%	0.34	0%
52	0.31	90%	40%	187.35	0.59	0.28	46%	0.34	0%
53	0.31	90%	40%	187.54	0.58	0.27	46%	0.34	0%
55	0.31	90%	40%	187.68	0.59	0.28	47%	0.34	0%
56	0.31	90%	40%	187.72	0.59	0.28	47%	0.34	0%
57	0.31	90%	40%	187.58	0.59	0.28	47%	0.34	0%
74	0.50	50%	25%	105.84	0.61	0.11	18%	0.17	47%
78	0.50	50%	25%	106.57	0.61	0.11	18%	0.18	46%

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
84	0.50	70%	25%	141.16	0.65	0.15	22%	0.21	28%
85	0.50	70%	25%	141.44	0.65	0.15	22%	0.21	26%
88	0.50	70%	25%	141.01	0.65	0.15	22%	0.21	25%
89	0.50	70%	25%	141.01	0.65	0.15	22%	0.21	26%
93	0.50	90%	40%	187.61	0.69	0.19	27%	0.26	9%
94	0.50	90%	40%	187.69	0.70	0.20	28%	0.27	8%
95	0.50	90%	40%	187.59	0.70	0.20	28%	0.27	8%
97	0.50	90%	40%	187.34	0.70	0.20	28%	0.27	9%
98	0.50	90%	40%	187.62	0.70	0.20	28%	0.26	7%
99	0.50	90%	40%	187.11	0.69	0.19	27%	0.26	10%
112	0.70	50%	25%	106.46	0.76	0.06	8%	0.13	87%
118	0.70	70%	25%	141.68	0.78	0.08	9%	0.15	81%
119	0.70	70%	25%	141.64	0.78	0.08	10%	0.15	79%
123	0.70	90%	40%	187.53	0.81	0.11	13%	0.18	67%
124	0.70	90%	40%	187.56	0.81	0.11	13%	0.19	66%
125	0.70	90%	40%	187.55	0.80	0.10	12%	0.18	72%

D-3: ITT Results for Study A (for Group 3)

This table continues on the next page.

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
5	0.31	25%	5%	47.50	0.38	0.07	18%	0.11	49%
9	0.31	25%	5%	47.40	0.38	0.07	18%	0.11	48%
13	0.31	25%	5%	47.40	0.39	0.08	19%	0.12	46%
21	0.31	50%	25%	106.07	0.47	0.16	33%	0.21	3%
25	0.31	50%	25%	105.81	0.47	0.16	33%	0.20	3%
29	0.31	50%	25%	106.62	0.47	0.16	33%	0.20	4%
36	0.31	70%	25%	141.26	0.52	0.21	39%	0.26	0%
40	0.31	70%	25%	140.84	0.52	0.21	39%	0.26	0%
44	0.31	70%	25%	140.95	0.52	0.21	39%	0.26	0%
50	0.31	90%	40%	187.33	0.59	0.28	47%	0.34	0%
54	0.31	90%	40%	187.20	0.58	0.27	46%	0.34	0%
58	0.31	90%	40%	187.37	0.59	0.28	47%	0.34	0%
63	0.50	25%	5%	47.32	0.55	0.05	9%	0.11	83%
67	0.50	25%	5%	47.58	0.55	0.05	9%	0.11	83%
75	0.50	50%	25%	106.62	0.61	0.11	17%	0.17	46%
79	0.50	50%	25%	106.35	0.62	0.12	18%	0.18	44%
86	0.50	70%	25%	141.22	0.65	0.15	22%	0.21	24%
90	0.50	70%	25%	141.47	0.64	0.14	22%	0.21	26%

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
96	0.50	90%	40%	187.27	0.69	0.19	27%	0.26	8%
100	0.50	90%	40%	186.99	0.70	0.20	27%	0.26	9%
105	0.70	25%	5%	47.59	0.73	0.03	4%	0.10	93%
113	0.70	50%	25%	106.36	0.76	0.06	7%	0.13	85%
120	0.70	70%	25%	141.39	0.78	0.08	9%	0.15	82%
126	0.70	90%	40%	187.33	0.80	0.10	12%	0.18	70%

D-4: ITT Results for Study A (for Group 4)

This table continues on the next page.

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
6	0.31	25%	5%	47.25	0.38	0.07	18%	0.11	49%
7	0.31	25%	5%	47.65	0.38	0.07	18%	0.12	49%
8	0.31	25%	5%	47.62	0.38	0.07	18%	0.11	50%
10	0.31	25%	5%	47.53	0.38	0.07	18%	0.11	48%
11	0.31	25%	5%	47.14	0.38	0.07	18%	0.11	48%
12	0.31	25%	5%	47.39	0.39	0.08	19%	0.12	47%
14	0.31	25%	5%	47.78	0.39	0.08	19%	0.12	47%
15	0.31	25%	5%	47.39	0.39	0.08	19%	0.12	45%
16	0.31	25%	5%	47.45	0.39	0.08	19%	0.12	45%
22	0.31	50%	25%	106.39	0.47	0.16	33%	0.20	3%
23	0.31	50%	25%	106.28	0.47	0.16	33%	0.21	2%
26	0.31	50%	25%	106.06	0.47	0.16	33%	0.20	3%
27	0.31	50%	25%	106.03	0.47	0.16	33%	0.20	3%
30	0.31	50%	25%	105.80	0.47	0.16	33%	0.21	3%
31	0.31	50%	25%	106.05	0.47	0.16	33%	0.21	4%
37	0.31	70%	25%	141.10	0.51	0.20	39%	0.26	0%
41	0.31	70%	25%	141.59	0.52	0.21	39%	0.26	0%

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
45	0.31	70%	25%	141.06	0.51	0.20	39%	0.26	0%
64	0.50	25%	5%	47.44	0.55	0.05	8%	0.10	84%
65	0.50	25%	5%	46.99	0.55	0.05	9%	0.11	82%
66	0.50	25%	5%	47.48	0.55	0.05	9%	0.11	84%
68	0.50	25%	5%	47.45	0.55	0.05	8%	0.10	86%
69	0.50	25%	5%	47.81	0.55	0.05	9%	0.10	83%
70	0.50	25%	5%	47.61	0.55	0.05	9%	0.11	82%
76	0.50	50%	25%	105.80	0.62	0.12	18%	0.18	46%
77	0.50	50%	25%	106.02	0.61	0.11	18%	0.17	44%
80	0.50	50%	25%	106.52	0.61	0.11	18%	0.17	47%
81	0.50	50%	25%	106.34	0.61	0.11	18%	0.17	46%
87	0.50	70%	25%	141.03	0.65	0.15	22%	0.21	25%
91	0.50	70%	25%	141.36	0.65	0.15	22%	0.21	26%
106	0.70	25%	5%	47.51	0.73	0.03	3%	0.10	96%
107	0.70	25%	5%	47.36	0.73	0.03	3%	0.10	92%
108	0.70	25%	5%	47.76	0.73	0.03	4%	0.10	92%
114	0.70	50%	25%	106.27	0.76	0.06	7%	0.13	86%
115	0.70	50%	25%	105.92	0.76	0.06	7%	0.13	88%
121	0.70	70%	25%	141.28	0.78	0.08	9%	0.15	81%

D-5: ITT Results for Study A (for Group 5)

Scenario	Study A			Study A					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
2	0.31	25%	5%	47.39	0.39	0.08	19%	0.12	47%
3	0.31	25%	5%	47.45	0.39	0.08	19%	0.12	45%
4	0.31	25%	5%	47.61	0.38	0.07	18%	0.11	49%
18	0.31	50%	25%	106.33	0.47	0.16	33%	0.20	3%
19	0.31	50%	25%	106.60	0.47	0.16	33%	0.21	3%
33	0.31	70%	25%	141.62	0.51	0.20	39%	0.26	0%
60	0.50	25%	5%	47.39	0.56	0.06	9%	0.11	79%
61	0.50	25%	5%	47.71	0.55	0.05	9%	0.10	83%
62	0.50	25%	5%	47.36	0.55	0.05	9%	0.11	82%
72	0.50	50%	25%	106.05	0.61	0.11	18%	0.17	48%
73	0.50	50%	25%	106.40	0.62	0.12	18%	0.17	45%
83	0.50	70%	25%	141.30	0.65	0.15	22%	0.21	25%
102	0.70	25%	5%	47.32	0.72	0.02	3%	0.09	94%
103	0.70	25%	5%	47.25	0.73	0.03	3%	0.10	93%
104	0.70	25%	5%	47.46	0.73	0.03	4%	0.10	93%
110	0.70	50%	25%	106.47	0.76	0.06	7%	0.13	89%
111	0.70	50%	25%	106.63	0.76	0.06	7%	0.13	87%
117	0.70	70%	25%	141.28	0.78	0.08	9%	0.15	82%
128	0.95	25%	5%	47.71	0.96	0.01	0%	0.09	96%
129	0.95	25%	5%	47.40	0.95	0.00	0%	0.09	96%
130	0.95	25%	5%	47.45	0.96	0.01	0%	0.09	96%
132	0.95	50%	25%	106.28	0.96	0.01	0%	0.10	96%
133	0.95	50%	25%	106.27	0.96	0.01	0%	0.09	96%
135	0.95	70%	25%	141.91	0.96	0.01	0%	0.09	95%

D-6: ITT Results for Study B (for Group 1)

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
1	0.31	0.25	0.05	47.68	0.39	0.08	19%	0.12	50%
17	0.31	0.50	0.25	106.46	0.47	0.16	33%	0.21	4%
32	0.31	0.70	0.25	141.19	0.51	0.20	39%	0.25	0%
46	0.31	0.90	0.40	187.73	0.59	0.28	47%	0.34	0%
59	0.50	0.25	0.05	47.15	0.55	0.05	9%	0.11	83%
71	0.50	0.50	0.25	106.29	0.61	0.11	18%	0.17	45%
82	0.50	0.70	0.25	141.45	0.65	0.15	22%	0.21	24%
92	0.50	0.90	0.40	187.31	0.69	0.19	27%	0.26	11%
101	0.70	0.25	0.05	47.71	0.73	0.03	3%	0.10	93%
109	0.70	0.50	0.25	106.17	0.76	0.06	7%	0.13	87%
116	0.70	0.70	0.25	141.70	0.78	0.08	9%	0.15	81%
122	0.70	0.90	0.40	187.22	0.80	0.10	12%	0.18	72%
127	0.95	0.25	0.05	47.52	0.95	0.00	-1%	0.09	96%
131	0.95	0.50	0.25	106.52	0.96	0.01	0%	0.10	95%
134	0.95	0.70	0.25	141.43	0.96	0.01	0%	0.09	95%
136	0.95	0.90	0.40	187.55	0.96	0.01	1%	0.10	96%

D-7: ITT Results for Study B (for Group 2)

This table continues on the next page.

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
20	0.50	0.25	0.05	47.48	0.55	0.05	9%	0.11	84%
24	0.70	0.25	0.05	47.49	0.73	0.03	4%	0.10	92%
28	0.95	0.25	0.05	47.85	0.96	0.01	0%	0.09	96%
34	0.50	0.25	0.05	47.47	0.55	0.05	9%	0.11	82%
35	0.50	0.50	0.25	106.38	0.62	0.12	18%	0.18	43%
38	0.70	0.25	0.05	47.53	0.73	0.03	3%	0.10	92%
39	0.70	0.50	0.25	106.40	0.76	0.06	7%	0.13	85%
42	0.95	0.25	0.05	47.61	0.95	0.00	0%	0.09	95%
43	0.95	0.50	0.25	105.86	0.96	0.01	0%	0.09	96%
47	0.50	0.25	0.05	47.40	0.55	0.05	8%	0.10	85%
48	0.50	0.50	0.25	106.53	0.61	0.11	18%	0.17	45%
49	0.50	0.70	0.25	141.30	0.65	0.15	22%	0.21	27%
51	0.70	0.25	0.05	47.39	0.73	0.03	4%	0.10	92%
52	0.70	0.50	0.25	106.36	0.76	0.06	7%	0.13	87%
53	0.70	0.70	0.25	141.26	0.78	0.08	10%	0.15	80%
55	0.95	0.25	0.05	47.30	0.96	0.01	0%	0.10	95%
56	0.95	0.50	0.25	105.65	0.96	0.01	0%	0.10	93%
57	0.95	0.70	0.25	141.25	0.96	0.01	1%	0.10	96%
74	0.70	0.25	0.05	47.85	0.73	0.03	4%	0.10	95%
78	0.95	0.25	0.05	47.60	0.95	0.00	0%	0.09	96%
84	0.70	0.25	0.05	47.78	0.73	0.03	3%	0.09	94%
85	0.70	0.50	0.25	106.42	0.76	0.06	7%	0.13	87%
88	0.95	0.25	0.05	47.35	0.95	0.00	-1%	0.09	95%

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
89	0.95	0.50	0.25	105.64	0.96	0.01	0%	0.09	97%
93	0.70	0.25	0.05	47.51	0.73	0.03	3%	0.10	94%
94	0.70	0.50	0.25	105.81	0.76	0.06	7%	0.13	87%
95	0.70	0.70	0.25	141.50	0.78	0.08	9%	0.15	81%
97	0.95	0.25	0.05	47.63	0.96	0.01	0%	0.09	95%
98	0.95	0.50	0.25	106.69	0.95	0.00	-1%	0.09	95%
99	0.95	0.70	0.25	141.14	0.96	0.01	0%	0.09	95%
112	0.95	0.25	0.05	47.47	0.96	0.01	0%	0.09	95%
118	0.95	0.25	0.05	47.72	0.95	0.00	0%	0.09	96%
119	0.95	0.50	0.25	107.08	0.96	0.01	0%	0.09	95%
123	0.95	0.25	0.05	47.90	0.95	0.00	-1%	0.09	95%
124	0.95	0.50	0.25	106.68	0.96	0.01	0%	0.10	95%
125	0.95	0.70	0.25	141.02	0.96	0.01	0%	0.10	95%

D-8: ITT Results for Study B (for Group 3)

This table continues on the next page.

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
5	0.50	0.25	0.05	47.41	0.55	0.05	9%	0.11	83%
9	0.70	0.25	0.05	47.36	0.73	0.03	3%	0.09	94%
13	0.95	0.25	0.05	47.63	0.95	0.00	0%	0.09	95%
21	0.50	0.50	0.25	106.32	0.61	0.11	18%	0.17	46%
25	0.70	0.50	0.25	106.29	0.76	0.06	8%	0.13	87%
29	0.95	0.50	0.25	106.49	0.96	0.01	0%	0.10	95%
36	0.50	0.70	0.25	140.34	0.64	0.14	21%	0.20	28%
40	0.70	0.70	0.25	141.38	0.78	0.08	9%	0.15	80%
44	0.95	0.70	0.25	141.54	0.95	0.00	0%	0.09	95%
50	0.50	0.90	0.40	187.32	0.70	0.20	28%	0.26	9%
54	0.70	0.90	0.40	187.46	0.80	0.10	12%	0.18	71%
58	0.95	0.90	0.40	187.80	0.96	0.01	1%	0.10	95%
63	0.70	0.25	0.05	47.49	0.73	0.03	3%	0.10	93%

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
67	0.95	0.25	0.05	47.32	0.95	0.00	0%	0.09	96%
75	0.70	0.50	0.25	106.50	0.76	0.06	8%	0.14	86%
79	0.95	0.50	0.25	106.00	0.96	0.01	0%	0.09	95%
86	0.70	0.70	0.25	141.26	0.78	0.08	10%	0.16	78%
90	0.95	0.70	0.25	141.10	0.96	0.01	0%	0.10	95%
96	0.70	0.90	0.40	187.55	0.81	0.11	12%	0.18	69%
100	0.95	0.90	0.40	187.35	0.96	0.01	1%	0.10	95%
105	0.95	0.25	0.05	47.49	0.95	0.00	0%	0.09	96%
113	0.95	0.50	0.25	106.37	0.96	0.01	0%	0.09	94%
120	0.95	0.70	0.25	140.71	0.96	0.01	1%	0.10	96%
126	0.95	0.90	0.40	187.26	0.96	0.01	1%	0.10	95%

D-9: ITT Results for Study B (for Group 4)

This table continues on the next page.

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
6	0.50	0.50	0.25	106.59	0.62	0.12	18%	0.18	46%
7	0.50	0.70	0.25	141.35	0.65	0.15	22%	0.21	25%
8	0.50	0.90	0.40	187.73	0.70	0.20	28%	0.27	8%
10	0.70	0.50	0.25	105.82	0.77	0.07	8%	0.14	84%
11	0.70	0.70	0.25	141.58	0.78	0.08	9%	0.15	80%
12	0.70	0.90	0.40	187.47	0.80	0.10	12%	0.18	71%
14	0.95	0.50	0.25	106.01	0.95	0.00	-1%	0.09	95%
15	0.95	0.70	0.25	141.05	0.96	0.01	0%	0.10	95%
16	0.95	0.90	0.40	187.48	0.96	0.01	1%	0.10	95%
22	0.50	0.70	0.25	141.20	0.64	0.14	21%	0.21	29%
23	0.50	0.90	0.40	187.32	0.70	0.20	28%	0.26	9%
26	0.70	0.70	0.25	141.16	0.78	0.08	10%	0.16	78%
27	0.70	0.90	0.40	187.50	0.81	0.11	13%	0.18	69%
30	0.95	0.70	0.25	141.21	0.96	0.01	0%	0.09	96%
31	0.95	0.90	0.40	187.65	0.96	0.01	1%	0.10	94%
37	0.50	0.90	0.40	187.46	0.69	0.19	27%	0.26	8%
41	0.70	0.90	0.40	187.40	0.81	0.11	13%	0.19	66%
45	0.95	0.90	0.40	187.63	0.96	0.01	0%	0.10	95%
64	0.70	0.50	0.25	106.09	0.76	0.06	7%	0.13	86%

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
65	0.70	0.70	0.25	141.50	0.78	0.08	9%	0.15	81%
66	0.70	0.90	0.40	187.52	0.81	0.11	12%	0.18	69%
68	0.95	0.50	0.25	106.28	0.95	0.00	-1%	0.09	95%
69	0.95	0.70	0.25	140.87	0.96	0.01	0%	0.10	95%
70	0.95	0.90	0.40	187.53	0.97	0.02	1%	0.11	95%
76	0.70	0.70	0.25	141.21	0.78	0.08	9%	0.14	83%
77	0.70	0.90	0.40	187.39	0.80	0.10	12%	0.18	70%
80	0.95	0.70	0.25	141.18	0.95	0.00	-1%	0.09	96%
81	0.95	0.90	0.40	187.86	0.96	0.01	1%	0.10	95%
87	0.70	0.90	0.40	187.29	0.81	0.11	12%	0.18	68%
91	0.95	0.90	0.40	187.55	0.96	0.01	1%	0.10	95%
106	0.95	0.50	0.25	105.98	0.96	0.01	0%	0.10	94%
107	0.95	0.70	0.25	140.93	0.96	0.01	0%	0.09	95%
108	0.95	0.90	0.40	187.39	0.96	0.01	1%	0.10	95%
114	0.95	0.70	0.25	141.28	0.96	0.01	0%	0.10	96%
115	0.95	0.90	0.40	187.09	0.97	0.02	1%	0.10	95%
121	0.95	0.90	0.40	187.90	0.96	0.01	1%	0.10	95%

D-10: ITT Results for Study B (for Group 5)

Scenario	Study B			Study B					
	True HR	Crossover proportion		Average crossover over simulated datasets	Mean estimated ITT HR	Absolute bias	Proportional bias	MSE of estimated ITT HR	Coverage of the estimated ITT HR
		Poor	Good						
2	0.31	0.50	0.25	106.00	0.46	0.15	33%	0.20	3%
3	0.31	0.70	0.25	141.40	0.52	0.21	39%	0.26	0%
4	0.31	0.90	0.40	187.63	0.59	0.28	47%	0.34	0%
18	0.31	0.70	0.25	141.23	0.51	0.20	39%	0.26	0%
19	0.31	0.90	0.40	187.37	0.59	0.28	47%	0.34	0%
33	0.31	0.90	0.40	187.80	0.59	0.28	47%	0.34	0%
60	0.50	0.50	0.25	105.96	0.62	0.12	18%	0.18	44%
61	0.50	0.70	0.25	141.20	0.64	0.14	22%	0.21	27%
62	0.50	0.90	0.40	187.26	0.70	0.20	27%	0.26	9%
72	0.50	0.70	0.25	141.15	0.65	0.15	22%	0.22	24%
73	0.50	0.90	0.40	187.73	0.69	0.19	27%	0.26	9%
83	0.50	0.90	0.40	187.77	0.70	0.20	28%	0.27	7%
102	0.70	0.50	0.25	106.20	0.76	0.06	8%	0.13	85%
103	0.70	0.70	0.25	141.25	0.78	0.08	9%	0.15	80%
104	0.70	0.90	0.40	187.41	0.81	0.11	12%	0.18	67%
110	0.70	0.70	0.25	141.12	0.78	0.08	9%	0.15	82%
111	0.70	0.90	0.40	187.27	0.81	0.11	13%	0.18	66%
117	0.70	0.90	0.40	187.64	0.80	0.10	12%	0.17	69%
128	0.95	0.50	0.25	105.98	0.96	0.01	0%	0.10	95%
129	0.95	0.70	0.25	141.17	0.96	0.01	0%	0.10	94%
130	0.95	0.90	0.40	187.52	0.96	0.01	1%	0.10	95%
132	0.95	0.70	0.25	141.22	0.96	0.01	0%	0.10	95%
133	0.95	0.90	0.40	187.66	0.96	0.01	1%	0.10	95%
135	0.95	0.90	0.40	187.76	0.96	0.01	0%	0.10	96%

D-11: RPSFTM-adjusted analysis for Study A (Group 1)

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
1	0.31	0.32	0.01	0%	0.06	86%
17	0.31	0.32	0.01	0%	0.06	92%
32	0.31	0.32	0.01	0%	0.07	93%
46	0.31	0.32	0.01	-2%	0.08	96%
59	0.50	0.52	0.02	2%	0.08	90%
71	0.50	0.52	0.02	2%	0.09	93%
82	0.50	0.52	0.02	1%	0.09	93%
92	0.50	0.52	0.02	1%	0.11	95%
101	0.70	0.70	0.00	-1%	0.07	97%
109	0.70	0.70	0.00	-1%	0.07	98%
116	0.70	0.70	0.00	-2%	0.08	98%
122	0.70	0.70	0.00	-2%	0.09	97%
127	0.95	0.95	0.00	-1%	0.09	95%
131	0.95	0.95	0.00	-1%	0.11	96%
134	0.95	0.96	0.01	-1%	0.13	94%
136	0.95	0.96	0.01	-1%	0.15	95%

D-12: RPSFTM-adjusted analysis for Study A (Group 2)

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
20	0.31	0.32	0.01	1%	0.07	91%
24	0.31	0.32	0.01	0%	0.06	91%
28	0.31	0.32	0.01	1%	0.07	91%
34	0.31	0.32	0.01	0%	0.07	94%
35	0.31	0.32	0.01	1%	0.08	93%
38	0.31	0.32	0.01	0%	0.07	94%
39	0.31	0.32	0.01	-1%	0.07	93%
42	0.31	0.32	0.01	0%	0.07	93%
43	0.31	0.32	0.01	0%	0.07	95%
47	0.31	0.32	0.01	0%	0.08	96%
48	0.31	0.33	0.02	1%	0.08	96%
49	0.31	0.32	0.01	-1%	0.08	97%
51	0.31	0.32	0.01	0%	0.08	97%
52	0.31	0.32	0.01	-2%	0.08	95%
53	0.31	0.32	0.01	-1%	0.08	95%
55	0.31	0.32	0.01	0%	0.08	97%
56	0.31	0.33	0.02	0%	0.08	96%
57	0.31	0.32	0.01	0%	0.08	96%

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
74	0.50	0.52	0.02	1%	0.09	93%
78	0.50	0.52	0.02	1%	0.09	91%
84	0.50	0.52	0.02	1%	0.10	93%
85	0.50	0.52	0.02	2%	0.09	94%
88	0.50	0.52	0.02	1%	0.10	92%
89	0.50	0.52	0.02	1%	0.09	95%
93	0.50	0.52	0.02	1%	0.10	96%
94	0.50	0.52	0.02	1%	0.11	95%
95	0.50	0.52	0.02	1%	0.11	95%
97	0.50	0.52	0.02	1%	0.11	95%
98	0.50	0.52	0.02	1%	0.11	95%
99	0.50	0.52	0.02	0%	0.11	95%
112	0.70	0.70	0.00	-1%	0.08	97%
118	0.70	0.70	0.00	-2%	0.08	97%
119	0.70	0.70	0.00	-1%	0.08	98%
123	0.70	0.70	0.00	-1%	0.10	98%
124	0.70	0.70	0.00	-1%	0.10	97%
125	0.70	0.70	0.00	-3%	0.09	97%

D-13: RPSFTM-adjusted analysis for Study A (Group 3)

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
5	0.31	0.32	0.01	1%	0.06	87%
9	0.31	0.32	0.01	1%	0.06	87%
13	0.31	0.32	0.01	2%	0.06	88%
21	0.31	0.32	0.01	1%	0.07	91%
25	0.31	0.32	0.01	0%	0.06	92%
29	0.31	0.32	0.01	1%	0.06	93%
36	0.31	0.32	0.01	1%	0.07	94%
40	0.31	0.32	0.01	1%	0.07	93%
44	0.31	0.32	0.01	1%	0.08	92%
50	0.31	0.32	0.01	-1%	0.08	97%
54	0.31	0.32	0.01	-1%	0.08	96%
58	0.31	0.32	0.01	0%	0.08	96%
63	0.50	0.52	0.02	2%	0.08	91%
67	0.50	0.52	0.02	2%	0.08	90%
75	0.50	0.51	0.01	1%	0.08	93%
79	0.50	0.52	0.02	2%	0.09	92%
86	0.50	0.52	0.02	2%	0.10	94%

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
90	0.50	0.51	0.01	1%	0.09	95%
96	0.50	0.52	0.02	0%	0.10	96%
100	0.50	0.52	0.02	1%	0.11	95%
105	0.70	0.70	0.00	0%	0.07	96%
113	0.70	0.70	0.00	-1%	0.08	97%
120	0.70	0.70	0.00	-2%	0.08	98%
126	0.70	0.70	0.00	-2%	0.10	97%

D-14: RPSFTM-adjusted analysis for Study A (Group 4)

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
6	0.31	0.32	0.01	1%	0.06	87%
7	0.31	0.32	0.01	0%	0.06	84%
8	0.31	0.32	0.01	0%	0.06	86%
10	0.31	0.32	0.01	1%	0.06	87%
11	0.31	0.32	0.01	1%	0.06	85%
12	0.31	0.32	0.01	1%	0.06	85%
14	0.31	0.32	0.01	2%	0.06	84%
15	0.31	0.32	0.01	1%	0.06	86%
16	0.31	0.32	0.01	2%	0.06	85%
22	0.31	0.32	0.01	1%	0.07	92%
23	0.31	0.32	0.01	0%	0.06	92%
26	0.31	0.32	0.01	1%	0.06	92%
27	0.31	0.32	0.01	1%	0.06	92%
30	0.31	0.32	0.01	1%	0.07	91%
31	0.31	0.32	0.01	-1%	0.07	90%
37	0.31	0.32	0.01	0%	0.07	94%
41	0.31	0.32	0.01	0%	0.07	94%
45	0.31	0.32	0.01	0%	0.07	93%
64	0.50	0.52	0.02	2%	0.08	91%
65	0.50	0.52	0.02	2%	0.08	89%
66	0.50	0.52	0.02	2%	0.08	91%
68	0.50	0.52	0.02	2%	0.08	90%
69	0.50	0.52	0.02	2%	0.08	90%
70	0.50	0.52	0.02	1%	0.08	90%
76	0.50	0.52	0.02	2%	0.09	93%
77	0.50	0.52	0.02	2%	0.09	92%
80	0.50	0.52	0.02	1%	0.09	93%
81	0.50	0.52	0.02	2%	0.09	93%

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
87	0.50	0.52	0.02	1%	0.09	94%
91	0.50	0.52	0.02	2%	0.10	93%
106	0.70	0.70	0.00	-1%	0.06	98%
107	0.70	0.70	0.00	-1%	0.07	96%
108	0.70	0.71	0.01	0%	0.07	97%
114	0.70	0.70	0.00	-1%	0.07	98%
115	0.70	0.70	0.00	-1%	0.07	98%
121	0.70	0.70	0.00	-1%	0.08	98%

D-15: RPSFTM-adjusted analysis for Study A (Group 5)

Scenario	Study A					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
2	0.31	0.32	0.01	1%	0.06	85%
3	0.31	0.32	0.01	2%	0.06	85%
4	0.31	0.32	0.01	1%	0.06	86%
18	0.31	0.32	0.01	-1%	0.06	91%
19	0.31	0.32	0.01	0%	0.06	92%
33	0.31	0.32	0.01	-1%	0.07	93%
60	0.50	0.52	0.02	3%	0.09	89%
61	0.50	0.52	0.02	2%	0.08	90%
62	0.50	0.52	0.02	2%	0.08	90%
72	0.50	0.52	0.02	2%	0.09	94%
73	0.50	0.52	0.02	2%	0.09	93%
83	0.50	0.52	0.02	2%	0.10	94%
102	0.70	0.69	-0.01	-2%	0.06	97%
103	0.70	0.70	0.00	-1%	0.07	98%
104	0.70	0.70	0.00	0%	0.07	96%
110	0.70	0.70	0.00	-1%	0.07	97%
111	0.70	0.70	0.00	-1%	0.07	98%
117	0.70	0.70	0.00	-2%	0.08	98%
128	0.95	0.95	0.00	-1%	0.10	96%
129	0.95	0.95	0.00	-1%	0.09	96%
130	0.95	0.95	0.00	-1%	0.10	96%
132	0.95	0.95	0.00	-1%	0.11	95%
133	0.95	0.95	0.00	-2%	0.11	96%
135	0.95	0.95	0.00	-2%	0.12	95%

D-16: RPSFTM-adjusted analysis for Study B (Group 1)

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
1	0.31	0.32	0.01	2%	0.06	86%
17	0.31	0.32	0.01	0%	0.07	92%
32	0.31	0.32	0.01	-1%	0.06	94%
46	0.31	0.33	0.02	1%	0.08	96%
59	0.50	0.52	0.02	1%	0.08	89%
71	0.50	0.52	0.02	2%	0.09	95%
82	0.50	0.52	0.02	2%	0.10	94%
92	0.50	0.52	0.02	0%	0.10	96%
101	0.70	0.70	0.00	-1%	0.07	96%
109	0.70	0.70	0.00	-1%	0.07	97%
116	0.70	0.70	0.00	-1%	0.08	97%
122	0.70	0.70	0.00	-2%	0.09	98%
127	0.95	0.95	0.00	-1%	0.09	95%
131	0.95	0.95	0.00	-1%	0.12	94%
134	0.95	0.95	0.00	-2%	0.12	95%
136	0.95	0.95	0.00	-2%	0.14	96%

D-17: RPSFTM-adjusted analysis for Study B (Group 2)

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
20	0.50	0.52	0.02	2%	0.08	91%
24	0.70	0.70	0.00	0%	0.07	97%
28	0.95	0.95	0.00	-1%	0.10	96%
34	0.50	0.52	0.02	2%	0.08	89%
35	0.50	0.52	0.02	2%	0.09	93%
38	0.70	0.70	0.00	-1%	0.07	96%
39	0.70	0.70	0.00	-1%	0.08	97%
42	0.95	0.95	0.00	-1%	0.10	95%
43	0.95	0.95	0.00	-1%	0.11	96%
47	0.50	0.52	0.02	2%	0.08	90%
48	0.50	0.52	0.02	1%	0.09	94%
49	0.50	0.52	0.02	1%	0.10	94%
51	0.70	0.70	0.00	-1%	0.07	97%
52	0.70	0.70	0.00	-1%	0.08	97%
53	0.70	0.70	0.00	-1%	0.08	97%
55	0.95	0.96	0.01	0%	0.10	95%
56	0.95	0.95	0.00	-2%	0.11	94%
57	0.95	0.95	0.00	-1%	0.12	95%

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
74	0.70	0.70	0.00	-1%	0.07	96%
78	0.95	0.95	0.00	-1%	0.09	96%
84	0.70	0.70	0.00	-1%	0.06	98%
85	0.70	0.70	0.00	-2%	0.07	98%
88	0.95	0.95	0.00	-1%	0.10	95%
89	0.95	0.95	0.00	-1%	0.11	96%
93	0.70	0.70	0.00	-1%	0.06	97%
94	0.70	0.70	0.00	-1%	0.07	98%
95	0.70	0.70	0.00	-1%	0.08	98%
97	0.95	0.95	0.00	-1%	0.10	94%
98	0.95	0.94	-0.01	-2%	0.11	94%
99	0.95	0.95	0.00	-2%	0.12	95%
112	0.95	0.95	0.00	-1%	0.10	95%
118	0.95	0.95	0.00	-1%	0.09	96%
119	0.95	0.95	0.00	-2%	0.11	95%
123	0.95	0.95	0.00	-1%	0.09	95%
124	0.95	0.95	0.00	-1%	0.11	95%
125	0.95	0.95	0.00	-2%	0.12	95%

D-18: RPSFTM-adjusted analysis for Study B (Group 3)

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
5	0.50	0.52	0.02	2%	0.08	91%
9	0.70	0.70	0.00	-1%	0.06	97%
13	0.95	0.95	0.00	-1%	0.10	94%
21	0.50	0.52	0.02	2%	0.09	94%
25	0.70	0.70	0.00	-1%	0.07	99%
29	0.95	0.95	0.00	-1%	0.11	96%
36	0.50	0.51	0.01	1%	0.09	95%
40	0.70	0.70	0.00	-1%	0.08	97%
44	0.95	0.94	-0.01	-3%	0.11	95%
50	0.50	0.52	0.02	1%	0.11	96%
54	0.70	0.70	0.00	-2%	0.09	97%
58	0.95	0.95	0.00	-2%	0.14	94%
63	0.70	0.70	0.00	-1%	0.07	96%
67	0.95	0.95	0.00	-1%	0.09	96%
75	0.70	0.70	0.00	-1%	0.08	97%
79	0.95	0.95	0.00	-2%	0.11	95%
86	0.70	0.70	0.00	-1%	0.09	96%

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
90	0.95	0.95	0.00	-2%	0.12	95%
96	0.70	0.70	0.00	-2%	0.10	98%
100	0.95	0.95	0.00	-2%	0.14	96%
105	0.95	0.95	0.00	-1%	0.10	95%
113	0.95	0.95	0.00	-2%	0.11	94%
120	0.95	0.95	0.00	-1%	0.12	96%
126	0.95	0.95	0.00	-2%	0.15	96%

D-19: RPSFTM-adjusted analysis for Study B (Group 4)

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
6	0.50	0.52	0.02	2%	0.09	92%
7	0.50	0.52	0.02	2%	0.10	93%
8	0.50	0.52	0.02	1%	0.11	95%
10	0.70	0.71	0.01	0%	0.08	98%
11	0.70	0.70	0.00	-2%	0.08	97%
12	0.70	0.70	0.00	-2%	0.09	98%
14	0.95	0.94	-0.01	-2%	0.10	96%
15	0.95	0.95	0.00	-1%	0.13	95%
16	0.95	0.95	0.00	-2%	0.14	95%
22	0.50	0.51	0.01	0%	0.09	94%
23	0.50	0.52	0.02	1%	0.11	96%
26	0.70	0.70	0.00	-1%	0.09	97%
27	0.70	0.70	0.00	-1%	0.10	97%
30	0.95	0.94	-0.01	-2%	0.11	96%
31	0.95	0.95	0.00	-2%	0.14	95%
37	0.50	0.52	0.02	1%	0.10	96%
41	0.70	0.71	0.01	-1%	0.10	97%
45	0.95	0.95	0.00	-2%	0.14	94%
64	0.70	0.70	0.00	-1%	0.07	98%
65	0.70	0.70	0.00	-2%	0.08	98%
66	0.70	0.70	0.00	-1%	0.10	97%
68	0.95	0.94	-0.01	-3%	0.10	95%
69	0.95	0.95	0.00	-1%	0.12	95%
70	0.95	0.96	0.01	-1%	0.15	95%
76	0.70	0.69	-0.01	-2%	0.07	98%
77	0.70	0.70	0.00	-2%	0.09	97%
80	0.95	0.94	-0.01	-3%	0.11	95%
81	0.95	0.95	0.00	-2%	0.14	95%

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
87	0.70	0.70	0.00	-2%	0.10	97%
91	0.95	0.95	0.00	-2%	0.14	96%
106	0.95	0.95	0.00	-1%	0.12	95%
107	0.95	0.95	0.00	-2%	0.12	95%
108	0.95	0.95	0.00	-2%	0.14	96%
114	0.95	0.95	0.00	-1%	0.12	96%
115	0.95	0.96	0.01	-1%	0.15	95%
121	0.95	0.95	0.00	-2%	0.14	95%

D-20: RPSFTM-adjusted analysis for Study B (Group 5)

Scenario	Study B					
	True HR	Mean estimated RPSFTM HR	Absolute bias	Proportional bias	MSE of estimated RPSFTM HR	Coverage of the estimated RPSFTM HR
2	0.31	0.32	0.01	0%	0.06	92%
3	0.31	0.32	0.01	0%	0.07	94%
4	0.31	0.32	0.01	-1%	0.08	96%
18	0.31	0.32	0.01	0%	0.07	94%
19	0.31	0.32	0.01	0%	0.08	96%
33	0.31	0.32	0.01	0%	0.08	96%
60	0.50	0.52	0.02	2%	0.09	92%
61	0.50	0.51	0.01	1%	0.09	94%
62	0.50	0.52	0.02	1%	0.10	96%
72	0.50	0.52	0.02	2%	0.10	92%
73	0.50	0.52	0.02	1%	0.10	96%
83	0.50	0.52	0.02	2%	0.11	96%
102	0.70	0.70	0.00	-1%	0.08	98%
103	0.70	0.70	0.00	-1%	0.08	97%
104	0.70	0.70	0.00	-2%	0.10	98%
110	0.70	0.70	0.00	-2%	0.08	98%
111	0.70	0.71	0.01	-1%	0.10	98%
117	0.70	0.70	0.00	-2%	0.09	98%
128	0.95	0.95	0.00	-1%	0.12	94%
129	0.95	0.95	0.00	-2%	0.13	94%
130	0.95	0.95	0.00	-2%	0.14	96%
132	0.95	0.95	0.00	-2%	0.12	95%
133	0.95	0.95	0.00	-2%	0.14	95%
135	0.95	0.95	0.00	-2%	0.13	95%

D-21: Results of the IC using both ITT HRs (for Group 1)

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
1	1.00	1.00	0.00	-2%	0.16	94%
17	1.00	1.01	0.01	-1%	0.17	94%
32	1.00	1.02	0.02	0%	0.17	95%
46	1.00	1.00	0.00	-2%	0.15	95%
59	1.00	1.01	0.01	0%	0.15	95%
71	1.00	1.00	0.00	-1%	0.14	96%
82	1.00	1.00	0.00	-1%	0.14	95%
92	1.00	1.02	0.02	0%	0.16	95%
101	1.00	1.01	0.01	-1%	0.15	94%
109	1.00	1.01	0.01	-1%	0.14	95%
116	1.00	1.01	0.01	-1%	0.14	96%
122	1.00	1.01	0.01	-1%	0.14	96%
127	1.00	1.01	0.01	-1%	0.14	96%
131	1.00	1.01	0.01	-1%	0.14	95%
134	1.00	1.02	0.02	0%	0.14	96%
136	1.00	1.01	0.01	-1%	0.14	95%

D-22: Results of the IC using both ITT HRs (for Group 2)

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
20	0.62	0.85	0.23	26%	0.36	42%
24	0.44	0.64	0.20	30%	0.29	25%
28	0.33	0.49	0.17	32%	0.23	18%
34	0.62	0.94	0.32	33%	0.45	20%
35	0.62	0.85	0.23	25%	0.35	44%
38	0.44	0.71	0.27	37%	0.37	10%
39	0.44	0.68	0.24	34%	0.34	16%
42	0.33	0.54	0.22	39%	0.29	5%
43	0.33	0.54	0.22	39%	0.29	4%
47	0.62	1.08	0.46	41%	0.61	3%
48	0.62	0.97	0.35	35%	0.49	13%
49	0.62	0.92	0.30	31%	0.43	23%
51	0.44	0.81	0.37	44%	0.48	1%
52	0.44	0.78	0.33	42%	0.45	4%
53	0.44	0.75	0.31	40%	0.42	4%
55	0.33	0.62	0.29	46%	0.38	0%
56	0.33	0.62	0.30	47%	0.39	1%
57	0.33	0.61	0.29	46%	0.38	1%

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
74	0.71	0.84	0.13	14%	0.24	79%
78	0.53	0.65	0.12	17%	0.21	67%
84	0.71	0.90	0.18	19%	0.30	64%
85	0.71	0.86	0.15	15%	0.26	75%
88	0.53	0.68	0.16	22%	0.25	52%
89	0.53	0.68	0.15	21%	0.25	56%
93	0.71	0.96	0.25	24%	0.38	43%
94	0.71	0.92	0.21	21%	0.34	55%
95	0.71	0.90	0.19	20%	0.31	60%
97	0.53	0.73	0.21	27%	0.31	31%
98	0.53	0.74	0.21	27%	0.31	33%
99	0.53	0.73	0.20	27%	0.30	35%
112	0.74	0.81	0.07	7%	0.17	90%
118	0.74	0.82	0.08	9%	0.19	89%
119	0.74	0.82	0.09	9%	0.19	88%
123	0.74	0.86	0.12	13%	0.23	81%
124	0.74	0.85	0.11	12%	0.23	82%
125	0.74	0.84	0.10	11%	0.22	85%

D-23: Results of the IC using both ITT HRs (for Group 3)

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
5	0.62	0.70	0.08	10%	0.18	90%
9	0.44	0.53	0.09	15%	0.17	77%
13	0.33	0.41	0.08	18%	0.14	66%
21	0.62	0.77	0.15	18%	0.25	68%
25	0.44	0.61	0.17	27%	0.25	36%
29	0.33	0.49	0.16	32%	0.23	17%
36	0.62	0.81	0.19	22%	0.31	54%
40	0.44	0.67	0.23	32%	0.32	18%
44	0.33	0.55	0.22	39%	0.30	5%
50	0.62	0.85	0.23	26%	0.35	42%
54	0.44	0.74	0.29	39%	0.39	6%
58	0.33	0.62	0.29	46%	0.37	1%
63	0.71	0.77	0.05	5%	0.16	91%
67	0.53	0.58	0.06	8%	0.13	90%
75	0.71	0.81	0.09	10%	0.20	88%
79	0.53	0.65	0.12	18%	0.21	67%

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
86	0.71	0.84	0.12	13%	0.24	79%
90	0.53	0.68	0.15	21%	0.24	53%
96	0.71	0.87	0.15	16%	0.27	71%
100	0.53	0.73	0.20	26%	0.30	35%
105	0.74	0.77	0.04	3%	0.14	95%
113	0.74	0.80	0.07	7%	0.17	91%
120	0.74	0.81	0.07	8%	0.18	90%
126	0.74	0.84	0.10	11%	0.22	84%

D-24: Results of the IC using both ITT HRs (for Group 4)

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
6	0.62	0.70	0.08	10%	0.18	90%
7	0.62	0.53	0.09	15%	0.17	77%
8	0.62	0.41	0.08	18%	0.14	66%
10	0.44	0.77	0.15	18%	0.25	68%
11	0.44	0.61	0.17	27%	0.25	36%
12	0.44	0.49	0.16	32%	0.23	17%
14	0.33	0.81	0.19	22%	0.31	54%
15	0.33	0.67	0.23	32%	0.32	18%
16	0.33	0.55	0.22	39%	0.30	5%
22	0.62	0.85	0.23	26%	0.35	42%
23	0.62	0.74	0.29	39%	0.39	6%
26	0.44	0.62	0.29	46%	0.37	1%
27	0.44	0.77	0.05	5%	0.16	91%
30	0.33	0.58	0.06	8%	0.13	90%
31	0.33	0.81	0.09	10%	0.20	88%
37	0.62	0.65	0.12	18%	0.21	67%
41	0.44	0.84	0.12	13%	0.24	79%
45	0.33	0.68	0.15	21%	0.24	53%
64	0.71	0.87	0.15	16%	0.27	71%
65	0.71	0.73	0.20	26%	0.30	35%
66	0.71	0.77	0.04	3%	0.14	95%
68	0.53	0.80	0.07	7%	0.17	91%
69	0.53	0.81	0.07	8%	0.18	90%
70	0.53	0.84	0.10	11%	0.22	84%
76	0.71	0.70	0.08	10%	0.18	90%
77	0.71	0.53	0.09	15%	0.17	77%
80	0.53	0.41	0.08	18%	0.14	66%

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
81	0.53	0.77	0.15	18%	0.25	68%
87	0.71	0.61	0.17	27%	0.25	36%
91	0.53	0.49	0.16	32%	0.23	17%
106	0.74	0.81	0.19	22%	0.31	54%
107	0.74	0.67	0.23	32%	0.32	18%
108	0.74	0.55	0.22	39%	0.30	5%
114	0.74	0.85	0.23	26%	0.35	42%
115	0.74	0.74	0.29	39%	0.39	6%
121	0.74	0.62	0.29	46%	0.37	1%

D-25: Results of the IC using both ITT HRs (for Group 5)

Scenario	IC calculated from the true underlying HRs	Both ITT HR				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
2	1.00	0.84	-0.16	-22%	-0.04	74%
3	1.00	0.76	-0.24	-35%	-0.13	49%
4	1.00	0.66	-0.34	-55%	-0.24	17%
18	1.00	0.92	-0.08	-12%	0.05	91%
19	1.00	0.80	-0.20	-27%	-0.08	66%
33	1.00	0.88	-0.12	-16%	0.01	84%
60	1.00	0.91	-0.09	-12%	0.04	88%
61	1.00	0.87	-0.13	-18%	-0.01	80%
62	1.00	0.80	-0.20	-27%	-0.09	63%
72	1.00	0.95	-0.05	-7%	0.08	93%
73	1.00	0.90	-0.10	-14%	0.02	86%
83	1.00	0.94	-0.06	-9%	0.06	92%
102	1.00	0.96	-0.04	-6%	0.08	94%
103	1.00	0.94	-0.06	-8%	0.07	92%
104	1.00	0.92	-0.08	-11%	0.04	88%
110	1.00	0.98	-0.02	-3%	0.11	95%
111	1.00	0.95	-0.05	-7%	0.07	93%
117	1.00	0.98	-0.02	-4%	0.10	95%
128	1.00	1.00	0.00	-1%	0.14	95%
129	1.00	1.00	0.00	-1%	0.14	95%
130	1.00	1.00	0.00	-2%	0.13	95%
132	1.00	1.01	0.01	-1%	0.14	95%
133	1.00	1.00	0.00	-1%	0.13	95%
135	1.00	1.00	0.00	-1%	0.13	95%

D-26: Results of the IC using RPSFTM HR for Study A and ITT HR for Study B (for Group 1)

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
1	1.00	0.84	-0.16	-24%	0.00	74%
17	1.00	0.70	-0.30	-50%	-0.16	47%
32	1.00	0.63	-0.37	-65%	-0.23	36%
46	1.00	0.55	-0.45	-93%	-0.32	27%
59	1.00	0.95	-0.05	-8%	0.10	92%
71	1.00	0.85	-0.15	-21%	-0.02	81%
82	1.00	0.81	-0.19	-28%	-0.05	74%
92	1.00	0.76	-0.24	-37%	-0.09	70%
101	1.00	0.97	-0.03	-5%	0.11	95%
109	1.00	0.93	-0.07	-10%	0.06	93%
116	1.00	0.90	-0.10	-13%	0.04	92%
122	1.00	0.88	-0.12	-17%	0.02	90%
127	1.00	1.01	0.01	-1%	0.14	96%
131	1.00	1.00	0.00	-2%	0.15	95%
134	1.00	1.01	0.01	-2%	0.16	95%
136	1.00	1.00	0.00	-3%	0.18	94%

D-27: Results of the IC using RPSFTM HR for Study A and ITT HR for Study B (for Group 2)

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
20	0.62	0.59	-0.03	-10%	0.09	91%
24	0.44	0.44	0.00	-5%	0.08	93%
28	0.33	0.34	0.01	0%	0.08	93%
34	0.62	0.59	-0.03	-10%	0.08	94%
35	0.62	0.53	-0.09	-22%	0.02	85%
38	0.44	0.44	0.00	-4%	0.09	94%
39	0.44	0.42	-0.02	-10%	0.07	92%
42	0.33	0.34	0.01	0%	0.08	95%
43	0.33	0.34	0.01	-1%	0.08	95%
47	0.62	0.59	-0.03	-10%	0.11	96%
48	0.62	0.54	-0.08	-22%	0.04	91%
49	0.62	0.50	-0.12	-30%	0.00	85%
51	0.44	0.45	0.00	-4%	0.10	97%
52	0.44	0.42	-0.02	-11%	0.08	94%
53	0.44	0.41	-0.03	-13%	0.07	95%

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
55	0.33	0.34	0.01	-1%	0.09	96%
56	0.33	0.34	0.02	0%	0.10	96%
57	0.33	0.34	0.01	-2%	0.09	96%
74	0.71	0.71	0.00	-3%	0.11	94%
78	0.53	0.55	0.02	1%	0.11	93%
84	0.71	0.72	0.00	-2%	0.13	95%
85	0.71	0.69	-0.03	-7%	0.09	94%
88	0.53	0.55	0.02	1%	0.12	94%
89	0.53	0.55	0.02	1%	0.11	95%
93	0.71	0.72	0.00	-3%	0.14	95%
94	0.71	0.69	-0.02	-8%	0.11	95%
95	0.71	0.68	-0.04	-10%	0.09	94%
97	0.53	0.55	0.02	0%	0.13	96%
98	0.53	0.55	0.03	1%	0.13	95%
99	0.53	0.54	0.02	-1%	0.13	95%
112	0.74	0.74	0.00	-2%	0.11	97%
118	0.74	0.74	0.00	-2%	0.11	97%
119	0.74	0.74	0.00	-1%	0.11	97%
123	0.74	0.75	0.01	-1%	0.13	97%
124	0.74	0.74	0.00	-2%	0.13	97%
125	0.74	0.73	-0.01	-4%	0.12	96%

D-28: Results of the IC using RPSFTM HR for Study A and ITT HR for Study B (for Group 3)

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
5	0.62	0.58	-0.04	-10%	0.06	90%
9	0.44	0.44	0.00	-3%	0.08	91%
13	0.33	0.34	0.01	1%	0.07	90%
21	0.62	0.53	-0.09	-21%	0.01	84%
25	0.44	0.42	-0.02	-9%	0.06	93%
29	0.33	0.34	0.01	0%	0.07	93%
36	0.62	0.51	-0.11	-27%	-0.01	82%
40	0.44	0.42	-0.03	-11%	0.06	93%
44	0.33	0.34	0.02	0%	0.09	93%
50	0.62	0.46	-0.16	-40%	-0.05	77%
54	0.44	0.40	-0.04	-15%	0.05	94%
58	0.33	0.34	0.01	-1%	0.09	97%
63	0.71	0.72	0.00	-2%	0.12	93%

L

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
67	0.53	0.55	0.02	2%	0.10	93%
75	0.71	0.68	-0.04	-8%	0.07	93%
79	0.53	0.55	0.02	2%	0.11	94%
86	0.71	0.67	-0.04	-10%	0.07	92%
90	0.53	0.54	0.02	0%	0.11	95%
96	0.71	0.65	-0.07	-15%	0.05	92%
100	0.53	0.54	0.02	0%	0.12	96%
105	0.74	0.74	0.01	-1%	0.11	96%
113	0.74	0.74	0.00	-2%	0.11	97%
120	0.74	0.73	-0.01	-3%	0.10	97%
126	0.74	0.73	-0.01	-4%	0.12	96%

D-29: Results of the IC using RPSFTM HR for Study A and ITT HR for Study B (for Group 4)

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
6	0.62	0.53	-0.09	-22%	0.00	76%
7	0.62	0.50	-0.12	-30%	-0.03	64%
8	0.62	0.46	-0.16	-39%	-0.07	48%
10	0.44	0.42	-0.02	-9%	0.05	89%
11	0.44	0.42	-0.02	-9%	0.05	87%
12	0.44	0.40	-0.04	-13%	0.03	86%
14	0.33	0.34	0.01	1%	0.08	90%
15	0.33	0.34	0.01	0%	0.07	90%
16	0.33	0.34	0.01	0%	0.07	89%
22	0.62	0.51	-0.11	-27%	-0.02	77%
23	0.62	0.46	-0.16	-39%	-0.07	61%
26	0.44	0.41	-0.03	-11%	0.05	91%
27	0.44	0.40	-0.04	-15%	0.03	90%
30	0.33	0.34	0.01	0%	0.08	93%
31	0.33	0.33	0.01	-2%	0.08	90%
37	0.62	0.46	-0.16	-39%	-0.06	68%
41	0.44	0.40	-0.04	-16%	0.04	91%
45	0.33	0.34	0.01	-1%	0.08	94%
64	0.71	0.68	-0.03	-7%	0.07	92%
65	0.71	0.67	-0.04	-9%	0.06	91%
66	0.71	0.65	-0.07	-13%	0.03	85%
68	0.53	0.55	0.02	2%	0.10	92%
69	0.53	0.54	0.02	1%	0.10	93%

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
70	0.53	0.54	0.01	0%	0.09	93%
76	0.71	0.68	-0.04	-8%	0.08	92%
77	0.71	0.65	-0.06	-12%	0.04	91%
80	0.53	0.55	0.02	1%	0.11	94%
81	0.53	0.54	0.02	0%	0.10	95%
87	0.71	0.64	-0.07	-14%	0.04	90%
91	0.53	0.54	0.02	0%	0.11	95%
106	0.74	0.74	0.00	-1%	0.10	96%
107	0.74	0.74	0.00	-2%	0.10	96%
108	0.74	0.74	0.00	-2%	0.10	96%
114	0.74	0.73	0.00	-3%	0.10	97%
115	0.74	0.73	-0.01	-3%	0.09	98%
121	0.74	0.73	-0.01	-3%	0.10	98%

D-30: Results of the IC using RPSFTM HR for Study A and ITT HR for Study B (for Group 5)

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
2	1.00	0.70	-0.30	-48%	-0.17	38%
3	1.00	0.63	-0.37	-63%	-0.25	18%
4	1.00	0.55	-0.45	-87%	-0.35	4%
18	1.00	0.62	-0.38	-67%	-0.25	27%
19	1.00	0.55	-0.45	-89%	-0.34	10%
33	1.00	0.55	-0.45	-92%	-0.33	13%
60	1.00	0.86	-0.14	-20%	-0.01	78%
61	1.00	0.81	-0.19	-26%	-0.06	66%
62	1.00	0.75	-0.25	-36%	-0.13	48%
72	1.00	0.80	-0.20	-28%	-0.06	69%
73	1.00	0.76	-0.24	-36%	-0.12	57%
83	1.00	0.75	-0.25	-37%	-0.12	60%
102	1.00	0.92	-0.08	-11%	0.03	92%
103	1.00	0.91	-0.09	-12%	0.02	89%
104	1.00	0.88	-0.12	-16%	0.00	85%
110	1.00	0.90	-0.10	-13%	0.03	91%
111	1.00	0.87	-0.13	-17%	-0.01	86%
117	1.00	0.88	-0.12	-17%	0.01	89%
128	1.00	1.00	0.00	-2%	0.14	95%
129	1.00	1.00	0.00	-2%	0.14	94%
130	1.00	1.00	0.00	-2%	0.13	95%

Scenario	IC calculated from the true underlying HRs	Adjusted Study A & ITT Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
132	1.00	1.00	0.00	-2%	0.15	95%
133	1.00	0.99	-0.01	-3%	0.14	94%
135	1.00	0.99	-0.01	-3%	0.15	94%

D-31: Results of the IC using ITT HR for Study A and RPSFTM HR for Study B (for Group 1)

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
1	0.21	15%	0.45	77%	0.21	15%
17	0.50	31%	0.83	48%	0.50	31%
32	0.68	38%	1.04	31%	0.68	38%
46	0.87	44%	1.31	30%	0.87	44%
59	0.09	6%	0.26	91%	0.09	6%
71	0.20	15%	0.39	84%	0.20	15%
82	0.27	19%	0.51	76%	0.27	19%
92	0.39	25%	0.67	66%	0.39	25%
101	0.06	3%	0.20	94%	0.06	3%
109	0.10	7%	0.26	94%	0.10	7%
116	0.13	9%	0.29	92%	0.13	9%
122	0.17	12%	0.35	91%	0.17	12%
127	0.02	0%	0.15	96%	0.02	0%
131	0.02	0%	0.17	94%	0.02	0%
134	0.04	1%	0.19	95%	0.04	1%
136	0.03	1%	0.21	95%	0.03	1%

D-32: Results of the IC using ITT HR for Study A and RPSFTM HR for Study B (for Group 2)

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
20	0.62	1.21	0.21	15%	0.45	77%
24	0.44	1.50	0.50	31%	0.83	48%
28	0.33	1.68	0.68	38%	1.04	31%
34	0.62	1.87	0.87	44%	1.31	30%
35	0.62	1.09	0.09	6%	0.26	91%
38	0.44	1.20	0.20	15%	0.39	84%
39	0.44	1.27	0.27	19%	0.51	76%
42	0.33	1.39	0.39	25%	0.67	66%

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
43	0.33	1.06	0.06	3%	0.20	94%
47	0.62	1.10	0.10	7%	0.26	94%
48	0.62	1.13	0.13	9%	0.29	92%
49	0.62	1.17	0.17	12%	0.35	91%
51	0.44	1.02	0.02	0%	0.15	96%
52	0.44	1.02	0.02	0%	0.17	94%
53	0.44	1.04	0.04	1%	0.19	95%
55	0.33	1.03	0.03	1%	0.21	95%
56	0.33	0.92	0.30	31%	0.45	29%
57	0.33	0.67	0.22	32%	0.32	22%
74	0.71	0.49	0.17	33%	0.24	24%
78	0.53	1.01	0.39	37%	0.55	12%
84	0.71	1.02	0.40	37%	0.58	19%
85	0.71	0.74	0.30	39%	0.41	8%
88	0.53	0.75	0.30	39%	0.42	13%
89	0.53	0.55	0.22	39%	0.30	8%
93	0.71	0.55	0.22	39%	0.31	14%
94	0.71	1.16	0.54	45%	0.72	2%
95	0.71	1.16	0.54	45%	0.75	4%
97	0.53	1.16	0.54	45%	0.75	8%
98	0.53	0.84	0.40	46%	0.52	1%
99	0.53	0.85	0.40	46%	0.54	3%
112	0.74	0.84	0.40	46%	0.53	4%
118	0.74	0.62	0.30	46%	0.39	1%
119	0.74	0.63	0.31	47%	0.41	3%
123	0.74	0.62	0.30	46%	0.40	7%
124	0.74	0.88	0.16	17%	0.28	74%
125	0.74	0.65	0.13	18%	0.21	69%

D-33: Results of the IC using ITT HR for Study A and RPSFTM HR for Study B (for Group 3)

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
5	0.62	0.75	0.13	16%	0.25	78%
9	0.44	0.55	0.11	18%	0.19	71%
13	0.33	0.41	0.08	19%	0.14	68%
21	0.62	0.92	0.30	31%	0.44	36%
25	0.44	0.67	0.23	33%	0.32	28%
29	0.33	0.50	0.17	33%	0.25	31%
36	0.62	1.03	0.41	38%	0.59	22%
40	0.44	0.75	0.31	39%	0.42	14%

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
44	0.33	0.56	0.23	40%	0.32	16%
50	0.62	1.16	0.54	45%	0.77	15%
54	0.44	0.85	0.41	47%	0.55	7%
58	0.33	0.63	0.30	47%	0.41	9%
63	0.71	0.80	0.08	9%	0.20	89%
67	0.53	0.59	0.06	9%	0.14	89%
75	0.71	0.88	0.17	17%	0.29	77%
79	0.53	0.66	0.13	18%	0.23	72%
86	0.71	0.94	0.22	22%	0.37	66%
90	0.53	0.69	0.17	22%	0.28	62%
96	0.71	1.01	0.29	27%	0.46	56%
100	0.53	0.75	0.22	27%	0.35	54%
105	0.74	0.78	0.04	4%	0.15	95%
113	0.74	0.82	0.08	8%	0.20	91%
120	0.74	0.82	0.09	9%	0.21	90%
126	0.74	0.86	0.12	12%	0.27	86%

D-34: Results of the IC using ITT HR for Study A and RPSFTM HR for Study B (for Group 4)

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
6	0.62	0.75	0.13	15%	0.27	82%
7	0.62	0.75	0.13	15%	0.28	83%
8	0.62	0.75	0.13	14%	0.29	88%
10	0.44	0.55	0.11	18%	0.19	76%
11	0.44	0.56	0.12	19%	0.20	74%
12	0.44	0.56	0.12	19%	0.21	79%
14	0.33	0.41	0.09	19%	0.15	70%
15	0.33	0.41	0.09	19%	0.15	74%
16	0.33	0.41	0.09	19%	0.16	77%
22	0.62	0.93	0.31	31%	0.48	42%
23	0.62	0.92	0.30	30%	0.48	54%
26	0.44	0.67	0.23	32%	0.33	34%
27	0.44	0.67	0.23	32%	0.35	41%
30	0.33	0.50	0.18	33%	0.26	32%
31	0.33	0.50	0.17	33%	0.27	42%
37	0.62	1.02	0.40	37%	0.60	35%
41	0.44	0.74	0.30	39%	0.42	24%
45	0.33	0.55	0.23	39%	0.33	25%
64	0.71	0.80	0.08	9%	0.20	92%
65	0.71	0.81	0.09	9%	0.21	92%

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
66	0.71	0.80	0.09	9%	0.22	93%
68	0.53	0.59	0.07	9%	0.16	87%
69	0.53	0.59	0.06	9%	0.16	90%
70	0.53	0.59	0.06	8%	0.17	90%
76	0.71	0.90	0.18	19%	0.32	76%
77	0.71	0.90	0.18	18%	0.33	79%
80	0.53	0.66	0.13	18%	0.24	73%
81	0.53	0.66	0.13	18%	0.25	77%
87	0.71	0.94	0.22	22%	0.38	73%
91	0.53	0.69	0.17	22%	0.28	68%
106	0.74	0.78	0.04	3%	0.16	95%
107	0.74	0.78	0.04	3%	0.17	94%
108	0.74	0.79	0.05	3%	0.19	94%
114	0.74	0.81	0.08	7%	0.20	91%
115	0.74	0.81	0.07	7%	0.21	93%
121	0.74	0.83	0.10	9%	0.24	90%

D-35: Results of the IC using ITT HR for Study A and RPSFTM HR for Study B (for Group 5)

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
2	1.00	1.24	0.24	16%	0.49	82%
3	1.00	1.24	0.24	16%	0.52	84%
4	1.00	1.25	0.25	16%	0.54	90%
18	1.00	1.50	0.50	30%	0.83	54%
19	1.00	1.51	0.51	30%	0.86	69%
33	1.00	1.66	0.66	37%	1.07	52%
60	1.00	1.09	0.09	5%	0.27	92%
61	1.00	1.10	0.10	6%	0.30	93%
62	1.00	1.10	0.10	6%	0.31	95%
72	1.00	1.20	0.20	14%	0.42	84%
73	1.00	1.22	0.22	15%	0.46	86%
83	1.00	1.28	0.28	19%	0.53	81%
102	1.00	1.04	0.04	2%	0.19	97%
103	1.00	1.06	0.06	3%	0.21	97%
104	1.00	1.06	0.06	3%	0.24	96%
110	1.00	1.10	0.10	7%	0.26	93%
111	1.00	1.10	0.10	6%	0.27	95%
117	1.00	1.13	0.13	10%	0.31	93%
128	1.00	1.02	0.02	-1%	0.17	95%
129	1.00	1.02	0.02	0%	0.19	94%

Scenario	IC calculated from the true underlying HRs	ITT Study A & Adjusted Study B				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
130	1.00	1.02	0.02	-1%	0.20	95%
132	1.00	1.03	0.03	0%	0.19	95%
133	1.00	1.03	0.03	0%	0.20	96%
135	1.00	1.03	0.03	0%	0.20	95%

D-36: Results of the IC using both RPSFTM HRs (for Group 1)

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
1	1.00	1.01	0.01	-4%	0.24	87%
17	1.00	1.03	0.03	-3%	0.30	92%
32	1.00	1.04	0.04	-2%	0.31	95%
46	1.00	1.02	0.02	-7%	0.33	97%
59	1.00	1.02	0.02	-1%	0.20	91%
71	1.00	1.01	0.01	-2%	0.20	95%
82	1.00	1.02	0.02	-3%	0.24	94%
92	1.00	1.04	0.04	-2%	0.30	97%
101	1.00	1.01	0.01	-1%	0.15	97%
109	1.00	1.01	0.01	-1%	0.16	97%
116	1.00	1.01	0.01	-1%	0.18	98%
122	1.00	1.02	0.02	-1%	0.21	98%
127	1.00	1.01	0.01	-1%	0.16	96%
131	1.00	1.01	0.01	-2%	0.18	95%
134	1.00	1.03	0.03	0%	0.21	96%
136	1.00	1.03	0.03	-2%	0.24	95%

D-37: Results of the IC using both RPSFTM HRs (for Group 2)

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
20	0.62	0.63	0.01	-3%	0.15	90%
24	0.44	0.46	0.01	-1%	0.10	93%
28	0.33	0.34	0.02	1%	0.08	93%
34	0.62	0.63	0.01	-3%	0.15	94%
35	0.62	0.64	0.02	-3%	0.17	94%
38	0.44	0.46	0.02	0%	0.11	96%
39	0.44	0.46	0.02	0%	0.12	95%
42	0.33	0.34	0.01	0%	0.09	95%
43	0.33	0.34	0.02	0%	0.09	95%

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
47	0.62	0.64	0.02	-3%	0.17	96%
48	0.62	0.64	0.02	-3%	0.18	96%
49	0.62	0.63	0.01	-4%	0.18	96%
51	0.44	0.46	0.02	0%	0.13	97%
52	0.44	0.46	0.02	-2%	0.13	95%
53	0.44	0.46	0.02	-1%	0.13	97%
55	0.33	0.34	0.01	-1%	0.09	97%
56	0.33	0.35	0.02	0%	0.11	95%
57	0.33	0.34	0.02	-1%	0.10	96%
74	0.71	0.74	0.03	1%	0.15	95%
78	0.53	0.55	0.02	2%	0.11	94%
84	0.71	0.75	0.03	2%	0.16	95%
85	0.71	0.75	0.04	2%	0.17	97%
88	0.53	0.55	0.03	2%	0.12	94%
89	0.53	0.55	0.03	1%	0.13	95%
93	0.71	0.75	0.03	1%	0.18	96%
94	0.71	0.76	0.04	1%	0.20	96%
95	0.71	0.76	0.04	1%	0.20	97%
97	0.53	0.55	0.03	1%	0.14	96%
98	0.53	0.56	0.03	2%	0.15	96%
99	0.53	0.55	0.03	1%	0.15	95%
112	0.74	0.74	0.01	-1%	0.11	97%
118	0.74	0.74	0.00	-2%	0.12	97%
119	0.74	0.75	0.02	-1%	0.14	97%
123	0.74	0.75	0.01	-1%	0.14	97%
124	0.74	0.75	0.01	-1%	0.15	97%
125	0.74	0.74	0.01	-2%	0.15	96%

D-38: Results of the IC using both RPSFTM HRs (for Group 3)

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
5	0.62	0.62	0.00	-3%	0.12	90%
9	0.44	0.46	0.02	1%	0.10	91%
13	0.33	0.34	0.02	2%	0.08	91%
21	0.62	0.63	0.01	-2%	0.14	94%
25	0.44	0.46	0.02	0%	0.10	96%
29	0.33	0.34	0.01	0%	0.08	93%
36	0.62	0.64	0.02	-2%	0.17	94%
40	0.44	0.47	0.03	1%	0.13	95%

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
44	0.33	0.35	0.02	2%	0.10	92%
50	0.62	0.63	0.01	-4%	0.18	97%
54	0.44	0.47	0.03	0%	0.14	97%
58	0.33	0.35	0.02	0%	0.11	96%
63	0.71	0.75	0.03	2%	0.15	94%
67	0.53	0.55	0.02	2%	0.11	93%
75	0.71	0.74	0.03	1%	0.16	95%
79	0.53	0.56	0.03	3%	0.13	94%
86	0.71	0.75	0.04	1%	0.18	96%
90	0.53	0.55	0.03	1%	0.13	95%
96	0.71	0.75	0.03	0%	0.19	97%
100	0.53	0.56	0.03	1%	0.15	95%
105	0.74	0.75	0.01	0%	0.12	96%
113	0.74	0.75	0.01	-1%	0.13	97%
120	0.74	0.74	0.00	-2%	0.13	96%
126	0.74	0.75	0.01	-2%	0.16	97%

D-39: Results of the IC using both RPSFTM HRs (for Group 4)

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
6	0.62	0.63	0.01	-2%	0.14	92%
7	0.62	0.63	0.01	-4%	0.15	91%
8	0.62	0.63	0.01	-4%	0.16	92%
10	0.44	0.46	0.02	0%	0.10	93%
11	0.44	0.47	0.03	2%	0.12	92%
12	0.44	0.47	0.03	1%	0.12	94%
14	0.33	0.35	0.02	2%	0.09	91%
15	0.33	0.34	0.02	1%	0.09	91%
16	0.33	0.35	0.02	2%	0.09	91%
22	0.62	0.64	0.02	-1%	0.16	94%
23	0.62	0.63	0.01	-4%	0.16	95%
26	0.44	0.46	0.02	0%	0.11	95%
27	0.44	0.46	0.02	0%	0.12	95%
30	0.33	0.35	0.02	1%	0.09	92%
31	0.33	0.34	0.02	-1%	0.10	92%
37	0.62	0.64	0.02	-4%	0.17	95%
41	0.44	0.46	0.02	-1%	0.12	96%
45	0.33	0.35	0.02	0%	0.10	95%
64	0.71	0.75	0.03	2%	0.15	95%

Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
65	0.71	0.75	0.04	3%	0.17	95%
66	0.71	0.75	0.04	2%	0.17	95%
68	0.53	0.56	0.03	3%	0.12	91%
69	0.53	0.55	0.03	2%	0.12	94%
70	0.53	0.55	0.02	0%	0.13	93%
76	0.71	0.76	0.05	3%	0.18	95%
77	0.71	0.76	0.04	2%	0.19	96%
80	0.53	0.56	0.03	2%	0.13	94%
81	0.53	0.56	0.03	2%	0.14	95%
87	0.71	0.75	0.03	1%	0.18	97%
91	0.53	0.56	0.03	1%	0.14	95%
106	0.74	0.75	0.01	-1%	0.12	96%
107	0.74	0.75	0.01	-1%	0.13	96%
108	0.74	0.76	0.02	-1%	0.15	96%
114	0.74	0.75	0.01	-2%	0.13	97%
115	0.74	0.75	0.01	-2%	0.14	97%
121	0.74	0.75	0.01	-2%	0.15	97%

D-40: Results of the IC using both RPSFTM HRs (for Group 5)

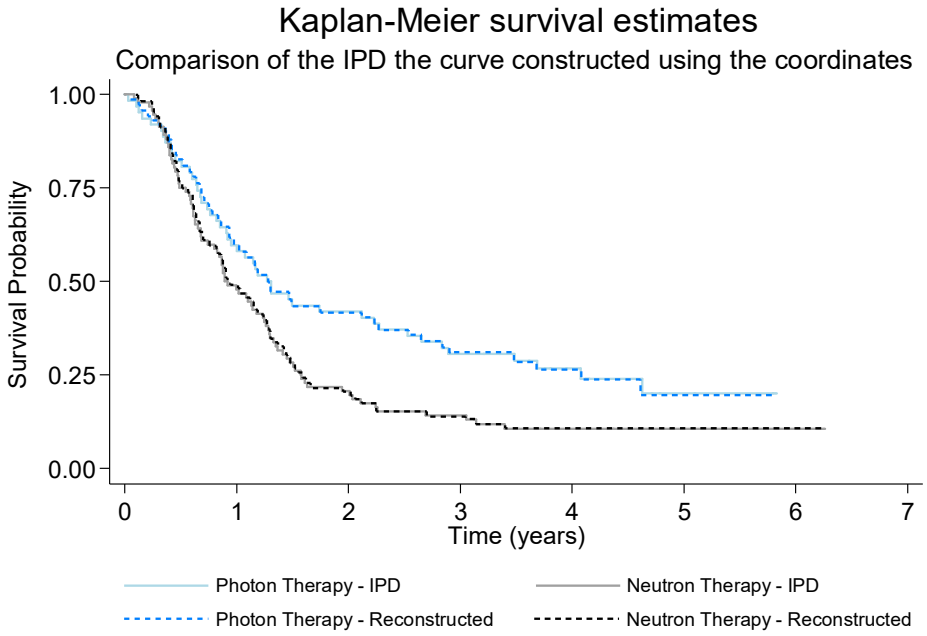
Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
2	1.00	1.04	0.04	-2%	0.28	91%
3	1.00	1.04	0.04	-2%	0.30	92%
4	1.00	1.05	0.05	-2%	0.32	94%
18	1.00	1.03	0.03	-4%	0.29	94%
19	1.00	1.03	0.03	-4%	0.31	96%
33	1.00	1.03	0.03	-5%	0.33	95%
60	1.00	1.02	0.02	-1%	0.21	93%
61	1.00	1.03	0.03	-1%	0.23	94%
62	1.00	1.03	0.03	-1%	0.24	96%
72	1.00	1.01	0.01	-3%	0.22	94%
73	1.00	1.04	0.04	-1%	0.26	97%
83	1.00	1.03	0.03	-2%	0.26	95%
102	1.00	1.00	0.00	-2%	0.14	99%
103	1.00	1.02	0.02	-1%	0.17	98%
104	1.00	1.02	0.02	-1%	0.19	97%
110	1.00	1.01	0.01	-1%	0.17	98%
111	1.00	1.01	0.01	-2%	0.17	98%
117	1.00	1.02	0.02	-1%	0.19	98%

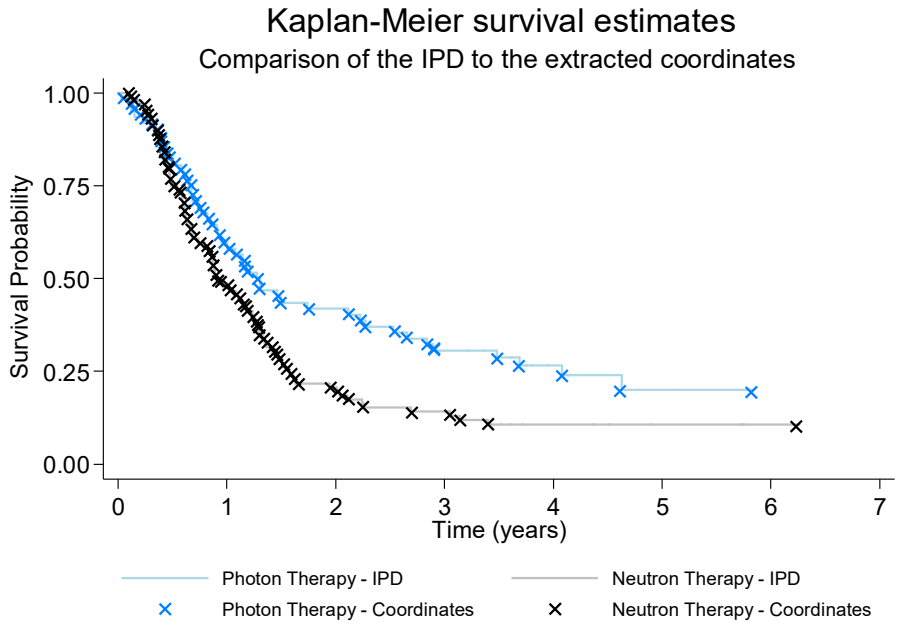
Scenario	IC calculated from the true underlying HRs	Both Adjusted				
		IC HR	Absolute bias	Proportional bias	MSE	Coverage
128	1.00	1.01	0.01	-1%	0.17	95%
129	1.00	1.02	0.02	-1%	0.19	94%
130	1.00	1.02	0.02	-1%	0.20	95%
132	1.00	1.02	0.02	-1%	0.19	95%
133	1.00	1.02	0.02	-2%	0.20	96%
135	1.00	1.02	0.02	-2%	0.21	95%

Appendix E: Illustrative Example: Coordinates and Model Fitting

E-1: Comparison of extracted coordinates to the IPD

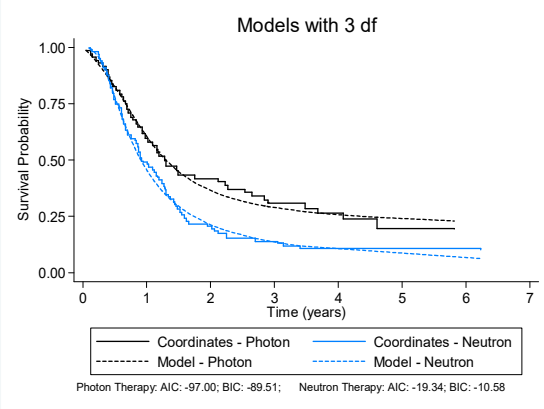
(top: extracted coordinates compared to IPD; bottom: Kaplan-Meier curve formed by joining up the coordinates)

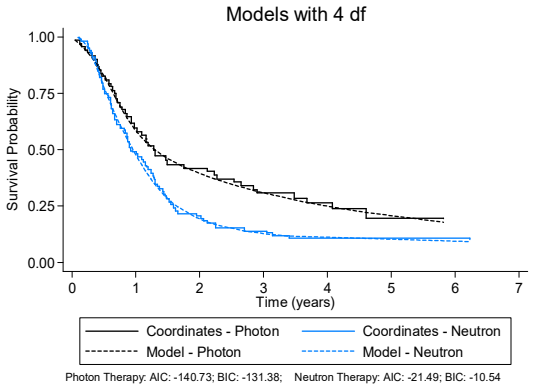


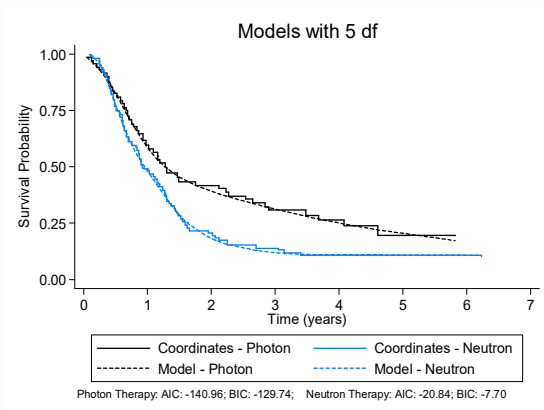


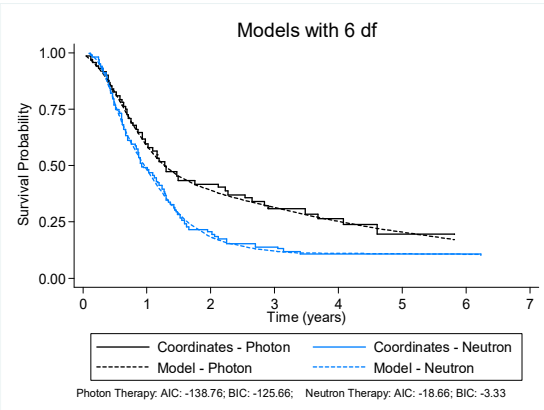
E-2: Comparison of models with different degrees of freedom

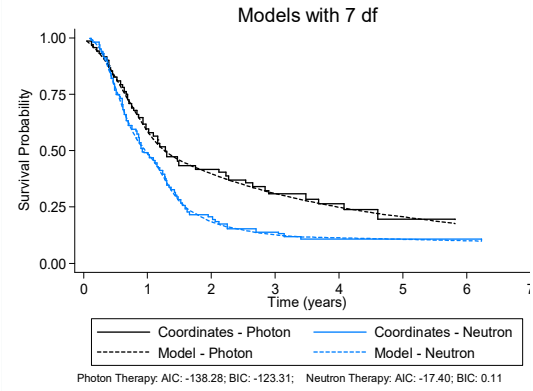
Model fit – showing model fit for 3, 4, 5, 6, 7, 8, 10 degrees of freedom. The knot locations used applied knots based on evenly spaced percentiles.

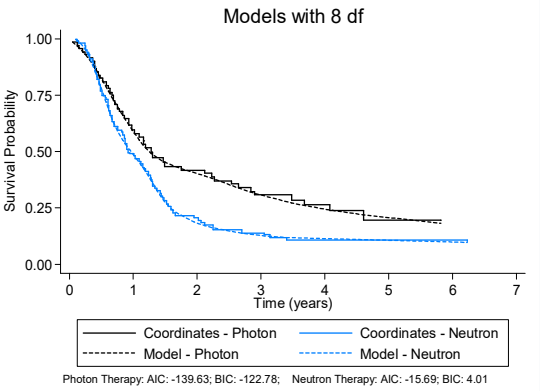


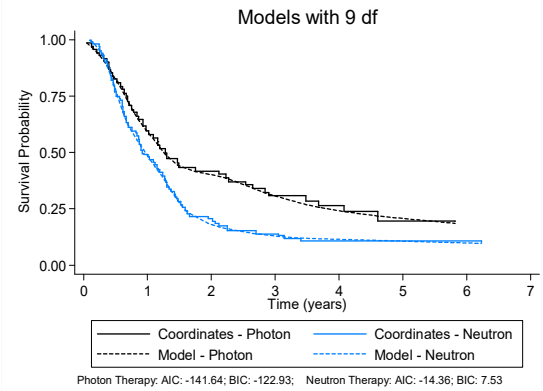


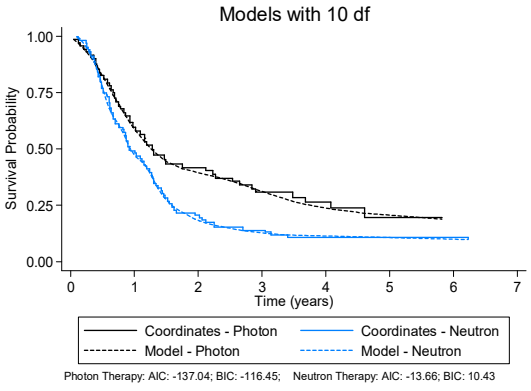












E-3: Model fit statistics

The follow table shows the model fit statistics using the AIC and BIC for the models fitted with different degrees of freedom.

Best models	Photon Therapy				Neutron Therapy			
	DF in model	AIC value	DF in model	BIC value	DF in model	AIC value	DF in model	BIC value
1	9	-141.6	4	-131.4	4	-21.5	3	-10.6
2	5	-141.0	5	-129.7	5	-20.8	4	-10.5
3	4	-140.7	6	-125.7	3	-19.3	5	-7.7
4	8	-139.6	7	-123.3	6	-18.7	6	-3.3
5	6	-138.8	9	-122.9	7	-17.4	7	0.1
6	7	-138.3	8	-122.8	8	-15.7	8	4.0
7	10	-137.0	10	-116.5	9	-14.4	9	7.5
8	3	-97.0	3	-89.51	10	-13.7	10	10.4

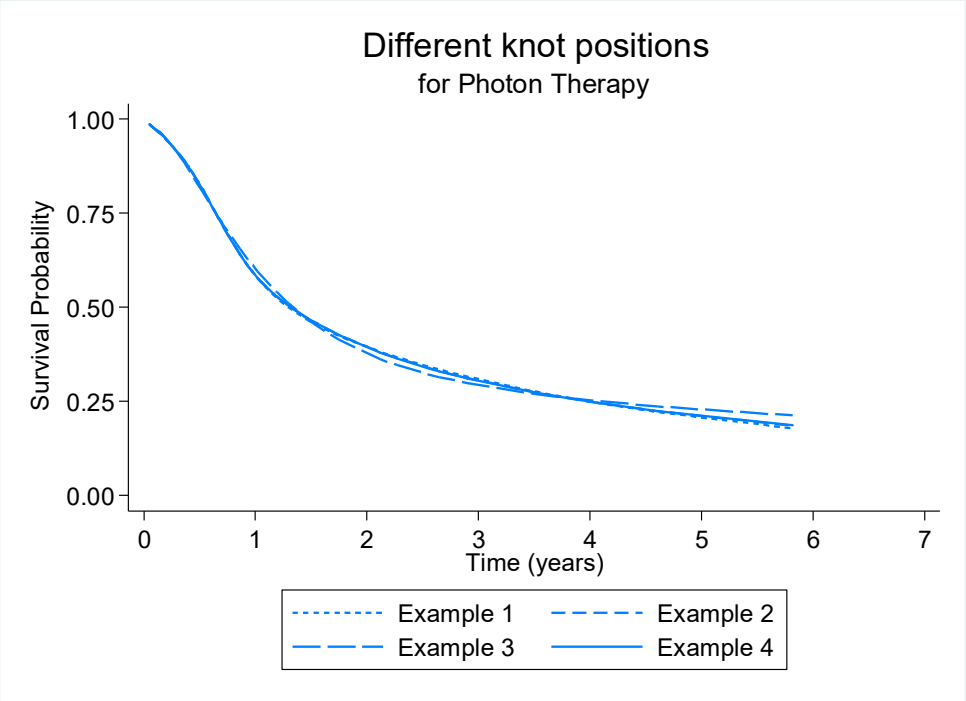
E-4: Discussion on the 'best model'

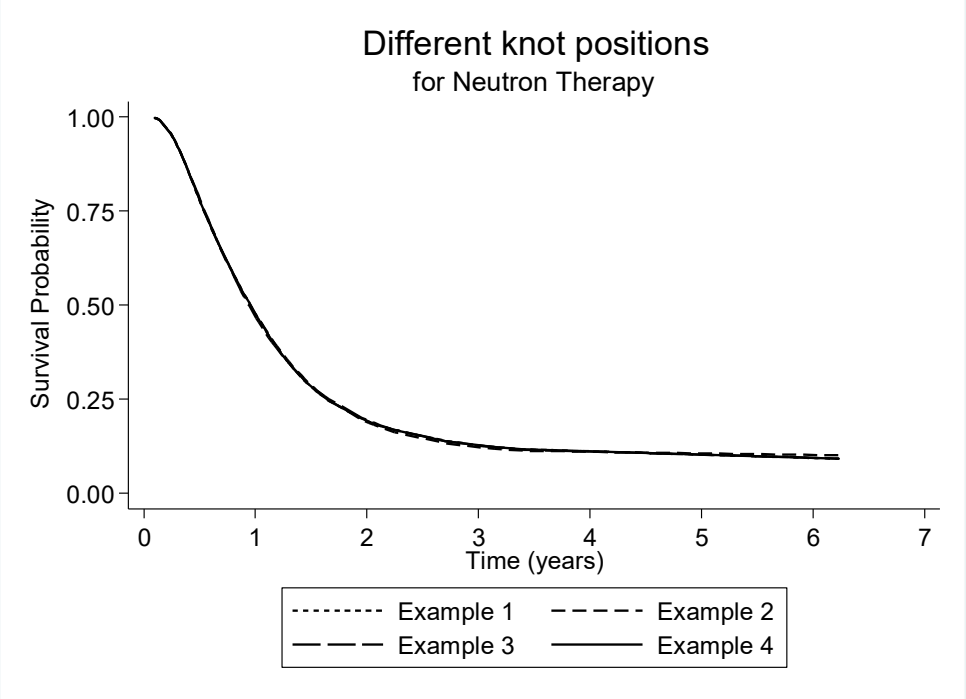
All the information from the table in D-3 was considered, along with visual inspection of the curves (D-2). Whilst the photon therapy, the AIC indicates the model with nine degrees of freedom to be the best, there are very little differences between this model and the models with four or five degrees of freedom. This is noticed within the BIC, which suggests that the extra complexity of the model with nine degrees is not necessary, highlighting the model with only four as the preferred. Since visual inspection shows the model with four degrees of freedom to fit well, and given relatively small gains from using more complex model, four degrees of freedom were selected for the final model.

For the neutron therapy group, the models with a lower number of degrees of freedom were selected both for the AIC and BIC. The model with four degrees of freedom is considered the best model from the AIC and second best by the BIC; this last criterion selecting the model with three degrees as being the best. However, on visual inspection the model with four degrees of freedom appears to fit the data more closely, and hence on this occasion the added complexity is justified.

Therefore, both treatment arms used models with four degrees of freedom.

E-5: Examination of different knot locations for 4 df





A variety of different knot positions were used. Based on visual inspection, for the Neutron Therapy group the change in knot positions made very different to the final curve, and most of the curves overlaid each other. For the photon group, one example, where the interior knot locations (i.e. knots 2, 3 and 4) were evenly distributed between

the boundary knots (fixed from the default values), gave some slightly more noticeable departures.

Example 1: Default knot positions – knot positions based on evenly spaced percentiles (e.g. at 20^{th}, 40^{th}, 60^{th} and 80^{th} percentile – double check)
Example 2: Default knot positions for the alternative therapy (e.g. Neutron therapy knot positions for the photon therapy group)
Example 3: Boundary knots were fixed and interior knots evenly distributed between them (calculated on the log-time scale)
Example 4: Knot positions were based on both sets of time coordinates to give a consistent estimate for both models

E-6: Replicating reported results for models with different dfs

Dataset	Results from Initial Analysis											
	Number of events				Median survival time				Hazard ratio		Restricted mean survival time at 3 ½ years			
	Photon		Neutron		Photon		Neutron				Photon		Neutron	
Original Individual Patient Data	47	(75.8%)	82	(89.1%)	1.27	(0.91, 2.27)	0.89	(0.74, 1.24)	1.57	(1.09, 2.25)	1.80	(1.47, 2.13)	1.29	(1.07, 1.50)
Risk table censoring distribution & models with:														
3 df	46.1	(74.4%)	83.6	(90.9%)	1.33	(0.89, 2.11)	0.95	(0.75, 1.16)	1.51	(1.05, 2.18)	1.75	(1.43, 2.06)	1.30	(1.10, 1.52)
5 df	47.3	(76.3%)	84.5	(91.8%)	1.40	(0.88, 2.45)	0.96	(0.75, 1.21)	1.55	(1.01, 2.42)	1.78	(1.46, 2.11)	1.22	(1.03, 1.43)
6 df *	47.3	(76.3%)	83.9	(91.2%)	1.31	(0.94, 1.86)	0.91	(0.72, 1.13)	1.61	(1.11, 2.37)	1.78	(1.46, 2.11)	1.25	(1.07, 1.46)
7 df	47.3	(76.3%)	82.2	(89.3%)	1.32	(0.90, 2.10)	0.92	(0.71, 1.15)	1.70	(1.17, 2.53)	1.78	(1.47, 2.11)	1.28	(1.08, 1.49)
8 df	47.3	(76.3%)	82.1	(89.2%)	1.31	(0.90, 2.07)	0.94	(0.72, 1.17)	1.64	(1.12, 2.41)	1.79	(1.47, 2.11)	1.28	(1.08, 1.49)
9 df	47.4	(76.5%)	82.1	(89.2%)	1.32	(0.90, 2.14)	0.96	(0.72, 1.21)	1.57	(1.09, 2.32)	1.78	(1.47, 2.10)	1.28	(1.08, 1.49)
10 df	47.3	(76.3%)	82.2	(89.3%)	1.31	(0.91, 2.22)	0.96	(0.71, 1.21)	1.57	(1.08, 2.30)	1.78	(1.47, 2.11)	1.28	(1.08, 1.49)

* This analysis was only conducted and averaged over 1999 datasets; for one dataset the maximum survival time for the Neutron therapy group did not exceed 3½ years, thus this dataset was excluded from

E-7: Secondary analysis for models with different dfs

Dataset	Results from Secondary Analysis									
	Coefficients from the time-dependent Cox model:				Hazard ratios from the Piecewise Cox model:					
	Main effects		Time dependent effect		0 and 6 months		6 and 12 months		12+ months	
Original Individual Patient Data	0.482	(-0.115, 0.850)	0.198	(-0.220, 0.617)	1.41	(0.69, 2.90)	1.39	(0.72, 2.68)	1.79	(1.04, 3.08)
Risk table censoring distribution & models with:										
3 df	0.529	(0.145, 0.923)	0.278	(-0.220, 0.743)	1.29	(0.59, 3.00)	1.79	(0.96, 3.57)	1.71	(0.99, 3.07)
5 df	0.539	(0.164, 0.941)	0.064	(-0.236, 0.608)	1.43	(0.66, 3.33)	1.48	(0.79, 2.94)	2.10	(1.21, 3.80)
6 df *	0.524	(0.142, 0.929)	0.175	(-0.229, 0.660)	1.37	(0.63, 3.19)	1.45	(0.77, 2.86)	2.00	(1.15, 3.56)
7 df	0.486	(0.107, 0.885)	0.170	(-0.313, 0.629)	1.31	(0.60, 3.00)	1.43	(0.76, 2.84)	1.89	(1.08, 3.39)
8 df	0.482	(0.108, 0.875)	0.165	(-0.323, 0.627)	1.30	(0.59, 2.98)	1.47	(0.79, 2.94)	1.84	(1.08, 3.30)
9 df	0.476	(0.102, 0.872)	0.161	(-0.321, 0.615)	1.30	(0.59, 3.00)	1.48	(0.80, 2.99)	1.80	(1.04, 3.28)
10 df	0.476	(0.104, 0.878)	0.166	(-0.321, 0.616)	1.34	(0.62, 3.07)	1.51	(0.80, 3.01)	1.77	(1.02, 3.15)

Appendix F: Reproducibility Study

F-1: Instructions for data extraction (Method A: Guyot; Method B: Simulation)

Where to take points: Methods

Method A.

1. Set up the dataset
2. Take points at the event times (i.e. at the bottom of each step in the curve – see Figure 1). Where there are dotted or thick lines, you may have to guess roughly where the event occurred.

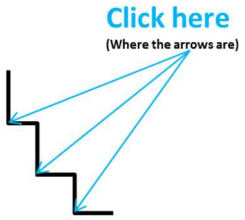


Figure 1: Where to take a point for an event time

NOTE: You might well have less points than the number of events but you certainly should not have more points than events

Method B.

1. Set up the dataset
2. Take points according to the following instructions:
    a. Aim to capture the shape of the curve – you don't need to click at every event time (the bottom of each step – as shown in the Figure 1 above)
    b. Take points throughout the entire time span
    c. Take at least 60 points and no more than 500 – you can see the number of points selected on the right hand side of the screen next to 'points'
    d. Take more points, where and when there are many events / many steps or in the curve (usually at the start of the timescale). This can also include a very sharp drop at the start of the curve.
    e. Towards the end of the time span (in the tail of the distribution), only click in the midpoint of the step as shown in Figure 2
    f. Ensure there is an estimate for the survival before any time after time = 0, where the 'numbers at risk' are given
    g. Avoid only taking points at evenly spaced intervals

h. Treatment groups from the same trial do not need to have the same number of points Do not try to click on every single point on the curve

i. Where there is an exceptionally sharp drop at the start of the curve, take several points at different survival times along this drop (this may not necessarily be at different time points).

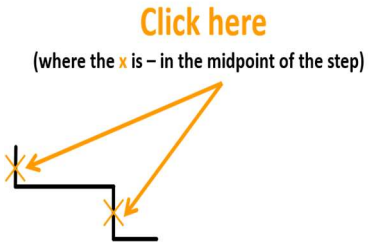


Figure 2. Where to take points in the tail

j. When only 60 to 80 points have been taken, approximately 15 - 20 of these should be within the first interval of the numbers at risk table (if given). They do not have to be at event times.

F-2: Instructions for data extraction (Method C: Simulation; Method D: Guyot)

Where to take points: Methods

1. Set up the dataset
2. Take points according to the following instructions:
    a. Aim to capture the shape of the curve – you don't need to click at every event time (the bottom of each step – as shown in the Figure 2 below)
    b. Take points throughout the entire time span
    c. Take at least 60 points and no more than 500 – you can see the number of points selected on the right hand side of the screen next to 'points'
    d. Take more points, where and when there are many events / many steps or in the curve (usually at the start of the timescale). This can also include a very sharp drop at the start of the curve.
    e. Towards the end of the time span (in the tail of the distribution), only click in the midpoint of the step as shown in Figure 1
    f. Ensure there is an estimate for the survival before any time after time = 0, where the 'numbers at risk' are given
    g. Avoid only taking points at evenly spaced intervals
    h. Treatment groups from the same trial do not need to have the same number of points Do not try to click on every single point on the curve
    i. Where there is an exceptionally sharp drop at the start of the curve, take several points at different survival times along this drop (this may not necessarily be at different time points).
    j. When only 60 to 80 points have been taken, approximately 15 - 20 of these should be within the first interval of the numbers at risk table (if given). They do not have to be at event times.

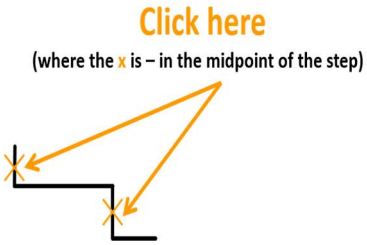


Figure 1. Where to take points in the tail

Method D.

1. Set up the dataset
2. Take points at the event times (i.e. at the bottom of each step in the curve – see Figure 2). Where there are dotted or thick lines, you may have to guess roughly where the event occurred.

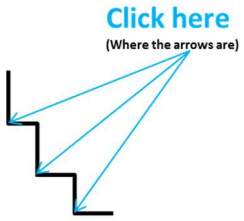


Figure 2: Where to take a point for an event time

NOTE: You might well have less points than the number of events but you certainly should not have more points than events

F-3: Results for the individual participants

Example 1						
Method:	**Guyot**			**Simulation approach**		
Participant	**Hazard ratio**	**Restricted mean survival time**		**Hazard ratio**	**Restricted mean survival time**	
		Control group	**New group**		**Control group**	**New group**
IPD	***0.516***	***22.667***	***25.610***	***0.516***	***22.667***	***25.610***
1	0.636	25.622	27.756	0.580	22.774	25.680
2	0.657	25.781	27.879	0.453	22.589	26.101
3	0.884	25.659	27.765	0.388	22.737	28.556
4	0.634	25.535	27.735	0.524	22.736	25.565
5	0.497	26.144	28.68	0.532	22.767	25.586
6	0.875	25.740	28.247	0.524	22.638	25.547
Average	***0.697***	***25.747***	***28.012***	***0.496***	***22.707***	***26.172***
Range	*0.497 – 0.884*	*25.535 – 26.144*	*27.735 – 28.689*	*0.388 - 0.580*	*22.589 - 22.775*	*25.547 - 28.556*
Example 2						
Method:	**Guyot**			**Simulation approach**		
Participant	**Hazard ratio**	**Restricted mean survival time**		**Hazard ratio**	**Restricted mean survival time**	
		Control group	**New group**		**Control group**	**New group**
IPD	***0.580***	***4.307***	***6.147***	***0.580***	***4.307***	***6.147***
1	0.563	4.498	6.318	0.580	4.330	5.986
2	0.579	4.657	6.294	0.585	4.359	6.044
3	0.581	4.572	6.286	0.589	4.258	6.007
4	0.578	4.467	6.208	0.577	4.277	6.055
5	0.588	7.108	7.909	0.552	4.612	6.308
6	0.622	4.698	6.098	0.571	4.186	5.891
Average	***0.585***	***5.000***	***6.519***	***0.576***	***4.337***	***6.049***
Range	*0.563 – 0.622*	*4.467 – 7.108*	*6.098 – 7.909*	*0.552 - 0.589*	*4.186 - 4.612*	*5.891 - 6.308*
Example 3						
Method:	**Guyot**			**Simulation approach**		
Participant	**Hazard ratio**	**Restricted mean survival time**		**Hazard ratio**	**Restricted mean survival time**	
		Control group	**New group**		**Control group**	**New group**
IPD	***0.881***	***12.585***	***14.138***	***0.881***	***12.585***	***14.138***
1	0.892	12.648	14.234	0.895	12.491	14.044
2	0.886	12.527	14.134	0.899	12.535	14.092
3	0.875	12.818	14.383	0.874	12.534	14.157
4	0.911	12.865	13.954	0.873	12.573	14.041
5	0.924	14.098	15.030	0.890	12.344	14.119
6	0.650	13.152	18.639	0.906	12.556	14.114
Average	***0.856***	***13.018***	***15.062***	***0.889***	***12.506***	***14.094***
Range	*0.650 – 0.924*	*12.527 – 14.098*	*13.954 – 18.639*	*0.873 - 0.906*	*12.344 - 12.573*	*14.041 - 14.157*

F-4: Results for the individual participants, stratified by method extraction order

Example 1							
Method:		**Guyot**			**Simulation approach**		
Extracting order	**Participant**	**Hazard ratio**	**Restricted mean survival time**		**Hazard ratio**	**Restricted mean survival time**	
			Control group	**New group**		**Control group**	**New group**
-	***IPD***	***0.516***	***22.667***	***25.610***	***0.516***	***22.667***	***25.610***
Guyot first	1	0.636	25.622	27.756	0.580	22.774	25.680
	3	0.884	25.659	27.765	0.388	22.737	28.556
	6	0.875	25.740	28.247	0.524	22.638	25.547
	Average	***0.798***	***25.674***	***27.923***	***0.497***	***22.716***	***26.594***
Simulation first	2	0.657	25.781	27.879	0.453	22.589	26.101
	4	0.634	25.535	27.735	0.524	22.736	25.565
	5	0.497	26.144	28.680	0.532	22.767	25.586
	Average	***0.596***	***25.820***	***28.098***	***0.503***	***22.697***	***25.751***
Regardless of order	***Average***	***0.697***	***25.747***	***28.012***	***0.496***	***22.707***	***26.172***
Example 2							
Method:		**Guyot**			**Simulation approach**		
Extracting order	**Participant**	**Hazard ratio**	**Restricted mean survival time**		**Hazard ratio**	**Restricted mean survival time**	
			Control group	**New group**		**Control group**	**New group**
-	***IPD***	***0.516***	***22.667***	***25.610***	***0.516***	***22.667***	***25.610***
Guyot first	1	0.563	4.498	6.318	0.580	4.33	5.986
	3	0.581	4.572	6.286	0.589	4.258	6.007
	6	0.622	4.698	6.098	0.571	4.186	5.891
	Average	***0.589***	***4.589***	***6.234***	***0.580***	***4.258***	***5.961***
Simulation first	2	0.579	4.657	6.294	0.585	4.359	6.044
	4	0.578	4.467	6.208	0.577	4.277	6.055
	5	0.588	7.108	7.909	0.552	4.612	6.308
	Average	***0.582***	***5.411***	***6.804***	***0.571***	***4.416***	***6.136***
Regardless of order	***Average***	***0.585***	***5.000***	***6.519***	***0.576***	***4.337***	***6.049***
Example 3							
Method:		**Guyot**			**Simulation approach**		
Extracting order	**Participant**	**Hazard ratio**	**Restricted mean survival time**		**Hazard ratio**	**Restricted mean survival time**	
			Control group	**New group**		**Control group**	**New group**
-	***IPD***	***0.881***	***12.585***	***14.138***	***0.881***	***12.585***	***14.138***
Guyot first	1	0.892	12.648	14.234	0.895	12.491	14.044
	3	0.875	12.818	14.383	0.874	12.534	14.157
	6	0.650	13.152	18.639	0.906	12.556	14.114
	Average	***0.806***	***12.873***	***15.752***	***0.892***	***12.527***	***14.105***
Simulation first	2	0.886	12.527	14.134	0.899	12.535	14.092
	4	0.911	12.865	13.954	0.873	12.573	14.041
	5	0.924	14.098	15.03	0.89	12.344	14.119
	Average	***0.907***	***13.163***	***14.373***	***0.887***	***12.484***	***14.084***
Regardless of order	***Average***	***0.856***	***13.018***	***15.062***	***0.889***	***12.506***	***14.094***

Appendix G: Calculation of the censoring distribution for TAnDEM

G-1: Initial calculation of censorings

Time at start of the interval	NAR	Survival at end of the interval	Change in NAR over interval	Prob. of surviving	Est. no. of cens. over interval
0	104	0.3901	68	0.3901	0.0000
5	36	0.2270	14	0.5818	0.0000
10	22	0.0993	13	0.4375	0.8696
15	9	0.0603	4	0.6071	0.0000
20	5	0.0390	1	0.6471	0.0000
25	4	0.0213	2	0.5455	0.2353
30	2	0.0035	1	0.1667	0.0000
35	1	0.0035	1	1.0000	1.0000
40	0	0.0035	0	1.0000	0.0000
Total					2.1049

G-2: Calculation of censoring distribution parameters, after applying the scale factor

Time at start of the interval	NAR	Change in NAR over interval	Est. no. of cens.	Interval length	Prob. of cens.	Interval specific haz.
0	104	68	0.0000	5	1.0000	0.0000
5	36	14	0.0000	5	1.0000	0.0000
10	22	13	3.1482	5	0.8156	0.0408
15	9	4	0.0000	5	1.0000	0.0000
20	5	1	0.0000	5	1.0000	0.0000
25	4	2	0.8518	5	0.7514	0.0572
30	2	1	0.0000	5	1.0000	0.0000
35	1	1	1.0000	5	0.0000	0.0000
40	0	0	0.0000	5	0.0000	0.0000
Total			5.0000			

Appendix H: Further discussion points for the simulation method

H-1: Distinct subsets of the Illustrative Example

The 2000 simulated datasets, produced for the Illustrative Example, were partitioned into 200-dataset subsets and averaged over. The results are given in the table below.

Distinct ordering:	Average log-HR	Average HR
1	0.43	1.53
2	0.47	1.60
3	0.45	1.56
4	0.46	1.58
5	0.48	1.61
6	0.44	1.55
7	0.42	1.53
8	0.45	1.57
9	0.48	1.62
10*	0.47	1.60

* This subset only contained 199 datasets

As can be seen the average estimates (for 200 datasets) ranged between 1.53 and 1.62, with some of the estimates being quite noticeably different from the reported IPD value of 1.57.

Appendix I: Development of 'Illness-Death' modelling approach

I-1: Generalising the formula for post-progression censoring

In Chapter 5, the following expression (Equation (5-12)) was introduced

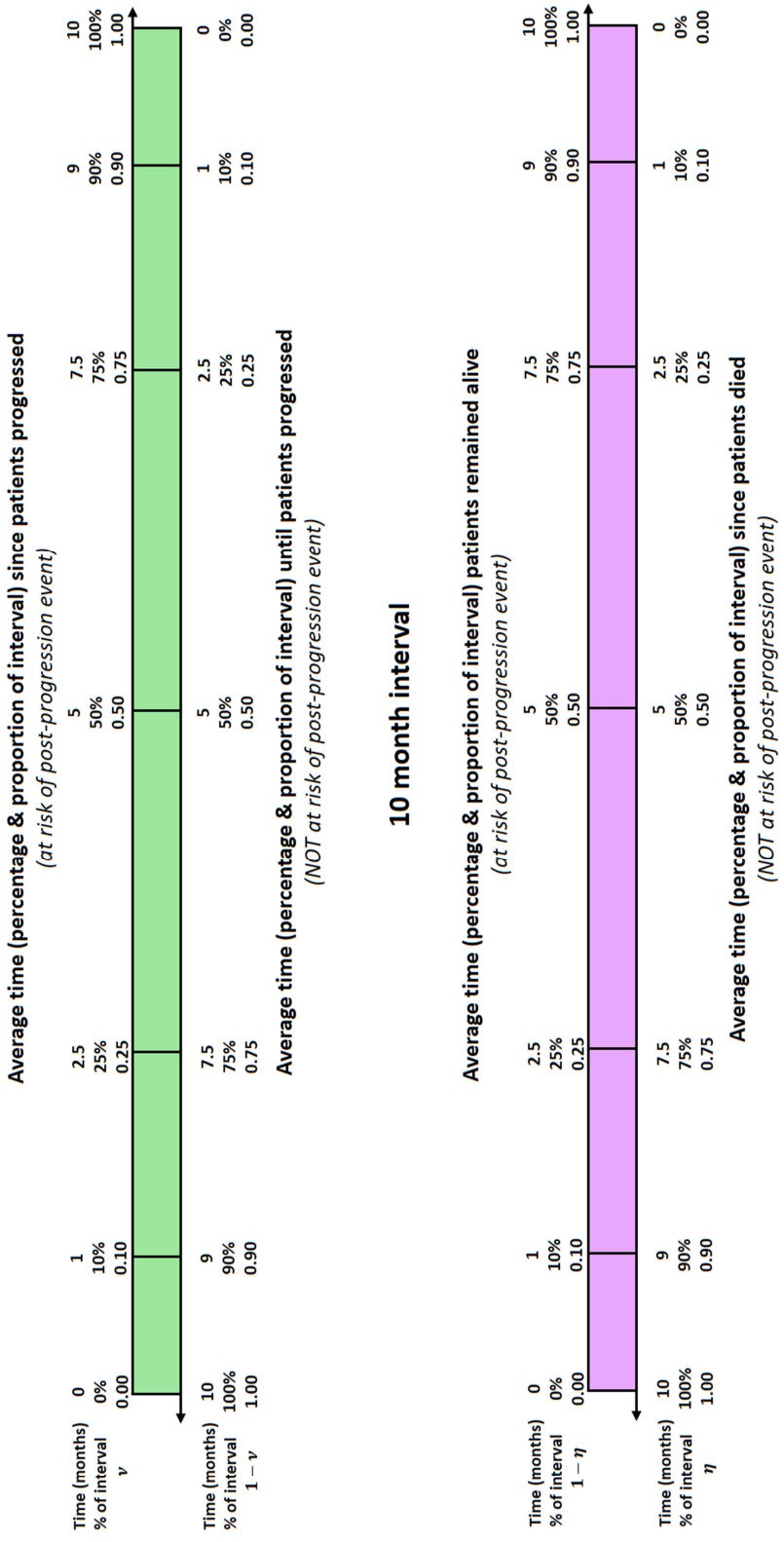
$$p_{PPCi} = 1 - \frac{c_{PPSi}}{(n_{PPSi} + \alpha_i \, d_{TTPi}) - \frac{1}{2}(y^*_{PPSi} - c_{PPSi})} \tag{5-16}$$

This can be generalised further such that the proportion

$$p_{PPCi} = 1 - \frac{c_{PPS_j}}{n_{PPS_j} + \nu_j \, d_{TTP_j} - \eta_j \, c_{PPS_j}} \tag{I1}$$

Example values and their interpretation are given below in Appendix I-2.

I-2: Example values of ν, η for a 10-month interval.



The values of ν, η can be different over each partition. More often though, $\nu_j = \eta_j = \frac{1}{2}$ has been chosen.

Appendix J: References for the studies included in the case study (Chapter 6)

J-1: BMS 099

Khambata-Ford S, Harbison CT, Hart LL, Awad M, Xu LA, Horak CE, et al. Analysis of potential predictive markers of cetuximab benefit in BMS099, a phase III study of cetuximab and first-line taxane/carboplatin in advanced non-small-cell lung cancer. *Journal of Clinical Oncology* 2010;28(6):918–27.

Lynch TJ, Patel T, Dreisbach L, McCleod M, Heim WJ, Hermann RC, et al. Cetuximab and first-line taxane/ carboplatin chemotherapy in advanced non-small cell lung cancer: results of the randomized multicenter phase III trial BMS099. *Journal of Clinical Oncology* 2010;28(6):911–7.

J-2: CHEN

Chen YM, Tsai CM, Fan WC, Shih JF, Liu SH, Wu CH, et al. Phase II randomized trial of erlotinib or vinorelbine in chemonaive, advanced, non-small cell lung cancer patients aged 70 years or older. *Journal of Thoracic Oncology* 2012;7 (2):412–8.

J-3: ENSURE

Wu Y-L, Zhou C, Liam CK, Wu G, Liu X, Zhong Z, et al. First-line erlotinib versus gemcitabine/cisplatin in patients with advanced EGFR mutation-positive non-small-cell lung cancer: analyses from the phase III, randomized, open-label, ENSURE study. *Annals of Oncology* 2015;26(9):1883–9.

Wu Y-L, Zhou C, Wu G, Liu X, Zhong Z, Lu S, et al. Quality of life (QOL) analysis from ENSURE, a phase 3, open-label study of first-line erlotinib versus gemcitabine/cisplatin in Asian patients with epidermal growth factor receptor (EGFR) mutation positive (MUT+) non-small cell lung cancer (NSCLC). *Journal of Thoracic Oncology* 2014;9: S37.

J-4: EURTAC

De Marinis F, Rosell R, Vergnenegre A, Massuti B, Felip E, Gervais R, et al. Erlotinib vs chemotherapy (CT) in advanced non-small cell lung cancer (NSCLC) patients with epidermal growth factor receptor (EGFR) activating mutations − the EURTAC

Phase II randomized trial interim results. *European Journal of Cancer* 2011;47: S597.

Rosell R, Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *The Lancet Oncology* 2012;13(3):239–46.

de Marinis F, Vergnenegre A, Passaro A, Dubos-Arvis C, Carcereny E, Drozdowskyj A, et al. Erlotinib-associated rash in patients with EGFR mutation-positive non-smallcell smallcell lung cancer treated in the EURTAC trial. *Future Oncology* 2015;11(3):421–9.

J-5: FASTACT 2

Mok T, Wu YL, Thongprasert S, Yu C, Zhang J, Ladrera L, et al. A randomized placebo-controlled phase III study of intercalated erlotinib with gemcitabine/platinum in first-line advanced non-small cell lung cancer (NSCLC): FASTACT-II. *Journal of Clinical Oncology* 2012;30(May 20 Suppl):7519.

Wu YL, Lee JS, Thongprasert S, Yu CJ, Zhang L, Ladrera G, et al. Intercalated combination of chemotherapy and erlotinib for patients with advanced stage non-small-cell lung cancer (FASTACT-2): a randomised, double-blind trial. *The Lancet Oncology* 2013;14(8):777–86.

J-6: TORCH

Di Maio M, Leighl NB, Gallo C, Feld R, Ciardiello F, Butts C, et al. Quality of life analysis of TORCH, a randomized trial testing first-line erlotinib followed by second-line cisplatin/gemcitabine chemotherapy in advanced non-small cell lung cancer. *Journal of Thoracic Oncology* 2012;7(12): 1830–44.

Gridelli C, Butts C, Ciardiello F, Feld R, Gallo C, Perrone F. An international, multicenter, randomized phase III study of first-line erlotinib followed by second-line cisplatin/ gemcitabine versus first-line cisplatin/gemcitabine followed by second-line erlotinib in advanced non-small-cell lung cancer: treatment rationale and protocol dynamics of the TORCH trial. *Clinical Lung Cancer* 2008;9(4):235–8.

Gridelli C, Ciardiello F, Gallo C, Feld R, Butts C, Gebbia V, et al. First-line erlotinib followed by second-line cisplatin-gemcitabine chemotherapy in advanced non-small cell lung cancer: the TORCH randomized trial. *Journal of Clinical Oncology* 2012;30(24):3002–11.

Tsao M, Gallo S, Saieg C, Santos M, Gebbia GDC, Perrone V, et al. Biomarkers of torch trial on first-line erlotinib followed by second-line chemotherapy in advanced nonsmall cell lung cancer patients. International Association for the Study of Lung Cancer. 3rd European Lung Cancer Conference, Geneva. 2012 Switzerland April 18-21. 2012.

J-7: GTOWG

Reck M, Von Pawel J, Fischer Jr, Kortsik C, von Eiff M, Koester W, et al. Erlotinib versus carboplatin/vinorelbine in elderly patients (age 70 or older) with advanced nonsmall cell lung carcinoma (NSCLC): a randomised phase II study of the German Thoracic Oncology Working Group. *Journal of Clinical Oncology* 2010;28:15s.

J-8: First-SIGNAL

Han JY, Park K, Kim SW, Lee DH, Kim HY, Kim HT, et al. First-SIGNAL: first-line single-agent Iressa versus gemcitabine and cisplatin trial in never-smokers with adenocarcinoma of the lung. *Journal of Clinical Oncology* 2012;30(10):1122–8.

J-9: TOPICAL

Lee SM, Khan I, Upadhyay S, Lewanski C, Falk S, Skailes G, et al. First-line erlotinib in patients with advanced non-small cell lung cancer unsuitable for chemotherapy (TOPICAL): a double-blind, placebo-controlled phase III trial. *The Lancet Oncology* 2012;13(11):1161–70.

J-10: INTACT 1 & INTACT 2

Bell DW, Lynch TJ, Haserlat SM, Harris PL, Okimoto RA, Brannigan BW, et al. Epidermal growth factor receptor mutations in non-small cell lung cancer: molecular analysis of the IDEAL/INTACT gefitinib studies. *Journal of Clinical Oncology* 2005;23:8081–92.

Giaccone G, Herbst R, Manegold C, Scagliotti G, Rosell R, Miller V, et al. Gefitinib in combination with gemcitabine and cisplatin in advanced non-small cell lung cancer: a phase III trial - INTACT 1. *Journal of Clinical Oncology* 2004;22:777–84.

Herbst RS, Giaccone G, Schiller JH, Natale RB, Miller V, Manegold C, et al. Gefitinib in combination with paclitaxel and carboplatin in advanced non-small cell lung cancer: a phase III trial - INTACT 2. *Journal of Clinical Oncology* 2004;22:785–94.

J-11: NEJ002 (referred to as NEJSG in Greenhalgh (2016))

Fukuhara T, Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, et al. Factors associated with a poor response to gefitinib in the NEJ002 study: smoking and the L858R mutation. *Lung Cancer* 2015;88:181–6.

Inoue A, Kobayashi K, Maemondo M, Sugawara S, Oizumi S, Isobe H, et al. Updated overall survival results from a randomized phase III trial comparing gefitinib with carboplatin-paclitaxel for chemo-naïve non-small cell lung cancer with sensitive EGFR gene mutations (NEJ002). *Annals of Oncology* 2012;24(1):54.

Kinoshita I, Inoue A, Kobayashi K, Maemondo M, Sugawara S, Oizumi S. Phase III Study of Gefitinib versus Chemotherapy by Carboplatin (CBDCA) plus Paclitaxel (TXL) as First-line Therapy for Non-small Cell Lung Cancer (NSCLC) with EGFR Mutations: North East Japan Gefitinib Study Group Trial 002 (NEJ002). *Respirology* 2009;14:

Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for non-small cell lung cancer with mutated EGFR. *The New England Journal of Medicine* 2010;362(25):2380–8.

Oizumi S, Kobayashi K, Inoue A, Maemondo M, Sugawara S, Yoshizawa H, et al. Quality of life with gefitinib in patients with EGFR-mutated non-small cell lung cancer: quality of life analysis of North East Japan Study Group 002 Trial. *The Oncologist* 2012;17(6):863–70.

Satoh H, Inoue A, Kobayashi K, Maemondo M, Oizumi S, Isobe H, et al. Low-dose gefitinib treatment for patients with advanced non-small cell lung cancer harboring sensitive epidermal growth factor receptor mutations. *Journal of Thoracic Oncology* 2011;6(8):1413–7.

J-12: IPASS

Fukuoka M, Wu YL, Thongprasert S, Sunpaweravong P, Leong SS, Sriuranpong V, et al. Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/ paclitaxel in clinically selected patients with advanced nonsmall cell lung cancer in Asia (IPASS). *Journal of Clinical Oncology* 2011;29(21):2866–74.

Ichinose Y, Nishiwaki Y, Ohe Y, Yamamoto N, Negoro S, Duffield E, et al. Analyses of Japanese patients recruited in IPASS, a phase III, randomized, open-label, first-line study of gefitinib vs carboplatin/paclitaxel in selected patients with advanced non-small cell lung cancer. *Journal of Thoracic Oncology* 2009;4:S443.

Mok T, Wu YL, Thongprasert S, Yang CH, Chu D, Saijo N, et al. Phase III randomised open-label first-line study of gefitinib vs carboplatin/paclitaxel in clinically selected patients with advanced non-small cell lung cancer (IPASS). *Annals of Oncology* 2008;19:1.

Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *The New England Journal of Medicine* 2009;361(10):947–57.

Ohe Y, Ichinose Y, Nishiwaki Y, Yamamoto N, Negoro S, Duffield E, et al. Phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in selected patients with advanced non-small cell lung cancer (IPASS): evaluation of recruits in Japan. *Journal of Clinical Oncology* 2009;27.

Thongprasert S, Duffield E, Saijo N, Wu YL, Yang JC, Chu DT, et al. Health-related quality-of-life in a randomized phase III first-line study of gefitinib versus carboplatin/ paclitaxel in clinically selected patients from Asia with advanced NSCLC (IPASS). *Journal of Thoracic Oncology* 2011;6(11):1872–80.

Wu Y, Fukuoka L, Mok M, Saijo TSK, Thongprasert N, Yang S, et al. Tumor response and health-related quality of life in clinically selected patients from Asia with advanced non-small-cell lung cancer treated with first-line gefitinib: Post hoc analyses from the IPASS study. *Lung Cancer* 2013; 81:280–7.

Wu Y, Mok T, Chu D, Han B, Liu X, Zhang L, et al. Evaluation of clinically selected patients with advanced nonsmall cell lung cancer recruited in China in a phase III, randomized, open-label, first-line study in Asia of gefitinib versus carboplatin/paclitaxel (IPASS). *Journal of Clinical Oncology* 2009;27.

Wu YL, Chu DT, Han B, Liu X, Zhang L, Zhou C, et al. Phase III, randomized, open-label, first-line study in Asia of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non-small cell lung cancer: evaluation of patients recruited from mainland China. *Asia- Pacific Journal of Clinical Oncology* 2012;8:232–43.

Yang CH, Fukuoka M, Mok T, Wu YL, Thongprasert S, Saijo N, et al. Final overall survival (OS) results from a phase III, randomised, open-label, first-line study of gefitinib v carboplatin/paclitaxel in clinically selected patients with advanced non-small cell lung cancer in Asia (IPASS). *Annals of Oncology* 2010;21.

J-13: WJTOG3405

Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Seto T, et al. Updated overall survival results of WJTOG 3405, a randomized hase III trial comparing gefitinib (G) with cisplatin plus docetaxel (CD) as the first-line treatment for patients with non-small cell lung cancer harbouring mutations of the epidermal growth factor receptor (EGFR). Journal of Clinical Oncology. 2012; Vol. 30, issue 15 (S1):7521.

Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Tsurutani J, et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *The Lancet Oncology* 2009;11(2):121–8.

Yoshioka H, Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, et al. Final overall survival results of WJTOG3405, a randomized phase 3 trial comparing gefitinib (G) with cisplatin plus docetaxel (CD) as the first-line treatment for patients with non-small cell lung cancer (NSCLC) harboring mutations of the epidermal growth factor receptor (EGFR). *Journal of Clinical Oncology* 2014;32(15 (May Suppl)):8117.

J-14: Yu 2014

Yu H, Zhang J, Wu X, Luo Z, Wang H, Sun S, et al. A phase II randomized trial evaluating gefitinib intercalated with pemetrexed/platinum chemotherapy or pemetrexed/ platinum chemotherapy alone in unselected patients with advanced non-squamous non-small cell lung cancer. *Cancer Biology & Therapy* 2014;15(7):832–9.

J-15: OPTIMAL

Chen G, Feng JF, Zhou C, Wu Y-L, Liu XQ, Wang C, et al. Quality of life (QoL) analyses from OPTIMAL (CTONG-0802), a phase III, andomised, open-label study of first-line erlotinib versus chemotherapy in patients with advanced EGFR mutation-positive non-small-cell lung cancer (NSCLC). *Annals Oncology* 2013;24(6):1615–22.

Wu YL, Zhou C, Chen G, Feng J, Liu X, Wang C, et al. First biomarker analyses from a phase III randomised openlabel first-line study of erlotinib versus carboplatin plus gemcitabine in Chinese patients with advanced non-small cell lung cancer (NSCLC) with EGFR activating mutations (OPTIMAL, CTONG0802). *Annals of Oncology* 2010;21.

Zhou C, Wu Y, Liu L, Wang X, Chen C, Feng G, et al. Overall survival (OS) results from OPTIMAL (CTONG0802), a phase III trial of erlotinib (E) versus carboplatin plus gemcitabine (GC) as first-line treatment for Chinese patients with EGFR mutation-positive advanced non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology* 2012;30(15 (May Suppl)):7520.

Zhou C, Wu YL, Chen G, Feng J, Liu X, Wang C, et al. Efficacy results from the randomised phase III OPTIMAL (CTONG 0802) study comparing first-line erlotinib versus carboplatin plus gemcitabine in Chinese advanced non small cell lung cancer (NSCLC) patients with EGFR activating mutations. *Annals of Oncology* 2010;21:6.

Zhou C, Wu YL, Chen G, Feng J, Liu X, Wang C, et al. Final overall survival results from a randomised, phase III study of erlotinib versus chemotherapy as first-line treatment of EGFR mutation-positive advanced non smallcell lung cancer (OPTIMAL, CTONG-0802). *Annals of Oncology* 2015;26:1877–83.

Zhou C, Wu YL, Chen G, Feng J, Liu XQ, Wang C, et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation positive nonsmall cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. *The Lancet Oncology* 2011;12(8):735

J-16: LUX-Lung 3

Boehringer Ingelheim. Submission of evidence on the use of afatinib in adult patients with locally advanced or metastatic non-small cell lung cancer (NSCLC) with activating Epidermal Growth Factor Receptor (EGFR) mutation (s).

O'Byrne KJ, Sequist LV, Schuler M, Yamamoto N, Hirsh V, Mok T, et al. LUX-Lung 3: Symptom and health-related quality of life results from a randomized phase III study in 1st-line advanced NSCLC patients harbouring EGFR mutations. 11th Annual British Thoracic Oncology Group Conference, Dublin Ireland, January 23-25. 2013:S11.

Sequist LV, Yang JCH, Yamamoto N, O'Byrne K, Hirsh V, Mok T, et al. Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutations. *Journal of Clinical Oncology* 2013;31:1–11.

Yang JC, Schuler M, Yamanoto N, O'Byrne K, Hirsch V, Mok T, et al. LUX-Lung 3: a randomized, open-label, phase III study of afatinib vs cisplatin/pemetrexed as first line treatment for patients with advanced adenocarcinoma of the lung harboring EGFR-activating mutations. *Journal of Clinical Oncology* 2012;30:LBA7500.

J-17: LUX-Lung 6

Geater SL, Xu C-R, Zhou C, Hu C-P, Feng J, Lu S, et al. Symptom and quality of life improvement in LUX-Lung 6: An open-label phase III study of afatinib versus cisplatin/ gemcitabine in Asian patients with EGFR mutation-positive advanced non-small-cell lung cancer. *Journal of Thoracic Oncology* 2015;10(6):883–9.

Geater SL, Zhou C, Hu C-P, Feng JF, Lu S, Huang Y, et al. LUX-Lung 6: Patient-reported outcomes (PROs) from a randomized open-label, phase III study in first-line advanced NSCLC patients harboring epidermal growth factor receptor (EGFR) mutations. *Journal of Clinical Oncology* 2013;31(15 (May Suppl)):8016.

Wu Y-L, Zhou C, Hu C-P, Feng J, Lu S, Huang Y, et al. Afatinib versus cisplatin plus gemcitabine for firstline treatment of Asian patients with advanced non-smallcell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. *The Lancet Oncology* 2014;15(2):213–22.

Yang C-H J, Wu Y-L, Schuler M. Afatinib versus cisplatinbased chemotherapy for EGFR mutation-positive lung adenocarcinoma (LUX-Lung 3 and LUX-Lung 6):

analysis of overall survival data from two randomised, phase 3 trials. *The Lancet Oncology* 2015;14:1173–8.

J-18: FLEX

O'Byrne J, Gatzemeier U, Bondarenko I, Barrios C, Eschbach C, Martens UM, et al. Molecular biomarkers in non-small cell lung cancer: a retrospective analysis of data from the phase 3 FLEX study. The Lancet Oncology 2011; 12:795–805.

Pirker R, Pereira JR, Szczesna A, von Pawel J, Krzakowski M, Ramlau R, et al. Cetuximab plus chemotherapy in patients with advanced non-small-cell lung cancer (FLEX): an open-label randomised phase III trial. The Lancet 2009; 373(9674):1525–31.

Bibliography

Batson S, Greenall G, Hudson P. Review of the Reporting of Survival Analyses within Randomised Controlled Trials and the Implications for Meta-Analysis. PLoS ONE 2016. 11(5): e0154870. https://doi.org/10.1371/journal.pone.0154870

Bell H, Latimer N, Amonkar M, Casey M. PRM192 Adjusting for treatment crossover in a trametinib metastatic melanoma RCT: Identifying the appropriate method. ISPOR 17th Annual European Congress, Amsterdam, The Netherlands. 2014

Bell H, Latimer N, Amonkar M, Swann S. RM1 Adjusting for treatment switching in RCTS – identifying, analysing and justifying appropriate methods: a case study in metastatic melanoma. ISPOR 18th Annual European Congress, Milan, Italy. 2015

Bellera CA, MacGrogan G, Debled M, de Lara T et al. Variables with time-varying effects and the Cox model : some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Medical Research Methodology 2010, 10(20).

Bennett I, Paracha N, Abrams K, Ray J. Accounting for Uncertainty in Decision Analytic Models Using Rank Preserving Structural Failure Time Modeling: Application to Parametric Survival Models. Value Health. 2018. 21(1):105-109. doi: 10.1016/j.jval.2017.07.008

Beyersmann, J. Latouche, A, Buchholz, A, Schumacher, M. Simulating competing risks data in survival analysis. Statist Med. 2009: 28: 956-971

Boucher RH, Abrams KR, Crowther MJ, Lambert PC, Morden, JP, Wailoo A, Latimer N. Adjusting for treatment switching in clinical trials when only summary data are available – an evaluation of potential methods. 06/11/2013 Poster

Presentation (ISPOR 16th Annual European Meeting, Dublin, Ireland). Value in Health 2013a:16(7)

Boucher RH. 2013b. Methods for dealing with treatment switching with summary data. MSc. University of Leicester

Boucher R, Abrams K, Lambert P. PRM107 Simulating individual patient level data using an illness-death modelling framework in order to adjust for treatment switching when only summary data are available. ISPOR 18th Annual European Congress, Milan, Italy. 2015

Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. Stat Med. 2002 Sep 15;21(17):2449-63.

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997 Jun;50(6):683-91.

Caldwell DM, Ades AE, Higgins JP: Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005, 331(7521):897-900. 3.

Carpenter JR, Kenward MG: Multiple Imputation and its Application. Wiley. 2014.

Cella D. Grunwald V, Nathan P, Doan J, Dastani H, Taylor F et al. Quality of life in patients with advanced renal cell carcinoma given nivolumab versus everolimus in CheckMate 025: a randomised, open-label, phase 3 trial. The Lancet Oncology. 2016. 17(7)pp994-1003

Collett D: Modelling survival data in medical research. Chapman and Hall 2003.

Conley RB, Al Naber J, Messner DA, Henshall C, Huang C, Rosner GL. PRM176 Developing best practices for managing treatment switching in the design,

conduct, analysis, and reporting of oncology drug clinical trials. ISPOR 21st Annual International Meeting, Washington DC, USA. 2016

Cox DR. Regression Models and Life Tables. Journal of the Royal Statistical Society, Series B. 1972. 34(2);pp187-220.

Data Protection Act. 2018. Available at; http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted.

Dickman, P., Coviello, E. Estimating and modelling relative survival. The Stata Journal. 2015: 15(1); pp186 – 215.

DigitizeIt. Digitizer software - digitize a scanned graph or chart into (x,y)-data. 2013 .URL www.digitizeit.de

Dimopoulos MA, Chen C, Spencer A, Niesvizky R, Attal M, Stadtmauer EA, Petrucci MT, Yu Z, Olesnyckyi M, Zeldis JB, Knight RD, Weber DM. Long-term follow-up on overall survival from the MM-009 and MM-010 phase III trials of lenalidomide plus dexamethasone in patients with relapsed or refractory multiple myeloma. Leukemia 2009. 23, 2147–2152;

Dimopoulos MA, Spencer A, Attal M, Prince HM, Harousseau JC, Dmoszynska A et al. Multiple Myeloma (010) study investigators. Lenalidomide plus dexamethasone for relapsed or refractory multiple myeloma. N Engl J Med 2007; 357: 2123–2132.

Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in Medicine 8(5):551–561, May 1989.

Errington RD, Ashby D. Gore, SM., Abrams, KA, Myint S, Bonnett DE, Blake, SW, Saxton, TE. High energy neutron treatment for pelvic cancers: study stopped because of increased mortality. 1991 BMJ 302 pp 1045-1051

Fisher LD, Dixon DO, Herson J, Frankowski RK, Hearron MS, Peace KE. Intention to treat in clinical trials. In: Peace KE, editor. Statistical issues in drug research and development. New York: Marcel Dekker; 1990. pp. 331–50.

General Data Protection Regulation. 2018. Available at; https://gdpr-info.eu/

Gershlick AH, Stephens-Lloyd A, Hughes S, Abrams KR, Stevens SE, Uren NG, et al. Rescue Angioplasty after Failed Thrombolytic Therapy for Acute Myocardial Infarction. N Engl J Med. 2005;353:2758-2768

Gurskyte L, Muresan B, Kulakova M, Postma M, Ouwens MJ, Heeg B. PCN64 Review of NICE HTA submissions including methodologies adjusting for overall survival in the presence of treatment switching ISPOR 21st Annual European Congress, Barcelona, Spain. 2018Greenhalgh J, Dwan K, Boland A, Bates V, Vecchio F, Dundar Y,Jain P, Green JA. First-line treatment of advanced epidermal growth factor receptor (EGFR) mutation positive non-squamous non-small cell lung cancer. Cochrane Database of Systematic Reviews 2016, Issue 5. Art. No.: CD010383. DOI: 10.1002/14651858.CD010383.pub2.

The Green Park Collaborative. 2016. Best Practices for the Design, Implementation, Analysis, and Reporting of Oncology Trials with High Rates of Treatment Switching A Guidance Document from the Green Park Collaborative

Gupta, SK. Intention-to-treat concept. Perspect Clin Res. A review. 2011; 2(3): 109-112.

Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012 Feb 1;12:9.

Hauschild A, Grobb J, Demidov L, Jouary T, Gutzmer R et al. An update on overall survival (OS) and follow-on therapies in BREAK-3, a phase III, randomized trial: Dabrafenib (D) vs. Dacarbazine (DTIC) in patients (pts) with BRAF

V600E mutation-positive Metastatic Melanoma (MM). Annals of Oncology 2014. 25(Supp 4) p378.

Henshall C, Latimer NR, Sansom L & Ward RL. Treatment switching in cancer trials: Issues and proposals. International Journal of Technology Assessment in Health Care. 2016 32(3), 167-174.

Hernandez-Villafuerte K, Fischer A & Latimer NR. Challenges and methodologies in using progression free survival as a surrogate for overall survival in oncology. International Journal of Technology Assessment in Health Care, 2018. 34(3), 300-316.

Hinchliffe, SR, Scott, DA, Lambert PC. Flexible parametric illness-death models. 2013 Stata Journal 13(4): pp. 759-775.

Hotta, K., Suzuki, E., Di Maio, M., Chiodini, P., et al. Progression-free survival and overall survival in phase III trials of molecular-targeted agents in advanced non-small-cell lung cancer. 2013: 79(1): pp.20 – 26

Hoyle M, Henley, W. Improved curve fits to summary survival data: application to economic evaluation of health technologies. BMC Medical Research Methodology 2011, 11:139 doi:10.1186/1471-2288-11-139.

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Benefit assessment in studies with allowed treatment switching. 2014. URL https://www.iqwig.de/en/events/iqwig-in-dialogue/iqwig-in-dialogue-2014.6046.html

Ishak KJ, Muszbek N, Altincatal A, Sarri G, Schlichting M, Zhou J. PCP45 The role of crossover and treatment switching in indirect treatment comparison in immuno-oncology. ISPOR 21st Annual European Congress, Barcelona, Spain. 2018

Jansen JP. Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. BMC Medical Research Methodology 2012, 12:152

Jonsson, L, Sandin R, Ekman, M, Ramsberg J, Charbonneau, C, Huang, X. et al. Analyzing overall survival in randomized controlled trials with crossover and implications for economic evaluation. Value In Health. 2014: 17 pp. 707-713.

Kaufman, B., Mackey, J., Clemens, M., Bapsy, P., Vaid, A., Wardley, A, 2009. Trastuzumab plus anastrozole versus anastrozole alone for the treatment of postmenopausal women with human epidermal growth factor receptor 2-positive, hormone receptor-positive metastatic breast cancer: results from the randomized phase III TAnDEM study. Journal of Clinical Oncology, 27(33), pp. 5529.

Lambert, PC., Dickman, PW., Nelson, CP., Royston, P. Estimating the crude probability of death due to cancer and other causes using relative survival models. Statistics in Medicine 2010: 29 (7-8); pp885-895.

Latimer NR, White IR, Abrams KR & Siebert U. Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times?. Statistical Methods in Medical Research. 2018a

Latimer, N. The (mis)use of treatment switching adjustment methods in health technology assessment – busting some myths! ISPOR 21st Annual European Congress, Barcelona, Spain. 2018b

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. Statistical Methods in Medical Research. 2018c. 27(3), 765-784.

Latimer NR, Henshall C, Siebert U & Bell H (2016) Treatment Switching: statistical and decision making challenges and approaches. International Journal of Technology Assessment in Health Care. 2016. 32(3), 160-166.

Latimer NR. Treatment switching in oncology trials and the acceptability of adjustment methods. Expert Review of Pharmacoeconomics & Outcomes Research. 2015. 15(4), 561-564.

Latimer N, Abrams K. 2014. NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching

Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting survival time estimates to account for treatment switching in randomised controlled trials – a simulation study. Health Economics & Decision Science (HEDS) Discussion Paper (DP 13/06), School of Health & Related Research (ScHARR), University of Sheffield, March 2013. Available at; http://www.shef.ac.uk/polopoly_fs/1.259501!/file/HEDSDP1306.pdf

Latimer N. 2012. Crossover adjustment methods in the context of economic evaluation. PhD. University of Sheffield. Available at: http://etheses.whiterose.ac.uk/3720/1/Thesis_Final_with_corrections.pdf

Law MG, Kaldor JM: Survival analyses of randomized clinical trials adjusted for patients who switch treatments. Statistics in Medicine 1996, 15:2069{2076.

Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010. 340:b5087. Available at; doi: https://doi.org/10.1136/bmj.b5087

Loyes T, Goethebeur E: A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. Biometrics 2003, 59:100-105.

Lu G, Ades AE: Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004, 23(20):3105-24. 4.

Lumley T: Network meta-analysis for indirect treatment comparisons. Stat Med 2002, 21(16):2313-24.

Maervoet J, Skaltsa K, Ivanescu C, Van Engen A. NI4 Do health technology agencies accept methods for dealing with treatment switching? ISPOR 17th Annual European Meeting, Amsterdam, Netherlands. 2014

McArthur, GA. Chapman, PB., Robert C, Larkin J., Haanen, JB., Drummer, R. Ribas, A et al. Safety and effcacy of vemurafenib in BRAFV600E and BRAFV600K mutation-positive melanoma (BRIM-3): extended follow-up of a phase 3, randomised, open-label study. Lancet Oncol. 2014; 15: 323–32. Available at; http://dx.doi.org/10.1016/ S1470-2045(14)70012-9

Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. BMC Med Res Methodol. 2011 Jan 11;11:4.

Morden, J. 2009. Methods for dealing with treatment switching. MSc. University of Leicester

Motzer RJ, Escudier B, Oudard S. Hutson TE, Porta C, Bracarda S. Efficacy of everolimus in advanced renal cell carcinoma: a double-blind, randomised, placebo-controlled phase III trial. 2008, 372(9637) pp449-456

National Institute for Health and Care Excellence (NICE). 2007. TA129 Bortezomib monotherapy for relapsed multiple myeloma.

National Institute for Health and Care Excellence (NICE). 2009a. TA169 Renal cell carcinoma – sunitinib: guidance.

National Institute for Health and Care Excellence (NICE). 2009b. TA171 Multiple Myeloma – lenalidomide : full guideline.

National Institute for Health and Care Excellence (NICE). 2011. TA215 Renal cell carcinoma (first line metastatic) - pazopanib: guidance

National Institute for Health and Care Excellence (NICE). 2012. TA257 Breast cancer (metastatic hormone-receptor) - lapatinib and trastuzumab (with aromatase inhibitor): guidance.

National Institute for Health and Care Excellence (NICE). 2014. TA321 Dabrafenib for treating unresectable or metastatic BRAF V600 mutation positive melanoma: guidance

National Institute of Health and Care Excellence (NICE). 2016. Nivolumab for previously treated advanced renal cell carcinoma

Othus M, Barlogie B, Leblanc ML, Crowley JJ. Cure models as a useful statistical tool for analysing survival. Clin Cancer Res. 2012; 18(14)3731-6

Pelissier SM, Gourgou-Bourgade S, Bonnetain F, Kramar, A. Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals. JCO 2008; 26(22)

Pohar Perme, M., J. Stare, and J. Estève. 2012. On estimation in relative survival. Biometrics 68: 113–120.

Reck M, Von Pawel J, Fischer Jr, Kortsik C, von Eiff M, Koester W, et al. Erlotinib versus carboplatin/vinorelbine in elderly patients (age 70 or older) with advanced nonsmall cell lung carcinoma (NSCLC): a randomised phase II study of the German Thoracic Oncology Working Group. Journal of Clinical Oncology 2010;28:15s.

Richardson PG, Sonneyeld P, Schuster MW, Irwin D, Stadtmauer EA, Facon T, Harousseau JL, Ben-Yehuda D, Lonial S, Goldschmidt H, Reece D, San-Miguel JF, Blade J, Boccadoro M, Dalton WS, Boral AL, Esseltine DL, Porter JB, Schenkein D, Anderson KC, APEX Investigators. Bortezomib or High-Dose Dexamethasone for Relapsed Multiple Myeloma. New England Journal of Medicine 2005;352:2487-98.

Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000; 56(3):779-788.

Robins JM, Tsiatis AA: Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics-Theory and Methods 1991, 20(8):2609{2631.

Robins JM. Structural Nested Failure Time Models. Andersen PK, Keiding N, editors. Survival Analysis. 4372-4389. 1998. Chichester, UK, John Wiley and Sons. The Encyclopedia of Biostatistics. Armitage, P. and Colton, T.

Royston P. Tools to simulate realistic censored survival-time distributions. Stata Journal 2012, 12(4) pp639-654

Royston, P., and M. K. B. Parmar. 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 21: 2175–2197.

Seymour MT, Maughan TS, Ledermann JA, Topham C, James R, Gwyther SJ, et al. Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial. Lancet. 2007;370:9582:143-52.

Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clark MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials, 2005. 2, pp. 209-217

Sternberg CN, Davis ID, Mardiak J, Szczylik C, Lee E, Wagstaff J, Barrios CH, Salman P, Gladkov OA, Kavina A, Zarba JJ, Chen M, McCann L, Pandite L, Roychowdhury DF, Hawkins RE. Pazopanib in Locally Advanced or Metastatic Renal Cell Carcinoma: Results of a Randomized Phase III Trial. Journal of Clinical Oncology 2010. 28(6) pp 1061-1068

Thorland K, Jansen, JP, Mills E. W6: Does it make sense to expand an evidence synthesis of randomized trials with observational real world data? ISPOR 16th Annual European Meeting, Dublin, Ireland. 2013.

Van Engen A, Latimer NR, Bohler YB, Holmstrom S. IP4: Should we adjust overall survival estimates for treatment switching in oncology? ISPOR 17th Annual European Meeting, Amsterdam, Netherlands. 2014.

Walker, A., White, I., Babiker, A. Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. Statistics in medicine, 2004: 23(4), pp. 571.

Wan, X.,Peng L, Li, Y., A Review and Comparison of Methods for Recreating Individual Patient Data from Published Kaplan-Meier Survival Curves for Economic Evaluations: A Simulation Study. PLOS. 2015. DOI: 10.1371/journal.pone.0121353

Watkins CL. PRM270 Practical considerations in the application of statistical methods for treatment switching. ISPOR 18th Annual European Congress, Milan, Italy. 2015

Weber DM, Chen C, Niesvizky R, Wang M, Belch A, Stadtmauer EA, Siegel D, Borrello I, et al. Lenalidomide plus Dexamethasone for Relapsed Multiple Myeloma in North America. New England Journal of Medicine 2007;357:2133-42.

White IR, Babiker AG, Walker S, Darbyshire JH: Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. Statistics in Medicine 1999, 18(19):2617{2634.

White IR, Walker S, Babiker A: strbee: Randomization-based efficacy estimator. The Stata Journal. 2002, 2 (Number 2): 140-150.

Williamson PR, Smith CT, Hutton JL, Marson AG. Aggregate data meta-analysis with time-to-event outcomes. Statistics in Medicine. 2002,21: 3337–3351

Zhang Y, Wright J, Luke S, Ryan J, van Engen A. PRM120 Do health technology agencies accept methods for dealing with treatment switching and immature OS data? ISPOR 19th Annual European Congress, Vienna, Austria. 2016