
# Reference Based Sensitivity Analysis for Time-to-Event Data

ANDREW DAVID ATKINSON

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy of the University of London

April 2019

Department of Medical Statistics
Faculty of Epidemiology and Population Health
LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Funded by: No funding received

# Declaration

I, Andrew David Atkinson, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature

Date

# Acknowledgements

I would like to take the opportunity to thank my supervisors James Carpenter and Mike Kenward of LSHTM for providing the impetus, guidance and support for carrying out the work. In particular, James accompanied me throughout the last 6 years, always on hand with encouragement. Despite retiring and moving to Scotland, Mike has continued to offer sage advice and guidance for the work, so special thanks also to him.

On a similar note I would like to thank Suzie Cro, now of Imperial College, for providing an excellent blueprint for the theoretical calculations in Chapter 4, and for painstakingly checking each line of the workings, including those in the appendices.

I would like to thank Tim Clayton and Stuart Pocock of LSHTM for allowing me to use the RITA-2 data set, particularly to Tim for providing cleaned data, and being on hand for my questions.

The analysis in Chapter 5 was partially funded from the Swiss National Science Foundation project number 324730_149792. Accordingly, my heartfelt thanks go to the principle investigator of the project, Hansjakob Furrer of the University Hospital in Bern, the Opportunistic Infections working group of COHERE, and all the cohorts within COHERE for allowing us to use their data. Thanks to Marcel Zwahlen of the University of Bern for numerous challenging discussions of the emulated trial, and to Jonathan Sterne of Bristol University, and Miguel Hernan of Harvard University, for taking the time to review the material in the final chapter.

My sincerest gratitude to Jonas Marschall and Hansjakob Furrer of the University Hospital in Bern, and John Van Den Anker, Marc Pfister and Julia Bielicki of the University Children's Hospital in Basel, for giving me the chance to change direction and work in this challenging field.

Thank you to all the patients for allowing us to use their data in the analyses.

And finally of course, wholehearted thanks to my family Louise, Jennifer and Kate for their continued patience, support and understanding through the numerous ups and downs.

# Abstract

The analysis of time-to-event data typically makes the censoring at random assumption, i.e. that — conditional on covariates in the model — the distribution of event times is the same, whether they are observed or unobserved. When patients who remain in follow-up are compliant with the trial protocol, then analysis under this assumption can be considered to address a *de-jure* ("while on treatment strategy") type of estimand.

In such cases, we may well wish to explore the robustness of our inference to more pragmatic, *de-facto*, ("treatment policy strategy"), assumptions about the behaviour of patients post-censoring. This is particularly the case when censoring occurs if patients change, or revert, to the usual (i.e. reference) standard of care.

Recent work has shown how such questions can be addressed for trials with continuous outcome data and longitudinal follow-up, using reference based multiple imputation. Such an approach has two advantages: (i) it avoids the user specifying numerous parameters describing the distribution of patient's post-withdrawal data, and (ii) it is, to a good approximation, information anchored, so that the proportion of information lost due to missing data under the primary analysis is held constant across the sensitivity analyses.

We develop similar approaches in the survival context, proposing a class of reference based assumptions appropriate for time-to-event data. We explore the extent to which sensitivity analyses using the multiple imputation estimator (with Rubin's variance formula) is information anchored, demonstrating this using theoretical results and simulation studies. The methods are illustrated using data from a randomized clinical trial comparing medical therapy with angioplasty in patients with angina.

Causal inference methods are established as the gold standard for analysing observational ("big") data. In a final step, we show that reference based methods can also be applied in this context by using sensitivity analysis in an investigation of the risk of opportunistic infections in a cohort of HIV positive individuals.

# Contents

# List of Tables

# List of Figures

# Glossary

| | |
|---|---|
| CABG | Coronary Artery Bypass Graft |
| CAR | Censoring at Random |
| cART | Combination Antiretroviral Therapy |
| CCAR | Censoring Completely at Random |
| CIR | Copy Increments in Reference |
| CNAR | Censoring Not at Random |
| COHERE | Collaboration of Observational HIV Epidemiological Research Europe |
| CPH | Cox Proportional Hazards |
| CR | Copy Reference |
| CROI | Conference on Retroviruses and Opportunistic Infections |
| EH/I | Extreme Hazard / Decrease |
| EH/I | Extreme Hazard / Increase |
| EM | Expectation-Maximisation |
| EMA | European Medicines Agency |
| FDA | US Food and Drug Administration |
| G-T | Grambsch-Therneau (test) |
| GBC | German Breast Cancer |
| HCF | Hazard Carried Forward |
| HIV | Human Immunodeficiency virus |

| | |
|---|---|
| HR | Hazard Ratio |
| HTB | Hazard Tracks Back |
| ICH | International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use |
| IDU | Intravenous Drug User |
| IE | Immediate Event |
| IPW | Inverse Probability Weighting |
| IQR | Interquartile Range |
| ITT | Intention to treat |
| IV | Instrumental Variable |
| IWHOD | International Workshop on HIV and Hepatitis Observational Databases |
| J2A | Jump to Active |
| J2R | Jump to Reference |
| LMCF | Last Mean Carried Forward |
| LOCF | Last Observation Carried Forward |
| MAR | Missing at Random |
| MCAR | Missing Completely at Random |
| MI | Multiple Imputation |
| MMRM | Mixed Model Repeated Measures |
| MNAR | Missing Not at Random |
| MSM | Marginal Structural Model |
| MVN | Multivariate Normal |
| NRC | US National Research Council |
| NRI | Non-random Intervention |
| OD | Opportunistic Disease |

| | |
|---|---|
| PCP | Pneumocystis Pneumonia |
| PH | Proportional Hazards |
| RCT | Randomised Controlled Trial |
| RMST | Restricted Mean Survival Time |
| RNA | Ribonucleic acid |
| VL | Viral Load |

# Chapter 1

# Introduction

*But, Mousie, thou art no thy-lane [alone],*
*In proving foresight may be vain;*
***The best-laid schemes o' mice an' men***
*Gang aft agley [askew],*
*An' lea'e us nought but grief an' pain,*
*For promis'd joy!*

*"To a Mouse, on Turning Her up in Her Nest with the Plough"*, Robert Burns, 1785

## 1.1 Missing data in a clinical trials

However carefully clinical trials are designed and planned, some baseline patient characteristic and — more typically — outcome data are often missing. This might occur when a patient is lost to follow-up, which could be, for example, due to non-compliance with the study protocol, or stopping an assigned treatment due to experiencing adverse effects. Of course, preventive measures, good design and consistent follow-up processes should be pursued to minimise the amount of missing data — since these would make many of the issues discussed here obsolete (LaVange and Permutt (2016) and Chapter 2 of O'Kelly and Ratitch (2014)).

In this thesis we focus primarily on missing outcome information, rather than baseline data, and in particular, such data in a time-to-event setting. Whatever the reason for the data being miss-

ing, review articles of trials suggest that perhaps 90% of trials have some kind of missing data (Wood *et al.*, 2004; Powney *et al.*, 2014; Bell *et al.*, 2014; Fiero *et al.*, 2016). In observational data settings missing data also arise, often for the same types of reasons as in a clinical trial. As we might expect, in epidemiological studies the picture regarding the levels of missing data is rather similar (Eekhout *et al.*, 2012).

Missing data cause unavoidable ambiguity in the analysis of data from clinical trials since any such analysis relies on untestable assumptions about the missing data. If a contextually implausible assumption concerning the missing data is made, the estimated treatment effect and associated variance will be biased, leading to potentially misleading inferences, which can directly influence patient care (Sterne *et al.*, 2009; White and Carlin, 2010; Ibrahim *et al.*, 2012; Jakobsen *et al.*, 2017). Hence, it is important to be clear about the assumptions being made about the missing data, and the subsequent impact of these assumptions made on the conclusions drawn. Typically, we choose a standard set of assumptions about the missing data for the primary analysis of a trial, and then investigate a number of other plausible scenarios concerning the missing data through a series of further *sensitivity analyses*. Since the observed data are consistent with different clinical interpretations, the results from the sensitivity analyses are compared with those from the primary analysis. If they are in line with one another, we may conclude that for the sensitivity analysis scenarios investigated, the outcome from the primary analysis is robust to contextually plausible departures from the assumption concerning the missing data mechanism defined for the primary analysis. If this is not the case, and the results change following the sensitivity analysis, then the investigators should report the conditions under which the results may change, along with the relative likeliness of these circumstances occurring. These steps provide more confidence in the results, especially when regulators are considering new treatments for approval.

Despite the ubiquity of missing data, until relatively recently most primary Randomized Controlled Trial (RCT) analyses either used only data from patients with complete data, that is, those with fully observed data, or in a longitudinal setting, used methods such as "last observation carried forward". While both these approaches may lead to unbiased results under certain causes of missing data, and are certainly simple to implement, they are at best inefficient.

A *complete case* analysis may also lead to less variability in treatment estimates, with its associated knock-on effect for the confidence intervals in the results from the trial. Of course, using just the subset of patients with data complete also reduces power, this issue becoming

aggravated as the number of covariates with missing data increases (page 43 of Molenberghs and Kenward (2007)).

As an alternative to just using the complete cases, we may use all the observed data, also known as an "available cases" analysis. A typical example of this is when considering longitudinal data in which all observed follow-up data at any visit is included in the analysis, irrespective of whether data from other visits for a specific subject were missing. The analysis is then based on defining a model for all the observed data and using this for inference, often employing the likelihood function or posterior distribution. This is the main focus of the text by Little and Rubin (2002), and is exemplified by the mixed model repeated measures approach (MMRM) presented in Molenberghs and Kenward (2007). However, we do not consider these approaches further here.

There are a number of possibilities for "filling in", or imputing, the missing data. We adopt the taxonomy for missing data methods from Little and Rubin (pages 19 and 60 of Little and Rubin (2002)). In terms of *imputation-based* procedures, perhaps the most obvious process would be to impute the *unconditional mean* of the respective covariate which is missing.

However, using the mean to impute has the unfortunate consequence that whilst we do not increase the information in the data, we are increasing the number of subjects in the analysis, so that the sample variance actually *decreases*. This is undesirable since we would like the imputation process to mirror the loss of information from having missing data. A variation on mean imputation is when the conditional mean is used to impute missing values, sometimes called *regression imputation*. In this case, the missing value is predicted conditional on the observed outcome and covariate values for the patients. *Stochastic imputation* follows the same approach, but adds a small amount of error to each imputed value to help with solving the lack of variability mentioned above.

For longitudinal data, a commonly used method is *last observation carried forward* (LOCF) . In this case, the last observed value is used to impute missing values for later visits without measurements. This has often been assumed to be a conservative approach, but it may equally be anti-conservative. This is the crux of the issue with LOCF — it is sensitive to the clinical context. As pointed out by Molenberghs and Kenward for the example of treatments for Alzheimer's diseases, "the goal is to prevent the patient from worsening. Thus, in a one year trial where a patient drops out after one week, carrying the last observation forward implicitly assumes no further worsening. This is obviously not conservative" (page 53 of Molenberghs

| Review article | Year | % of articles with missing data | Complete case analysis | single imputation methods | robust methods |
|---|---|---|---|---|---|
| Wood *et al.* | 2004 | 89% | 65% | 20% | 3% |
| Eekhout *et al.* | 2012 | 92% | 81% | 14% | 13% |
| Powney *et al.* | 2014 | 91% | 32% | 14% | 22% |
| Bell *et al.* | 2014 | 95% | 45% | 27% | 27% |
| Fiero *et al.* | 2016 | 93% | 55% | 8% | 29% |

Table 1.1.1: Summary of review articles on missing data

and Kenward (2007)). Despite many examples and the consistent message that using LOCF can lead to biased results (Mallinckrodt *et al.*, 2004; Carpenter *et al.*, 2003; Beunckens *et al.*, 2005), it is often still used as the simplest alternative, particularly in analyses of observational cohort data. There are also other single imputation methods, mostly developed in other settings such as survey analysis. For example, *hot deck single imputation* fills in missing values with those from people with similar characteristics (Little and Rubin, 2002). *Predictive mean matching* is similar to this — values are imputed by finding the "nearest-neighbour" to the individual with missing values, and using this donor's observed values as substitutes (van Buuren, 2012). All such single imputation methods generally further aggravate problems because the analysis cannot distinguish between actual and imputed values, and so underestimate the variance.

Multiple imputation (MI) , the method we use predominantly for the work presented here, essentially builds on stochastic imputation by incorporating additional variability into the imputation process. Being Bayesian in nature MI assumes estimates from fitting a model to the observed data are normally distributed and uses a draw from this distribution to inject variability into the imputed data. In a further step following imputation, an additional component is added to the variance calculation to ensure that it is suitably inflated to reflect the information lost from the missing data (MI is defined in more detail later in this chapter). We note at this point that under missing at random (MAR) , the maximum likelihood based approaches (e.g. MMRM) mentioned above, and those involving MI will end up with essentially the same results (up to Monte Carlo error). This can of course be used as a useful cross-check of the MI process prior to investigating more complex missingness mechanisms, assuming a closed form solution for the likelihood function is available.

Table 1.1.1 briefly summarises five review articles of major medical journals showing that despite high levels of missingness in trials, very few used statistically valid methods such as mul-

tiple imputation or likelihood based approaches to suitably account for missing data. Nonetheless, the increase in the number of studies in the period 2004 to 2014 using such methods is striking. This trend mirrors the increase in availability of standard software using more rigorous methods during this period, which has allowed specialists and non-specialists alike to perform more robust missing data analyses (Rezvan *et al.*, 2015).

This trend is encouraging, showing that the adoption of more reliable methods is possible, if supported with software implementations. Interestingly, the review by Eekhout *et al.* referred to in Table 1.1.1 focussed on epidemiological studies, but the results are very similar to the other reviews concerning missing data in trials.

Table 1.1.1 also highlights the importance of defining methods for handling missing data that are not only valid in a statistical sense, but which are also convenient in terms of ease of use: Adoption of new methods is often directly related to simplicity of implementation. This is the first of three key requirements which need to be taken into account when defining new statistical methods:

| |
|---|
| The first key facet when considering new sensitivity analysis methods is their *practicality*, that is, their ease of implementation and use. |

We will define three such key facets in this introductory chapter, and due to their importance, we will refer back to them in the remainder of the thesis as motivation regarding the proposed sensitivity analysis approaches.

There is, however, a potential downside from the uptake of new missing data methods driven by increased use of readily available software. Software implementations often implement a default set of assumptions regarding the missing data. Whilst the standard assumptions often correspond to the most natural starting point, and are certainly the most straightforward to perform quickly, there is a tendency for the user to accept the premise of these standard assumptions without much reflection and consideration of potential alternatives. The missing data analysis using the standard set of assumptions often stops at this point, without exploration of what would have happened with other, perhaps more tenable scenarios, in the context of a specific trial. Framing these other scenarios so that they are i. *clinically plausible*, ii. *accessible*

in terms of the assumptions made, and iii.) relatively easy to be implemented, is the focus of this thesis.

To help us think about missing data, and the different assumptions that might be applicable for such data, it is often helpful to consider the potential relationship between the observed data and missing data. A common framework for such assumptions was proposed by Little and Rubin (e.g. page 12 of Little and Rubin (2002)), and these may be used to provide the foundation for the assumptions underlying the primary and subsequent sensitivity analyses with regards to missing data. Little and Rubin proposed these definitions for different types of missing data, and these have also been adopted for medical settings.

Let $\boldsymbol{Y} = (y_{i,j})$ be a $(n \times K)$ rectangular data set with the $i$th row $y_i = y_{i,1}, \ldots, y_{i,K}$ where $y_{i,j}$ is the value of the $y_j$th variable for subject $i$. Define the missing data matrix $\boldsymbol{M} = (m_{ij})$, such that $m_{i,j} = 1$ if $y_{i,j}$ is missing and $m_{i,j} = 0$ if $y_{i,j}$ is observed. $\boldsymbol{M}$ defines the *pattern* of missing data.

The missing data mechanism may be defined by the conditional distribution of $\boldsymbol{M}$ given $\boldsymbol{Y}$, say $f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\phi})$, where $\phi$ are the unknown parameters of this distribution.

Now, if missingness does not depend on the values of the data $\boldsymbol{Y}$, *missing* or *observed*, so that

$$f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\phi}) = f(\boldsymbol{M}|\boldsymbol{\phi}) \quad \text{for all } \boldsymbol{Y}, \boldsymbol{\phi}, \tag{1.1.1}$$

then the data are *missing completely at random* (MCAR) . The missingness in this case does not depend on the data values at all.

Now, let $\boldsymbol{Y}_{obs}$ be the observed data, and $\boldsymbol{Y}_{mis}$ be the values that are missing.

If the missingness depends only on the $\boldsymbol{Y}_{obs}$, but not on the $\boldsymbol{Y}_{mis}$, then the missing data mechanism is said to be *missing at random* (MAR),

$$f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\phi}) = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\phi}) \quad \text{for all } \boldsymbol{Y}_{mis}, \boldsymbol{\phi}. \tag{1.1.2}$$

We have suppressed the covariates in these expressions, but MAR implies that the missingness process is dependent on both the observed outcome data and any covariates (baseline or time varying). MAR is the most commonly applied assumption for the missing data process in the

analysis of RCT and observational data.

Finally, if the missing data mechanism depends on the values of Y, observed *and* missing, then it is said to be *missing not at random* (MNAR) . For clarity,

$$f(\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{\phi}) = f(\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{\phi}) \quad \text{for all } \boldsymbol{Y},\boldsymbol{\phi}. \tag{1.1.3}$$

Example 1

A simple example to illustrate this is as follows (adapted from page 12 of Little and Rubin (2002)). The CD4 count is a biomarker used to track disease progression in the study of the Human Immunodeficiency Virus (HIV) . Regular measurement of the CD4 count is an important diagnostic tool for clinicians, but turning up for measurement visits is thought to be dependent on certain risk factors. So, for example intravenous drug users (IDUs) are thought to have a higher risk of not turning up regularly.

Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ be a random sample of CD4 counts from patients in a specific month, and define $\boldsymbol{M} = (m_1, \ldots, m_n)$ to be the vector of missingness indicators, with $X$ denoting an indicator variable for whether the patient is an IDU ($X = 1$) or not ($X = 0$). Furthermore, suppose the joint distribution of the outcome and missingness $f(y_i, m_i)$ is independent between subjects, then,

$$f(\boldsymbol{Y},\boldsymbol{M}|X,\boldsymbol{\theta},\boldsymbol{\phi}) = f(\boldsymbol{Y}|X,\boldsymbol{\theta})f(\boldsymbol{M}|X,\boldsymbol{Y},\boldsymbol{\phi}) = \prod_{i=1}^{n} f(y_i|X,\boldsymbol{\theta}) \prod_{i=1}^{n} f(m_i|y_i,x_i,\boldsymbol{\phi}),$$

where $f(y_i|x_i,\boldsymbol{\theta})$ is the density of $y_i$ with unknown distributional parameters $\boldsymbol{\theta}$, and $f(m_i|y_i,x_i,\boldsymbol{\phi})$ is the density of a Bernoulli distribution for the missingness indicator $m_i$, such that the probability $y_i$ is missing is $Pr(m_i = 1|y_i,x_i,\boldsymbol{\phi})$.

If missingness is independent of the CD4 count $Y$, so that $Pr(m_i = 1|y_i,\boldsymbol{\phi}) = \boldsymbol{\phi}$, then the missing data mechanism is MCAR. We are making the assumption that the patients not turning up for their measurement visits is a chance occurrence — tantamount to saying "anyone can forget a doctor's visit".

Now, if the missingness is random after conditioning on whether the patient is an IDU or not, then the missingness is at random (MAR). In this case, we are making the assumption that not turning up at a visit is a random occurrence within each strata of $X$, but that, for example, IDUs may have a higher risk of not turning up. Contrastingly, if $Pr(m_i = 1|y_i, x_i, \boldsymbol{\phi}) = f(y_i, x_i, \boldsymbol{\phi})$, that is, a missing visit is dependent on both observed and missing values of $y_i$ and $x_i$, then the missing data mechanism is MNAR. In this case, we suspect that not turning up for a visit is dependent on being an IDU, and the patient's disease status, as measured by the CD4 count — this could indeed be a plausible assumption for this example. ■

Furthermore, if we consider longitudinal measurements $y_{ij}$ for patient $i$ at timepoints $j$, as explified by the CD4 counts for each patient introduced above, then the missingness *pattern* of the measurement is often also of interest. *Monotone* missingness is a pattern in which if $y_{ij}$ is missing, then all subsequent measurements are also missing for that patient. We assume monotonone missingness for the time-to-event data which we consider in this thesis. If the missingness pattern is non-monotone, so there is intermittent missingness, then special methods often have to be applied (recent examples of which are Sun *et al.* (2018) and Perkins *et al.* (2018)).

Rubin introduced an additional definition for *ignorability*. Harel states

> "Rubin (1976) introduced the concepts regarding how to find the minimum condition under which the missingness process does not need to be modeled (in likelihood or Bayes) — in other words, when standard MI is valid. For that to occur, two assumptions must hold. First, the MAR or MCAR assumption must be valid. Second, the parameter estimates used for imputation and those estimated in the analysis model must be independent (distinct). Together, these 2 assumptions imply *ignorability*, which means that the missingness model necessary under MNAR can be ignored and the observational data will be sufficient", (italics added) (Harel *et al.*, 2018).

These definitions provided the basis for discussion of the missing data assumptions underpinning the primary and sensitivity analysis scenarios for a trial.

The next section reviews terminology which clarifies the relationship between the underlying

assumptions concerning the missing data within a trial, and how they are framed in terms of the clinical end point.

## 1.2 Estimands

With any clinical trial it is important to define the *estimand* of interest. According to the latest European Medicines Agency (EMA) addendum to the guideline on statistical principles for clinical trials regarding estimands and sensitivity analysis in clinical trials, the estimand

> "is the target of estimation to address the scientific question of interest posed by the trial objective." (CHMP, 2018)

To put this definition into context, the estimand is the quantity of interest whose true value we would like to determine. An estimator is a method for estimating the estimand. An estimate is an approximation of the estimand that comes from the use of a specific estimator.

In the language of causal inference, which we encounter later in Chapter 5, estimands are defined in terms of their *potential* outcomes. Thus, a causal estimand in a randomised controlled trial quantifies the effect of the treatment relative to the control, but also introduces a counterfactual component. Thus, in the causal literature we are interested in estimating what would have happened to the *same* subjects under *different* treatment conditions. Since patients are randomised to an active treatment or the control in such a setting, we are not able observe the same subject under both the treatment and control — we are only able to observe a subject's observed response to taking the active treatment (say), but not the control, and vice versa.

The definition of the estimand determines which data are used in the primary analysis. This includes a non-ambiguous definition regarding which data are considered missing. For example, data which are observed but not directly applicable in the primary analysis because they have been collected after treatment switching. Complementing the definition of the estimand are the statistical methods (e.g. multiple imputation) we use for estimation and inference. In addition, we may well need to make some further primary analysis assumptions — for example, that the data is missing at random — to perform the primary analysis.

The ICH E9 addendum goes on to describe the four attributes of an estimand:

- "the population, that is, the patients targeted by the scientific question."
- "the variable (or endpoint), to be obtained for each patient, that is, required to address the scientific question."
- "the specification of how intercurrent events are reflected in the scientific question of interest." Intercurrent events are "events that occur after treatment initiation and either preclude observation of the variable or affect its interpretation". So, for example, for time-to-event data censoring would considered an intercurrent event.
- "the population-level summary for the variable which provides the treatment effect of interest" (CHMP, 2018).

We use *sensitivity analysis*, focussed on this same estimand, to investigate the *sensitivity* of inference for the specific set of primary analysis assumptions relating to the missing data. In this way, we are able to explore the impact of the untestable assumptions underlying the primary analysis. In line with current thinking, we differentiate between *de-jure* and *de-facto* estimands (Carpenter *et al.*, 2013; Akacha *et al.*, 2017) to clarify the assumptions underpinning the primary and sensitivity analyses. Briefly, and again in the language of the ICH E9 addendum, *de-jure* equates to a "while on treatment" estimand usually associated with treatment *efficacy*, whereas *de-facto* would be considered a "treatment policy" type of estimand, frequently related to treatment *effectiveness*.

*De-jure* estimands

For a specific estimand we define a "*deviation from the study protocol relevant to the estimand*" Carpenter and Kenward (2012) — that is, a violation of the protocol such that post-deviation data can no longer directly be used for inference regarding the estimand. It is difficult to make sweeping statements as regards to what constitutes a deviation, since this will be trial specific. However, typical examples of a deviation relevant to a *de-jure* estimand would be unblinding, non-compliance with treatment, withdrawal from treatment and loss to follow-up. In contrast, for a *de-facto* estimand, non-compliance with treatment and withdrawal from treatment might not be considered a deviation (page 246 of Carpenter *et al.* (2014)). From these typical examples of deviation, we can see that the resulting post-deviation data sets may contain slightly different numbers of patients, and/or number of visits for each patient in a longitudinal trial.

An estimate pertaining to a *de-jure* estimand might assume that, post deviation, patients con-

tinue to follow their randomised arm defined in the study protocol. In the context of estimating treatment effects, as opposed to evaluating safety, the *de-jure* estimand addresses questions of *efficacy*, as if the assigned treatments were taken as specified in the protocol. For a safety endpoint, the *de-jure* estimand is typically of primary interest. For example, this might determine whether, under ideal compliance conditions, there are a significant number of (serious) adverse events. Accordingly, the assumptions underpinning the estimate of a *de-jure* estimand for the primary analysis may actually be counterfactual.

*De-facto* estimands

*De-facto* estimands, on the other hand, apply to the treatment effect based on the original randomisation. In this case, we are measuring the effect of being in a particular treatment group, irrespective of subsequent compliance, and are not measuring treatment compliance itself. *De-facto* estimands are therefore concerned with questions of *effectiveness*, that is, the treatment effect we might expect in practice if the treatment were used in the conceptual target population at large (of course provided they behave as in the clinical trial). For a safety endpoint a *de-facto* estimand typically would be less appropriate. For example, in a placebo controlled trial, a *de-facto* estimand would typically be a conservative estimate of the treatment effect. If the treatment effect is not statistically significant, then a *de-facto* estimand would be inappropriate as a safety endpoint since "one could naively conclude that a treatment is safe because the ITT [intention to treat, equivalent in this case to a *de-facto* estimand] effect is null, even if treatment causes serious adverse effects. The explanation may be that many subjects stopped taking the treatment before developing adverse effects", (Toh and Hernan, 2008). This example emphasises the unifying nature of the *de-jure* and *de-facto* definitions for estimands, applicable for both treatment effect and safety related outcomes.

In our settings, the *de-facto* estimand is that which usually relates to the sensitivity analysis scenarios which we wish to investigate. Of course, with no protocol deviations, *de-facto* and *de-jure* estimands are equivalent.

At this point it is worthwhile to point out that it is not necessarily always the case that the primary assumption is *de-jure* in a trial. The primary and sensitivity analysis assumptions could assume different *de-facto* behaviour, such as in a pragmatic trial. This is the case in the illustrative application provided in Chapter 2 in the context of the RITA-2 trial — an example of an estimand following a "treatment policy strategy" (page 7 of CHMP (2018)).

This vocabulary establishes a framework for the analysis which includes the:

1. Estimand — encompassing the decision of whether *de-jure* or *de-facto* applies, and associated definitions for what constitutes "deviation", after which we assume data are missing.
2. Primary analysis — including assumptions regarding the missing data, statistical methods and inference.
3. Sensitivity analyses about the missing data — including statistical methods and inference.

Alongside progress made in conceptualising the way we think about missing data in trials, guidelines have also been published for addressing the issues raised by missing data in this context, specifically relating to policy, regulatory process and methodology. The next section reviews current guidelines regarding sensitivity analyses.

## 1.3 Regulatory Framework

The European Medicines Agency (EMA) published a key document in 2010 detailing guidelines on missing data in confirmatory clinical trials, which highlights issues associated with analysis of primary efficacy endpoints when patients are followed up longitudinally (CHMP, 2010). Focussing specifically on sensitivity analysis, the EMA states:

*"Sensitivity analysis should show how different assumptions influence the results obtained"*, CHMP (2010).

A 2010 Food and Drug Adminstration (FDA) mandated report by the US National Research Council (NRC) on the prevention and treatment of missing data in clinical trials goes into more detail, documenting guidelines, methods and providing recommendations on the prevention and treatment of missing data in clinical trials NRC (2010). Recommendation 15 of the NRC report echoes this, stating:

*"Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about missing data mechanisms should be a mandatory component of reporting"*, NRC (2010).

Underlining the importance of sensitivity analysis, Recommendation 18 of the same report goes on to say that:

*"There remain several important areas where progress is particularly needed, namely: (1) methods for sensitivity analysis and principled decision making based on the results from sensitivity analyses . . . "*

More recently, the proposed addendum to the ICH E9 (2017) guideline clarified vocabulary and presented tangible examples for framing sensitivity analysis in the context of clinical trials, stating in §A.5.2.2:

*"Missing data require particular attention in a sensitivity analysis because the assumptions underlying any method may be hard to justify and impossible to test."*

In summary, since missing data introduce ambiguity into inference for trial estimands, sensitivity analysis is desirable, if not mandatory, to explore the robustness of the conclusions to a range of plausible assumptions.

With this in mind, the missing at random (MAR) assumption would seem to be the natural starting point for a sensitivity analysis, since this implies that the conditional distribution of later follow-up data given earlier follow-up data are the same, whether or not we see the later data. Since we make essentially this assumption when we apply the results from the trial data to the broader population, this is the logical point of embarkation for subsequent sensitivity analyses.

Whilst there has been significant progress made in defining sensitivity analysis methods (for example, part V onwards in Molenberghs and Kenward (2007), chapter 8 onwards in Daniels and Hogan (2008), and chapter 7 in O'Kelly and Ratitch (2014), and the references therein), there is a lag in providing practical and accessible methods. Indeed, the NRC singles out *"methods for assessing and limiting the impact of informative censoring for time-to-event outcomes"* as an area in need of further research (NRC, 2010). This statement provided the key impetus to start work on the PhD in 2012.

The focus of the thesis is to develop and adapt sensitivity analysis approaches defined for longitudinal data with a continuous outcome to the time-to-event setting. In the next section we introduce and discuss time-to-event data, and in particular, the specialities associated with this type of data when performing sensitivity analysis.

## 1.4 Statistical methods for analysing time-to-event data

Survival analysis is often used to model time-to-event data in clinical and observational studies. However, event times are sometimes not observed, and these are referred to as *censored* at the patient's last observation. This happens for many reasons, *e.g.* withdrawal from treatment due to adverse effects, loss to follow up, or because the scheduled end of funded follow-up of the study is reached before the event occurs. We consider exclusively right censored data since this is the most commonly occurring type of time-to-event data in a trial setting. In the interests of completeness, first we briefly recap the standard definitions and terms used in survival analysis.

*Definition of right censoring*

Let $i$ denote subjects and let $\boldsymbol{T}, \boldsymbol{C}$ be, respectively, random variables denoting the event and censoring time. We observe $y_i = \min(t_i, c_i)$, with $t_i \in \boldsymbol{T}$ and $c_i \in \boldsymbol{C}$, and define $\boldsymbol{R}$ to be a vector of censoring indicators, $r_i$, for each subject such that

$$r_i = \begin{cases} 1 & \text{if } c_i \leq t_i \\ 0 & \text{if } c_i > t_i \end{cases}$$

*Definition of survival function*

Let $\boldsymbol{T}$ be a positive continuous random variable with density function $f(t)$ and cumulative distribution function $F(t)$, then the probability that the time-to-event is larger than a time $t$ is the survival function $S(t)$,

$$S(t) = Pr(T > t) = \int_t^\infty f(u)du = 1 - F(t).$$

$S(t)$ is a monotonically decreasing function.

*Definition of the hazard function*

The hazard function $h(t)$ is defined to be the event *rate* at time $t$ conditional on survival up until

time $t$ or later. So, if we suppose that a subject has survived up to time $t$, but will not survive until a short time later, denoted by $dt$, then,

$$h(t) = \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt} = \lim_{dt \to 0} \frac{Pr(t \leq T \leq t + dt)}{dt Pr(S > t)} = \frac{f(t)}{S(t)},$$

where $f(t)$ is the density function of $F(t)$, $f(t) = F'(t) = \frac{d}{dt}F(t)$. Since $f(t) = -\frac{d}{dt}S(t)$, there is the following relationship between the survival function and the hazard, $h(t) = -\frac{d}{dt}\log(S(t))$. From these definitions, we obtain the following expression for the cumulative hazard $H(t)$,

$$H(t) = \int_{-\infty}^{x} h(u)du = -\log\ S(t),$$

or equivalently, $S(t) = \exp(-H(t))$.

Having established the definitions and properties for a typical survival analysis, we return to considering censoring in more detail.

As with missing data, censored patients cannot be ignored; they have important information to convey, and this additional information has to be included in the analysis. There are clear parallels between censored data and missing longitudinal visit information, since in both cases, we are often aware of the time point at which a patient was still present in the trial, that is their last known visit, and their status at this time, and thereafter no further information is available.

Accordingly, censoring may be considered as a type of missing data process. The definitions from the introductory remarks to this chapter from Little and Rubin (2002), and methodologies from the field of missing data analysis, can also be used, albeit with appropriate minor modifications and nuances.

Rather than referring to the observed variables being "missing", the definitions are altered to reflect the censoring i.e. *Censoring Completely at Random* (CCAR) , *Censoring at Random* (CAR) , *Censoring not at Random* (CNAR) . Essentially, the definitions remain the same as with the missing data case — therefore, for example, a censoring at random mechanism means that the censoring and event time distribution are independent, conditional on the observed outcome and covariates.

To understand what these expressions mean in practical terms, we consider a likelihood based

approach. Let us assume we have a random sample of $N$ patients, and following the definitions just introduced, we have data couples $(t_i, c_i)$. We actually observe $(y_i, r_i)$ for each of the $i = 1, \ldots, N$ patients, and analogously define the cumulative distribution function for the censoring times, $C$, as $G(t)$ with density function $g(t)$. If we assume the event time and censoring time distributions are independent, then we can write the likelihood function as:

$$L(\theta, \phi; y, r) = \prod_{i=1}^{n} \{[f(y_i; \theta)]^{r_i} [S(y_i; \theta)]^{(1-r_i)}\} \{[g(y_i; \phi)]^{(1-r_i)} [S(y_i; \phi)]^{(r_i)}\}, \tag{1.4.1}$$

for $y_i \in Y$, and where $\theta$ and $\phi$ are the parameters of the event time and censoring distribution respectively. Assuming our primary interest is in the event time distribution, then assuming independence between event and censoring distributions $f(\cdot)$ and $g(\cdot)$, and $\theta$ and $\phi$ do not have any common parameters, so we only have to consider the first half of this expression:

$$L(\theta; y, R) = \prod_{i=1}^{n} \{[f(y_i; \theta)]^{r_i} [S(y_i; \theta)]^{(1-r_i)}\}, \tag{1.4.2}$$

or substituting in the expression for the hazard,

$$L(\theta; y, R) = \prod_{i=1}^{n} \{[h(y_i; \theta)]^{r_i} [S(y_i; \theta)]\}. \tag{1.4.3}$$

With the above definition, the censoring and event time processes are not linked. Here, we have suppressed the baseline covariates in the expression, but had we not, assuming this expression holds irrespective of any subgroups of patients, then we would have Censoring Completely at Random (CCAR). If the event and censoring times are independent, conditional on the covariates, then this would imply the censoring process is at random (CAR). If the event and censoring time processes are not independent, then censoring is not at random (CNAR), also known as "informative" censoring.

Again, analogously to missing data, when analysing a trial with censored data we might proceed by performing the primary analysis under the standard censoring assumption, typically censoring at random. We would then carry out a pre-specified sensitivity analysis under another set of assumptions, typically in which censoring was *informative* (CNAR).

We have now prepared the foundations for our discussion of sensitivity analysis for time-to-event data. We begin by providing an often used general classification of approaches to sensitivity analysis modelling, before going on to focus on the methods we have used.

## 1.5 Sensitivity analysis approaches

### 1.5.1 Introduction

Our goal is to estimate a treatment effect, typically modelled using proportional hazards. To do this, we need to model a patient's time-to-event, conditional on treatment and other appropriate, contextually relevant covariates.

Most modelling approaches for investigating departures from censoring at random involve either a *selection* based mechanism where the missing data is explicitly defined, or alternatively, where different conditional distributions are defined for the missing data, based on properties of the observed variables, leading to the explicit modelling of *patterns* of missingness (Hogan and Laird, 1997b).

More formally, and using the notation introduced in the last section, given the joint distribution $P(Y, R)$, for event times $Y$ and censoring indicators $R$, we can re-formulate the joint distribution in terms of either a selection or pattern mixture mechanism (Hogan and Laird (1997a), cf. page 17 of Carpenter and Kenward (2012)):

$$Pr(r_i|y_i)Pr(y_i) = Pr(y_i, r_i) = Pr(y_i|r_i)Pr(r_i) \tag{1.5.1}$$

where the middle term of this expression is the joint distribution of event and censoring given the covariates. Covariates have been suppressed in the above, but of course are allowed. The equalities in the expression underline that we may, in principle, specify a missingness mechanism in either modelling paradigm, although as Carpenter and Kenward point out "even in apparently simple settings, explicitly calculating the selection implication of a pattern mixture model, or vice versa, can be awkward" (page 18 of Carpenter and Kenward (2012)).

### 1.5.2 Selection models

The left hand side of equation (1.5.1) expresses the joint distribution as a *selection model*, that is, a product of the density of the censoring process, conditional on the event time distribution and covariates, and the marginal distribution of the event times given the covariates. A schematic overview of selection modelling methods is presented in Figure 1.1 of Molenberghs and Kenward (2007), along with requisite theory and examples (particularly Chapters 15 and 19).

Informative censoring for time-to-event data has been the subject of much research using selection modelling approaches. Scharfstein *et al.* (1999) initially proposed a semi-parametric selection model, and subsequently refined their methodology in a number of papers (Scharfstein *et al.*, 2001; Shardell *et al.*, 2008; Scharfstein and Robins, 2002; Rotnitzky *et al.*, 2002; Scharfstein *et al.*, 2018). Interestingly, the section on time-to-event data in the NRC report mentioned in section 1.3 only mentions this methodology for sensitivity analysis (page 105 of NRC (2010)).

Siannis *et al* build on this work, developing "local sensitivity analysis" for time-to-event data (Siannis, 2004; Siannis *et al.*, 2005; Siannis, 2011). The methods approximate the effect of limited small dependencies between censoring and failure by adding a perturbation term to the maximum likelihood expression used when assuming CAR. This approach avoids having to explicitly model the joint distribution of censoring and failure. Sensitivity parameters are restricted to a small range of values, outside of which the approximation may no longer be adequate.

Bradshaw *et al.* (2010) take a slightly different approach to investigate non-ignorably missing covariates using a full Bayesian approach, extending earlier formulations of survival analysis for CAR data (e.g. Ibrahim *et al.* (2001)). The authors note that although selection models can be sensitive to miss-specification (e.g. Herring *et al.* (2004)), the inclusion of some of the covariates indicative of missingness help to improve model fit and convergence.

Whilst there is substantial methodological literature on selection models, they are often less well used in practice. This is because they require more specialist modelling skills, are not often implemented in commercial software, and the selection model parameters are quite difficult to interpret.

### 1.5.3 Pattern mixture models

The right hand side of equation (1.5.1) defines the joint distribution in *pattern mixture terms* — a product of the probability distribution of the event times within each censoring pattern given the covariates, and the marginal probability of each censoring pattern occurring, given the covariates. The theory and background for pattern mixture models are discussed, for example, in Chapter 8.4 of Daniels and Hogan (2008), Chapter 10 of Carpenter and Kenward (2012) and in Chapter 7 of O'Kelly and Ratitch (2014).

For time-to-event data, a recent paper by Jackson *et al.* (2014) explores sensitivity analysis under departure from CAR for the Cox proportional hazards model using multiple imputation (MI) combined with bootstrapping to generate the imputed data sets. Their pattern mixture modelling approach builds on the concept that censoring introduces a "shock" to the patient hazard (adopted from Letué (2008)). An explicit sensitivity analysis parameter is introduced into the model allowing newly imputed event times for censored patients to either reflect an improvement or a deterioration in their post-censoring condition. This is the same principle used for so-called "*delta*" ($\delta$) sensitivity analysis methods in the missing data literature (for example, Leacy *et al.*, 2017; Tompsett *et al.*, 2018, and references therein). Such $\delta$ methods are often implemented to conduct sensitivity analyses, and therefore we explain the principles behind the method in more detail.

Again, we let $\boldsymbol{Y}$ be an independent positive random variable denoting the event time process with censoring indicator $\boldsymbol{R}$, with covariate dependent hazard function $h(y_i|x_i)$, for fully observed covariates, $x_i \in \boldsymbol{X}$. A new event time $y_i$ for a patient censored is generated by augmenting the hazard rate under CAR by a sensitivity parameter $\delta$,

$$h(y_i|x_i) = \begin{cases} h_{CAR}(y_i|x_i) & \text{if} \quad r_i = 1 \\ \exp(\delta) h_{CAR}(y_i|x_i) & \text{if} \quad r_i = 0 \end{cases}$$

and imputing an event time from the corresponding inverted cumulative hazard function using the method of Bender *et al.* (2005).

As the parameter $\delta$ is varied, so the robustness of the conclusions to departures from CAR can be investigated. Jackson *et al.* varied $\delta$ in the range $[-3, -2, -1, \ldots, 10]$, and then compared the results with those from the primary analysis under CAR. A variation of the $\delta$ method, the

so-called "tipping point" analysis, changes the $\delta$ parameter until the treatment difference is no longer statistically significant, assuming this was the case for the outcome from the primary analysis. If the primary analysis did not result in a statistically significant treatment difference, then alternatively, the $\delta$ can be adjusted until the treatment difference *becomes* significant. In either case, it is then up to the trial team to decide if the "tipping point" represents a clinically plausible multiplier of, for example, the baseline hazard.

This highlights one of the main drawbacks of such $\delta$ methods, namely, choosing a meaningful range of parameters for $\delta$, and then benchmarking them in some way to the concrete clinical setting. Such decisions require iterative discussions within the trial team and are often difficult to conclude satisfactorily, especially when considering $\delta$ multipliers of a hazard or odds ratio. Gilbert *et al.* (2013) suggest developing standard bounds, and increments between these bounds in which to vary $\delta$. Carpenter and Kenward (2012) have proposed that the sensitivity parameter is sampled from from a normal distribution, rather than taking a pre-defined range of values.[1]

In an observational data setting, Brinkhof *et al.* (2010) adopt a novel solution to the problem of dimensioning the $\delta$ sensitivity analysis parameter. In their analysis, as imputation model they embed $\delta$ into the parametric Weibull model for the hazard:

$$h(t_i|C_i, X_i) = \begin{cases} \exp(X_i^\beta)\gamma t^{\gamma-1} & \text{if} \quad t_i < C_i \\ \exp(\delta)\exp(X_i^\beta)\gamma t^{\gamma-1} & \text{if} \quad t_i \geq C_i, \end{cases}$$

where $t > 0$, and $\gamma$ is the usual shape parameter of the distribution. As analysis model they fitted the Kaplan-Meier product limit estimate to estimate 1-year survival. Their sensitivity analysis approach was used to explore the robustness of inference concerning mortality in HIV positive patients lost to follow-up in sub-Saharan Africa. A meta-analysis of five *other* Southern Africa observational studies was used to dimension $\delta$ appropriately.

This solution to the dimensioning problem of course assumes that similar studies are available to define a suitable range for $\delta$. When this is not the case, then dimensioning $\delta$ is often difficult, as pointed out in a recent study involving observational data from a Southern African HIV cohort carried out by Leacy *et al.*. They note in their discussion that "... we encountered some difficulty in selecting an appropriate range of delta values" for their sensitivity analysis (Leacy

[1]Interestingly, in this case we lose information by assuming $\delta$ is sampled from a distribution, rather than being fixed at a specific value, and this means that the information anchoring principle defined later may no longer hold.

*et al.*, 2017).

In a different context, Mason *et al.* (2017a) leverage the Bayesian approach and elicit expert opinion concerning $\delta$, again using a pattern mixture model implemented using multiple imputation. However, this has proved controversial due to the difficulty in eliciting priors in a controlled manner (Heitjan, 2017).

Therefore, although the local sensitivity analysis methods from selection modelling and the $\delta$ methods using pattern mixture models are elegant and relatively straightforward to implement, they raise questions as to the definition of a meaningful *range* for the sensitivity parameter in the context of the specific clinical trial, and this might explain why, to date, their use has been relatively limited in trials. As Daniels and Hogan point out (quoting from Scharfstein *et al.* (1999)) when defining key guidelines for such sensitivity analysis methods (Daniels and Hogan, 2008):

*"...the biggest challenge in conducting sensitivity analyses is the choice of one or more sensitivity parameterized functions whose interpretation can be communicated to patient matter experts with sufficient clarity..."*

In terms of the $\delta$ method there appears to be no "golden ticket" to resolving the dimensioning issue.

In summary, the complexity in defining sensitivity analyses to reflect *clinically plausible* scenarios, that also utilise appropriately understandable (e.g. to non-statisticians) measures of uncertainty regarding the parameters, represents a significant hurdle to the adoption of these methods.

| |
|---|
| This represents the second key facet when considering new sensitivity analysis methods — their clinical plausibility, including the ability to contextualise them to the trial team and other key stakeholders. |

There is a third type of modelling approach, "shared parameter models", which are less well known mainly due to their relative complexity compared to pattern mixture and selection models.

### 1.5.4 Shared parameter models

The final type of modelling approach is known collectively as *shared parameter* models, or *frailty* models in a time-to-event data setting. These models include latent random effects shared between both factors in the joint distribution (see, for example, Chapter 17 of Molenberghs and Kenward (2007)). Using the same notation as above, assuming $y_i$ and $r_i$ are conditionally independent, given frailty (random) effects $b_i$, a shared parameter model can be expressed as:

$$Pr(y_i, r_i) \;=\; \int Pr(y_i|r_i, b_i)Pr(r_i|b_i)f(b_i)db_i, \tag{1.5.2}$$

with the shared parameter $b_i$ being a latent effect following an unestimable, user specified distribution, which drives both the event and missingness process.

A brief overview of these methods and associated examples is provided in Chapter 17 of Molenberghs and Kenward (2007). Early adoption of such approaches proved difficult due to the lack of commercially available software. More recently these models have found widespread popularity due to software implementations both in R (Rizopoulos (2012)) and Stata (Lambert and Royston, 2009; Crowther *et al.*, 2013).

Latent class models are an extension of shared parameter models which "capture unmeasured heterogeneity between the subjects through a latent variable" (page 432 of Molenberghs and Kenward (2007)). Following fitting of the model, classification according to the latent groups is possible. This provides a rather elegant pattern mixture based sensitivity analysis of the outcome conditional on these groups (Muthen *et al.*, 2011; Beunckens *et al.*, 2008; Proust-Lima *et al.*, 2014).

In terms of applying these approaches for time-to-event data, there are now several examples. Bivariate and frailty models for explicitly linking the censoring and failure mechanisms are investigated in the papers by Emoto and Matthews (1990) and Huang and Wolfe (2002). Thiebaut *et al.* (2005) analysed clustered survival data with dependent censoring using frailty models to define the propensity for failure assuming patients in the same cluster share a common unobserved frailty, rather like mixed effects models for continuous data. The model allows for different types of censoring, some of which may be informative.

Relevant for our later illustrative example in Chapter 5, Taffé *et al.* (2008) proposed a joint

modelling approach involving the time since infection, the CD4 trajectory and the drop-out process. More recently, Li and Su proposed a joint model for informative drop out with a longitudinal biomarker fitted to data from HIV observational cohort (Li and Su, 2018). We revisit this type of data in our example application in Chapter 5. These approaches are undoubtedly at the cutting edge of methodological research for studies involving HIV cohort data. However, as pointed out by Li and Su, quoting Chapter 8 of Daniels and Hogan (2008) "research for sensitivity analysis strategies under the shared parameter framework is very limited and it is not clear how to perform sensitivity analysis without changing the inferences on the observed data". The interpretation of the last part of the sentence is a little opaque, but we assume it is referring to the additional requirement to choose an appropriate distribution for $f(b_i)$ in shared parameter models, and the influence this has on the results, which makes sensitivity analysis using such an approach considerably more complex.

We chose a different approach to sensitivity analysis, based on pattern mixture models implemented using multiple imputation, which we feel is potentially more practical in the sense of our definition earlier in this chapter, making the assumptions made for the sensitivity analysis more accessible, which in turn helps to frame the scenarios in such a way that they are clinically plausible. Here, by "accessible" we mean that the relevant assumptions for the clinical context can be made transparently.

The next section introduces the final piece of the jigsaw in terms of defining the key requirements when considering the appropriateness of new sensitivity analysis methods.

## 1.6 Information Anchoring principle

Cro *et al.* proposed the *information anchoring* principle which we present here because of its importance for the ideas we develop in subsequent chapters. We begin by transposing their definition to the survival context.

Consider a clinical trial in which time-to-event data is collected from patients, denoted by $Y$, in order to estimate a treatment effect $\theta$. We denote those patients experiencing the event by $Y_{obs}$, and those censored by $Y_{cens}$. We make a primary set of assumptions, for example, that all censored patients are "censored at random" (CAR), meaning that, in a frequentist sense, the censoring mechanism can be fully accounted for by conditioning on the covariates of the $Y_{obs}$

patients with events. The estimate of $\theta$ under this primary assumption is denoted by $\hat{\theta}_{obs,CAR}$.

Furthermore, let us assume that we are able to observe a realisation of the event times for the censored patients, $Y_{cens,CAR}$, under the primary assumption of CAR. Of course, this is a hypothetical construct, but it will help to frame the definition of information anchoring.

Taken together, the observed data, $Y_{obs}$, and the realisation of the event times for the censored patients, $Y_{cens,CAR}$, we obtain a *full* data set under the primary assumption. We define $\hat{\theta}_{full,primary}$ to be the corresponding estimate of $\theta$ after fitting the primary analysis model to this full data set.

For the sensitivity analysis, we make a different set of assumptions concerning the distribution of post-censoring data, that is, scenarios in which censoring is assumed to be informative (i.e. censored *not* at random).

Defined analogously to the primary analysis, for the sensitivity analysis we have $\hat{\theta}_{obs,sensitivity}$ and $\hat{\theta}_{full,sensitivity}$, whereby "full" is defined again from our hypothetical construct of $Y_{cens,sens}$, but this time under a specific set of assumptions for the sensitivity analysis.

Furthermore, we define the observed information about $\theta$ under the primary and sensitivity analyses by $I(\dots)$. Since there is less information when there is censored data, then we would expect the following (Cro *et al.*, 2018):

$$\frac{I(\hat{\theta}_{full,primary})}{I(\hat{\theta}_{obs,primary})} > 1, \tag{1.6.1}$$

and,

$$\frac{I(\hat{\theta}_{full,sensitivity})}{I(\hat{\theta}_{obs,sensitivity})} > 1. \tag{1.6.2}$$

The principle of information anchored sensitivity analyses compares these two ratios:

$$\frac{I(\hat{\theta}_{full,primary})}{I(\hat{\theta}_{obs,primary})} = \frac{I(\hat{\theta}_{full,sensitivity})}{I(\hat{\theta}_{obs,sensitivity})}, \tag{1.6.3}$$

so that the proportion of information lost due to missing data is constant across primary *and* sensitivity analyses. If equation (1.6.3) above holds then we say that the sensitivity analysis is *information anchored* with regards to the primary analysis.

| |
|---|
| This represents the third and final key facet when considering new sensitivity analysis methods — their *information anchoring* properties, so that the proportion of information lost due to missing data is held constant across primary *and* sensitivity analyses. |

If a sensitivity analysis method is information anchored, even approximately, then we can be confident that the method itself is not injecting (equation (1.6.4)) or taking away (equation (1.6.5)) information,

$$\frac{I(\hat{\theta}_{full,primary})}{I(\hat{\theta}_{obs,primary})} > \frac{I(\hat{\theta}_{full,sensitivity})}{I(\hat{\theta}_{obs,sensitivity})} \quad \text{— information negative, taking away information,} \tag{1.6.4}$$

$$\frac{I(\hat{\theta}_{full,primary})}{I(\hat{\theta}_{obs,primary})} < \frac{I(\hat{\theta}_{full,sensitivity})}{I(\hat{\theta}_{obs,sensitivity})} \quad \text{— information positive, injecting information.} \tag{1.6.5}$$

If the results from the primary and sensitivity analysis are clinically equivalent, we can conclude that the results are relatively robust to plausible departures from the assumptions regarding the censoring mechanism made for the primary analysis (e.g. CAR). If they do not, we need to reflect carefully, and may need to be much more cautious in our interpretations of the results from the trial.

In either case, if the information anchoring principle holds, we have created a level playing field for the primary and sensitivity analysis, and we can be confident that at least the comparison *itself* can be relied upon.

## 1.7 Summary of motivation for thesis

We have now established the cornerstones for evaluating new sensitivity analysis approaches. That is, in terms of their:

- **Practicality** — their ease of implementation and use.
- **Clinical plausibility** — including the ability to contextualise them to the trial team.
- **Information anchoring** properties — so that the proportion of information lost due to missing data is held constant across primary *and* sensitivity analyses.

The goal of this thesis is to extend and develop reference-based sensitivity analysis, originally proposed by Carpenter *et al.* (2013) in the longitudinal continuous data setting, to time-to-event data. In the next section we introduce the multiple imputation procedure, then in section 1.9 we set out the roadmap for achieving this goal.

## 1.8 Multiple Imputation

There is now a vast body of literature reviewing methods for handling missing data in a statistically robust manner (e.g. Little and Rubin, 2002; Allison, 2002; Molenberghs and Kenward, 2007). The relative practicality of using multiple imputation (MI), compared to the more specialised knowledge required for direct likelihood or Expectation-Maximisation (EM) based methods, makes it attractive to analysts. The book by Carpenter and Kenward presents a practical guide to MI for various applications, including methods for time-to-event data (cf. Chapters 8.1 and 8.2 of Carpenter and Kenward (2012)). The draw of multiple imputation is that it provides a computationally practical approach which utilises all the information available in the data set under both missing/censoring at random and missing/censoring not at random assumptions. An additional attraction is that the original primary analysis model, also known as the "substantive" model, is fitted to the imputed datasets.

Other missing data methods, for example, those based on inverse probability weighting, weighted generalised estimating equations and doubly robust estimation, continue to be developed (e.g.

Liang and Zeger, 1986; Robins *et al.*, 1995; Bang and Robins, 2005; Carpenter *et al.*, 2007; Tsiatis *et al.*, 2011; Daniel and Kenward, 2012), particularly in the context of causal inference techniques for modelling observational data (reviewed in more detail in Chapter 5). However, MI remains the dominant tool in practice, and therefore it is natural to seek to use it to perform sensitivity analysis in a principled and statistically rigorous way.

We now describe how MI may be used to impute events times for censored patients. Of course, this is not necessary assuming censoring is at random, since maximum likelihood methods will provide the same results (up to Monte Carlo error), assuming a closed form expression for the likelihood function. However, when censoring is not at random, MI is by far the most practical solution in terms of implementation. Furthermore, for such pattern mixture approaches we can tailor imputation in each pattern to reflect different CNAR scenarios. For example, in a typical scenario we might make the CAR assumption for those administratively censored at the end of the study, but make a CNAR assumption for patients lost to follow-up on one (or both) arms.

At this point it is perhaps important to re-iterate that we focus on multiply imputing event times for censored patients, rather than for missing covariate data. Therefore, we assume that either there is no missing covariate data, or if there is missing data, it can also be included in the imputation process in the appropriate way.

The main steps for multiply imputing new events times for censored patients are as follows (Carpenter *et al.*, 2013):

- MI1: Under CAR, a draw is taken from the parameters of the Bayesian posterior distribution of the survival function. This is done as follows:

  We fit an appropriate model for the survival time to the observed data using maximum likelihood. We draw estimates for the parameters by assuming that they asymptotically have a multivariate normal sampling distribution (cf. page 179 of Carpenter and Kenward). In this way, we attempt to approximate a full Bayes model.

- MI2: For each censored patient, the draws from the posterior distribution are used to construct the post-censoring survival function *for this patient*.

  Comment: We note here that later for this step we will manipulate the posterior distribution to provide the different scenarios for the sensitivity analyses. There are a number of

options for defining the post-censoring hazard. These are described in more detail in the next chapter.

- MI3: A "new" event time is imputed by sampling from the survival function, making sure that this new event time is greater than or equal to the time the patient was originally censored. This process is repeated for each censored patient.
- MI4: Steps MI1 to MI3 are repeated using a new draw of the parameters from the fitted imputation model, resulting in a number of imputed data sets. The analysis model for the time-to-event data is then fitted to each of the multiply imputed data sets, and the resulting point and variance estimates are combined using a set of rules originally defined by Rubin (e.g. Little and Rubin (2002)).

We now define Rubin's rules in more detail. Denote the point and variance parameter estimates from fitting the imputed data set $k$, to the analysis model as $\hat{\beta}_k$ and $\hat{\sigma}_k^2$ respectively. Rubin's rules for inference are defined for the MI estimator of $\beta$ as:

$$\hat{\beta}_{MI} \;=\; \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k, \tag{1.8.1}$$

with variance estimator

$$\hat{V}_{MI} \;=\; \hat{W} + \left(1 + \frac{1}{K}\right)\hat{B}, \tag{1.8.2}$$

where

$$\hat{W} \;=\; \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2, \tag{1.8.3}$$

and

$$\hat{B} \;=\; \frac{1}{(K-1)} \sum_{k=1}^{K} (\hat{\beta}_k - \hat{\beta}_{MI})^2. \tag{1.8.4}$$

Rubin's variance estimator, as defined in equation (1.8.2), is bounded below by the variance of the treatment estimator had the missing data actually been observed, and inflated by the between imputation variance, to capture the loss of information due to the missing data (the $\hat{B}$ component in equation (1.8.4)).

In step MI4 in the above algorithm, it is important to note that we assume the estimates resulting from fitting the analysis model to each of the multiply imputed data sets are normally distributed. As Carpenter and Kenward point out, and this is particularly important to be aware of for time-to-event data, "quantities like odds ratios and hazard ratios should be log transformed before the MI procedure [that is, Rubin's rules are]... applied" (page 48 of Carpenter and Kenward (2012)).

The final step, using "Rubin's rules", combines the estimates from each of the multiply imputed data sets to produce valid point and variance estimates, both for Bayesian and frequentist inference. In his 1994 paper, Meng comments that "multiple imputation is motivated from the Bayesian perspective, yet ... its primary application area ... [is] traditionally dominated by frequentist analyses" (Meng, 1994). This provokes a discussion of the properties of MI from a frequentist perspective, since this is crucial to the understanding and usage of multiple imputation. Formal arguments regarding the frequentist properties of MI are presented in Carpenter and Kenward Chapter 2.5 "Frequentist Inference" (Carpenter and Kenward, 2012).

Briefly summarised, Rubin provides some conditions for MI to have good frequentist properties:

> "Despite being Bayesian in nature, provided some subtle conditions hold, Rubin's combination rules also provide valid frequentist inference, in that they provide an estimator which is asymptotically unbiased and an accompanying estimate of variance which can be used to construct confidence intervals with coverage equal to that specified", (Cro, 2016).

Rubin outlines the requirements for this as follows:

1. "Draw imputations following the Bayesian paradigm as repetitions from a

Bayesian posterior distribution of the missing values under the chosen models for non-response and data, or an approximation to this posterior distribution that incorporates appropriate between imputation variability.

2. Choose models of non-response appropriate for the posited response mechanism.

3. Choose models for the data that are appropriate for the complete-data statistics likely to be used — if the model for the data is correct, then the model is appropriate for all complete-data statistics" (from page 110 of Molenberghs and Kenward (2007), quoting pages 126-127 of Rubin (1987)).

However, as Carpenter and Kenward point out: "How useful a guide these three conditions are in practice is hard to say. Apart from the simplest settings it is difficult to justify these rigorously" (page 63 of Carpenter and Kenward (2012)). Meng extended them with a more mathematical formulation, including the introduction of a new definition for "congeniality". *Congeniality* means that the procedure for analyzing multiply imputed data sets can be derived from (is "congenial" to) the model adopted for multiple imputation (Meng, 1994). He goes on to explain: "when an analysis procedure is congenial to the imputation model, the inference from the repeated-imputation combining rules with infinitely many imputations agrees . . . with the (desired) incomplete data analysis under the analysts procedure."

With reference to this definition, in her PhD thesis S. Cro expands on this statement "we interpret this as the imputation and analysis model must have the same content and structure and so be formed around the *same assumptions* to be congenial." (own italics). We will revisit this fundamental point later in Chapter 3 when we discuss Rubin's variance estimator in relation to our proposed methods for sensitivity analysis.

Conversely of course, when the analysis procedure does not correspond to the imputation model, it is *uncongenial*. Meng provides concrete examples of this: "uncongeniality occurs at least in the following three cases: First, the imputation model is largely unknown to the analyst, [*no longer usually the case, especially in medical applications*] who also has limited or no access to the imputer's extra resources. Second, different purposes of imputing missing observations and of substantive analyses suggest that different models can better accommodate their different needs. Third, *several models are considered for imputation or for analysis, such as when conducting a sensitivity study of underlying model assumptions*." (italics added).

A common example of uncongenial MI is when the imputation model contains more covariates than the analysis model. Interestingly, this is common practice and is often recommended in texts (for example, White *et al.* (2011), and the imputation guidelines section 2.10 of Carpenter and Kenward (2012)). We encountered an example of this earlier in section 1.5.3 in which Brinkhof *et al.* fitted an adjusted Weibull proportional hazards model as imputation model, but the Kaplan-Meier product limit estimate was used as analysis model.

While the rigour in Meng's 1994 paper frames this discussion perfectly in a mathematical sense, his formal definitions are not intuitive, at least at first glance. Luckily, he goes on to provide a common sense footing for discussions surrounding MI, explaining that an *uncongenial* setting "essentially mean[s] that the analysis procedure does not correspond to the imputation model ... [it] can lead to bias and discrepancies between the long run sampling variance and that obtained by applying Rubin's rules". He goes on to explain:

> "in cases where the imputer does have such extra information [that is, a *richer* imputation model in terms of covariates, compared to the substantive model], the decomposition [Rubin's rules] provides a conservative estimate of the sampling variance of the repeated-imputation estimator" (Meng, 1994),

which bring us to the main dilemma associated with MI. A key attraction of the method is the relative simplicity of Rubin's general variance formula, and this is what marks MI out from other methods, but at the same time this has been the target of criticism. Concretely, S. Cro points out that,

> "when the substantive model and imputation model do not satisfy this condition [congeniality], they are described as uncongenial. The validity of Rubin's variance estimator is *not guaranteed* when this is the case", (italics added), (Cro, 2016).

This final point was the crux of much of the methodological controversy as MI began to be increasingly used in practice (see, for example, Nielson (2003) regarding the efficiency of Rubin's variance estimator, along with the rebuttal of numerous arguments against the use of MI in Rubin (1996)). This criticism of the overestimation of the variance using Rubin's estimator is apportioned to the "existence of an extra cross term in the decomposition" (Meng, 1994), referring to the between imputation term $\hat{B}$ in equation (1.8.4).

The discussion has since coalesced around two issues, i.) the relative *inefficiency* of Rubin's estimator in uncongenial settings, and ii.) whether there are viable alternatives to Rubin's rules for estimation.

For the first point, it is irrefutable that the MI variance estimator is conservative in some uncongenial settings (page 66 of Carpenter and Kenward (2012)). Robins and Wang confirm this: "in certain settings the variance estimator ... proposed by Rubin will be inconsistent with upward bias, resulting in conservative confidence intervals whose expected length is longer ... than necessary" (Robins and Wang, 2000). However, J. K. Kim estimated the exact bias of the multiple imputation variance estimator concluding that "the bias of Rubin's variance estimator is negligible for large sample sizes, but ... may be sizable for small sample sizes" (Kim, 2004).

Robins and Wang proposed an alternative variance estimator to Rubin's which "in contrast to the estimator proposed by Rubin, is consistent even when the imputation and analysis model are misspecified and incompatible with one another" (Robins and Wang, 2000). This would appear to be a potential solution — however, it turns out that despite having better variance properties than Rubin's variance estimator, their estimator falls short in terms of one of our key facets, namely *practicality*. The results from simulation studies performed by Hughes *et al.* confirm that "overall Rubin's multiple imputation variance estimator can fail in the presence of incompatibility and/or misspecification ... Robins and Wang's multiple imputation could provide more robust inferences" (Hughes *et al.*, 2014). However, they go on to note that:

> "A major disadvantage of Robins and Wang's method is that calculation of the imputation variance estimator is considerably more complicated than for Rubin's MI ... with a greater burden on both the imputer and the analyst .... To our knowledge, there is no generally available software implementing the Robins and Wang method. The analyst must make available *derivatives of the estimating equations* for use in calculation of variance estimates, and these become harder to calculate as the complexity of the analysis procedure increases. Also, the complexity of the calculations conducted by the imputer increases when there are multiple incomplete variables", (italics added), (Hughes *et al.*, 2014).

As Molenberghs and Kenward state in the preface of their book "a key prerequisite for a method to be embraced, no matter how important, is the availability of trustworthy and easy-to-use software" (Molenberghs and Kenward, 2007).

Liu and Peng also confirm the shortcoming of Rubin's variance estimate — "[the] conventional MI approach ... inflates the variance estimates, which results in an overly conservative test for the treatment effect" (Liu and Peng, 2016). They considered a full Bayesian approach for their sensitivity analysis, and found "more appropriate variance estimates from Bayesian MCMC". The authors go on to note that their model was easily implemented with SAS. As with Robins and Wang's estimator, it is perhaps important to question the practicality of such an approach for non-technical experts, both in terms of complexity and implementation time required.

Despite this "relatively warm" debate concerning the properties of Rubin's estimator, most authors stress a more pragmatic outlook. We return to Rubin, who makes the salient point that if the imputer's model is far from reality, then "all methods handling non-response are in trouble" (Rubin, 1996). Furthermore, Meng relativises the issues associated with uncongeniality — "it is vital to recognise that disagreements between the repeated-imputation analysis and the (best possible) incomplete-data analysis does not automatically invalidate the repeated imputation inference", going on to make the point that "in short, with sensible imputations and complete data procedures, it is generally wise for the analyst to use the standard combining rules [Rubin's rules], despite the presence of uncongeniality" (Meng, 1994).

We leave the final word on this topic to Carpenter and Kenward, who echo this sentiment — the "mildly conservative behaviour [of Rubin's variance] is an acceptable price to pay for the exceptional simplicity, flexibility and generality of the MI procedure" (Carpenter and Kenward, 2012).

Nonetheless, we need to be aware of the potential behaviour of Rubin's rules in uncongenial settings in the context of sensitivity analyses. We will revisit, and explore this issue in more depth, when we come to discuss the properties of the MI variance estimator for our sensitivity analysis method in Chapters 3, 4 and 5.

One of the advantages of multiple imputation is that we can readily modify the existing imputation model to explore the sensitivity of inferences to departures from CAR. This has the potential to provide a flexible approach for targeting clinically relevant estimands, such as those discussed by Mallinckrodt *et al.* (2017).

The next section presents a brief review of the sensitivity analysis literature focussing on the methods for time-to-event data.

## 1.9 Reference-based sensitivity analysis methods

We previously introduced the model based classification of sensitivity analysis with selection, pattern mixture, and shared parameter based approaches. Cro *et al.* (2018) recently proposed one which places more focus on what we have defined as a method's inherent *practicality*. They differentiate between two broadly defined types of sensitivity analysis. We have adopted these definitions since they help to clarify why our sensitivity analysis approach is novel. This also provides the rationale for considering the properties of the estimates from our approach in more detail.

In the first class of sensitivity analysis methods, referred to as "Class-1", for each sensitivity analysis scenario there are a set of assumptions, and an appropriate analysis is identified and performed consistent with these assumptions. Most of these methods require the analyst to make distributional assumptions regarding the missing data. The texts by Molenberghs and Kenward focussing on clinical studies, and Daniels and Hogan on longitudinal analysis in a Bayesian setting, review and propose such methods for sensitivity analysis, that is, those in which a parametric form for the post-deviation distribution is explicitly defined (Molenberghs and Kenward, 2007; Daniels and Hogan, 2008).

In contrast, in the second class of sensitivity analysis the primary analysis is retained in the sensitivity analysis ("Class-2"), but the statistical behaviour of the missing data are assumed to diverge from that assumed under the primary analysis model (Carpenter and Kenward, 2012; O'Kelly and Ratitch, 2014). These approaches combine a pattern mixture modelling paradigm with MI, imputing missing data by reference to an appropriately chosen group, or groups of patients from the observed data. Some of these approaches, which were pioneered by Little and Yau, are often referred to as "controlled" or "reference-based" methods (Little and Yau, 1996):

- "Controlled" refers to the fact that for these techniques the form of the imputation of the censored data involves specification of parameters *controlled* by the analyst — not estimated by the data.
- "Reference" thus called since it avoids specifying potentially lots of patterns for the missing data by instead making *reference* to other groups of patients.

For example, consider a two arm clinical trial with patients randomised to either a control treatment or an experimental treatment. The primary analysis might estimate the hazard ratio,

making an appropriate *de-jure* assumption, such as patients being censored at random when they deviate from the study protocol.

An associated sensitivity analysis scenario might make the assumption that those deviating on the treatment arm revert to the control treatment, and implement this by multiply imputing missing values on the treatment arm "by reference" to *observed data* on the control arm.

Besides being relatively accessible in terms of the assumptions made, reference based methods provide a concrete clinical context for the sensitivity analysis. These approaches are also comparatively straightforward to implement, requiring only a slight modification of standard multiple imputation techniques, as discussed later in the examples in Chapters 3, 4 and 5. Such approaches also avoid fully modelling the missing data process, which is often a rather complex, time consuming process requiring specialised statistical knowledge.

Notwithstanding the advantages associated with Class-2 sensitivity analysis methods, the assumptions for the primary analysis will not typically be consistent with the data generating mechanism assumed by the sensitivity analysis. If the sensitivity analysis assumes deviating patients take the control treatment as rescue medication, this would be contrary to, for example, a missing at random assumption for the primary analysis. This *inconsistency* (or uncongeniality in the vocabulary of Meng discussed in section 1.8) means that we are no longer able to automatically rely on using Rubin's MI rules. The behaviour of these rules in the presence of this type of inconsistency needs to be re-evaluated.

However, the argument we make takes a slightly different tack to that usually encountered in the literature. Rather than stating that due to uncongeniality, Rubin's rules no longer apply, we flip the argument around, establishing a set of properties (called here *facets*) which the variance estimator following multiple imputation should have, key amongst them being the principle of information anchoring, and determining if Rubin's variance estimator satisfies these properties.

Indeed, when using Class-2 sensitivity analysis methods, the *properties* of the point and variance estimators chosen for the primary analysis may change as we move to the sensitivity analysis. This may sound rather implausible, but is readily shown by considering a simple example.

Example 2

The example is based on that in Cro *et al.* (2018), re-worked for the survival setting.

Consider a study with $n = 100$ fully observed log normally distributed event times $T_i$, such that post log transformation, $\ln T_i = Y \sim \mathcal{N}(\mu, \sigma^2)$, with known variance $\sigma^2$. Furthermore, let us assume that we are interested in estimating the mean of the population, $\hat{\mu}$ from this sample by the average of the log event times. With fully observed data, that is, when there is no censoring then the *information* we know about $\mu$ is $\frac{n}{\sigma^2} = \frac{100}{\sigma^2}$.

Suppose now that, $n_d$ of the patient times are censored. We would like to perform a class-2 sensitivity analysis, so that our estimator (the sample average of the log event times) remains the same in the primary and sensitivity analysis. Now, let us assume that the primary analysis assumes the data are censored completely at random (CCAR). Our sensitivity analysis will assume that the censored values are from patients with the same mean, $\mu$, but a different variance, $\sigma^2_{censored}$.

For the *primary analysis*, since we make the CCAR assumption we may obtain valid inference by *either* calculating the mean of the $(100 - n_d)$ value, *or* multiply imputing new event times (e.g. using a Tobit model as imputation model). Whichever method we choose, we will end up with the same information about the mean, namely now, $\frac{(100-n_d)}{\sigma^2}$.

For the *sensitivity analysis* we multiply impute making an appropriate assumption regarding the post-censoring behaviour. Of course, our treatment estimate remains the mean, $\hat{\mu}$, but the statistical information about the mean is now a weighted average of the information from the observed data, $\frac{(100-n_d)}{\sigma^2}$, and the information from the assumed event time distribution for the censored patients $\frac{n_d}{\sigma^2_{censored}}$, that is $\frac{100^2}{\{(100-n_d)\sigma^2+n_d\sigma^2_{censored}\}}$.

We can see from this expression that the information about the mean depends on $\sigma^2_{censored}$, and therefore for the sensitivity analysis the analyst *controls* the information.

This phenomenon is illustrated in Figure 1.9.1, with the information we (as the analyst) define on the x-axis and the resulting information about the mean on the y-axis. Letting $n_d = 25$ and assuming $\sigma^2 = 1$, we can see that when $\sigma^2_{censored} < \sigma^2 = 1$, the information about the mean in the sensitivity analysis is greater than from the 100 observations (i.e. the sensitivity analysis is information *positive*), and when $1 \leq \sigma^2_{censored} \leq 2.3$ then the information is greater than in the $(n - n_d) = 75$ observations, whereas when $\sigma^2_{censored} > 2.3$ the information is less than in the observed $(n-n_d) = 75$ observations (i.e. the sensitivity analysis is information *negative*). ■

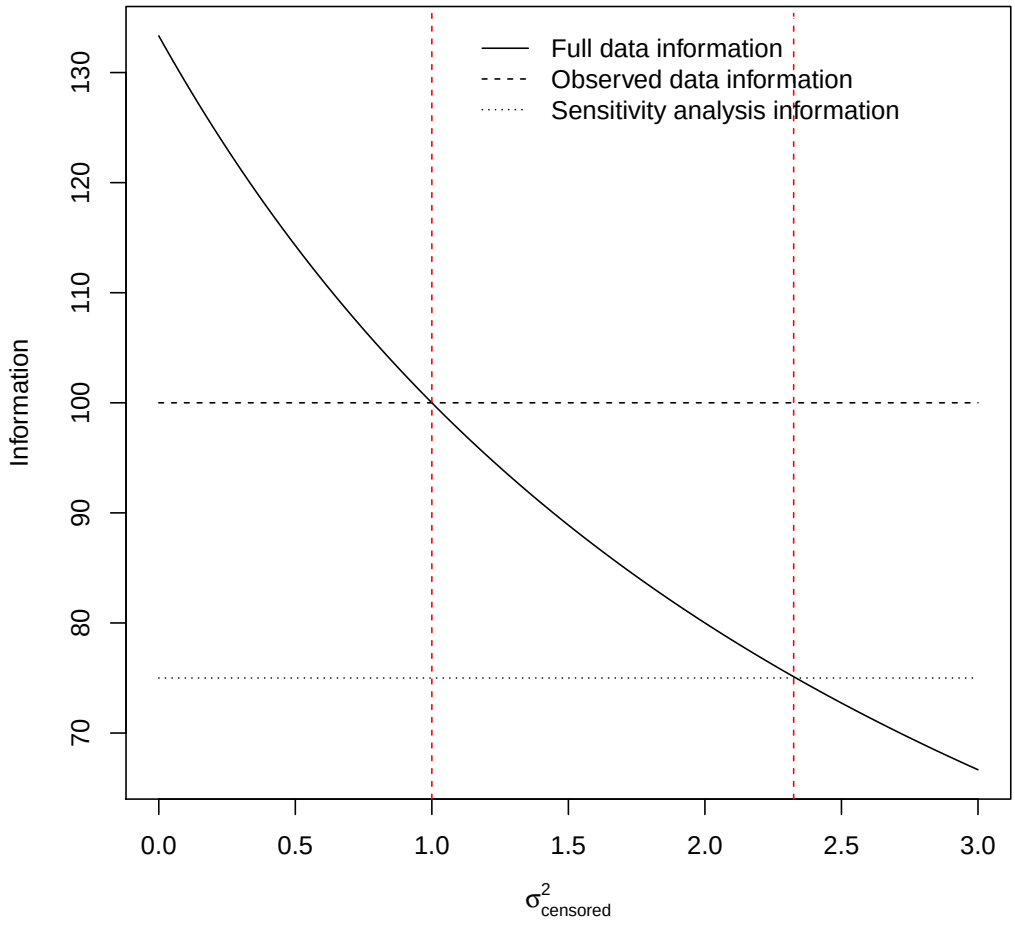


Figure 1.9.1: Information about the sample mean varies with $\sigma^2_{censoring}$ (derived from Cro *et al.* (2018)))

.

As this simple example shows, a logical choice for the variance estimator for the primary analysis might behave in an unexpected way under a specific sensitivity analysis scenario, if the choice of variance estimate for those censored is not made carefully.

In addition, with reference based methods we are essentially using some of the data twice; once when we utilise data from, for example, the reference arm to impute data in the active arm; and secondly, to estimate the treatment effect in the reference arm. Consequently, this naturally reduces the variability of the data in both arms, reducing the overall variability in the data in the sensitivity analysis relative to the primary analysis.

In summary, common to all sensitivity analysis methods, the aim is to assess the behaviour of the treatment estimate under alternative, clinically plausible, scenarios. To justify the use of the reference based approach, we need to explore the properties of the primary analysis estimators under the scenarios; their properties may well change as we move from the primary to the sensitivity analysis. Thus, a sensible variance estimator for the primary analysis may behave in an unexpected manner under certain sensitivity analysis scenarios. Indeed, there are examples in which the variance estimator with a reference based method *decreases* as the proportion of missing values increases (Cro *et al.*, 2018). Such counter-intuitive properties would undermine our confidence in the approach and of course would reward trialists for losing data! It is therefore important to quantify the amount of statistical information available in the sensitivity analysis *relative* to the primary analysis, to determine if the sensitivity analysis is injecting new information, or taking away information, relative to the primary analysis. Due to the potential inconsistency between analysis and postulated data generating mechanisms with class-2 sensitivity analysis approaches, we cannot rely on properties derived under the primary analysis assumptions being consistent under the sensitivity analysis assumptions.

We therefore need alternative criteria for the assessment of potential sensitivity analysis methods and, in particular, for the variance of the treatment estimate. This leads us naturally to the principle of *information anchoring* introduced previously in section 1.6. If it holds, it ensures that information is neither created or destroyed as we move from the primary to the sensitivity analysis, establishing a so-called "level playing field". This is important for regulators and industry, since it provides confidence in the results of sensitivity analyses conducted in this way.

In terms of the sensitivity analysis methods we propose here, it will be used as the key metric to determine if they provide trustworthy results which can be used with confidence. The next section sets out a roadmap for the remainder of the thesis which describe and demonstrate the

new methods.

## 1.10 Clinically relevant and accessible sensitivity analysis for time-to-event outcomes

In the continuous data with longitudinal follow-up setting, Carpenter *et al.* proposed that, once the estimand is defined, patients should be followed up until they deviate from the protocol in a way that is relevant to the estimand. This thesis builds on recent work in the survival context, proposing an analogous class of reference based assumptions appropriate for time-to-event data.

In Chapter 2, we show how each of the proposals in Carpenter *et al.* (2013) may be applied in the context of time-to-event data. This includes the proposals of Lu *et al.* (2015) and Lipkovich *et al.* (2016). We illustrate their *practicality* and clinical *plausibility* using both simulated data and a real data set, the German Breast Cancer (GBC) data (introduced in the next section).

With Class-2 reference based sensitivity analysis using MI, we need to better understand the behaviour of the estimates, since in this setting the imputation and analysis models are based on different sets of assumptions.

In Chapter 3, we therefore investigate the properties of Rubin's variance estimator in more detail. We use our proposals for reference based imputation for time-to-event data, and show how imputation and inference can be performed using Rubin's rules, demonstrating by simulation that these rules provide both unbiased estimates, and give inferences that are approximately information anchored relative to the primary analysis. For illustration, we consider a clinical trial in cardiovascular disease, the RITA-2 data (also introduced in the next section).

Chapter 4 builds on this empirical approach, by showing how, in certain circumstances, the principle of information anchoring can be shown to hold *generally* for reference based sensitivity analysis in a time-to-event setting. These theoretical results are then put to the test using a simulation study, with their application again illustrated using the RITA-2 data.

Chapters 3 and 4 together highlight that our proposal provides a solid foundation in terms of their statistical validity, establishing confidence in their use in the time-to-event setting. The groundwork for the new methods now confirmed, we then change tack slightly to consider the

challenges of using our methods for observational "big" data. Whilst recognising that randomised controlled trials are the gold standard for determining the effects of new treatment strategies, it is not always possible to implement such trials due to cost, timelines, such as when longer term effects are not yet available (for example, Garcia-Albeniz *et al.* (2017)), or ethical reasons. In such cases, observational cohort data has often provided opportunities to estimate possible effects, frequently with a view to providing focus for subsequent confirmatory clinical trials.

There is a wealth of literature concerning causal methods developed to overcome some of the confounding issues associated with using observational data in this way (for example, Hernan and Robins, 2018; Newsome *et al.*, 2017, refer also to Chapter 5.3 of this document).

Recently developed trial "emulation" approaches mimic the randomisation process of an RCT by adjusting fitted models to overcome potential bias when using observational data for estimation of treatment effects in a robust manner. If a trial is being "emulated" in this way then it seems natural to attempt to apply sensitivity analysis methods developed for an RCT setting to this observational data. We propose and illustrate a concrete example of this in an analysis of patients with pneumocystis pneumonia (PCP), an opportunistic disease (OD) contracted by individuals having a weakened immune system, and one of the most frequent AIDS defining diagnoses in resource rich countries. In Chapter 5 we show how, based on the concepts of causal inference methods, our approach may also be applied to an emulated trial, presenting an example using data from COHERE (again, introduced in the next section).

Finally, Chapter 6 discusses the relevance of the work in context with recent publications, and propose areas for further research.

Prior to presenting the work in detail, the next section introduces the data sets used to illustrate the approaches.

## 1.11 Motivating data sets

### 1.11.1 German breast cancer data

This data come from a comprehensive cohort study by the German Breast Cancer Study Group, consisting of 720 patients, recruited from July 1984 to December 1999, all of whom had primary node positive breast cancer (Schmoor *et al.*, 1996). The study compared the response to treatment of 448 women, 223 of whom received a lower dose of chemotherapy, and 225 received a higher one (Schumacher *et al.*, 1994), implemented using a factorial design. The primary analysis showed no benefit from the higher chemotherapy dose, but there was a significant effect for those patients taking additional hormonal treatment.

We focus on a subgroup of 448 patients, in which the effectiveness of three versus six cycles of chemotherapy, with and without additional hormonal treatment, were compared. Further details of the study are found in Appendix A, along with the papers of, for example, Sauerbrei and Royston (1999) and Sauerbrei *et al.* (1999).

### 1.11.2 The RITA-2 Study

The Second Randomized Intervention Treatment of Angina (RITA-2) (Henderson *et al.*, 1997, 2003) randomized 1018 eligible coronary artery disease patients from the UK and Ireland to receive either Percutaneous Transluminal Coronary Angioplasty (PTCA, n=504) or continued medical treatment (n=514). Those patients randomised to angioplasty received the intervention in the first three months. The primary endpoint of the study was a composite of all cause mortality and definite non-fatal myocardial infarction. After 7 years, there were 73 deaths (14.5%) on the PTCA arm and 63 (12%) on the medical arm (difference in proportions +2.2% [-2%, 6.4%], p = 0.21).

The study concluded that an initial policy of PTCA was associated with greater improvement in angina symptoms, with this effect being particularly present in patients with more severe angina, and that the increased risk of performing PTCA should be offset against these benefits.

### 1.11.3 Observational data from COHERE

The Collaboration of Observational HIV Epidemiological Research Europe (COHERE) is a collaborative group of 33 adult, paediatric, and mother/child HIV cohorts across Europe. The collaboration allows comparisons across age categories and provides a mechanism to rapidly compile datasets to address novel research questions that cannot be studied adequately in individual cohorts (http://www.cohere.org, http://www.eurocoord.net).

| Guideline | Conditions | Primary prohylaxis | Stopping rule |
|---|---|---|---|
| NIH | 1 | CD4 $<$ 200 cells/$\mu$L | $\geq$ 200 cells $\mu$L |
| EACS | 1 | CD4 $\leq$ 200 cells/$\mu$L | CD4 $>$ 200 cells/$\mu$L **OR** |
| | 2 | | CD4 100-200 cells/$\mu$L **AND** |
| | | | HIV-VL undetectable for 3 months |

Table 1.11.1: NIH and EACS guidelines for PCP prophylaxis (NIH, 2018; EACS, 2018)

Previous analyses of the COHERE cohort data suggested that primary PCP prophylaxis can be safely withdrawn in patients with CD4 counts of 100-200 cells/$\mu$L if HIV-RNA is suppressed (Mussini *et al.*, 2000; Qiros *et al.*, 2001; Mocroft *et al.*, 2010). Table 1.11.1 summarises the current guidelines which are, at least partially, based on the results from these studies. A more recent study added to these findings, indicating that PCP incidence off prophylaxis was below 1/100 person years for virologically suppressed individuals with a CD4 count above 100 cells per $\mu$L, and thus primary (and secondary) prophylaxis might not be needed in such cases (Furrer *et al.*, 2015). However, it remains to be determined if PCP prophylaxis might be fully withdrawn for patients with consistently suppressed HIV viral load (VL), irrespective of CD4 count.

## 1.12 Focus of the thesis

As highlighted in section 1.3, the NRC recognised the need for further research into sensitivity analysis methods for time-to-event data in 2010. Reference based methods, which to date have been well received in continuous data settings, are an obvious candidate for extension to time-to-event data.

We begin by defining the new reference based sensitivity analysis methods and demonstrating their practicality in terms of their ease of implementation and use. The aim is to help technical

and non-technical experts alike to gain an impression of their simplicity and accessibility. We go on to provide a sound theoretical footing for our methods so that both industry and regulators will have confidence in using the methods.

The motivation for the final application of the methods to observational data was born out of necessity. Working with such data, and being regularly confronted with both missing baseline and time varying covariate data, along with missing outcome information, presents a challenge for the analyst. Using multiple imputation to fill in data is commonplace, but using MI for longitudinal, or for time-to-event outcomes, or the combination of the two still presents a significant hurdle. The final chapter seeks to address this issue.

At this point, we take note of the current status of publications resulting from this work. Chapters 2 and 3 were submitted to Pharmaceutical Statistics in April 2018, and the manuscript received generally positive reviews (Atkinson *et al.*, 2018). We intend to submit the work in chapter 4 to a methodological journal early in 2019. The clinical part of the analysis in Chapter 5 was presented as a poster at the 22nd International Workshop on HIV and Hepatitis Observational Databases (IWHOD) in March 2018, along with a separate poster describing the methodological approaches taken. An abstract summarising these results was presented at the Conference on Retroviruses and Opportunistic Infections (CROI) in March 2019.

We begin in Chapter 2 by reviewing the methods defined by Carpenter *et al.*, extending them for use with time-to-event data.

# Chapter 2

# Reference based methods for time-to-event data

## 2.1 Introduction

We begin this chapter by describing how each of the proposals in Carpenter *et al.* (2013), which were developed for longitudinal data with a continuous outcome, can be mapped to the time-to-event setting. These methods piece together pre-deviation, or in this case pre-*censoring* data, with post-deviation/post-*censoring* distributions from other trial arms. Of course, in a survival analysis context the distributions often used are the survival or hazard function. Multiple imputation (MI) is then used to calculate appropriate estimates of the treatment effect and associated standard errors, these being derived in the normal way using Rubin's rules.

For each of the methods we present a schematic illustration with two panels. On the top panel we describe the possible effect of the method in the longitudinal data setting. On the bottom panel, we show what we might expect to see in the time-to-event setting. We then define a number of new proposals for methods which may be specifically appropriate for censored data.

The practical performance is then explored through application to a simulated data set. This includes the effect of using different post-censoring behaviours on the proportional hazards assumption, since these are the types of models we use throughout — they are the most frequently

used models for a survival analysis. To end the chapter, the methods are applied to the German Breast Cancer data set to demonstrate their usefulness with real data.

We seek to address a number of key questions in this chapter:

1. Which of the reference based sensitivity analysis methods developed for longitudinal data are suitable for use with time-to-event data?
2. Can the methods be applied in concrete clinical settings?
3. Are the methods *practical* and *clinically plausible* as defined in Chapter 1, that is, are they easy to implement, use and explain?

We begin by reviewing the methods of Carpenter *et al.* (2013) and define the analogue approaches for time-to-event endpoints.

## 2.2 Defining the post-deviation distribution in terms of other treatment arms

Consider a two arm trial, with patients randomly assigned to either an active treatment, or a reference treatment (e.g. placebo, or standard of care). Consider a time-to-event outcome, and suppose a number of patients in the active arm are censored. To keep the presentation simple we assume that no other censoring occurs. Following Carpenter *et al.* (2013), we describe a number of options for imputing the missing event times.

Let $i = 1, \ldots, n$ index patients and $t_i$ the event time. $t_i$ is only observed if $t_i < c_i$, where $c_i$ is the censoring time. Define

$$\begin{cases} x_i = 1 & \text{if patient } i \text{ is in the active group, and} \\ x_i = 0 & \text{if patient } i \text{ is in the reference group} \end{cases},$$

and, for times $t < c_i$ let the hazard at time $t$ for patient $i$ be $h(t; x_i, \beta) = h_0(t) \exp(\beta x_i)$, where $h_0(t)$ is the hazard in the reference group. We assume proportional hazards so that $\beta$ is the log hazard ratio of treatment.

For patient $i$, censored at $c_i$, we now define their hazard as follows:

$$h_i(t) = \begin{cases} h_0(t)\exp(\beta x_i) & t \leq c_i \\ h_{post,i}(t) & t > c_i \end{cases}, \tag{2.2.1}$$

where the index *post* denotes the post-censorship hazard.

Once we specify a form for $h_{post,i}$ we can apply multiple imputation to event times for all censored patients, then fit our substantive model to each imputed data set before combining the results for final inference using Rubin's rules.

In the next section, we describe how to impute the missing event times under censoring at random, that is, when we assume $h_{post,i}(t) = h_0(t)\exp(\beta x_i)$. In this case, our inferences should be equivalent (up to Monte-Carlo error) to those from maximum (partial) likelihood. We then go on to consider alternative reference based specifications for the post-censoring hazard, appropriate for investigating sensitivity analysis scenarios.

## 2.3 Imputation under CAR

Our multiple imputation approach follows that described in Chapter 8.1.3 of Carpenter and Kenward (2012). First, we need to choose our substantive model. Throughout this chapter, we develop the concepts using the Cox proportional hazards model. Imputing the missing event times under this model involves drawing proper imputations from the baseline hazard, $h_0(t)$. We do this by estimating the baseline cumulative hazard function using the Nelson-Aalen estimator (Nelson, 1972). We then utilise the resulting discrete step function to calculate the hazard at a specific point in time. This is a similar method to the one used, for example, by Jackson *et al.* (2014), although they use the Breslow estimate instead.

Imputation proceeds as follows:

1. Under censoring at random, fit the Cox Proportional Hazards (CPH) model to the observed data, obtaining the maximum likelihood estimates of the parameters $\hat{\beta}$ and associated covariance matrix, $\hat{\Sigma}$.

   For $k = 1, \ldots, K$ imputations

(a) Draw $\tilde{\beta} \sim N(\hat{\beta}, \hat{\Sigma})$, which are vectors of coefficients for the treatment and covariates for the imputation model.

(b) For each patient with censored data, draw the event time from $h_i(t; \tilde{\beta})$, by equating the conditional survivor function, $S(t_i|t_i > c_i, x_i, \tilde{\beta})$ to a uniform distribution and solving for $t_i$.

Under our CPH model, we draw $u_i \sim U[0, 1]$, and then estimate the baseline hazard using the Nelson-Aalen estimator of the cumulative hazard, or equivalently, use the Kaplan-Meier product limit estimator of the baseline survival function. The resulting step function can be used to estimate a new event time by using a reverse look up to find $u_i$, noting of course that the new time must be greater than or equal to the existing censoring time for the patient.

We note at this point that, in line with other authors (e.g. White and Royston, 2009), the above method does not take into account uncertainty in the Kaplan-Meier estimate. We avoid the additional implementation complexity this would entail here, and circumvent this issue completely in later chapters by using a parametric survival model as imputation model.

2. Fit the substantive model to each imputed dataset resulting in $K$ estimates of the log hazard ratio, and combine the results using Rubin's rules.

As usual with MI, there are two key steps for introducing variability into the imputed data: Firstly, the draws of parameter estimates from their asymptotic multivariate normal sampling distribution $N(\hat{\beta}, \hat{\Sigma})$, and secondly, the generation of a new survival time ($u_i$).

Of course, we do not, and usually would not, impute missing survival times under CAR, since we can write down the likelihood directly in this case, and then calculate the maximum likelihood estimators. Multiple imputation under CAR should give us the same results (up to Monte Carlo error), provided the imputation model is a good fit for the data. We use MI assuming CAR as a simple cross-check to validate the method and code, before embarking on the sensitivity analysis.

For the sensitivity analysis with continuous data, Carpenter *et al.* (2013) provide a number of suggestions for constructing the joint distribution of pre- and post-censoring data. Each technique describes a difference between the *de-jure* and *de-facto* behaviour post-censoring. The next section goes through each of these, in turn proposing an analogous approach for time-

to-event data.

## 2.4 Proposals for reference based imputation under Censored not at Random (CNAR)

### 2.4.1 Introduction

We now provide proposals for reference based imputation under CNAR. To keep the presentation simple, we focus on imputing censored outcomes in the intervention group ($x_i = 1$); although the approach is quite general. Without loss of generality we assume those censored on the reference arm are censored at random. For each method we define a different reference group for the post-censorship hazard, and briefly discuss its plausibility in applications.

The following sections describe each of the studied sensitivity analysis methods in detail. Each time, we start with the original definition presented by Carpenter *et al.* (2013) for the longitudinal data setting, and then extend it for the time-to-event domain.

### 2.4.2 Jump to Reference (J2R)

For the longitudinal data shown in the top panel of Figure 2.4.1, the last observation prior to deviating is at time $t$ = 3. Under Jump to Reference (J2R) , the joint distribution is constructed from the pre-deviation means from Treatment B, and the post-deviation means from *Treatment A* (the reference arm), both estimated from observed data in the respective groups assuming MAR. This results in imputed outcomes at times $t$ = 4, 5 and 6, denoted by the squares in the figure.

For time-to-event data, this is schematically illustrated in the bottom panel of Figure 2.4.1. The patient is censored at time $c = \log(t) = 7$, and then the J2R method imputes a new event time at time $T^*$, using the reference arm hazard for $t > c$. Note that the reference hazard is estimated from the reference arm assuming *censoring at random*.

When the active treatment has a lower hazard, jump-to-reference models a scenario in which a

censored patient from the active treatment experiences no further benefit, but instead reverts to the hazard in the control (reference) group. For example, this might occur when Treatment B is a higher dose of Treatment A — if a patient randomised to Treatment B has to discontinue the treatment due to increased toxicity, their dose (and hazard) then drops to that of the reference Treatment A.

As usual, once a patient's post-censoring hazard is specified, the event time is imputed by generating a new time $T^*$. Since we require the event time to be after the censoring time, the hazard under Jump to Reference is defined by:

$$h_{post,i}(t|t > c, x = 1) \;=\; h(t|t > c, x = 0) \;=\; h_0(t)\exp(\beta^T x) \;=\; h_0(t|t > c),$$

where $x = 1$ is the indicator variable for Treatment B, and again, we assume a proportional hazards model, so that the multiply imputed event times are generated from the baseline hazard.

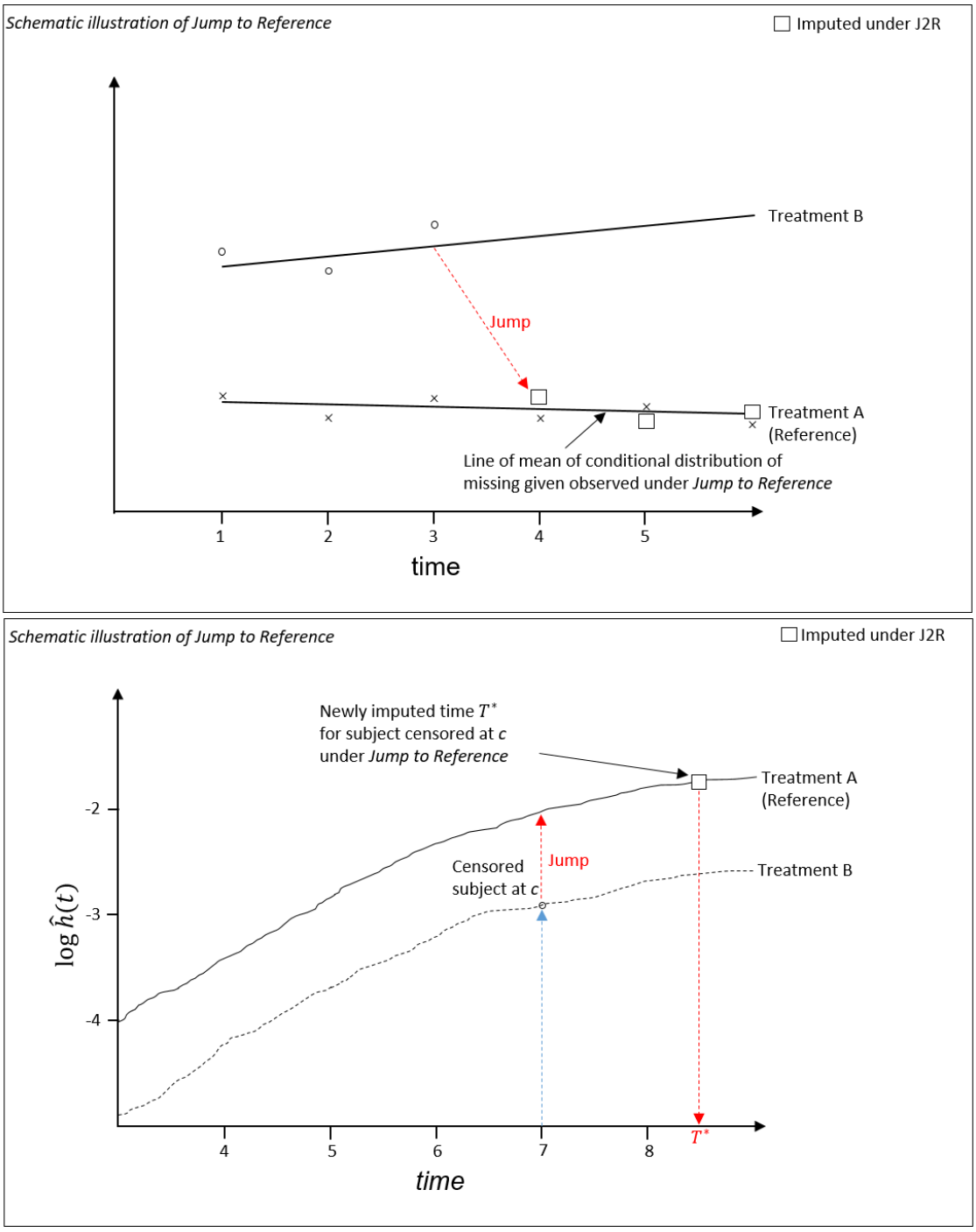


Figure 2.4.1: Top panel: longitudinal data Jump to Reference; bottom panel: time-to-event data Jump to Reference.

### 2.4.3 Last Mean Carried Forward / Hazard Carried Forward

For the longitudinal data shown in the top panel of Figure 2.4.2, with last mean carried forward (LMCF) , the patient is expected, on average, to get neither worse or better following deviation from protocol. The mean of the distribution remains constant at the value of the mean for the respective randomised treatment arm at the time of the last pre-deviation measurement i.e. at $t$ = 3 in the figure (the red dotted line). This results in imputed outcomes at $t$ = 4, 5 and 6, represented by the unfilled diamonds in the figure. It is important to note that this does not result in the post-deviation predicted means themselves being imputed, rather the last mean is used as the basis for imputing the post-deviation times.

For the time-to-event data in the bottom panel of Figure 2.4.2, an analogous concept called "Hazard Carried Forward" (HCF) has been defined. For HCF, we project the hazard forwards by first fitting an appropriate parametric model. This model is used to summarise the average hazard for all patients on the chosen trial arm with events prior to the censoring time $c = \log(t) = 7$. In the figure, this is represented by the red line up to the censoring time. A new event time $T^*$ is imputed based on extrapolating, or carrying forward, this "average" hazard, illustrated by the dotted red line on the figure. From this, we can impute the missing event time, represented by the diamond.

This scenario might be applicable when a patient's hazard remains constant post-censoring, analogous to LMCF for longitudinal data. Thus, for example, under HCF the patient's accumulated time on Treatment B might have a continued positive effect, even though the patient has discontinued treatment — Treatment B might have reached a certain critical level within the patient's body, and continues to have a prolonged positive effect.

Therefore, under this assumption for time-to-event data, when a patient in the active arm is censored at $c$, their post censorship hazard remains what it was at that $c$, i.e. $h_{post,i}(t) = h_i(c_i)$.

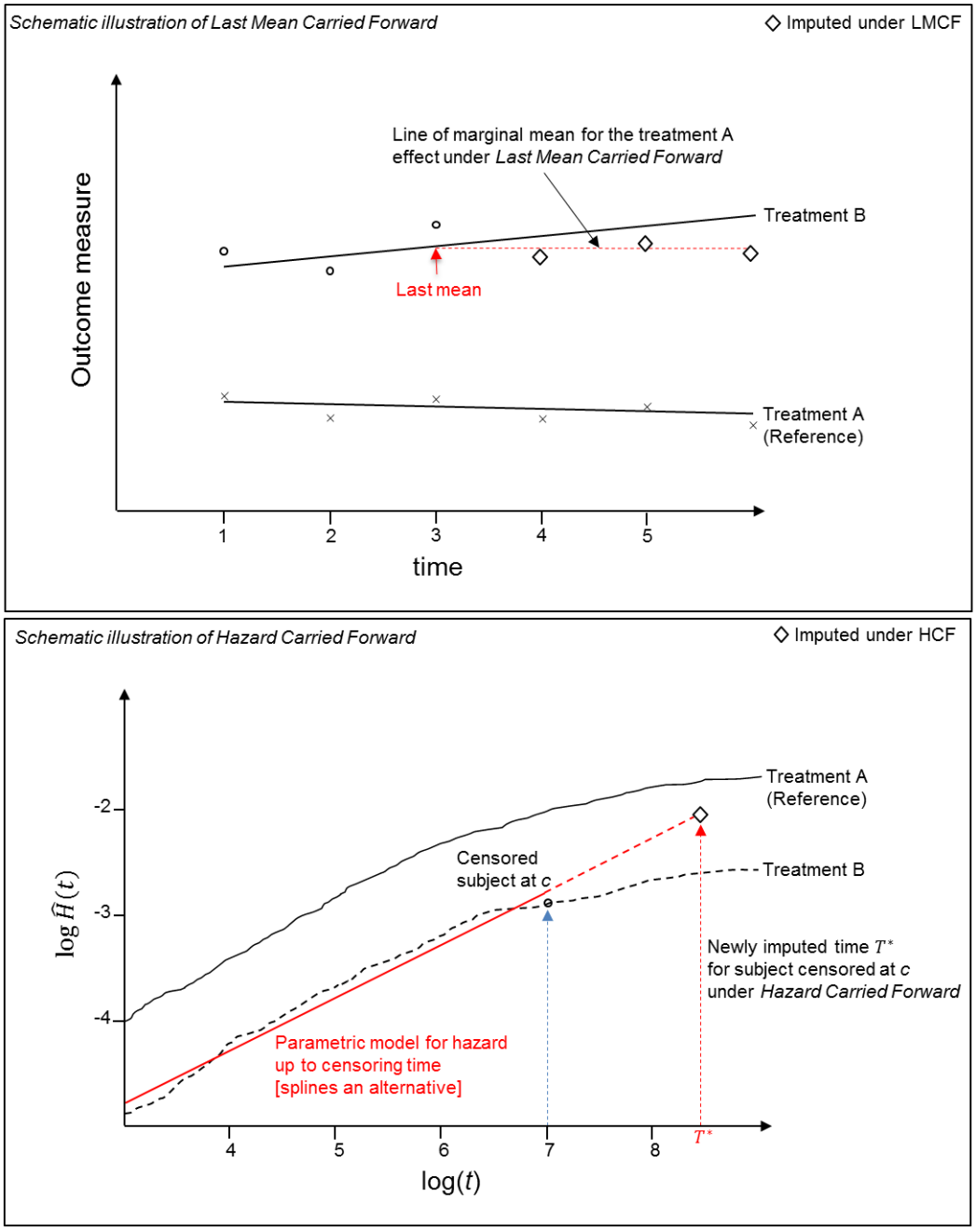


Figure 2.4.2: Top panel: longitudinal data Last Mean Carried Forward; bottom panel: time-to-event data Hazard Carried Forward.

Comment

A slight modification of this approach is to use the whole history of the hazard on the reference arm. The post-censoring hazard for a patient censored at time $c$ would then be defined in terms of a parametric (or spline) model, defined by the complete (or possibly only local) history of the pre-censoring hazard for the patient:

$$h_{post,i}(t) \;=\; \tilde{f}(h(t|t \leq c, x = 1)) \times t^*,$$

for some parametric function $\tilde{f}$ and imputed event time $t^*$. For example, linear extrapolation is shown in the bottom panel of Figure 2.4.2. For the parametric function we might also define $\tilde{f}$ to be a Weibull model, with hazard defined at time $t$ as

$$h(t) = \left(\frac{k}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{k-1},$$

where $k > 0$ is the shape parameter, and $\lambda > 0$ is the scale parameter of the distribution.

Thus, for a patient censored at time $c$, the hazard would be defined as:

$$h(c) = \left(\frac{k}{\lambda}\right)\left(\frac{c}{\lambda}\right)^{k-1},$$

and therefore, assuming the hazard for times $t^* > c$ follows this model, the cumulative hazard would be:

$$H(t|t^* > c, x = 1) = \left(\frac{k}{\lambda}\right)\left(\frac{c}{\lambda}\right)^{k-1}(t^* - c).$$

### 2.4.4 Copy Increments in Reference

With Copy Increments in Reference (CIR) , shown in the top panel of Figure 2.4.3, the post-deviation mean increments are copied from the reference group. This means that the "delta" treatment effect on Treatment A (the reference) is copied, and used for the post-deviation missing outcomes on Treatment B (denoted by unfilled crosses at times $t = 4$, 5 and 6). This might, for example, be seen in an Alzheimer's study in which treatment halts disease progress, but stopping treatment allows the disease to progress again (cf. p.251 of Carpenter and Kenward (2012)).

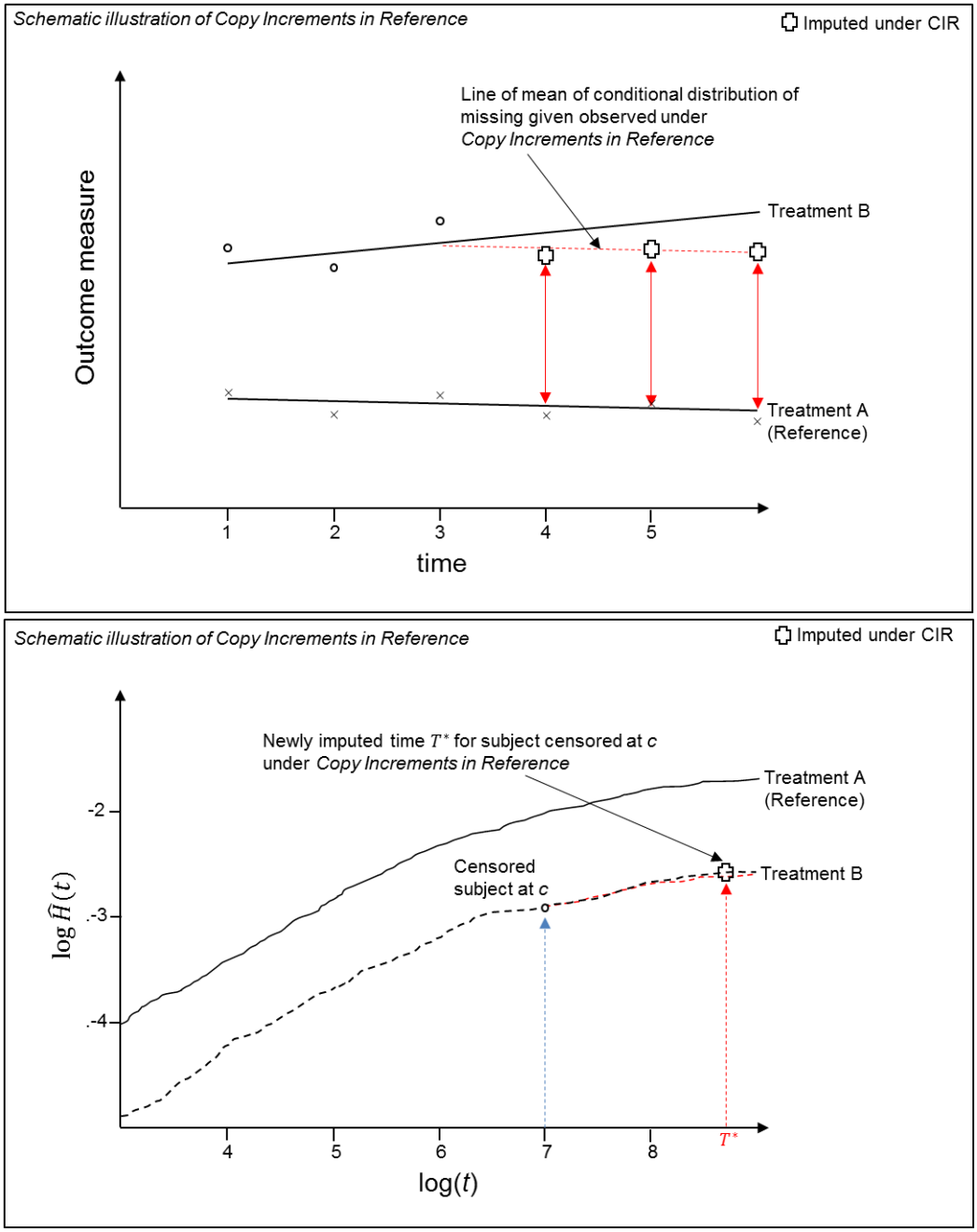


Figure 2.4.3: Top panel: longitudinal data Copy Increments in Reference; bottom panel: time-to-event data Copy Increments in Reference.

Assuming proportional hazards, for the time-to-event data in the bottom panel of Figure 2.4.3, under Copy Increments in Reference the post-censoring distribution follows the existing hazard rate for treatment B, with small fluctuations, represented by the unfilled cross in the figure.

Here, the post-censoring hazard copies the increments in the reference hazard, so that

$$h_{post,i}(t|t > c_i) = \frac{h_{act}(c_i)}{h_{ref}(c_i)} h_{ref}(t),$$

where $h_{act}$ refers to the hazard on the active arm and $h_{ref}$ to that on the reference arm.

The treatment lines run parallel to one another in the bottom panel of the figure because we assume the hazards are proportional, and therefore by copying post-censoring increments in the reference, the post-censoring hazard for a patient continues to be that of their randomised arm. Therefore, under proportional hazards, this is equivalent to censoring at random; under non-proportional hazards, it will of course differ.

Thus, Copy Increments in Reference has no useful counterpart with survival data, if the pre-deviation hazards are proportional. Indeed, the CIR method mapped to time-to-event data will lead to similar results as imputation, or standard (partial) likelihood analysis, under CAR.

### 2.4.5 Copy Reference

Under "Copy Reference" (CR) for longitudinal data, the deviating patient's whole outcome distribution, both pre- and post-deviation, is assumed to be exactly the same as for the reference. The imputation distribution uses the mean and variance-covariance matrix from the reference arm. The pre-deviation data from the patient on the treatment arm is *not* used in the estimation of reference arm distribution, only data from patients on the reference arm. This is illustrated in the top panel of Figure 2.4.4, where post-deviation imputed outcomes, denoted by hexagons, track back towards the Treatment A conditional mean. This models the case in which those deviating do not respond to Treatment B, or possibly never took it.

Copy Reference is defined analogously for time-to-event data. A random draw from the hazard of the reference arm is taken, and only accepted if the corresponding imputed time exceeds the censoring time for the patient; this is represented in bottom panel of Figure 2.4.4 by the unfilled hexagon.

The post-censoring hazard for a censored patient under CR is defined as if they were *always* on the reference treatment throughout:

$$h_{post,i}(t) \ = h_{ref}(t).$$

As with the longitudinal data definition, this models the case where a patient responds exactly as if they were on the reference, with no response to Treatment B.

Since we are now considering survival data, this method is equivalent to the "Jump to reference" approach in the longitudinal data setting. Also, under this method, we can choose for the patient to jump to the hazard in the reference group at any time $t$ during the follow-up, but $t = c$ is most natural.

In the remainder of this section, some new methods are defined which are not based on those from longitudinal data.

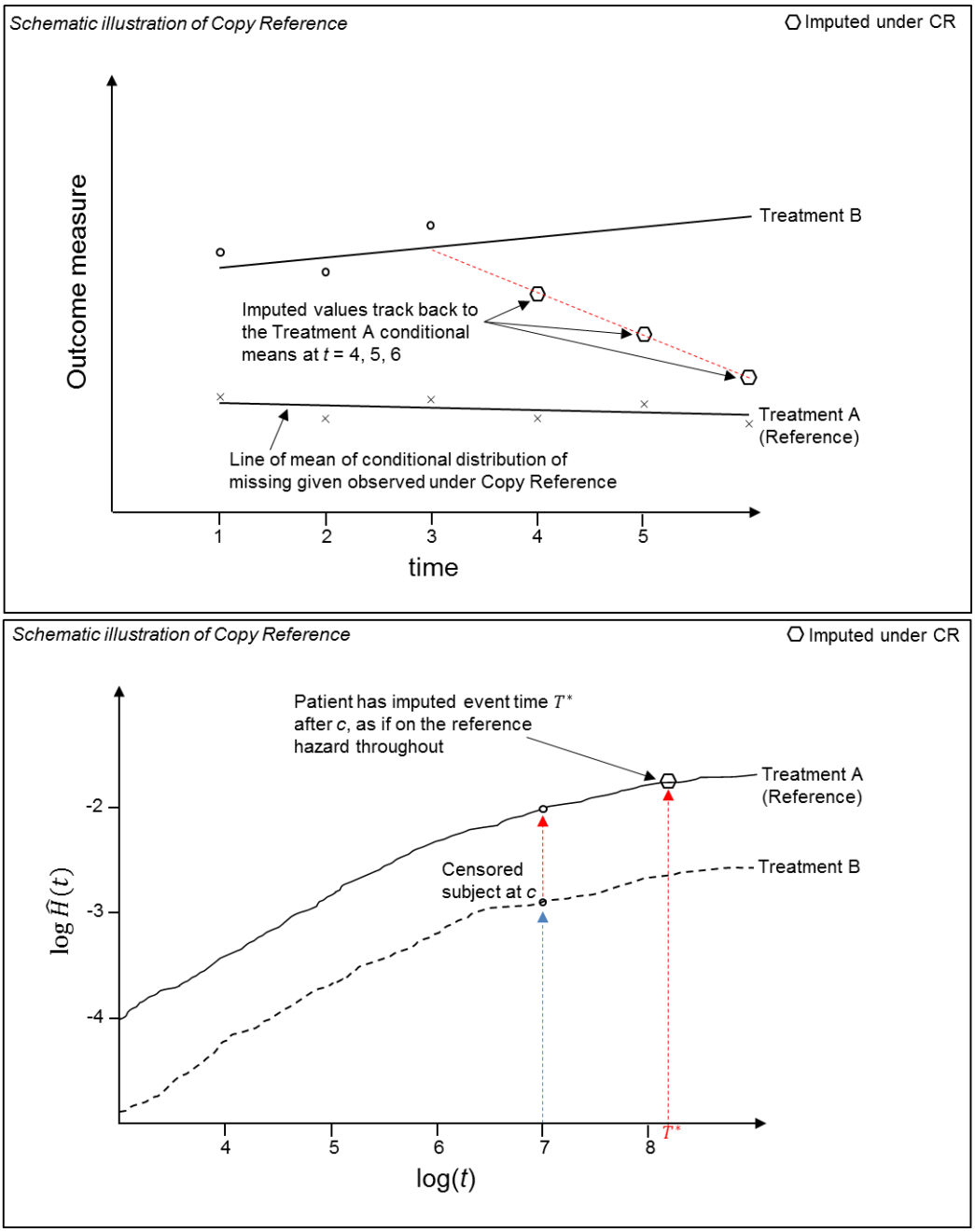


Figure 2.4.4: Top panel: longitudinal data Copy Reference; bottom panel: time-to-event data Copy Reference.

### 2.4.6 Immediate Event

The Immediate Event (IE) method imputes the next event time on *either* the reference or treatment arm, whichever is sooner, following patient censoring on the treatment arm, i.e. an "immediate" failure. The post-censoring hazard under IE is defined solely in terms of event times $T^*$, not the hazard rate:

$$T^* \;=\; \inf(t : t > c,\; \forall x \in (0, 1)).$$

In Figure 2.4.5, the next event after censoring at time $c$ is from the reference arm, denoted by the cross at the end of the red arrow, leading to the imputed time at $T^*$, denoted by the unfilled pentagon on the curve for Treatment B. In essence, this is rather similar to the hot-deck imputation methods mentioned in Chapter 1, in which the nearest neighbour is used as a substitute.

This might model the case where censoring is the result of severe complications which rapidly lead to the patient's death. Of course, this is an extreme case for a sensitivity analysis, but may be useful when considering "boundary" scenarios.

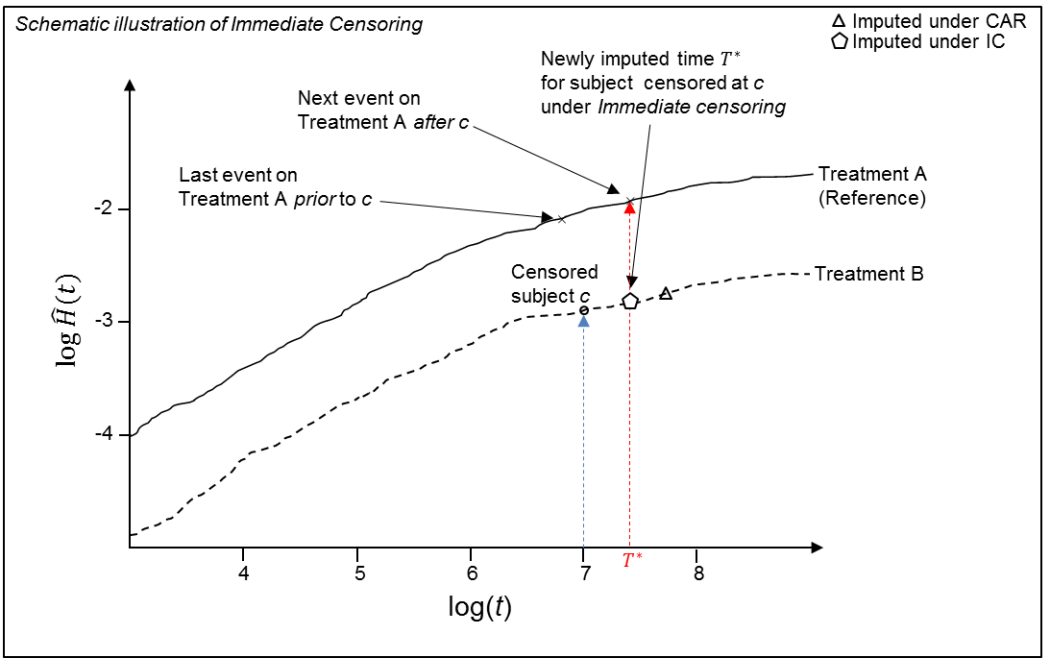


Figure 2.4.5: Immediate Event for time-to-event data.

### 2.4.7 Hazard Increases/Decreases to extremes

For these methods, the post-censoring hazard for Treatment B experiences an extreme increase (**E**xtreme **H**azard / **I**ncrease, EH/I), or an extreme decrease (**E**xtreme **H**azard / **D**ecrease, EH/D).

For example, this could model a case in which:

- toxicity causes the hazard to rise to a much high level (EH/I), or,
- the patient drops out of the study due to significant improvement in their health, with further treatment incurring unnecessary additional side effects (EH/D).

The post-censoring hazard rate for a patient on Treatment B censored at time $c$ is defined solely in terms of a pre-defined hazard $L$:

$$h(t|t > c, x = 1) \; = \; L.$$

In Figure 2.4.6, both hazard increasing and decreasing are illustrated by points *A* and *B*, with associated hazards denoted by red dotted lines, leading to the imputed event times, $T^*$ (the stars).

Again, as with the IE event, this might also be applicable when investigating boundary scenarios within the context of a sensitivity analysis.

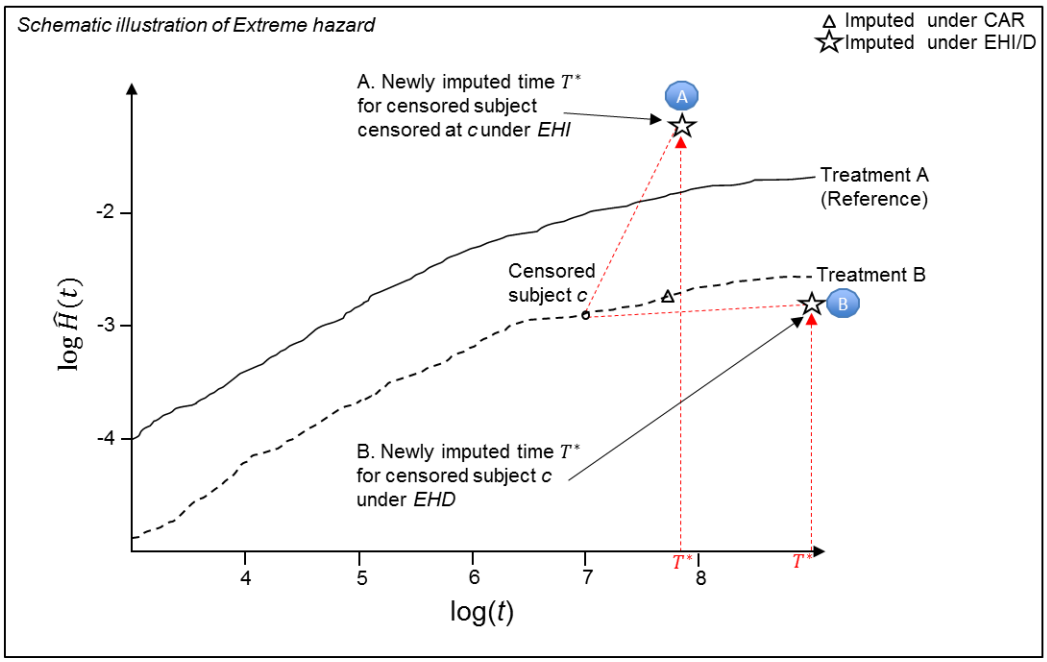


Figure 2.4.6: A. time-to-event data Extreme Hazard - Increasing (EH/I) and B. Extreme Hazard - Decreasing (EH/D).

### 2.4.8 Hazard Tracks Back to reference in time window

Figure 2.4.7 illustrates this method, in which the hazard for Treatment B tracks back to the reference hazard in a defined manner, and within a specific time window (denoted by $\omega$).

For example, we might use this approach to model a scenario in which a change in the hazard occurs after the patient is censored, increasing for a short period, after which the hazard stabilises. This scenario might, for example, be applicable for a new treatment which causes adverse effects causing some patients to drop-out, at which time the toxicity increases the risk for a short time, before side effects stabilise. Conversely, if the hazard decreases following censoring and then stabilises, then this might model a scenario in which there is a positive carry-over effect from the experimental treatment, following which the hazard goes back to its usual rate.

A number of possible options present themselves for defining the shape of the trajectory with which the hazard tracks back to the reference. For example, and referring back to Figure 2.4.7, the tracking mechanism might be linear (*A*); it could increase sharply initially, and then run at a tangent to the reference cumulative hazard (*B*), or it might run tangential to the cumulative hazard of the treatment, before increasing steeply towards the reference (*C*). Whichever trajectory is chosen, the hazard for a patient increases for the post-censoring time window $\omega$, following which it reverts back to the original Treatment B hazard.

The example in Figure 2.4.7 shows a Treatment B patient censored at time $c$. A new event time is imputed, based on a sample hazard rate for events in the time window ($\log(t) = 7,\ \log(t) = 7 + \omega$). A new event time on the linear trajectory is illustrated in the figure by the inverted triangle. If the options (*B*) or (*C*) above were used, then the interval would be defined using an appropriately defined equation for the trajectory.

With this approach, the hazard for Treatment B, $h_B(c)$, tracks back to the reference hazard, $h_A(c)$, in a defined manner, and within a specific time window, denoted by $\omega$ (so $h_A(c + \omega)$).

The hazard for a patient increases for the post-censoring time window $\omega$, following which it reverts back to the original Treatment B hazard. The post-censoring hazard under linear HTB is defined in terms the discrete hazard rates on the time window, $\omega$:

$$h(t|t > c, x = 1) = \begin{cases} h_B(c) \ + \ (t-c)\left[\frac{h_A(c+\omega)-h_B(c)}{\omega}\right] & \text{if } t \in (c, c+\omega) \\ h_B(c) & \text{if } t > c+\omega \end{cases}$$

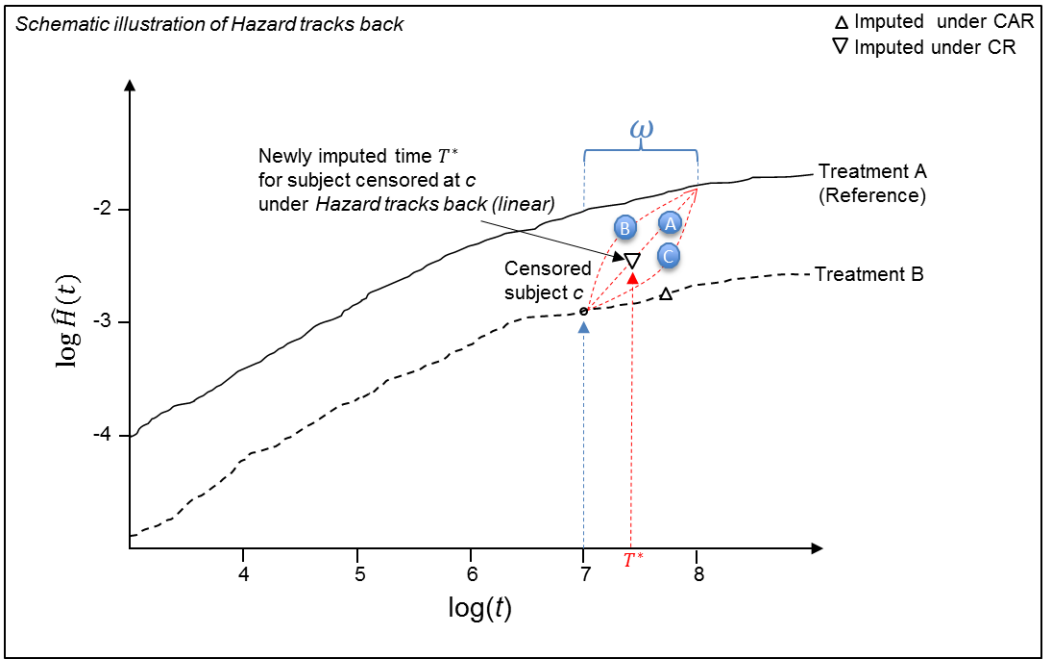


Figure 2.4.7: Hazard Tracks Back to reference in window $\omega$; A. Linear trajectory for the hazard; B. Hazard tracking back, tangentially to the reference Treatment A after an initial steep increase; C. Hazard tracking back, initially tangentially to Treatment B, followed by a steep increase.

### 2.4.9 Delta methods

We briefly introduced *delta* methods in section 1.5.3. With such methods, a patient's post censoring hazard, is a multiple $\delta$ of (for example) the hazard in the active arm prior to censoring, i.e.

$$h_{post}(t|t > c_i) = \delta \times h_{act}(t).$$

This method requires the definition of the sensitivity analysis parameter $\delta$, and its distribution.

With reference to Figure 2.4.8 the subject $c$ is censored at $t = 7$. The hazard of the treatment B group is $h_{act}(t)$, and to impute a new event time for subject $c$ we define a new hazard $\Delta h_{act}(t)$, where here, for example, we choose a hazard twice that of that for the Treatment B arm i.e. $2h_{act}(t)$. Using this new hazard, a new event time $T^*$ is multiply imputed. We note that the hazard in this case is much higher than that on the Treatment A arm (the reference). However, this need not necessarily be the case.

As for longitudinal data, the delta-method is an approach that requires the user to specify a sensitivity parameter. This has the potential advantage that a so-called "tipping point" analysis can also be performed, whereby $\Delta$ is moved away from 1 (i.e. CAR) until the conclusions change. Alternatively, we may seek expert opinion on $\Delta$, but this may be controversial (Mason *et al.*, 2017a; Heitjan, 2017; Mason *et al.*, 2017b).

The main advantage of the delta method is that it is rather straightforward to implement. However, as mentioned in the introductory remarks in section 1.5.3, the main drawback is that it is then difficult to interpret clinically. For example, one might perform a tipping point analysis and find that the treatment effects are similar in the primary and sensitivity analysis until the $\delta$ parameter is increased by (say) a factor of 2. But it is often difficult for the trial team to determine if this multiplier would be clinically plausible in a real situation — for example, is the doubling of the log hazard realistic? Equally, in a more complex scenario in which we define a different $\delta$ for each arm of the trial for the sensitivity analysis, the correlation between these different $delta$ values would also have to be derived, which would also be difficult to elicit meaningfully from experts.

Reference based methods avoid these issues since they translate seamlessly to clinically plausi-

ble scenarios. For this reason, whilst acknowledging their use in trials, we chose to focus on the reference based methods, since they are inherently aligned with our need for *clinical plausibility* and accessibility to the trial team stakeholders.

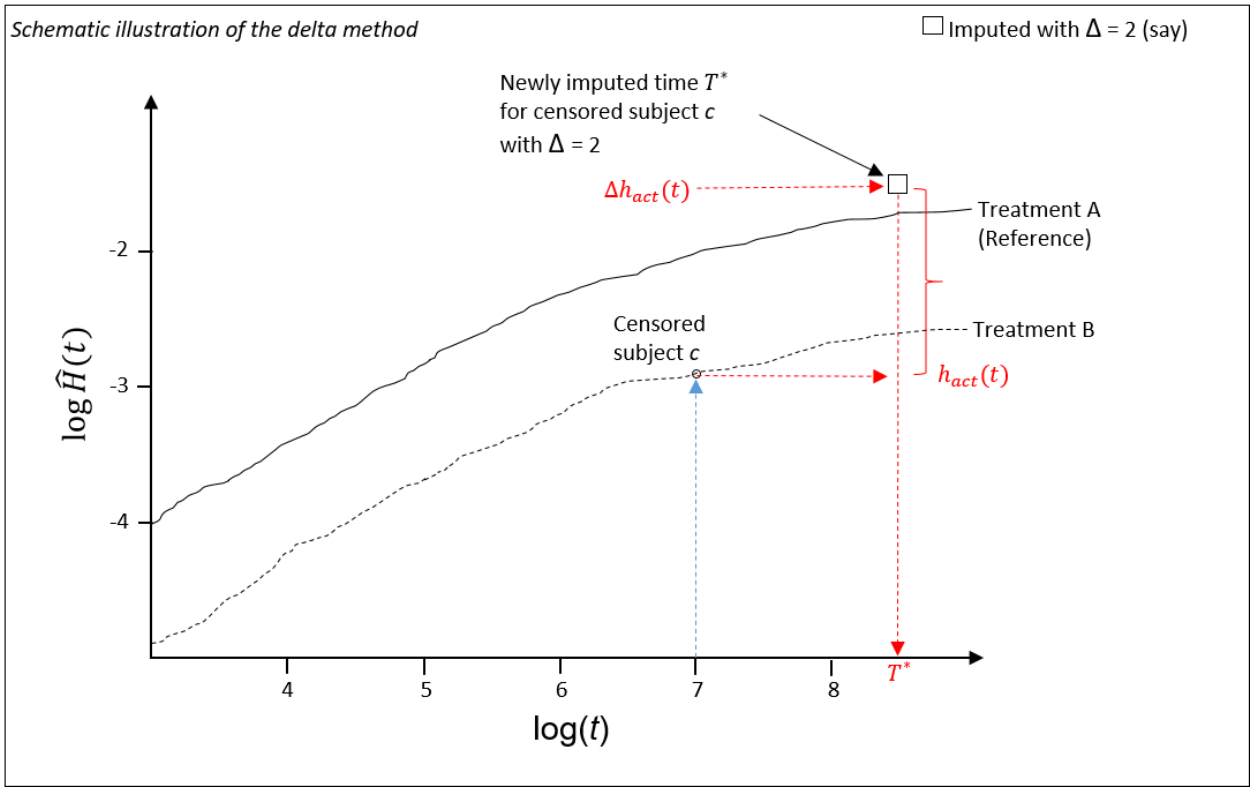


Figure 2.4.8: Delta method for time-to-event data data

## 2.5 Summary

We have now summarised how each of the methods set out by Carpenter *et al.* may be mapped to scenarios in which there is a time-to-event outcome.

In the next section we provide results from applying these methods to a simulated data set to validate their *practicality* and *clinical plausibility*.

## 2.6 Visualisation of the methods using simulated data

### 2.6.1 Introduction

To investigate the sensitivity analysis methods in a time-to-event setting, our strategy is to consider a simple clinical trial, with two arms, with patients either randomised to reference Treatment A, coded as $x = 0$, or new Treatment B, coded $x = 1$ . For each of the sensitivity analysis methods, censored event times in the reference treatment arm (A) are always imputed under CAR. However, the joint distribution for experimental treatment arm (B) is constructed using one of the different approaches (i.e. modelling potential CNAR scenarios).

The Cox proportional hazards model is fitted as substantive model, since this is the most common model for the analysis of time-to-event data from clinical trials (Cox, 1972). We simulated event times from an exponential distribution, with control arm hazard $h(t) = 0.001$, and hazard ratio $\exp(\beta) = 0.5$ using the approach described by Bender *et al.* (2005). For survival times $t$, a draw from the the survival distribution is given as:

$$S(t|x) = -\frac{log(U)}{\lambda_1 \exp(\beta x)}, \;\; \beta = \log\left(\frac{\lambda_2}{\lambda_1}\right),$$

where

$U$ is a variable generated from a uniform distribution on $[0, 1]$,
$\beta$ is the regression coefficient for the Cox proportional hazards model,
$\lambda_1$ is the baseline hazard for the reference Treatment A patient group,
$\lambda_2$ is the hazard for the Treatment B patient group, and,
$x$ is the binary treatment covariate.

Using this generating function, and fitting the Cox proportional hazards model leads to an estimate of approximately $\exp(\beta) = 0.5$, so that the hazard rate of the treatment group ($x = 1$) is half that of the reference group ($x = 0$).

A second uniformly distributed set of censoring times is generated for each patient. If this time is less than the original exponentially distributed event time, then the patient is defined to be censored, otherwise the patient is defined to have experienced the event. We applied this process

to generate event times for 1000 patients, equally split between both arms, uniformly censored at a specific rate. Data sets simulated censoring rates of 10%, 50% and 75% in our study.

Once the data sets for the three censoring levels were generated, each of the MI approaches for modelling CNAR were then applied. For each method and censoring level, 20 imputed data sets were generated, and Cox Proportional Hazards model fitted as analysis model. We also calculated and plotted the Kaplan-Meier product limit estimator for the survival function of each arm, so that we were able to visually identify the effect of the methods on the proportional hazards assumption.

A single simulated data set was used to demonstrate the application of each of the methods since practicality and clinical plausibility were the main focus of this initial study. The results presented here compare empirical versus expected behaviour of the point estimators.

The next chapter presents a more comprehensive simulation study where the focus is on the statistical properties of the variance estimates derived from applying the methods.

### 2.6.2 Results

Of the four existing methods defined in Section 2.4, Jump to Reference (J2R), Hazard Carried Forward (HCF), Copy Increments in Reference (CIR) and Copy Reference (CR), we found that, as expected, only "Jump to Reference" led to significantly different results when compared to imputing under CAR. For this reason, we only present the results from simulating the "Jump to Reference" method in detail, along with the results from the new approaches developed especially for time-to-event data.

Jump to Reference

The bottom right panel of Figure 2.6.1 shows the Nelson-Aalen cumulative hazard curves for the estimated cumulative hazard of both treatments. The reference arm does not have any censored patients, and those censored on the treatment are multiply imputed using the J2R approach. 50% of the patients are censored on the treatment arm.

Assuming proportional hazards, the curves in Figure 2.6.1 would run parallel to one another. As might we might expect, the cumulative hazard for Treatment B under J2R no longer runs parallel to that for the reference treatment A (bottom right panel in Figure 2.6.1). The convergence was also present at the 75% censoring level, but was not visible with only 10% censoring. This is exactly what might be expected, since under J2R, the treatment B hazard becomes "diluted" or "mixed" with the hazard of the reference arm (treatment A) for the censored patients.

We can predict the level of convergence between the two lines using a very simple calculation, which defines the new parameter estimate for the Cox model under J2R. On the reference treatment A, the hazard is assumed to be constant throughout ($\lambda_1$), whereas on Treatment B the hazard is dependent on the proportion of patients being censored at time $t$, and consequently imputed under J2R:

$$h_{post}(t|x=1) = c(t)\lambda_1 + (1-c(t))\lambda_2,$$

where
$c(t)$ is the proportion of patients censored on Treatment B at time $t$, $\lambda_1$ is the baseline hazard for the reference Treatment A patient group, and,
$\lambda_2$ is the hazard for the Treatment B patient group.

Therefore, the hazard ratio under J2R can be expressed as:

$$\beta_{J2R} = \frac{c(t)\lambda_1 + (1-c(t))\lambda_2}{\lambda_1} = c(t) + (1-c(t))\beta_{orig},$$

where
$\beta_{orig}$ is the hazard ratio for the model imputed under CAR (see the bottom left panel of Figure

2.6.1).

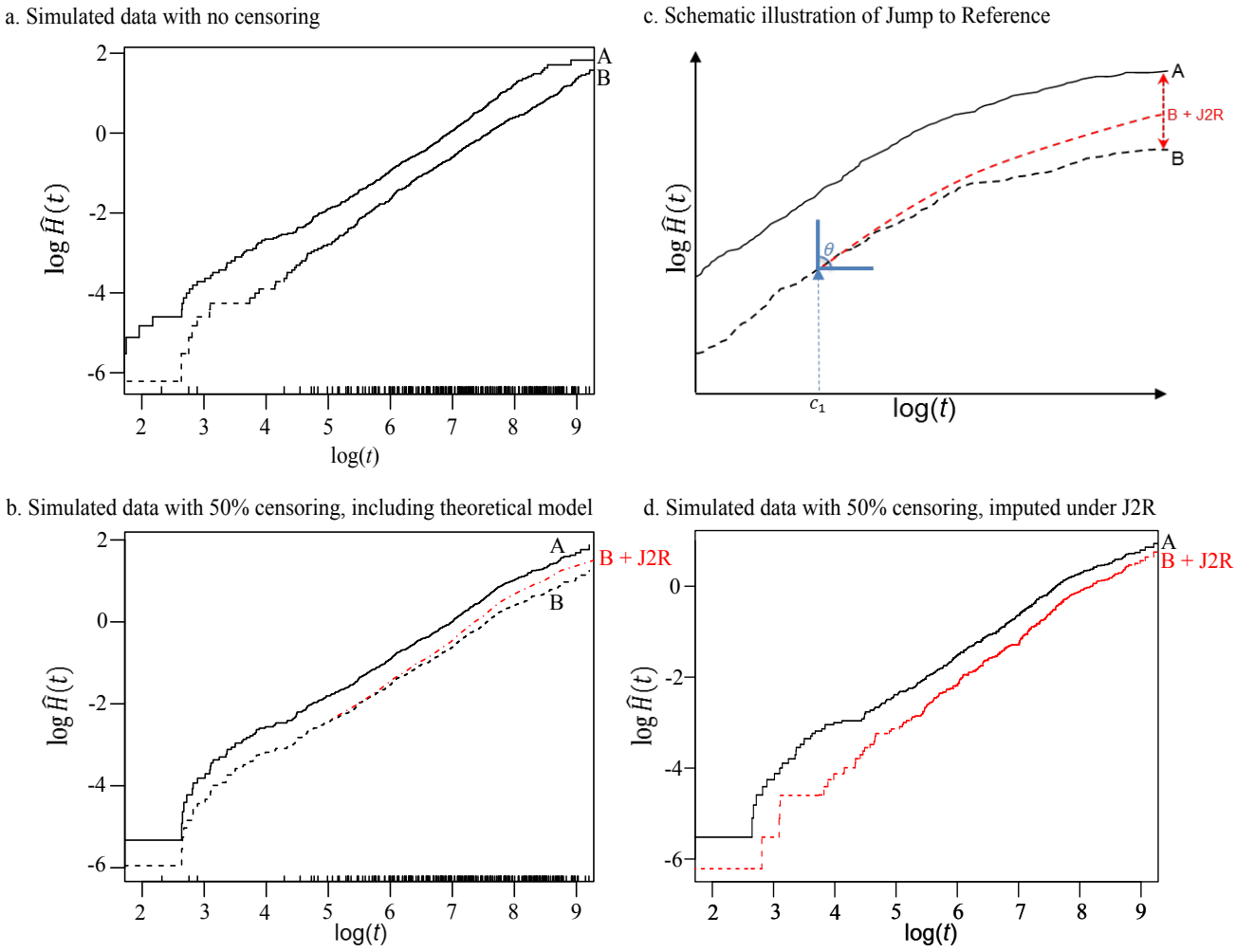


Figure 2.6.1: Comparison of empirical and theoretical results under Jump to Reference
a.) Top left panel: Simulated data with no censoring, plotted using the Kaplan-Meier estimate.
b.) Bottom left panel: Simulated data with 50% censoring, both treatments imputed under CAR, including the 50% censoring rug on the x-axis, and the theoretical model for $\beta_{J2R}$ (dot-dashed in red).
c.) Top right panel: Schematic illustration of the theoretical prediction of post-censoring proportional hazards under Jump to Reference (J2R, dashed red).
d.) Bottom right panel: Assessment of proportional hazards under Jump to Reference with 50% censoring; reference Treatment A imputed under CAR (solid black line); Treatment B imputed under J2R (red dotted line).

Immediate Event

We would expect this method to have a drastic effect on the cumulative hazard curve, assuming the censoring level is relatively high, making slope of the cumulative hazard curve even steeper. In the simulation study, the effect is clearly visible in the bottom panel of Figure 2.6.2, with the Treatment B curve crossing the reference curve when there are 50% censored patients on the Treatment B arm.

Although clinically rather unrealistic, the IE method could be used as a possible worst case scenario in the context of a sensitivity analysis.

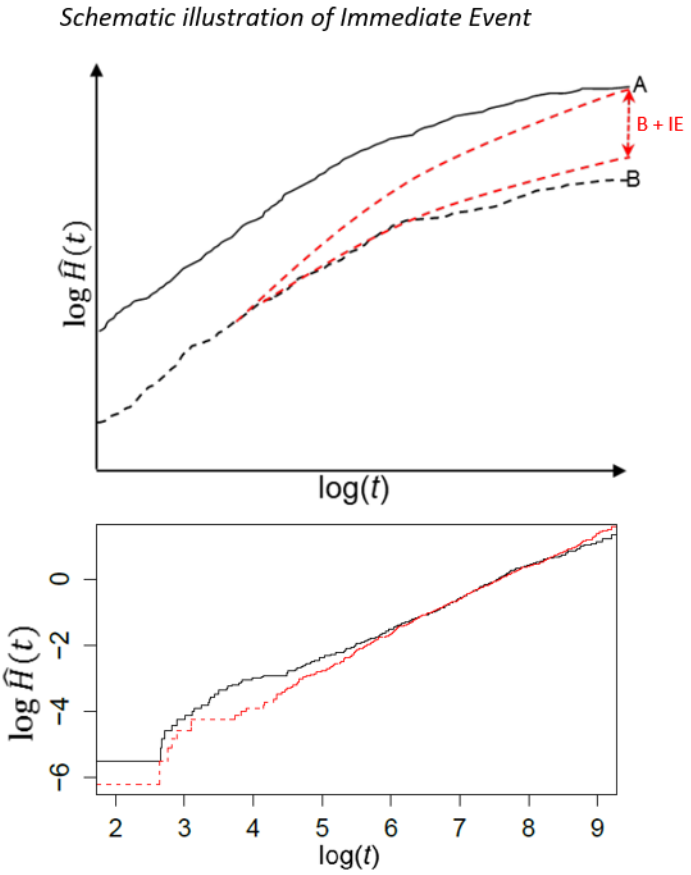


Figure 2.6.2:
Top panel: Schematic illustration of the theoretical effect post-censoring on proportional hazards under Immediate Event (IE), with a range possible depending on the censoring level (red dotted arrow).
Bottom panel: Simulation assessment of proportional hazards under Immediate Censoring with 50% censoring; reference Treatment A imputed under CAR (black solid line) ; Treatment B imputed under IC (red dotted line).

Extreme Hazard Increasing/Decreasing

Under Extreme Hazard Increasing (EH/I), the post-censoring hazard increases for patients on Treatment B (top panel of Figure 2.6.3). The bottom panel of this figure shows the simulated data, imputed under EH/I, with 50% censoring. As with the IE method, the log cumulative hazard curves for the treatment groups cross.

Under Extreme Hazard Decreasing (EH/D), the patient's hazard decreases from that at the time of censoring, to a much lower level. This was noticeable from the simulated data set under EH/D, even at the relatively low level of 10% censoring, with the effect becoming more pronounced as the censoring level increases (see Figure 2.6.4 for 50% censoring).

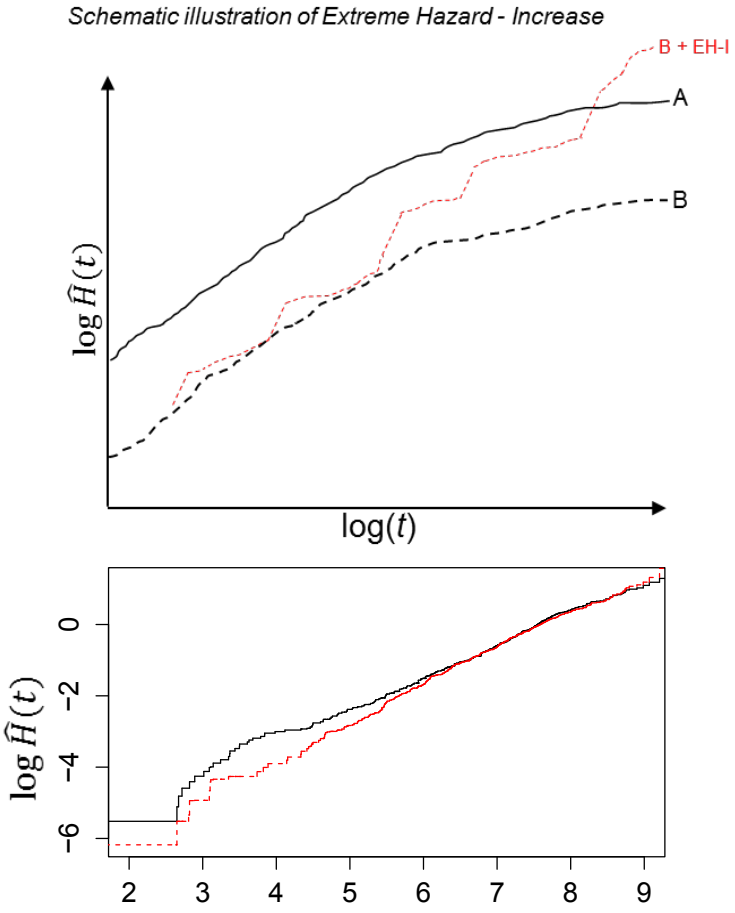


Figure 2.6.3:
Top panel: Schematic illustration of the theoretical prediction of post-censoring proportional hazards under Extreme Hazard Increasing (EH/I).
Bottom panel: Simulation assessment of proportional hazards under EH/I with 50% censoring; reference Treatment A imputed under CAR (black solid line); Treatment B imputed under EH/I (red dotted line).

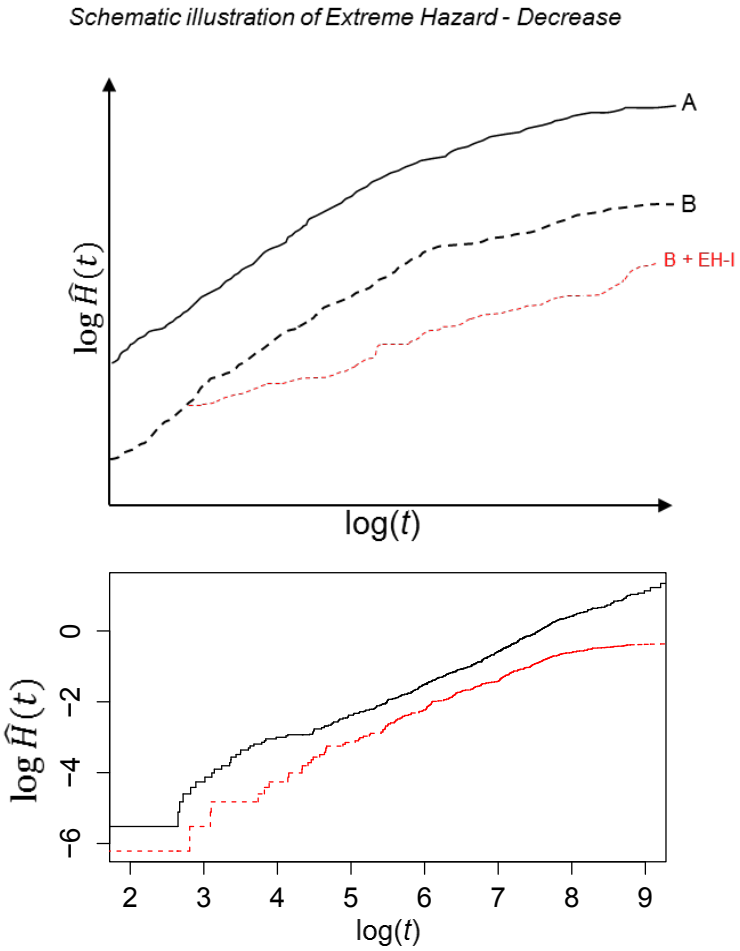


Figure 2.6.4:
Top panel: Schematic illustration of the theoretical prediction of post-censoring proportional hazards under Extreme Hazard Decreasing (EH/D).
Bottom panel: Simulation assessment of proportional hazards under EH/D with 50% censoring; reference Treatment A imputed under CAR (black solid line); Treatment B imputed under EH/D (red dotted line).

Hazard Tracks Back

For the HTB method, various window lengths were investigated, combined with different censoring levels, to try to quantify the influence of the window parameter when applying this method. Generally, there was evidence of interplay between the censoring level and window length (bottom half of Figure 2.6.5), especially for the longest time windows ($\omega = 2000, 5000$). If the window is short then the convergence of the curves is quicker than for longer windows. Furthermore, window length and the *distribution* of the censoring during the follow-up period is also important. For example, if the window is short and there is more censoring at the beginning of the follow-up period, then the convergence of the curves will be more predominant compared to the case in which the window length is short and the censoring is mostly in the later phase of the follow-up period.

A slight modification of this approach increases the hazard after each treatment, continues at this higher level for a short period, after which the hazard stabilises until the next dose of the treatment is given, when the hazard increases again (top half of Figure 2.6.5). This might model a certain subgroup of patients experiencing side effects immediately after each treatment.

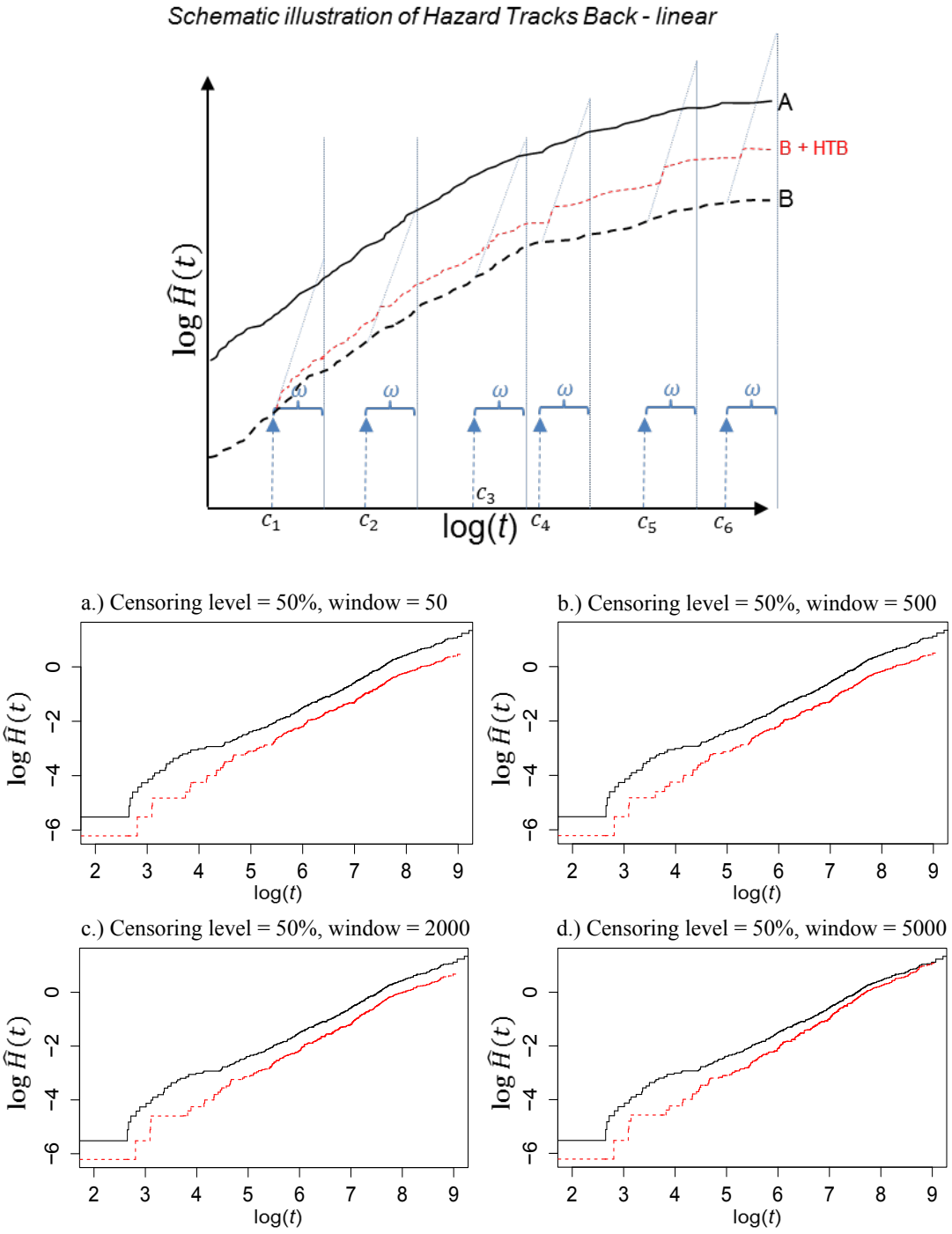


Figure 2.6.5:
Top panel: Schematic illustration of the theoretical prediction of post-censoring proportional hazards under Hazard Tracks Back to reference (HTB).
Bottom panel: Simulation assessment of proportional hazards under Hazard Tracks Back to reference with 50% censoring; reference Treatment A imputed under CAR (black solid line); Treatment B imputed under HTB (red dotted line).
Panels: a.) window size $\omega = 50$, b.) $\omega = 500$, c.) $\omega = 2000$, and d.) $\omega = 5000$.

## 2.7 Discussion

The results from this visualisation study provide an initial evaluation of the sensitivity analysis methods for time-to-event data in terms of their *practicality*, which encompasses both ease of implementation and use, and their potential applicability in terms of clinical plausibility.

The J2R method provides a way of changing the post-censoring hazard rate in a controlled manner, without radically altering the treatment effect. This method might well enable realistic clinical situations to be investigated in the context of a sensitivity analysis, such as when the reference treatment is the standard of care for the disease.

The EH/I and IE methods can be used for exploring abrupt changes in the post-censoring hazard. Both model less realistic clinical scenarios, but could be applicable in a setting in which sudden the onset of complications might be expected. To contrast this, the EH/D method models a type of "best case" scenario in which post-censoring a patient is expected to improve markedly.

Of the four methods defined by Carpenter *et al.*, "Jump to Reference" (J2R), "Last Mean Carried Forward" (LMCF), "Copy Increments in Reference" (CIR) and "Copy Reference" (CR), we found that only "Jump to Reference" led to point estimates of the hazard ratio which were significantly different (that is, not within the 95% confidence intervals) from those following multiple imputation under CAR (results not shown for HCF, CIR, CR). Of course, this was what might have been expected since we assume proportional hazards for the data.

Having illustrated the theory, we now consider what happens when the methods are applied to real data. In the next section, we provide more details of the German Breast Cancer data, which has a censoring level of 58%, and apply each of the sensitivity analysis approaches.

## 2.8 Application of the sensitivity methods to the German Breast Cancer data

### 2.8.1 Introduction

We now apply each of the sensitivity analysis methods to the German Breast Cancer (GBC) data which was briefly introduced in section 1.11.1. The main goal is to determine how robust the original conclusions from the study are to departures from CAR using the proposed sensitivity analysis methods. As a secondary goal we also consider the plausibility of the various sensitivity analysis methods in this setting to arrive at a final interpretation of the trial.

For the purposes of this survival analysis, an event is defined to occur with the first recurrence of the disease. Table 2.8.1 provides summary statistics of the event and censoring levels.

In terms of the chemotherapy treatment, 52% of the patients experience a recurrence of the disease with 3 cycles, compared to 48% with 6 cycles. Given the limited difference between the chemotherapy levels, it is unsurprising that the log-rank test results in a p-value of 0.5 for 3 versus 6 cycles. This confirms the results from the original study in which there was no discernable effect from the reduction of 6 to 3 cycles in terms of patient survival (cf. Schumacher *et al.* (1994)).

However, the instances of disease recurrence are higher without hormonal treatment (64%, median recurrence time of 1684 days), compared to those taking hormonal treatment (36%, median of 2030 days). This difference in survival rates is clearly visible in Figure 2.8.1 (log rank test, p-value of 0.1).

| Randomisation group | Hormonal treatment | Chemotherapy cycles | Total | Events | Censored |
|---|---|---|---|---|---|
| 1 | No | 3 | 133 | 61 (46%) | 72 (54%) |
| 2 | No | 6 | 138 | 60 (43%) | 78 (57%) |
| 3 | Yes | 3 | 90 | 37 (41%) | 53 (59%) |
| 4 | Yes | 6 | 87 | 31 (46%) | 56 (64%) |
| **Total** | | | **448** | **189 (42%)** | **259 (58%)** |

Table 2.8.1: Treatment combinations and their censoring levels.

Since there is a tangible treatment difference for those taking the hormonal treatment, we explore whether imputation of censored survival times under the various censoring not at random sensitivity analysis scenarios provides additional insights into hormonal treatment efficacy.

Patients were followed up regularly, with clinical examinations every 3 months during the last 2 years, every 3 months for the subsequent 3 years, and every 6 months in years 6 and 7. Not all patients adhered to the schedule, with 63 patients having follow-up times longer than a year, and several patients missing information for more than 2 years. Therefore, the censored patients are a mixture of those surviving until the end of the study (i.e. administrative censoring), and those lost to follow-up during the study. Of the latter group, no additional information was available as to the reasons for dropping out. It may be assumed that these could be due to lack of tolerance to the 6 cycle chemotherapy treatment, lack of adherence to the daily hormonal treatment, other non-adherence to study protocol reasons, or the full recovery, or death, of the patient.

A large proportion of all patients did not experience a recurrence of the disease before the end of the study (259, 58%), with 62% of those taking additional hormonal treatment being censored. By convention, it is usually assumed that such missing event times are Censored at Random (CAR).

We focus on demonstrating the *feasibility* of using new methods with real data, assuming censoring is at random on both arms for the primary analysis. For reference-based sensitivity analysis, we usually fix one arm to be the "reference" arm, and assume patients are censored at random on this arm. We then vary the assumptions concerning the other "active" arm of the study to investigate different informative censoring scenarios, either for all patients randomised to this arm, or a subset of the patients in which CNAR might be appropriate. This means that the *type of censoring* often drives the post-censoring assumptions on the active arm, often based on clinical judgement. For example, the CAR assumption would usually be considered plausible for those administratively censored at the end of the study on the active arm, and for those lost to follow-up on the active arm we might assume "jump to reference" for their post-censoring hazard.

For the GBC sensitivity analysis, we assume CAR for patients not on hormonal therapy, and CNAR for all those randomised to hormonal therapy. Given the potential reasons for censoring in this data, we concede that these assumptions may not be appropriate for this trial. However, the example is sufficient to explore the feasibility of our proposals. Further, the different meth-

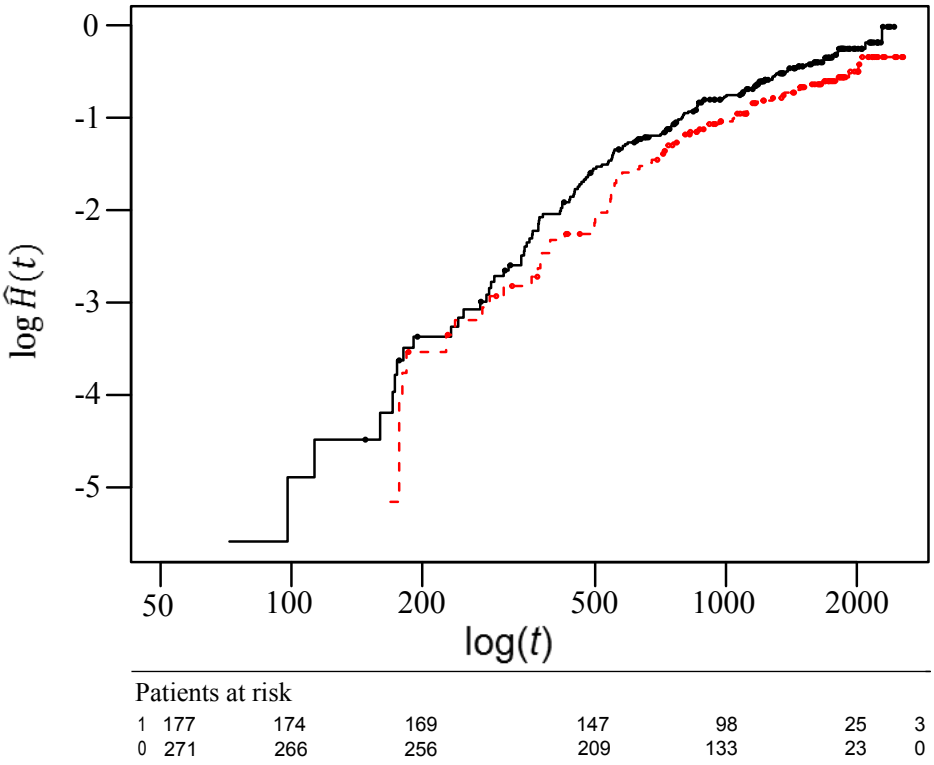


Figure 2.8.1: Log cumulative hazard against log time for the GBC data; no hormonal treatment (black solid line, no hormone therapy = 0) versus treatment with hormonal treatment (red dotted line, hormone therapy = 1); circles mark censored times; p = 0.1 (log rank test).

ods illustrated could be used as building blocks for modelling possible other post-censoring behaviour defined according to the reason of censoring (as envisaged for pattern mixture models).

### 2.8.2 Model for the data

To simplify the application of each of the methods, we focus on the treatment difference between those patients taking hormonal treatment versus those not, irrespective of chemotherapy treatment group. This is a valid analysis as the factorial design assumes, as is the case here, that there is no interaction between the treatments.

A Cox Proportional Hazards model was fitted to the data as analysis model with backward selection based on the AIC used to select relevant variables. This resulted in the following variables being included in the analysis model: Baseline tumour grade (*grad*), number of involved nodes (*npos*), and progesterone receptor level (*nprog*). This leads to the following hazard function:

$$h(t) = h_0(t) \exp(\beta_1 I[hther] + \beta_2 grad + \beta_3 npos + \beta_4 nprog)$$

where
$h_0(t)$ is the baseline hazard function,
$\beta_j$ are model coefficients for the Cox Proportional Hazards model fitted in the usual way using maximum likelihood ($j = 1, \ldots, 4$), and,
$I[hther]$ is an indicator function for hormonal treatment, taking value 0 for the reference arm, and 1 for those patients taking additional hormonal treatment.

As with the study using simulation data of the previous section, the multiple imputation method defined earlier in Section 2.2 is used to generate 20 imputed data sets. As recommended in Carpenter and Kenward chapter 8.1.3, we fit an imputation model including "...all the covariates necessary for CAR as well as those not involved but nevertheless predictable of survival". In this case, this meant including the same covariates as those included in the analysis model above, and additionally the indicator variable for the number of chemotherapy cycles.

Referring back to our discussion of congeniality in Chapter 1.8, we note that the differences between the multiple imputation and analysis models imply that in this case we have a practical example of uncongeniality.

The results from applying each of the new sensitivity analysis methods to the GBC data are

presented in the next section. Again, as with the simulated data, those methods which did not produce significantly different results compared with imputing under CAR are not shown.

### 2.8.3 Results from applying the sensitivity analysis methods to the GBC data

Table 2.8.2 on page 91 summarises the results from the investigation, highlighting those methods in which the parameter estimates were outside the confidence intervals of the original Cox model (superscripted exclamation mark), and cases where the proportional hazards assumption was violated according to the Grambsch-Therneau (G-T) test (Grambsch and Therneau, 1994)) (superscripted minus sign).

The results are consistent with the investigations using the simulated data set (see section 2.6.2). Given the censoring at random assumption used in the primary analysis, the estimated treatment effect and the patterns of censoring, we expected the J2R method to result in a limited treatment difference, and this was indeed the case (see Figure 2.8.2).

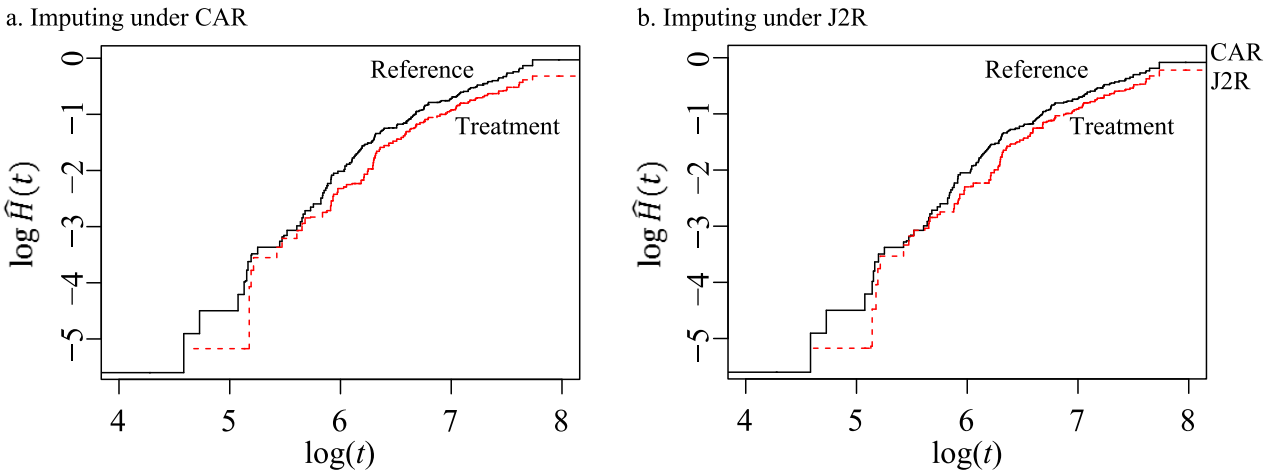


Figure 2.8.2: Left panel a.) Log cumulative hazard for reference (no hormone therapy) and treatment (hormonal therapy) arms following multiple imputation under CAR.
Right panel b.) Log cumulative hazard following multiple imputation under CAR for the reference arm and under J2R for the treatment arm.

As might be expected, the results from the IE, EH/I, EH/D and HTB methods (the latter with longer windows sizes $\omega$) led to extreme cases being modelled, with considerable deviation from proportional hazards. IE and EH/I have very similar profiles for the cumulative hazard of the treatment arm (refer to Figure 2.8.3), both being higher than the curve for the reference arm. Both convergence (EH/I) and divergence (EH/D) of the cumulative hazard curves are clearly visible in the figure.

For HTB, the shorter window lengths ($\omega = 50, 500$) did not produce a significant deviation from proportional hazards, and none of the parameters were outside the confidence intervals of those imputed under CAR.

In contrast, the parameter estimate for the hormonal treatment was significant at the 5% level for the longer window lengths ($\omega = 2000, 5000$), and there was some convergence of the cumulative hazard curves (cf. the bottom panels in Figure 2.8.4). Interestingly, there seems to be no additional effect from increasing the window length from 2000 to 5000 days.

| Sensitivity analysis method | Treatments and covariates | | | | Global G-T test |
|---|---|---|---|---|---|
| | Hormonal treatment | Tumour grade | No. involved nodes | Progesterone level | |
| Censoring at Random | $-0.258^{+}$<br>(0.141)<br>0.067 | $0.251^{-}$<br>(0.130)<br>0.054 | $0.051^{+}$<br>(0.008)<br>$< 0.001$ | $-0.002^{+}$<br>(0.001)<br>0.046 | + |
| Jump to Reference | $-0.128^{-}$<br>(0.144)<br>0.374 | $0.279^{-}$<br>(0.133)<br>0.036 | $0.052^{+}$<br>(0.008)<br>$< 0.001$ | $-0.003^{+}$<br>(0.001)<br>0.003 | - |
| Immediate Event | $0.897^{-!}$<br>(0.118)<br>$< 0.001$ | $0.234^{-}$<br>(0.108)<br>0.030 | $0.049^{+}$<br>(0.008)<br>$< 0.001$ | $-0.001^{-}$<br>(0.0004)<br>0.012 | - |
| Extreme Hazard / Increase | $0.739^{-!}$<br>(0.117)<br>$< 0.001$ | $0.240^{+}$<br>(0.109 )<br>0.028 | $0.049^{+}$<br>(0.008)<br>$< 0.001$ | $-0.001^{-}$<br>(0.0004)<br>$< 0.001$ | - |
| Extreme Hazard / Decrease | $0.619^{-!}$<br>(0.149)<br>$< 0.001$ | $0.244^{-}$<br>(0.130)<br>0.061 | $0.049^{+}$<br>(0.008)<br>$< 0.001$ | $-0.002^{+}$<br>(0.001)<br>0.046 | - |
| Hazard Tracks Back - $\omega = 50$ | $-0.337^{-}$<br>(0.148)<br>0.023 | $0.252^{-}$<br>(0.128)<br>0.049 | $0.049^{+}$<br>(0.008)<br>(0.008) | $-0.002^{-}$<br>(0.001)<br>0.046 | + |
| Hazard Tracks Back - $\omega = 500$ | $-0.257^{+}$<br>(0.145)<br>0.076 | $0.233^{-}$<br>(0.126)<br>0.064 | $0.048^{+}$<br>(0.008)<br>$< 0.001$ | $-0.002^{-}$<br>(0.001)<br>0.046 | + |
| Hazard Tracks Back - $\omega = 2000$ | $0.124^{-!}$<br>(0.144)<br>0.389 | $0.251^{-}$<br>(0.128)<br>0.050 | $0.049^{+}$<br>(0.008)<br>$< 0.001$ | $-0.002^{-}$<br>(0.001)<br>0.046 | - |
| Hazard Tracks Back - $\omega = 5000$ | $0.141^{-!}$<br>(0.143)<br>0.324 | $0.263^{-}$<br>(0.122)<br>0.031 | $0.049^{+}$<br>(0.008)<br>$< 0.001$ | $-0.002^{-}$<br>(0.001)<br>0.046 | - |

Table 2.8.2: Sensitivity methods applied to GBC data; parameter estimates for the model, with standard errors in (brackets) followed by p-values; each method used 20 imputations.
Superscript plus $(+)$ denotes the proportional hazards assumption holds under the respective post-censoring imputation method for the respective treatment/covariate (according to the Grambsch-Therneau (G-T) test).
Superscript minus $(-)$ indicates that the proportional hazards assumption does not hold, again according to the G-T test.
Superscript exclamation mark (!) means the parameter estimate from the particular imputation method is outside the 95% confidence interval of the parameter estimate for the model fitted to the data following imputation under CAR.

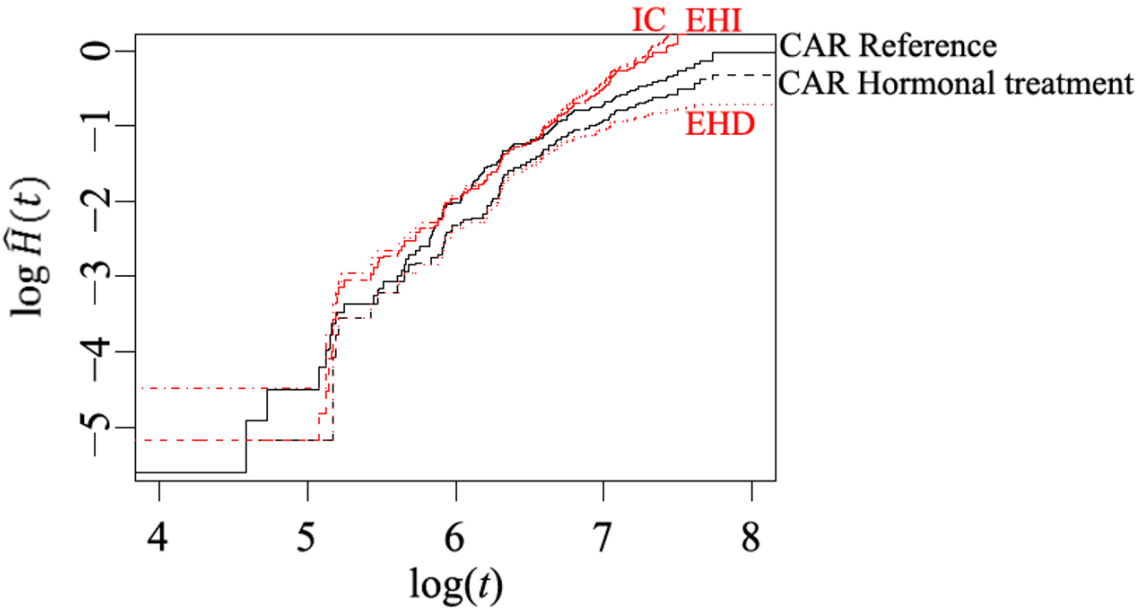


Figure 2.8.3: Log cumulative hazard for reference (no hormonal therapy) and treatment (hormone therapy) arms under CAR and EH/I, EH/D and IE; reference arm without hormonal treatment imputed under CAR (black solid line); hormonal treatment imputed under CAR (black dashed line); hormonal treatment imputed under EH/I (dashed), EH/D (dotted) and IC (dot-dashed) in red.

The investigation of the sensitivity analysis approaches using the GBC data provide additional important insights into the potential behaviour of the methods in terms of their practicality, highlighting their merits, especially in terms of their clinical plausibility.

Given that we are not aware of the exact reasons for censoring for the GBC data, we know only that those censored are a mixture of those administratively censored at the end of the study, those stopping treatment due to adverse effects, and those lost to follow-up for other reasons. Therefore, it is difficult to justify the plausibility of using the "Jump to Reference" approach for this data set, and accordingly, we present the results here as a proof of concept for the approaches only.

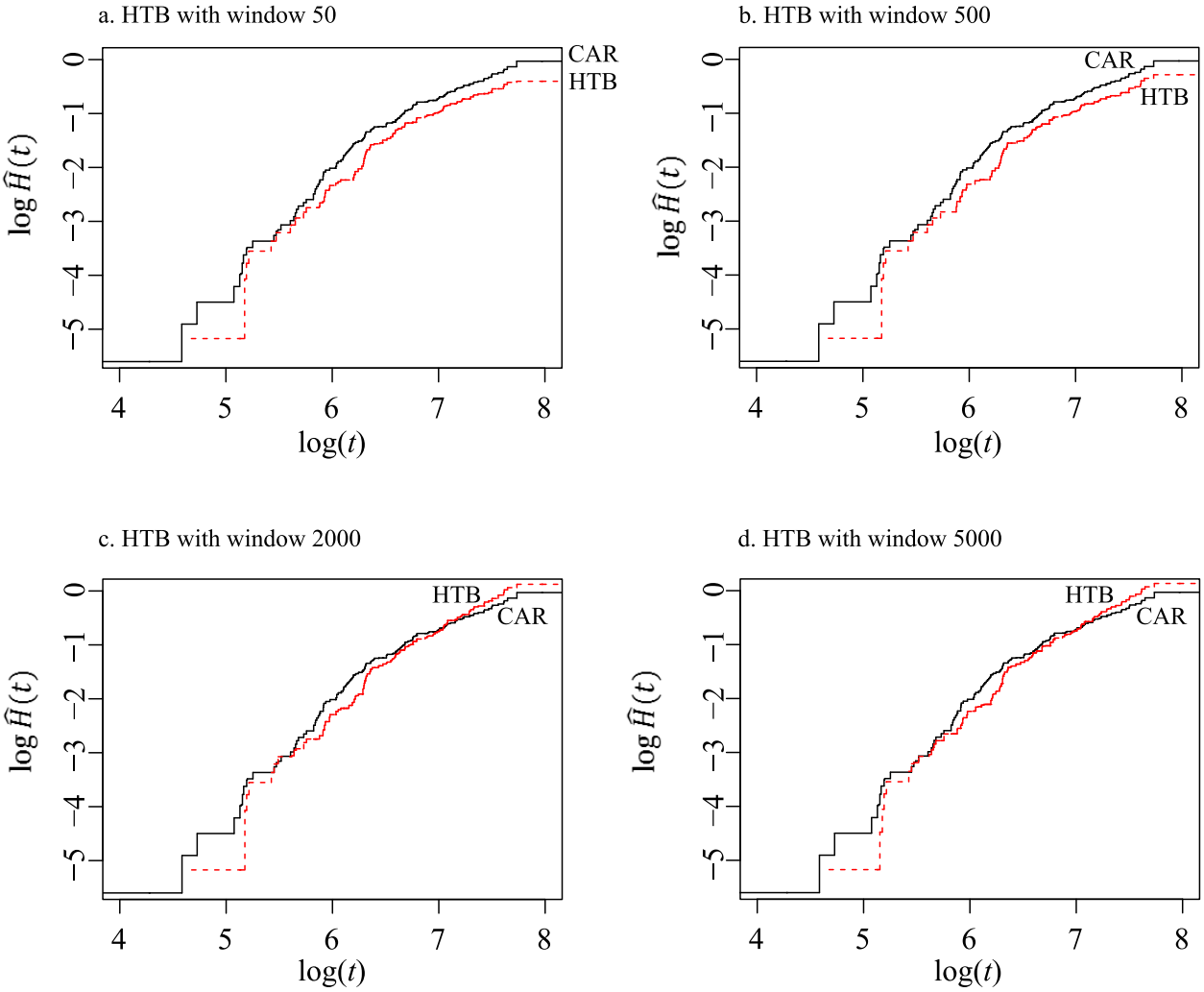


Figure 2.8.4: Log cumulative hazard for reference (no hormone therapy) and treatment (hormonal therapy) arms under CAR and HTB with varying window lengths; patients not taking hormonal treatment are imputed under CAR (solid black line); patient taking hormonal treatment are imputed under HTB (red dotted line);
Panels: a.) window length is 50, b.) 500, c.) 2000, d.) 5000.

## 2.9 Discussion of results

### 2.9.1 Evaluation of methods

The results from both the visualisation study and real data application demonstrate the feasibility of the methods originally used for longitudinal data being used in the time-to-event domain.

Furthermore, a number of new approaches were proposed, which are especially relevant for time-to-event data.

Table 2.9.1 summarises the results using three criteria to evaluate each of the methods. "Clinical plausibility" refers to the relevance of the particular method in a clinical trial, or observational study, for investigating realistic clinical scenarios. The second criteria in the table defines whether the method requires the user to specify a sensitivity analysis parameter. This is important both in terms of ease of use, but also for the acceptability of the method in practice, since the definition of the parameter is often the source of discussion (Leacy *et al.*, 2017). In the fourth column of the table, we provide an indication of the ease of implementation of each of the methods ("Practicality"), which includes an appraisal of the relative simplicity in explaining each approach, also important for the methods to be adopted.

| Sensitivity analysis method | Criteria | | |
|---|---|---|---|
| | Clinical plausibility | Parameter specification | Practicality |
| Censoring at Random | Y | N | Y |
| Jump to Reference | Y | N | Y |
| Immediate Event | ? | N | Y |
| Extreme Hazard Increasing | ? | Y | Y |
| Extreme Hazard Decreasing | ? | Y | Y |
| Hazard Tracks Back | ? | Y | N |

Table 2.9.1: Comparison of sensitivity analysis methods; Y (yes), N (no), ? (unclear)

The summary table above, which brings together the results from the visualisation study and application using the GBC data, all point towards the "Jump to Reference" method being the best option for investigating plausible departures from CAR. The other approaches would be suitable for modelling rather extreme scenarios ("Immediate Event"), or as with the "delta method", require the definition of additional parameters, which makes them potentially more difficult to define and explain.

### 2.9.2 The proportional hazards assumption

The proportional hazards assumption can be visually investigated by plotting the Schoenfeld residuals (Schoenfeld, 1982), or, as we used here, using a statisical test such as that proposed by Grambsch and Therneau (1994) – although in both cases this is often not definitive, and can be insensitive to certain forms of non-proportionality. The recent publication by
Keogh and Morris (2018) reviews and discusses methods for determining if the proportional hazards assumption holds, as does Ng'andu's comparison of methods for assessing proportional hazards (Ng'andu, 1997).

Many analyses of trials assume that hazards are proportional, although this is increasingly being challenged, for example, in the oncology setting (Royston and Parmar, 2011, 2013), with the restricted mean survival time replacing the hazard ratio as the measure of treatment effectiveness. Other methods, such as using piece-wise proportional hazards models, flexible parametric models (Lambert and Royston, 2009) or non-parametric methods (Zhao *et al.*, 2016) might also be considered for the analysis model if the proportional hazards assumption in unrealistic.

Even assuming that the primary analysis *uses* proportional hazards (PH), apart from CAR, the proposed methods imply a mixture of hazards in the active arm, which therefore, strictly speaking, violates the PH assumption. This is not just a facet of reference based sensitivity analysis approaches. Any sensitivity analysis involving the CNAR assumption will technically violate the proportional hazards assumption, since CNAR implies a mixture of hazards in each arm. Our simulation studies showed that, for most of the methods, a censoring level of 10% did not lead to the assumption being significantly undermined. At censoring levels of 50% and above, this was not the case, and the results from the investigation using the GBC data, which has a censoring level of 58%, seemed to confirm this.

We advocate taking a pragmatic approach. If proportional hazards is not considered reasonable either *a priori*, or following *posthoc* investigations using, for example, the methods outlined above, then alternative endpoints, respectively models, can be adopted for the primary analysis. In terms of the sensitivity analysis, our reference based methods are equally applicable even when the proportional hazards assumption no longer holds, although they may need to be adapted depending on the chosen end point or modelling approach taken.

## 2.10 Summary

We have considered each of the original proposals from Carpenter *et al.* (2013) and extended them for use with time-to-event data. We continued by exploring their behaviour using both a simulated data set and the GBC data. This has demonstrated that the methods can be applied in a survival analysis context. With substantial censoring, problems may arise when using the Cox Proportional Hazards model, and methods for coping with such situations have been outlined.

This brings to a close the first part of the PhD in which the focus was placed on the first two of the facets for sensitivity analysis which we defined in Chapter 1, namely *practicality* and their *clinical plausibility*.

An important, but often neglected, aspect of sensitivity analysis is that the analyst has control not only of the mean, but also the variance of the unobserved data. Relative to the primary analysis, it is therefore quite possible for a sensitivity analysis to increase, hold anchored, or decrease the statistical information about the treatment effect. If the information is held anchored as defined in section 1.6, then this provides confidence in the sensitivity analysis method. This important

(but often neglected) aspect needs to be investigated for our approach to be acceptable, particular in a regulatory setting.

In the following chapters, we consider the properties of Rubin's variance estimator following multiple imputation using reference-based sensitivity analysis with a time-to-event outcome. The aim is to demonstrate that reference based approaches not only provide unbiased estimates, but also conform to the information anchoring principle. Whilst we have investigated several different methods in this chapter, for the remainder of the PhD we focus on the most practically applicable method, Jump to Reference.

# Chapter 3

# Information anchoring for reference based sensitivity analysis with time-to-event data

## 3.1 Introduction

As highlighted in section 1.8 there has been some discussion of the use of Rubin's rules to estimate the variance following multiple imputation. Furthermore, in the case of reference based imputation, the issue is perhaps even more controversial, since, as Meng points out "a procedure that cannot be embedded into any Bayesian model should perhaps be avoided" (Meng, 1994). We interpret this to mean that if the MI process is congenial then it should be able to be implemented in a single step Bayesian procedure[1]. This is clearly not the case for reference based methods such as "Jump to Reference", since we would have to sample from two different hazard rates simultaneously. Of course, in code this might be possible, but the spirit of Meng's original statement would not not be upheld in this case.

In summary, we are certainly in an uncongenial setting when we use reference based methods, and as previously mentioned, this may lead to conservative variance estimators. The controversy surrounding the variance overestimation continues to bubble: S. Seaman *et al.* in their comments to the original paper by Carpenter and Kenward proposing the methods (Carpenter *et al.*, 2013) contend that

[1] Personal communication with James Carpenter, 25.5.18

> "under . . . 'Jump to Reference' etc. . . . , the imputer assumes more than the analyst, which is known to cause the RR [Rubins' Rules] variance estimator to overestimate the repeated sampling variance (Meng, 94)", Seaman *et al.* (2014).

More recently, Y. Tang defines the reasons for the potential issue:

> "As illustrated by Lu (2014) and Seaman et al. (2014) via simulation, Rubins (1987) variance estimator tends to overestimate the sampling variance of the MI estimator in the control-based imputation due to uncongeniality between the imputation and analysis models (Meng, 1994). Specifically, the imputation procedure assumes that the statistical behaviour of outcomes varies by pattern in the experimental arm [e.g. with Jump to Reference], but such an assumption is not made in the analysis of the imputed data, which are often analyzed by a standard method such as the primary analysis model", (Tang, 2018).

He goes on to specify this in more detail:

> "the joint distribution [of the outcome] $y_i$ among subjects with the same covariates are assumed to vary by pattern in the imputation model, but be identical in the analysis of the imputed data",

This summarises the issues in a nutshell for our reference based setting, before going on to claim that

> "the key finding is that the bias of the MI variance is generally small or negligible in the delta-adjusted PMM [pattern mixture model], but can be sizable in the control-based PMM", (Tang, 2018).

This provides the requisite motivation to investigate the properties of Rubin's estimator when applying our reference based methods in the time-to-event setting.

To reiterate our method briefly, the primary analysis model *is retained* in the sensitivity analysis, the data sets are multiply imputed and fitted to this model, and Rubin's rules applied to the

resulting estimates. If the proportional hazards assumption is used for the primary analysis, it follows that it is no longer strictly consistent with the data generating mechanism used for the sensitivity analysis, for example, Jump to Reference. This means that the usual justification for Rubin's MI rules, which we reviewed in section 1.8, no longer applies.

This, of course, provokes the following question — what are the properties of Rubin's estimator when using reference based imputation for *time-to-event data*? The current and next chapters specifically address this question.

In the context of longitudinal data, Carpenter *et al.* (2014) sketch that, because distributional information is borrowed under reference based methods, the standard likelihood calculation results in an artificial gain in statistical information about the treatment effect, relative to what we would expect to see if the missing data were able to be observed under the reference based assumption.

By constrast, they propose, and Cro *et al.* (2018) prove, that sensitivity analyses for continuous longitudinal data using Rubin's rules are — to a good approximation — *information anchored*. Referring back to the definition in section 1.6, this means that reference based imputation using Rubin's rules approximately preserves the fraction of information lost due to missing data across each of the assumptions. Whichever assumption is chosen for the primary analysis (typically missing or censoring at random), the information about the treatment effect lost due to missing data is constant across the primary and sensitivity analyses. This property underpins our confidence in using such Class-2 methods for sensitivity analysis.

In this chapter, we begin by presenting results from a simulation study investigating if information anchoring holds for "Jump to Reference" when applied to time-to-event data. We then show that these principles also apply for a real data using the RITA-2 clinical trial as illustrative example. In Chapter 4 we go on to derive analytic results that support this statement for certain specific time-to-event settings.

## 3.2 Simulation study

We simulate time-to-event data from a two arm trial, with active and reference (i.e. control) arms. Without loss of generality, we only censor patients in the active arm; all event times in the reference arm are observed.

We used the Cox Proportional Hazards model as imputation and analysis model in the last chapter. Imputing the missing events under this model involves drawing proper imputations from the baseline hazard, $h_0(t)$, which entails additional computational complications (Jackson *et al.*, 2014). Instead, this time we use the Weibull proportional hazards model as imputation and analysis model. This is sufficiently flexible for many applications; in other settings, an alternative would be to use flexible splines as the parametric model for the baseline hazard, again with proportional hazards (e.g. Lambert and Royston, 2009; Royston and Parmar, 2011, 2013).

The MI procedure is essentially that set out in section 2.3, but with a some slight changes. In step 1(a) we fit a Weibull model to the observed data, and in step 1(b), as before, we draw $u_i \sim U[0,1]$,, but this time rather than estimating the baseline hazard, since we have a parametric function we solve

$$S(t_i|t_i > c_i, x_i, \tilde{\beta}) = \frac{S(t_i; x_i, \tilde{\beta})}{S(c_i; x_i, \tilde{\beta})} = u_i;$$

which has a simple closed form solution. The rest of the procedure is as defined previously.

We simulated event times from an exponential distribution, with control arm hazard $h(t) = 0.01$, and hazard ratio $\beta$, again using the approach described by Bender *et al.* (2005). Data in the active arm were censored at random, and then imputed assuming (i) censoring at random and (ii) Jump to Reference. We varied the active arm censoring levels from 0% to 80%, and explored three different sample sizes: $n = 125, 250$ and $n = 500$ in each arm. For all the results presented below we used $K = 50$ imputations and $1000$ replications.

To each simulated dataset, we fitted the Weibull proportional hazards model,

$$\hat{h}_i(t) = \kappa t^{\kappa-1} \exp(\hat{\alpha} + \hat{\beta} x_i), \tag{3.2.1}$$

where $\kappa$ is the usual scale parameter of the distribution, and $\hat{\alpha}$ and $\hat{\beta}$ are the estimates from fitting the model. We focus on the treatment estimate $\hat{\beta}$.

For the first scenario, the hazard ratio used to generate the data is $\beta = 0.8$ (log hazard ratio $-0.22314$) with 250 patients in each arm, giving a power of $0.7$ when there is no censoring. Table 3.2.1 summarises the results.

Specifically, the second row of Table 3.2.1 shows the results when there is no censoring. The mean of the estimates of $\beta$ across the $S = 1000$ replications,

$$\widehat{\mathrm{E}}[\hat{\beta}] = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_s, \tag{3.2.2}$$

is $-0.22695$. Over the $S$ replications, the mean value of the asymptotic variance estimate, calculated as the inverse of the observed information,

$$\widehat{\mathrm{E}}[\hat{\mathrm{V}}_{inf}(\hat{\beta})] = \frac{1}{S}\sum_{s=1}^{S}\hat{\mathrm{V}}_{inf}(\hat{\beta}_s), \tag{3.2.3}$$

is $0.00797$, while, letting $\hat{\beta}_{\cdot} = \sum_{s=1}^{S}\hat{\beta}_s, /S$ be the usual empirical variance estimate,

$$\hat{\mathrm{V}}_{emp}(\hat{\beta}) = \frac{1}{(S-1)}\sum_{i=1}^{S}(\hat{\beta}_s - \hat{\beta}_{\cdot})^2, \tag{3.2.4}$$

is 0.00807. Therefore, we see that when there is no censoring, the mean of $\hat{\beta}_s$ over the $S = 1000$ replications is unbiased, and the theoretical and empirical variance estimates agree as expected.

We now explore what happens when data are censored at random in the active arm only. When this happens, we need to make an (untestable) assumption about the censored data. Here, we estimate the hazard ratio by multiple imputation under this assumption.

The top half of Table 3.2.1 shows the results when we assume data are censored at random and impute accordingly. We define three quantities from the multiple imputation estimates analogous to (3.2.2)–(3.2.4) above. These are, first the mean of the estimates across the $S$ replications,

$$\widehat{\mathrm{E}}[\hat{\beta}_{MI}] = \frac{1}{1000}\sum_{s=1}^{S}\hat{\beta}_{s,MI}, \tag{3.2.5}$$

second the mean of the "Rubin's rules" variance of these estimates,

$$\widehat{\mathrm{E}}[\hat{\mathrm{V}}_{RR}(\hat{\beta}_{MI})] = \frac{1}{S}\sum_{s=1}^{S}\hat{\mathrm{V}}_{RR}(\hat{\beta}_{s,MI}), \tag{3.2.6}$$

and third the empirical variance of the $S$ multiple imputation estimates,

$$\hat{\mathrm{V}}_{emp}(\hat{\beta}_{MI}) = \frac{1}{(S-1)}\sum_{i=1}^{S}(\hat{\beta}_{s,MI} - \hat{\beta}_{.,MI})^2, \tag{3.2.7}$$

where $\hat{\beta}_{.,MI} = \sum_{s=1}^{1000}\hat{\beta}_{s,MI}/S$.

To assess the information anchoring properties, in columns 3 and 5 of Table 3.2.1 the censored data is re-created (put back) under the current assumption before the quantities are calculated. In the top half of the table, we assume censoring at random. If they are re-created under this assumption, then we get a full dataset from the exponential data generating model. Therefore, in the top half of Table 3.2.1 the values in columns 3 and 4 only differ from each other by Monte-Carlo variation as the proportion of censoring increases. Likewise, columns 5 and 6 only differ by Monte-Carlo variation.

In column 7, we see — again as expected — that Rubin's rules variance of the imputation estimate increases as the proportion of censoring increases, and this agrees well with the empirical variance of the MI estimator.

Now consider the bottom half of Table 3.2.1. Here, when the data are censored, we assume "Jump to Reference". As above, in columns 3 and 5, we re-create (put back) the data under this assumption. Column 3 shows that the mean treatment effect attenuates as the proportion of censoring increases, and comparing with column 2 we see there is no systematic bias. Columns 5 and 6 show that when censored, data is recreated under the current assumption, the information-based and empirical variance estimates are similar, as expected, and do not vary markedly as the proportion of censoring increases.

Now consider column 8. This shows the empirical variance of the MI estimates. Because imputation under Jump to Reference borrows information from the reference arm, the empirical variance *declines* as the proportion of censoring increases. Furthermore, it is *less than* the variance we would see if the assumption held true and we saw the data (column 6). We therefore argue that the empirical variance in column 8 (and theoretical approximations to it) is not appro-

priate: using it would imply that by censoring 80% of the active arm, we *double* the statistical information about the treatment effect.

Instead, we advocate using Rubin's rules variance (column 7). We see that this increases as the proportion of censored data increases, reflecting the loss of information about the treatment effect.

To explore this further, as Figure 3.2.1 shows, the proportionate increase in variance (column 7 divided by column 5) under Censoring at Random using Rubin's rules approximates that under Jump to Reference, and this approximation is particularly good for lower proportions of censoring. As discussed above, this is what we call *information anchoring*. In other words, the proportion of information lost due to missing data is similar to that under the primary analysis assumption (CAR) and the sensitivity analysis assumption (J2R), at least up to a censoring level of 60% on the active arm.

These results are in line with the theory for continuous data (Cro *et al.*, 2018), which shows that the approximation of Rubin's rules to information anchoring improves as the treatment effect decreases. To explore this further, we now consider additional scenarios. Figure 3.2.2 shows results for a hazard ratio of 0.5 and 0.8, for sample sizes of 500 and 1000 patients in each arm.

In each panel, the horizontal line $-\times-$ is the variance of the log-hazard ratio when the censored data are recreated under Jump-to-Reference. That is to say they are derived in the same way as column 5 in Table 3.2.1. The $-\diamond-$ lines show the empirical variance of the multiple imputation estimator under Jump-to-Reference, and are derived in the same way as column 8 in Table 3.2.1. The $-\circ-$ line denotes the Rubin's rules variance of the multiple imputation estimator under Jump-to-Reference (cf column 7 in Table 3.2.1), with $-+-$ showing the information anchored variance (i.e. that calculated by re-arranging the expression in equation 1.6.3.

Consistent with Table 3.2.1, column 8, we see that under Jump-to-Reference the empirical variance of the MI estimator drops below that we would obtain if we actually observed data under this assumption. However, the Rubin's rules variance under CAR and Jump-to-Reference are very similar, especially as the hazard ratio approaches 1 (top panels of Figure 3.2.2), and for smaller proportions of censoring — both more likely in trials.

Thus, for reference based imputation of the type described here, the study suggests that, at least for simulated data, Rubin's rules provide unbiased estimates, and are approximately information anchored; that is the loss of information due to missing data is approximately constant across

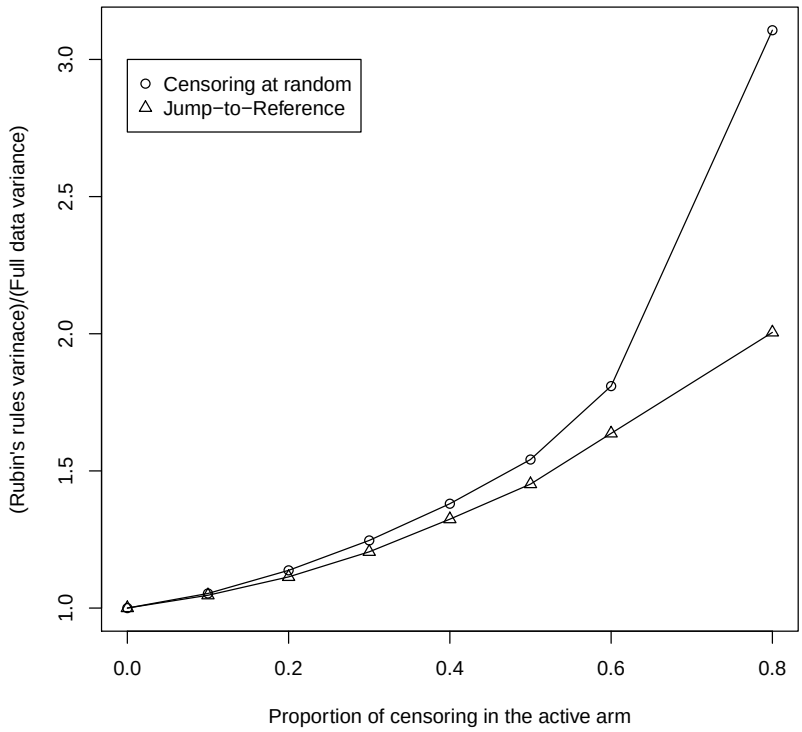


Figure 3.2.1: Proportionate increase in variance as censoring increases under (a) censoring at random and (b) Jump to Reference

the primary assumption about censoring and the sensitivity assumptions.

In the next section, we consider an application of Jump to Reference to the RITA-2 clinical trial in cardiovascular disease which was introduced in section 1.11.2.

Table 3.2.1: Simulation results: exponential data generating process, 250 patients in each arm, censoring in the active arm only; Weibull analysis and imputation model, $S = 1000$ replications. Explanations of column headings in the text.

| Column: 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Censoring % (active arm) | True $\beta$ | $\widehat{\mathrm{E}}[\hat{\beta}]$ (censored data re-created under current assumption) | $\widehat{\mathrm{E}}[\hat{\beta}_{MI}]$ | $\widehat{\mathrm{E}}[\hat{\mathrm{V}}_{inf}(\hat{\beta})]$ (censored data re-created under current assumption) | $\hat{\mathrm{V}}_{emp}(\hat{\beta})$ (censored data re-created under current assumption) | $\widehat{\mathrm{E}}[\hat{\mathrm{V}}_{RR}(\hat{\beta}_{MI})]$ | $\hat{\mathrm{V}}_{emp}(\hat{\beta}_{MI})$ |
| No censoring | −0.22314 | −0.22695 | | 0.00797 | 0.00807 | | |
| Analysis assuming Censoring At Random | | | | | | | |
| 10% | −0.22314 | −0.22679 | −0.22821 | 0.00797 | 0.00813 | 0.00850 | 0.00844 |
| 20% | −0.22314 | −0.22692 | −0.22933 | 0.00797 | 0.00801 | 0.00918 | 0.00912 |
| 30% | −0.22314 | −0.22690 | −0.23009 | 0.00796 | 0.00820 | 0.01006 | 0.00985 |
| 40% | −0.22314 | −0.22620 | −0.23086 | 0.00797 | 0.00784 | 0.01114 | 0.01093 |
| 50% | −0.22314 | −0.22726 | −0.23146 | 0.00797 | 0.00838 | 0.01244 | 0.01227 |
| 60% | −0.22314 | −0.22497 | −0.22866 | 0.00798 | 0.00798 | 0.01460 | 0.01456 |
| 80% | −0.22314 | −0.22627 | −0.23433 | 0.00798 | 0.00808 | 0.02507 | 0.02483 |
| Analysis assuming Jump-to-Reference | | | | | | | |
| 10% | −0.22608 | −0.20751 | −0.20833 | 0.00793 | 0.00784 | 0.00830 | 0.00703 |
| 20% | −0.18232 | −0.18727 | −0.18941 | 0.00792 | 0.00793 | 0.00882 | 0.00621 |
| 30% | −0.16127 | −0.16615 | −0.16807 | 0.00790 | 0.00796 | 0.00952 | 0.00536 |
| 40% | −0.13976 | −0.14452 | −0.14639 | 0.00790 | 0.00801 | 0.01046 | 0.00468 |
| 50% | −0.11778 | −0.12274 | −0.12559 | 0.00790 | 0.00819 | 0.01147 | 0.00424 |
| 60% | −0.09531 | −0.09508 | −0.09972 | 0.00793 | 0.00827 | 0.01298 | 0.00382 |
| 80% | −0.04879 | −0.04956 | −0.05521 | 0.00803 | 0.00817 | 0.01610 | 0.00350 |

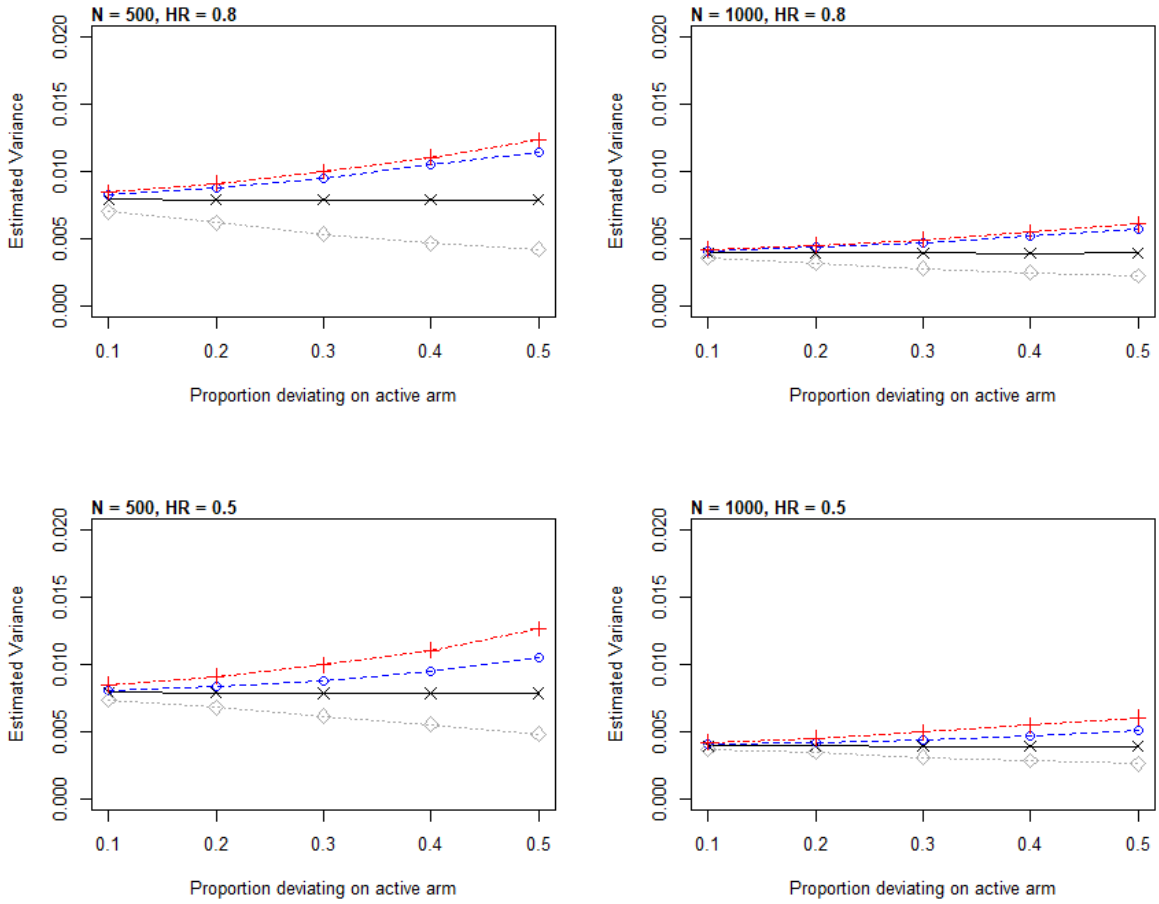


Figure 3.2.2: Simulation results: exploration of information anchoring for two sample sizes and two hazard ratios. For each scenario, as the proportion of active arm censoring increases, each panel shows the evolution of the variance of the estimated hazard ratio calculated in four ways: (i) $-+-$ information anchored variance; (ii) $-\circ-$ Rubin's MI variance under Jump to Reference; (iii) $-\times-$ $\widehat{\mathrm{E}}[\hat{\mathrm{V}}_{inf}(\hat{\beta})]$ when censored data are re-created under Jump to Reference; (iv) $-\diamond-$ $\hat{\mathrm{V}}_{emp}(\hat{\beta}_{MI})$ under Jump to Reference.

## 3.3 Reference based sensitivity analysis for the RITA-2 Study

We introduced the RITA-2 study in section 1.11.2. RITA-2 was a so-called pragmatic trial, so that although patients were initially randomised to PTCA or medical treatment, in the course of the follow-up patients received further procedures according to clinical need, and the trial was designed to compare a policy of beginning with medical treatment against a policy of beginning with PTCA. Subsequent non-random interventions (NRIs) were either PTCA, or when necessary, a coronary artery bypass graft (CABG). In the PTCA arm, 17.0% of patients had a second PTCA, while 12.7% had a CABG. By contrast, on the medical arm 27% had a non-randomised PTCA and 12.3% had a CABG. The main goal was to estimate *de-facto* (intention to treat) effects comparing the two strategies, PTCA versus medical therapy.

For the purposes of this illustration, all cause mortality is taken as the event, and we compare two analyses. The first is essentially the intention to treat (ITT) analysis of the original trial, where follow up is continued after NRIs. This may be thought of as the *de-facto* estimand.

In the second analysis, we censor medical arm patients at their first NRI, and seek to empirically demonstrate how the Jump to Reference approach can be used to emulate such a *de-facto* analysis. If our emulation of the *de-facto* analysis gives similar results to those in the original *de-facto* analysis — in this example where the data for it is actually available — this builds confidence that such emulations can be used in settings where the actual data are not observed.

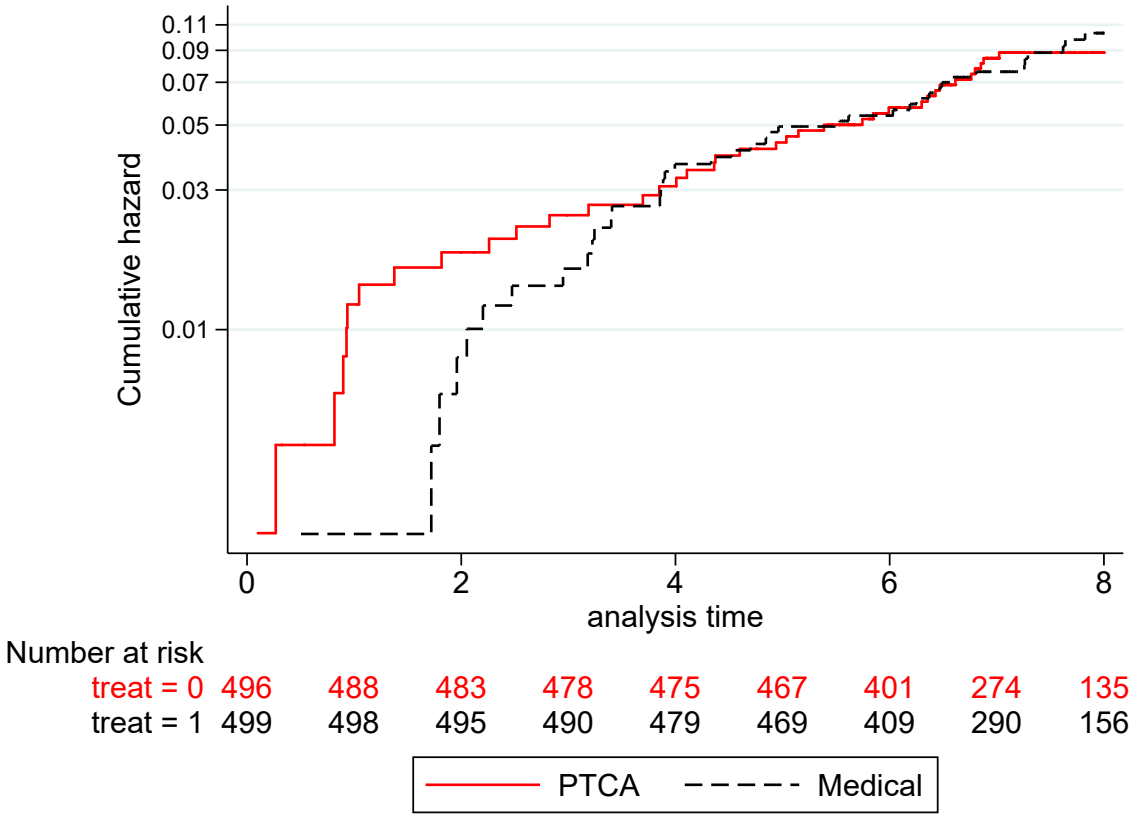


Figure 3.3.1: RITA-2 trial: Nelson-Aalen cumulative hazard survival plots for all cause mortality; patients censored at loss to follow-up.

Table 3.3.1: RITA-2 analysis: Estimated all cause mortality hazard ratios comparing PTCA with the medical intervention based on the original study data (top) and emulated "Jump to PTCA" *de-facto* scenario (bottom); hazard ratio $> 1$ indicating the risk is higher on the medical arm.

| Estimand | Hazard ratio (95% CI) | p-value |
|---|---|---|
| *De-facto* analysis of study data | 1.02 (0.67, 1.57) | 0.93 |
| Emulated *de-facto* analysis:<br><br>Medical arm patients are censored at their first non-randomised intervention and their event times are imputed under "Jump to PTCA arm". | 1.15 (0.75, 1.55) | 0.49 |

The analysis of all-cause mortality including outcome data from patients with NRIs can be regarded as one concerning a *de-facto* or "treatment policy type" of estimand (as defined on page 17 of the ICH E9 addendum (CHMP, 2018)).

The *de-facto* log-cumulative hazards for each arm are shown in Figure 3.3.1, and the treatment effect from an unadjusted Weibull proportional hazards model is shown in the top part of Table 3.3.1.

The ITT aspect of the original study is emulated using Jump to Reference. To do this we leave the PTCA arm data unchanged. For the medical arm data, we artificially censor patients at their first NRI, and then they "Jump to Reference", which in this context means "Jump to PTCA arm". The principles being the same, the reference arm is the intervention in this case. This again highlights the flexibility of our approach which allows different assumptions to be made to mimic realistic clinical outcomes: We have made the CAR assumption for all censored patients, *apart from* a subgroup of patients which were censored due to an NRI. We implement this using the multiple imputation approach described in section 3.2.

Specifically, the primary analysis model is an unadjusted Weibull model. For multiple imputation under "Jump to PTCA arm", the Weibull model is retained. In line with the recommendations from, for example, page 79 of Carpenter and Kenward (2012), we include all variables

potentially involved in the censoring process. We therefore include the following covariates: treatment, sex, age, BMI, systolic blood pressure, angina grade, and indicator variables for unstable angina, breathlessness grade, presence of a previous MI, activity level, treatment for hypertension, diabetes, smoking status, beta blockers, long acting nitrates, calcium antagonists, lipid-lowering drugs, aspirin, ace inhibitors, and number of diseased vessels. Multiply imputed event times exceeding the maximum study period of 8 years were censored administratively, in line with the assumptions used for the analysis in the original study.

The results of emulating the *de-facto* analysis by censoring medical arm patients at NRI and imputing under "Jump to PTCA arm" are shown in the bottom part of Table 3.3.1. The emulated *de-facto* results agree well with the actual *de-facto* analysis of the original study, with both p-values far from statistical significance. The solid red line in Figure 3.3.2 shows the estimated log cumulative hazard for the medical arm from fitting the Weibull model to the imputed data under "Jump to PTCA arm". As might be expected, it is initially closer to the medical arm, but as more patients on the medical arm have early NRIs, it tracks back to the PTCA arm. However, the model's proportional hazards assumption means that, in accommodating the early higher hazard in the medical arm, it under-shoots the PCTA arm between years 2–5. This is why the emulated *de-facto* hazard ratio is larger than the actual one in Table 3.3.1.

## 3.4 Summary

For longitudinal data with a continuous outcome, Cro *et al.* provided a theoretically based proof that using multiple imputation with Rubin's rules is aligned with the information anchoring principle (Cro *et al.*, 2018). With time-to-event data, the results of the simulation study presented in this chapter closely mirror those obtained in the longitudinal setting, suggesting that analogous theoretical results might hold with time-to-event data.

Rather unusually, the RITA-2 trial data allows us to compare the results of a *de-facto* analysis using the observed event times with an emulated *de-facto* analysis. The results are similar, providing empirical support for this approach in situations where, for whatever reason, data are censored but we wish to explore the robustness of our conclusions to the censoring at random assumption.

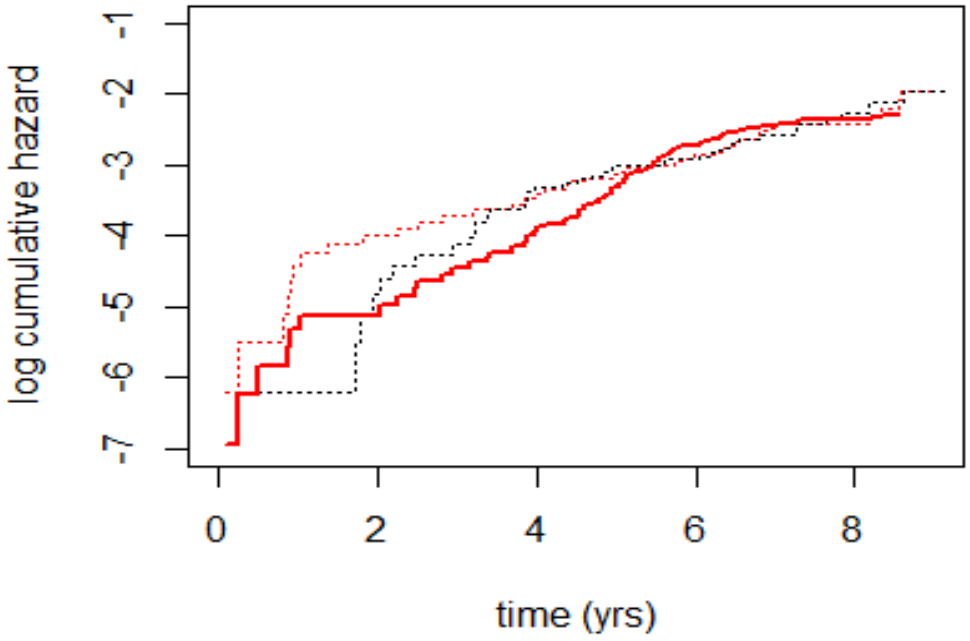


Figure 3.3.2: Plot of the log cumulative hazard against time with Nelson-Aalen estimates for the PTCA arm (upper dashed, red) and medical arm (lower dashed, black). The solid (red) line shows the estimated Weibull model log cumulative hazard for the medical arm when patients are censored at their first non-randomised intervention and "Jump to PTCA arm".

Conversely, it might be argued that this illustrative example, whilst providing evidence of the applicability of the method, might be deemed atypical, particularly of pharmacological trials. However, other authors have presented examples of similar approaches in more traditional settings (e.g. the open label, double blinded study in Lu *et al.* (2015)), and we maintain that the analysis presented here is unique in providing such a comparison of observed versus emulated *de-facto* behaviour.

A reviewer pointed out a potential improvement to the approach used for the RITA-2 study. To more adequately model the risk of PTCA, that is, post-operative improvement following successful surgery, followed by a steady increase in risk at time goes on, we could multiply impute new events for those with NRIs by jumping to the hazard of the reference from the point of randomisation (i.e. *time 0* in the trial) onwards, and "pasting" this hazard to the time at which the patient was censored. In this way, we more closely mirror the changing risk profile of patients undertaking surgery. Whilst we did not implement this for the general proof of concept of the illustrative example, we acknowledge the appropriateness of the proposed improvement, and also its excellent demonstration of the flexibility of pattern mixture models.

The presentation and results from the RITA-2 example serve as motivation for a further investigation of whether the information anchoring principle holds *generally* for time-to-event data, and this is the focus of the next chapter.

# Chapter 4

# Behaviour of Rubin's variance estimator for reference based sensitivity analysis with time-to-event data

## 4.1 Introduction

The last chapter demonstrated that, at least based on empirical data, the principle of information anchoring holds when using reference based sensitivity analysis for time-to-event data.

In this chapter, we take a slightly different tack, adopting a more analytical approach to determine whether this principle holds more *generally*. To make this tractable, specific distributional assumptions and other simplifications were made.

The PhD thesis of S. Cro, and recently published work by Cro *et al.*, provided a solid foundation and blueprint for the required methodological steps used in this chapter (Cro, 2016; Cro *et al.*, 2018).

We begin by describing our two arm clinical trial setting, including the distributional assumptions concerning the data on both arms. We make the same normality assumptions concerning the data generating process of both arms of the trial, and rely on the properties of the truncated normal distribution to take into account the censoring process. After reviewing general results

for this type of distribution, we firstly derive analytic expressions for the mean and variance when i.) there is no censoring, and ii.) when censoring is at random. We then go on to derive an expression under a censoring not at random assumption, taking the Jump to Reference approach as an example of this. Finally, our main theorem is presented in which we provide a bound on the difference between the information anchored variance under censoring at random and that under censoring not at random. We provide simulated results as validation of these analytical expressions, and again demonstrate their applicability using the RITA-2 data.

## 4.2 Clinical trial setting with time-to-event data

We consider a two arm clinical trial in which patients are randomised either to a new treatment or the control (i.e. reference) arm. The time from patient randomisation to when an event occurs, typically death or treatment failure, is the primary endpoint of the study.

Our aim is once more to extend previously derived theoretical results from the longitudinal data setting to time-to-event data. Cro *et al.* based their results on the bivariate normal distribution, and analogously, we assume that event times are bivariate log normally distributed, again consisting of two repeated measurements per patient. However, for our time-to-event setting, we define the first time point ($T_1$) to be the time of randomisation, and the second time point ($T_2$) to be the event, or censoring, time. Furthermore, we assume that, due to randomisation, the mean and variance of time $T_1$ is the same on both arms. Patients are right censored if they deviate from protocol — for example, if they stop taking the assigned treatment due to adverse effects, are lost to follow-up, or do not experience the event before the end of the study. In addition, and without loss of generality, in our setting we assume that patients are only censored on the treatment arm, so that those on the control (reference) arm always experience the outcome event of interest (i.e. patients are fully observed).

As treatment effect, we are interested in comparing the difference in mean log time-to-event at $T_2$ between the trial arms. We test if this difference is statistically significant at the 5% level against the null hypothesis of no difference using a standard t-test with pooled variance, making the *de-jure* assumption of CAR for those censored on the treatment arm. (For the rationale behind using this approach for a survival analysis, please refer to the comments at the end of this section).

For the $n_r$ patients on the control (reference) arm:

$$\begin{pmatrix} Y_{rj1} \\ Y_{rj2} \end{pmatrix} \sim N\left[ \begin{pmatrix} \mu_1 \\ \mu_{r2} \end{pmatrix}, \begin{pmatrix} \sigma_{r11} & \sigma_{r12} \\ \sigma_{r12} & \sigma_{r22} \end{pmatrix} \right], \; j = 1, \ldots, n_r,$$

where $r$ denotes the reference arm, $Y_{rji} = \ln T_{rji}$ are the $j = 1, \ldots, n_r$ normally distributed times (following the log transformation) on the reference arm, $i = 1$ is randomisation time $T_1$, and $i = 2$ denotes the event or censoring time $T_2$.

At $T_2$, $n_d$ of the $n_a$ patients on the treatment (active) arm are censored, with $n_o$ of the $n_a$ patients having an event ($n_o + n_d = n_a$). Let $O$ and $D$ define respectively, the set of indices for those patients with events (i.e. observed) and those censored (or deviating for some reason, $D$). Again, we assume a bivariate normal distribution:

$$\begin{pmatrix} Y_{aj1} \\ Y_{aj2} \end{pmatrix} \sim N\left[ \begin{pmatrix} \mu_1 \\ \mu_{a2} \end{pmatrix}, \begin{pmatrix} \sigma_{a11} & \sigma_{a12} \\ \sigma_{a12} & \sigma_{a22} \end{pmatrix} \right], \; j \in O$$

$$\begin{pmatrix} Y_{aj1} \\ Y_{aj2} \end{pmatrix} \sim N\left[ \begin{pmatrix} \mu_1 \\ \mu_{d2} \end{pmatrix}, \begin{pmatrix} \sigma_{a11} & \sigma_{d12} \\ \sigma_{d12} & \sigma_{d22} \end{pmatrix} \right], \; j \in D,$$

with $a$ denoting the active arm and $d$ denoting those deviating, with other subscripts being analogously defined as for the reference arm. Details and properties of the bivariate normal distribution are shown in Appendix B.

For the sensitivity analysis we make the *de-facto* assumption of "Jump to Reference" (J2R) for those censored on the active arm. This approach uses the hazard from the *reference* arm, in this case those taking the control treatment, to create new event times for censored patients on the treatment arm.

To briefly recap, under J2R a "new" event time $T_i^*$ for a patient $i$ on the active arm who is censored at time $T_i$ is calculated using the hazard of the reference arm such that,

$$h_{post,i}(t|t > T_i, \text{active}) \;:=\; h(t|t > T_i, \text{reference}),$$

and this is used as the basis for the multiple imputation process for the sensitivity analysis (as in section 2.4.2).

For the sensitivity analysis, without loss of generality, we continue to assume censoring is at random for those censored on the control arm. (Although the modularity of these methods allow other appropriate assumptions to be made for either, or both arms). Since we are using a reference-based sensitivity analysis method, we retain the primary analysis, the t-test, and fit this to the multiply imputed data sets created under the J2R assumption for post-censoring behaviour.

The goal is to confirm that the principle of information anchored sensitivity analyses holds for this type of approach. This means that we require the following equality to hold, at least approximately:

$$\frac{I(\hat{\theta}_{full,primary=CAR})}{I(\hat{\theta}_{obs,primary=CAR})} = \frac{I(\hat{\theta}_{full,sensitivity=J2R})}{I(\hat{\theta}_{obs,sensitivity=J2R})}, \tag{4.2.1}$$

so that the proportion of information lost due to missing data is constant across primary *and* sensitivity analyses.

In the next section, a formula for the information ratio $\frac{I(\hat{\theta}_{full,CAR})}{I(\hat{\theta}_{obs,CAR})}$ under the *de-jure* assumption of CAR is derived. This leads to a similar expression for the variance ratio under the *de-facto* assumption of J2R, $\frac{I(\hat{\theta}_{full,J2R})}{I(\hat{\theta}_{obs,J2R})}$. In a final step, we compare the two ratios in equation (4.2.1), providing an upper bound on the difference between them.

# Comments on modelling approach

In their recent publicaton Cro *et al.* demonstrated theoretical results based on a longitudinal data setting with continuous endpoint assuming bivariate normal data (Cro *et al.*, 2018). Using this as the natural starting point for our extension to the time-to-event setting, we assumed log normally distributed times, which is often, at least approximately, the case for time-to-event

data. Following log transformation to achieve normality, the most efficient estimator in such settings is the mean log time to the occurrence of the event, and accordingly, the t-test is the most appropriate choice to test the difference between the two arms at time $T_2$ (with these assumptions for the data generating process).

Notwithstanding the logic of the above argument, using the t-test to determine the mean log treatment difference for a survival analysis might be thought unconventional. The most common choice for the primary analysis model would usually be the Cox Proportional Hazards (CPH) model, with the hazard ratio over the total follow-up period defined as treatment effect for the trial. However, the CPH model inherently assumes that the hazards are proportional, even though, as previously mentioned, this is increasingly being challenged in many clinical settings (Royston and Parmar, 2011, 2013). The restricted mean survival time (RMST) is now frequently used instead of the hazard ratio as a preferable clinical endpoint. Although not always equivalent to the RMST, there are clear parallels between using the RMST and the mean log time-to-event used to calculate the endpoint in the clinical trial setting we defined in this section.

Of course, we are using the t-test since this allows us to rely, at least in part, on the results from Cro *et al.* (2018), but as the above argumentation shows, this is perhaps becoming a more common approach. The theoretically derived results in this chapter could also be attempted by considering a semi- of fully parametric modelling approach using the (partial) likelihood. However, this was outside the scope of the PhD, but could present an interesting avenue for potential further study.

## 4.3 Information anchoring under the *de-jure* assumptions

### 4.3.1 Variance estimation when data is fully observed

Let us assume that we are able to observe a realisation of the censored data, $Y_{cens}$ under the primary assumption of CAR, and we then consolidate this with the fully observed data, $Y_{obs}$, forming a full data set under the *de-jure* assumption of CAR. Fitting the primary analysis model to this data set leads to a treatment estimate $\hat{\theta}_{full,CAR}$, the subscript here making the CAR assumption explicit. We start by deriving the expression for $V(\hat{\theta}_{full,CAR})$, the variance of this estimate. Conditioning on $n_d$, the number of patients censored, the expected treatment effect at

time $T_2$ is a weighted average of the mean estimates at time $T_2$ from those censored, and those with events, compared to the reference arm mean ($\mu_{r2}$):

$$\hat{\theta}_{full,CAR} = \left(\frac{n_o}{n_a}\mu_{a2} + \frac{n_d}{n_a}\mu_{a2}\right) - \mu_{r2}$$

The variance of this estimate in the full data with no censoring is just the standard result assuming normal distribution results,

$$E[V(\hat{\theta}_{full,CAR})] = \frac{\hat{\sigma}^2_{22,r}}{n_r} + \frac{\hat{\sigma}^2_{22,a}}{n_a} =$$

$$\frac{\frac{1}{(n_r-1)}\sum_{j=1}^{n_r}(Y_{rj2} - \bar{Y}_{r2})^2}{n_r} + \frac{\frac{1}{(n_a-1)}\sum_{j=1}^{n_a}(Y_{aj2} - \bar{Y}_{a2})^2}{n_a} = \frac{2\sigma_{22}}{n}, \tag{4.3.1}$$

assuming equal numbers of patient in both arms of the study $n_r = n_a = n$, and equal variance in both arms $\sigma_{r22} = \sigma_{a22} = \sigma_{22}$.

### 4.3.2 Censoring on the active arm

Let us assume $n_d$ patients on the active arm are censored at a randomly defined, but fixed time point $\alpha$. We now borrow the notation and standard results from the (right-side) truncated normal distribution. The expected value for those values greater than $\alpha$ is:

$$E(\tilde{Y}_{a2j} \mid \tilde{Y}_{a2j} > \alpha) = \mu_{a2} + \sqrt{\sigma_{22}}\left[\frac{\phi(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}})}{1 - \Phi(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}})}\right] = \mu_{a2} + \sqrt{\sigma_{22}}\lambda, \tag{4.3.2}$$

for $j = 1, \ldots, n_d$ patients censored at $\alpha$; $\phi$ and $\Phi$ being the density and CDF of the standard normal distribution, respectively. The fraction part (in large square brackets) of this expression is known as the inverse Mills ratio (Greene, 2003), hereafter shorthanded as

$$\lambda = \left[ \frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{1-\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)} \right].$$

Note that the expression in equation 4.3.2 is just the "usual" expected value with an additional term $\sqrt{\sigma_{22}}\lambda$, treating $\lambda$ as a constant for specific values of $\mu_{a2}, \sigma_{22}$ and $\alpha$.

The standard expression for the variance of the truncated normal distribution may also be used[1]:

$$VAR(\tilde{Y}_{a2j} \mid \tilde{Y}_{a2j} > \alpha) = \sigma_{a2d} = \sigma_{22}\left[1 - \frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{1-\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\left[\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{1-\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)} - \left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)\right]\right]. \tag{4.3.3}$$

Analogously, if we consider the $n_o$ fully observed patients on the active arm we may define

$$E(\tilde{Y}_{a2j} \mid \tilde{Y}_{a2j} < \alpha) = \mu_{a2} - \sqrt{\sigma_{22}}\left[\frac{\phi(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}})}{\Phi(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}})}\right] = \mu_{a2} - \sqrt{\sigma_{22}}\dot{\lambda}. \tag{4.3.4}$$

for $j = 1, \ldots, n_o$ observed patients, with the variance defined as:

$$VAR(\tilde{Y}_{a2j} \mid \tilde{Y}_{a2j} < \alpha) = \sigma_{a2o} = \sigma_{22}\left[1 - \left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)} - \left(\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\right)^2\right]. \tag{4.3.5}$$

Without loss of generality, the truncation limit $\alpha$ is assumed to be greater than the mean throughout the analysis.

[1] In practice, the formula from Barr and Sherrill in Appendix C often provided a more accurate estimate for censoring levels around 10%.

### 4.3.3 Multiple imputation

In order to estimate the variance under the *de-jure* assumption of CAR following multiple imputation, $E[V(\hat{\theta}_{MI,CAR})]$, the first step is to define an appropriate imputation model. Multiple imputation relies on Bayesian argumentation, so the assumption is made that the posterior estimates $\hat{\beta}$ from the fitted model are normally distributed.

We assume the observed data dominates the posterior distribution, and, using inferential arguments set out on pages 56-60 of Carpenter and Kenward (2012), without any important loss of generality assume the variance is known.

To establish the properties of multiply imputed data on the active arm, a natural starting point is to fit a censored regression (Tobit) model to the observed data (Tobin, 1958; Greene, 2003), then based on estimates from this fitted model, impute new events for the censored patients, and finally derive the variance of the combined sets of observed and imputed data using Rubin's rules.

In more detail, the process is as follows:

1. Fit the Tobit regression model for the observed data on the active arm $Y_{aj2}$ on $Y_{aj1}$, including the censoring at $\alpha$:

$$Y_{aj2} = \hat{\beta}_0 + Y_{aj1}\hat{\beta}_1 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \hat{\sigma}_{2.1}), \quad i = 1, \ldots, n_o,$$

   resulting in the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ and estimate of the residual variance $\hat{\sigma}_{2.1}$.

2. Obtain a draw from the approximate Bayesian posterior distribution assuming non-informative priors by first drawing

$$\tilde{\sigma}_{2.1} = \frac{(n_o - 2)\hat{\sigma}_{2.1}}{X},$$

   where $X \sim \chi^2_{n_o-2}$. We assume the model estimates are bivariate normally distributed with mean $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ and covariate matrix:

$$\hat{\boldsymbol{V}} = \hat{\sigma}_{2.1} \begin{bmatrix} n_o & \sum_{i=1}^{n_o} \hat{Y}_{aj1} \\ \sum_{i=1}^{n_o} \hat{Y}_{aj1} & \sum_{i=1}^{n_o} \hat{Y}_{aj1}^2 \end{bmatrix}^{-1},$$

then taking a draw of MVN($\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{V}}$), resulting in a vector of estimates $(\tilde{\beta}_0, \tilde{\beta}_1)$.

3. Impute the censored observations by drawing from the resulting regression model using a set of the new estimates:

$$\tilde{Y}_{aj2} = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{Y}_{aj1} + \tilde{\epsilon}_i, \;\; \tilde{\epsilon}_i \sim \mathcal{N}(0, \tilde{\sigma}_{2.1}), \quad k = 1, \ldots, n_d,$$

4. Repeating these steps $K$ times results in $K$ complete data sets.

5. Fit the substantive model, the t-test, to each of the $k = 1, \ldots, K$ complete data sets in turn, resulting in estimates $\hat{\theta}_k$, $\hat{\sigma}_k^2$ for multiply imputed data set $k$, which we combine to form overall estimates using Rubin's rules. The MI estimate of $\theta$ is $\hat{\theta}_{MI,CAR} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k$, for $k = 1, \ldots K$. Rubin's variance estimator is defined as:

$$E[V(\hat{\theta}_{MI,CAR})] = E(\hat{W}) + \left(1 + \frac{1}{K}\right) E(\hat{B})$$

where

$$\hat{W} = \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2,$$

and

$$\hat{B} = \frac{1}{(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}_{MI}\right)^2.$$

We have now summarised the standard MI procedure usually followed for a primary analysis in which CAR was assumed.

### 4.3.4 Rubin's variance estimate under CAR

Now, to derive an estimate of Rubin's variance *analytically* we have to take a slightly different approach since there is no closed form solution to calculate the maximum likelihood estimators for the Tobit model.

The direction we take is to write down an expression for a typical multiply imputed event time, and then work from there to derive an expression for Rubin's variance estimate. To do this, we combine our knowledge of the observed data, the properties of the bivariate normal distribution (refer to Appendix B), and the standard results for the truncated normal distribution stated in section 4.3.2.

The imputation model for the $j$th of $n_d$ censored values from the $k$th of $K$ imputed data sets is defined as:

$$\tilde{Y}_{aj2,k} = \bar{\tilde{Y}}_{a2o,k} + \tilde{\beta}_k(Y_{aj1} - \bar{Y}_{a1o}) + \sqrt{\tilde{\sigma}_{22,k}}\dot{\tilde{\lambda}} + \sqrt{\tilde{\sigma}_{22,k}}\tilde{\lambda} + \tilde{\epsilon}_{j,k},\ j \in D,\ k = 1, \ldots, K \quad (4.3.6)$$

with

$$\tilde{\sigma}_{2.1,k}|Y_o, \hat{\sigma}_{2.1} \sim \frac{\hat{\sigma}_{2.1}(n_o - 2)}{\chi^2_{n_o-2}},$$

where $\hat{\sigma}_{2.1}$ is the estimate of the residual variance from the fitted Tobit model, or equivalently, using the properties of bivariate normality $\hat{\sigma}_{2.1} = \hat{\sigma}_{22} - \frac{\hat{\sigma}^2_{12}}{\hat{\sigma}_{11}}$ (refer to Appendix B for details);

$$\bar{\tilde{Y}}_{a2o,k}|Y_o, \tilde{\sigma}_{2.1,k} \sim \mathcal{N}(\bar{Y}_{a2o}, n_o^{-1}\tilde{\sigma}_{2.1,k}),$$

$$\tilde{\beta}_k|Y_o, \tilde{\sigma}_{2.1,k} \sim \mathcal{N}(r/q, q^{-1}\tilde{\sigma}_{2.1,k}),$$

where $r = \sum_{i=1}^{n_o}(Y_{aj1} - \bar{Y}_{a1o})(Y_{aj2} - \bar{Y}_{a2o})$ and $q = \sum_{i=1}^{n_o}(Y_{aj1} - \bar{Y}_{a1o})^2$; the coefficient $\hat{\beta}_k$ of the regression model is $\frac{\sigma_{12}}{\sigma_{11}}$ (again, using properties in Appendix B);

$$\sqrt{\tilde{\sigma}_{22,k}} \mid \hat{\sigma}_{22} \sim \frac{\sqrt{\hat{\sigma}_{22}}(n_a - 2)}{\chi_{n_a-2}},$$

where $\hat{\sigma}_{22}$ is the sampling variance at time $T_2$ on the active arm. We note at this point that we use the $\chi^2$-distribution here as an estimate of the sampling variance of the standard deviation $\sqrt{\hat{\sigma}_{22}}$; and finally,

$$\tilde{\epsilon}_{j,k}|\ \bar{\tilde{Y}}_{a2d,k},\ \tilde{\sigma}_{22,k} \sim Tr\mathcal{N}(0,\ \tilde{\sigma}_{a2d,k}\ ,\ a = (\alpha - \bar{\tilde{Y}}_{a2d,k})),$$

where the right-hand side of the expression denotes the truncated normal distribution with mean 0 and variance $\tilde{\sigma}_{a2d,k}$, truncated on the left-hand side at $a = (\alpha - \bar{\tilde{Y}}_{a2d,k})$; we use this re-location so that the mean of this expression is centred at zero, with the variance as we require, *and* we ensure that multiply imputed events are greater than the original censoring time for patient $j$.

To simplify the expression in equation (4.3.6), we separate the different sources of variability within the imputed data sets using new parameters $u_k$, $b_k$ and $w_k$. We re-write equation (4.3.6) as:

$$\tilde{Y}_{aj2,k} = \bar{Y}_{a2o} + u_k + (r/q + b_k)(Y_{aj1} - \bar{Y}_{a1o}) + \sqrt{\tilde{\sigma}_{22,k}}\dot{\tilde{\lambda}} + w_{k,\dot{\lambda}} + \sqrt{\hat{\sigma}_{22}}\tilde{\lambda} + w_{k,\lambda} + \epsilon_{j,k}, \quad (4.3.7)$$

for $j \in D,\ k = 1, \ldots, K$, with

$$r = \sum_{i \in O}(Y_{ai1} - \bar{Y}_{a1o})(Y_{ai2} - \bar{Y}_{a2o})$$

$$q = \sum_{i \in O}(Y_{ai1} - \bar{Y}_{a1o})^2$$

$$u_k | Y_o, \tilde{\sigma}_{2.1,k} \sim \mathcal{N}(0, n_o^{-1}\tilde{\sigma}_{2.1,k})$$

$$b_k|Y_o, \tilde{\sigma}_{2.1,k} \sim \mathcal{N}(0, q^{-1}\tilde{\sigma}_{2.1,k})$$

$$\tilde{\epsilon}_{j,k}|\ \bar{\tilde{Y}}_{a2d,k},\ \tilde{\sigma}_{22,k} \sim Tr\mathcal{N}(0,\ \tilde{\sigma}_{a2d,k}\ ,\ a = (\alpha - \bar{\tilde{Y}}_{a2d,k})),$$

$$w_{k,\dot{\lambda}}|Y_o, \tilde{\sigma}_{22,k} \sim N(0, VAR(w_{k,\dot{\lambda}})),$$

$$w_{k,\lambda}|Y_o, \tilde{\sigma}_{22,k} \sim N(0, VAR(w_{k,\lambda})),$$

with $VAR(w_{k,\lambda})$ for $N = n_a - 1$ defined by:

$$VAR(w_{k,\lambda}) = \lambda^2 \left(N - \left(\sqrt{2}\,\frac{\Gamma((N+1)/2)}{\Gamma(N/2)}\right)^2\right)\left(\sqrt{\frac{\sigma_{22}}{N}}\right)^2,$$

with mean of the $\chi^2$-distribution defined as $\sqrt{2}\,\frac{\Gamma((N+1)/2)}{\Gamma(N/2)}\left(\sqrt{\frac{\sigma_{22}}{N}}\right)$, for $N = n_a - 1$.[2] We define $VAR(w_{k,\dot{\lambda}})$ analogously substituting $\dot{\lambda}$ for $\lambda$ in the above formula.

We recall that the primary endpoint is the difference in means between the reference and active arms at time $T_2$. Following imputation, the estimate is defined as:

$$E(\hat{\theta}_{MI,CAR}) = E(\hat{\mu}_{a2,MI} - \hat{\mu}_{r2}),$$

(where $\hat{\mu}_{a2,MI} = \frac{1}{K}\sum_{i=1}^{k}\hat{\mu}_{a2,k}$), where $\mu_{r2}$ remains the same since we only have missing data on the active arm.

For imputed data set $k$ we have a weighted average of the observed and deviating data:

[2]The variance of the standard deviation of $\sigma_{22}$ is a $\chi^2$-distributed variable (refer to, for example, page 171 of Kenney and Keeping (1951)).

$$\hat{\theta}_k = \hat{\mu}_{a2,k} - \hat{\mu}_{r2} = \frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}\bar{Y}_{a2,k} - \bar{Y}_{r2} \tag{4.3.8}$$

with mean value for the deviators from the $k$th imputation defined by the expected value for a typical deviating patient. We average the $n_d$ patients of the form in equation (4.3.7):

$$\bar{Y}_{a2,k} = \frac{1}{n_d}\sum_{j\in D}\tilde{Y}_{aj2,k} = \bar{Y}_{a2o} + u_k + (r/q + b_k)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) + \sqrt{\bar{\tilde{\sigma}}_{22,k}}\bar{\dot{\lambda}} + w_{k,\dot{\lambda}} + \sqrt{\bar{\tilde{\sigma}}_{22,k}}\bar{\lambda} +$$

$$w_{k,\lambda} + \bar{\epsilon}_k, \quad j \in D,\ k = 1, \ldots, K \tag{4.3.9}$$

where the average deviating patient response $\bar{Y}_{a1d} = \frac{1}{n_d}\sum_{j\in D}\tilde{Y}_{aj1}$, and the error terms $\bar{\epsilon}_k = \frac{1}{n_d}\sum_{j\in D}\tilde{\epsilon}_{j,k}$. The terms in $\lambda$ and $\tilde{\sigma}_{22,k}$ are constant over the $n_d$ terms for the $k$th imputation, the bar superscript has been added for consistency of reading only.

Using the above expression, we now average over all $K$ imputed data sets:

$$\hat{\theta}_{MI} = \frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}\Big(\bar{Y}_{a2o} + \bar{u} + (r/q + \bar{b})(\bar{Y}_{a1d} - \bar{Y}_{a1o}) + \sqrt{\hat{\sigma}_{22}}\bar{\dot{\lambda}} +$$

$$\bar{w}_{k,\dot{\lambda}} + \sqrt{\hat{\sigma}_{22}}\bar{\lambda} + \bar{w}_{k,\lambda} + \bar{\epsilon}\Big) - \bar{Y}_{r2} \tag{4.3.10}$$

Taking expectations, noting $E(u) = 0$, $E(b) = 0$, $E(w_{k,\cdot}) = 0$ and $E(\bar{\epsilon}) = 0$, and replacing $r/q$ by $\beta$,

$$E(\hat{\theta}_{MI}) = \frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}\Big(\bar{Y}_{a2o} + \beta(\bar{Y}_{a1d} - \bar{Y}_{a1o}) + \sqrt{\hat{\sigma}_{22}}\dot{\lambda} + \sqrt{\hat{\sigma}_{22}}\lambda\Big) - \bar{Y}_{r2}.$$

At baseline (assuming randomisation), $\bar{Y}_{a1d} = \bar{Y}_{a1o}$, so the term in $\beta$ disappears. Using the definition of $E(Y|Y < \alpha)$ presented in equation (4.3.6), we substitute back in the population values to find:

$$E(\hat{\theta}_{MI,CAR}) = \mu_{a2} - \frac{n_o}{n_d}\sqrt{\sigma_{22}}\dot{\lambda} + \frac{n_d}{n_a}\sqrt{\sigma_{22}}\lambda - \mu_{r2} = \mu_{a2} - \mu_{r2} + \sqrt{\sigma_{22}}\left(\frac{n_d}{n_a}\lambda - \frac{n_o}{n_a}\dot{\lambda}\right) \quad (4.3.11)$$

which is the unbiased expression we might expect to obtain.

This has established our first result for the expectation of Rubin's MI estimator under CAR. We now focus on the variance of Rubin's MI estimator under CAR using equations (4.3.7) and (4.3.9) as building blocks.

As in equation (4.3.1) in section 4.3.1, we need to calculate the following expression,

$$E[V(\hat{\theta}_{MI,CAR})] = \frac{\hat{\sigma}^2_{22,r}}{n_r} + \frac{\hat{\sigma}^2_{22,a}}{n_a} =$$

$$\frac{\frac{1}{(n_r-1)}\sum_{j=1}^{n_r}(Y_{rj2} - \bar{Y}_{r2})^2}{n_r} + \frac{\frac{1}{(n_a-1)}\sum_{j=1}^{n_a}(Y_{aj2} - \bar{Y}_{a2})^2}{n_a},$$

but this time data are censored on the active arm, and the analysis has to take into account multiply imputed data. The first part of this expression for $n_r = n_a$ we calculate directly, since there is no missingness on the reference arm,

$$\frac{E[\hat{\sigma}^2_r]}{n} = \frac{\sigma_{22}}{n}. \quad (4.3.12)$$

For the second part of the expression pertaining to the active arm, we decompose the summation into observed and censored parts, substituting our new expressions for $\bar{Y}_{a2,k}$ and $\tilde{Y}_{aj2,k}$,

$$(n_a - 1)\hat{\sigma}^2_a = E\left(\sum_{j\in o}(Y_{aj2} - \hat{\mu}_{a2,k})^2\right) + E\left(\sum_{j\in d}(\hat{Y}_{aj2,k} - \hat{\mu}_{a2,k})^2\right) \quad (4.3.13)$$

Now, to calculate the above expression we need to consider both components of Rubin's variance estimator:

$$E(\hat{V}_{MI}) = E(\hat{W}) + \left(1 + \frac{1}{K}\right) E(\hat{B}) \tag{4.3.14}$$

where

$$\hat{W} = \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2,$$

with the within imputation variance estimator for imputation $k$ defined as $\hat{\sigma}_k^2$, and,

$$\hat{B} = \frac{1}{(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}_{MI}\right)^2,$$

which is the between imputation variance.

Now, referring back to the summation in equation (4.3.13), we need to evaluate the expression for the observed and censored parts. For the observed cases, substituting in equation (4.3.9), we obtain:

$$E\left[\sum_{j \in o} (Y_{aj2} - \hat{\mu}_{a2,k})^2\right] =$$

$$E\left[\sum_{j \in o} \Big( (Y_{aj2} - \bar{Y}_{a2o}) - \frac{n_d}{n_a} u_k - \frac{n_d}{n_a} \left(\frac{r}{q} + b_k\right) (\bar{Y}_{a1d} - \bar{Y}_{a1o}) - \frac{n_d}{n_a} \bar{\dot{\lambda}} \sqrt{\bar{\dot{\sigma}}_{22,k}} \right.$$

$$\left. - \frac{n_d}{n_a} w_{k,\dot{\lambda}} - \frac{n_d}{n_a} \bar{\lambda} \sqrt{\bar{\tilde{\sigma}}_{22,k}} - \frac{n_d}{n_a} w_{k,\lambda} - \frac{n_d}{n_a} \bar{\epsilon}_k \Big)^2 \right].$$

For the patients deviating, we use both equations (4.3.7) and (4.3.9), and again write out the full summation so that we can identify the terms:

$$E\left[\sum_{j \in d} (\hat{Y}_{aj2,k} - \hat{\mu}_{a2,k})^2\right] =$$

$$E\left[\sum_{j\in d}\Big(\bar{Y}_{a2o}+u_k+\left(\frac{r}{q}+b_k\right)(Y_{aj1}-\bar{Y}_{a1o})+\sqrt{\tilde{\sigma}_{22,k}}\dot{\tilde{\lambda}}+w_{k,\dot{\lambda}}+\sqrt{\hat{\sigma}_{22}}\tilde{\lambda}+w_{k,\lambda}+\epsilon_{j,k})-\right.$$

$$\left.\frac{n_o}{n_a}\bar{Y}_{a2o}-\frac{n_d}{n_a}\bar{Y}_{a2,k}\Big)^2\right]=$$

$$E\left[\sum_{j\in d}\Big(\bar{Y}_{a2o}+u_k+\left(\frac{r}{q}+b_k\right)(Y_{aj1}-\bar{Y}_{a1o})+\sqrt{\tilde{\sigma}_{22,k}}\dot{\tilde{\lambda}}+w_{k,\dot{\lambda}}+\sqrt{\hat{\sigma}_{22}}\tilde{\lambda}+w_{k,\lambda}+\epsilon_{j,k})-\frac{n_o}{n_a}\bar{Y}_{a2o}-\right.$$

$$\left.\frac{n_d}{n_a}(\bar{Y}_{a2o}+u_k+(r/q+b_k)(\bar{Y}_{a1d}-\bar{Y}_{a1o})+\sqrt{\bar{\tilde{\sigma}}_{22,k}}.\dot{\bar{\lambda}}+w_{k,\dot{\lambda}}+\sqrt{\bar{\tilde{\sigma}}_{22,k}}.\bar{\lambda}+w_{k,\lambda}+\bar{\epsilon}_k)\Big)^2\right]=$$

$$E\left[\sum_{j\in d}\Big((\hat{Y}_{aj2,k}-\bar{Y}_{a2d,k})+\right.$$

$$\left.\frac{n_o}{n_a}\left(u_k+\left(\frac{r}{q}+b_k\right)(\bar{Y}_{a1d}-\bar{Y}_{a1o})+\sqrt{\hat{\sigma}_{22}}\dot{\bar{\lambda}}+\bar{w}_{k,\dot{\lambda}}+\sqrt{\hat{\sigma}_{22}}\bar{\lambda}+\bar{w}_{k,\lambda}+\bar{\epsilon}_k\right)\Big)^2\right].$$

In the final derivation above, we have re-formulated in terms of the known result for $VAR(Y|Y > \alpha)$, and then added $\frac{n_o}{n_a}\bar{Y}_{a2d,k}$ to complete the square (i.e. so that we are still subtracting the full value of $\hat{\mu}_{a2,k}$ from the original summation).

For both observed and deviating parts of equation (4.3.13), it remains only to calculate these squared expressions term by term to derive $E(\hat{W})$. The workings are presented in Appendix D.

This results in an estimate for the within imputation variance component of the expression for Rubin's variance estimate in equation (4.3.14),

$$E(\hat{W}) = \frac{1}{K}\sum_{k_1}^{K} E[\hat{\sigma}_k^2] = (1-\pi_d)\sigma_{a2o} + \left(\pi_d + \frac{(1-\pi_d)}{n}\right)\sigma_{a2d} + (1-\pi_d)\pi_d\sigma_{22}\left(\dot{\lambda}+\lambda\right)^2 +$$

$$(1-\pi_d)\pi_d\left(VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right) + \frac{1}{n}\left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{n_o}\right] + \frac{\pi_d}{n}\sigma_{2.1}, \tag{4.3.15}$$

where we simplify $\pi_d = n_d/n_a$, $n_o/n_a = (1-\pi_d)$, and assume $n_a = n_r = n$.

We now move onto the second component of Rubin's variance, $E(\hat{B})$, the between imputation variance:

$$E(\hat{B}) = E\left[\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_{MI})^2\right],$$

where $\hat{\theta}_k$ is the weighted average in equation (4.3.8), which uses the expression for $\bar{Y}_{a2,k}$ in equation (4.3.9):

$$\hat{\theta}_k = \frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}(\bar{Y}_{a2o} + u_k + \left(\frac{r}{q} + b_k\right)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) +$$

$$\dot{\lambda}\sqrt{\bar{\bar{\sigma}}_{22,k}} + w_{k,\dot{\lambda}} + \lambda\sqrt{\bar{\bar{\sigma}}_{22,k}} + w_{k,\lambda} + \bar{\epsilon}_k) - \bar{Y}_{r2},$$

and we use equation (4.3.10) as $\hat{\theta}_{MI}$, the expected value Rubin's MI estimate under CAR:

$$\hat{\theta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k =$$

$$\frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}\Big(\bar{Y}_{a2o} + \bar{u} + (r/q + \bar{b})(\bar{Y}_{a1d} - \bar{Y}_{a1o}) + \sqrt{\hat{\sigma}_{22}}\bar{\dot{\lambda}} +$$

$$\bar{w}_{k,\dot{\lambda}} + \sqrt{\hat{\sigma}_{22}}\bar{\lambda} + \bar{w}_{k,\lambda} + \bar{\epsilon}\Big) - \bar{Y}_{r2}$$

Written out in full using these expressions:

$$E(\hat{B}) = E\left[\sum_{k=1}^{K}\left(\Big(\frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}(\bar{Y}_{a2o} + u_k + \left(\frac{r}{q} + b_k\right)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) +\right.\right.$$

$$\dot{\lambda}\sqrt{\bar{\tilde{\sigma}}_{22,k}} + w_{k,\dot{\lambda}} + \lambda\sqrt{\bar{\tilde{\sigma}}_{22,k}} + w_{k,\lambda} + \bar{\epsilon}_k) - \bar{Y}_{r2}\Big) -$$

$$\Big(\frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}(\bar{Y}_{a2o} + \bar{u} + \left(\frac{r}{q} + \bar{b}\right)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) +$$

$$\left.\left.\bar{\dot{\lambda}}\sqrt{\bar{\tilde{\sigma}}_{22,k}} + \bar{w}_{\dot{\lambda}} + \bar{\lambda}\sqrt{\bar{\tilde{\sigma}}_{22,k}} + \bar{w}_{\lambda} + \bar{\epsilon}) - \bar{Y}_{r2}\Big)\right)^2\right]$$

Again, this expression is evaluated term by term (refer to Appendix D), and simplified to obtain:

$$E[\hat{B}] = \pi_d^2\left(\frac{\sigma_{2.1}}{n_o} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda})\right), \tag{4.3.16}$$

which is an asymptotic expression assuming K $\to \infty$.

Using equations (4.3.15) and the above equation (4.3.16), we have both components for calculating $\hat{V}_{MI}$,

$$E(\hat{V}_{MI}) = E(\hat{W}) + \left(1 + \frac{1}{K}\right)E(\hat{B}) \approx$$

$$(1-\pi_d)\sigma_{a2o} + \left(\pi_d + \frac{(1-\pi_d)}{n}\right)\sigma_{a2d} + \pi_d(1-\pi_d)\sigma_{22}\left(\dot{\lambda} + \lambda\right)^2 +$$

$$(1-\pi_d)\pi_d\left(VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right) + \frac{1}{n}\left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{n_o}\right] + \frac{\pi_d}{n}\sigma_{2.1} +$$

$$\pi_d^2 \left( \frac{\sigma_{2.1}}{n_o} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) \right), \tag{4.3.17}$$

where the approximation in the first line is due to using the asymptotic result as K $\rightarrow \infty$ for $E(\hat{B})$. Further, we again let $n_a = n_r = n$, $\pi_d = \frac{n_d}{n_a}$, and $VAR(w_{k,\cdot})$ is the variance of the Mills ratio term over the $K$ imputed data sets with the appropriate form for both $\lambda$ and $\dot{\lambda}$.

Now, taken together, the first three terms in $\sigma_{a2o}, \sigma_{a2d}$ and $\sigma_{22}$ in the second line of equation (4.3.17) are approximately equal to the expected estimated variance from the patients on the active arm had there been no deviation. That is,

$$\frac{\sigma_{22}}{n} \approx \frac{1}{n} \left[ (1-\pi_d)\sigma_{a2o} + \sigma_{a2d} \left( \pi_d + \frac{(1-\pi_d)}{n} \right) + \pi_d(1-\pi_d)\sigma_{22} \left( \dot{\lambda} + \lambda \right)^2 \right] \tag{4.3.18}$$

This approximation is verified in the final section of Appendix D.

Using the above simplification, letting $(n_a - 1) \approx n_a$ and $K \rightarrow \infty$ we simplify to obtain,

$$E(\hat{V}_{MI}) \approx \frac{\sigma_{22}}{n} + \pi_d(1-\pi_d)\left( VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda}) \right) + \frac{1}{n}\left[ \frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{n_o} \right] + \frac{\pi_d}{n}\sigma_{2.1} + \pi_d^2 \left( \frac{\sigma_{2.1}}{n_o} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) \right). \tag{4.3.19}$$

To arrive at the pooled variance of the treatment difference under CAR following MI, we just add the expression above to the variance for the reference arm, $\frac{E[\hat{\sigma}_r^2]}{n} = \frac{\sigma_{22}}{n}$, and obtain,

$$E[V(\hat{\theta}_{MI,CAR})] = \frac{2\sigma_{22}}{n} + \pi_d(1-\pi_d)\left( VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda}) \right) + \frac{1}{n}\left[ \frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{n_o} \right] + \frac{\pi_d}{n}\sigma_{2.1} + \pi_d^2 \left( \frac{\sigma_{2.1}}{n_o} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) \right). \tag{4.3.20}$$

### 4.3.5 Information ratio under CAR

We now have the building blocks for the first result, concerning the information ratio $\frac{I(\hat{\theta}_{full,primary})}{I(\hat{\theta}_{obs,primary})}$ which, under the primary assumption of CAR is rewritten as $\frac{I(\hat{\theta}_{full,CAR})}{I(\hat{\theta}_{obs,CAR})}$. In equation (4.3.1) of section 4.3.1 we defined an estimate of $I(\hat{\theta}_{full,CAR})$ for the hypothetical case in which we fully observed the data under CAR. In the previous section, the expression for $E[V(\hat{\theta}_{MI,CAR})]$ was obtained, which is an estimate of $1/I(\hat{\theta}_{obs,CAR})$. Therefore, the required ratio may be estimated by calculating $\frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]}$.

**Lemma 1**: *The ratio of the information in the full data relative to that in the incomplete data assuming CAR following multiple imputation, using the asymptotic expressions for Rubin's variance estimator as $K$ tends to infinity, is bounded above by*

$$\frac{I(\hat{\theta}_{full,CAR})}{I(\hat{\theta}_{MI,CAR})} = \frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]} \lesssim 1 + \frac{\rho^2}{2} + (1-\rho^2)\left[\frac{1}{n_o} + \pi_d + \pi_d^2 + \pi_d^3 + \pi_d^4 + \ldots\right] + \frac{\pi_d}{2} + \pi_d C\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right], \tag{4.3.21}$$

assuming $n = n_a = n_r$, $\pi_d = \frac{n_d}{n}$, and $\rho^2 = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}$, which is the correlation between times $T_1$ and $T_2$ squared, with $C$ being the variance of $\sqrt{\sigma_{22}}$.[3]

*Proof:* Refer to Appendix E.

For the principle of information anchoring to hold, the ratio assuming CAR shown above should be, at least approximately, the same numerically as that for the sensitivity analysis following multiple imputation under the *de-facto* Jump to Reference assumption for censored patients.

[3] $C$ is the variance of the standard deviation of $\sigma_{22}$, a $\chi^2$-distributed variable, so that $C = \sigma^2\left[N - 1 - \left[\sqrt{2}\,\frac{\Gamma((N+1)/2)}{\Gamma(N/2)}\right]^2\right]$, with $N = n_a$. Refer to, for example, page 171 of Kenney and Keeping (1951).

## 4.4 Information anchoring under Jump to Reference

Under J2R, the $n_d$ censored patients obtain multiply imputed event times based on the reference-arm hazard. This has the effect of reducing the difference between the estimated event times on the two arms, since we now have $n_d$ additional observations generated under the hazard from the reference arm. We referred to this phenomenon in section 2.6.2 as the "dilution" or mixing effect, which led to the cumulative hazard curves converging (cf. bottom right panel of Figure 2.6.1 on page 76).

We find that, consistent with what we might expect to happen, the absolute difference in the expected value of the point estimate for the treatment difference following multiple imputation, reduces in size compared to the CAR case:

$$E(\hat{\theta}_{MI,J2R}) = \frac{n_o}{n_a}(\mu_{a2} - \mu_{r2}) - \frac{n_o}{n_a}\dot{\lambda}\sqrt{\sigma_{22}} + \frac{n_d}{n_a}\lambda\sqrt{\sigma_{22}},$$

where, this time, $\lambda$ is calculated using $\mu_{r2}$ instead of $\mu_{a2}$ in the Mills ratio term. Leaving aside the terms in the inverse mill ratio, which approximately cancel one another out, the treatment difference is reduced due to the "hazard dilution" effect.

**Lemma 2**: *The ratio of the information in the full data relative to the incomplete data under the de-facto assumption of J2R is:*

$$\frac{I(\hat{\theta}_{full,sensitivity})}{I(\hat{\theta}_{obs,sensitivity})} = \frac{E[V(\hat{\theta}_{MI,J2R})]}{E[V(\hat{\theta}_{full,J2R})]} \approx \frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]}, \tag{4.4.1}$$

with

$$E[V(\hat{\theta}_{MI,J2R})] = \left[\frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d}\right] +$$

$$+\pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) + 2\pi_d(1-\pi_d)\lambda_{r2}\sqrt{\sigma_{22}}\Delta_c + \pi_d(1-\pi_d)\Delta_c^2 +$$

$$\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}^2 + \frac{(1-\pi_d)^2}{n}\rho^2\sigma_{22} + \frac{3\pi_d(1-\pi_d)^2}{n^2}\sigma_{22}(1-\rho^2)+$$

$$\pi_d^2\left(\frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) + \left[\frac{1}{n} + \frac{(1+\pi_d)}{n^2\pi_d}\right]\sigma_{22}(1-\rho^2)\right), \tag{4.4.2}$$

and,

$$E[V(\hat{\theta}_{full,J2R})] = \frac{\sigma_{22}}{n} + \frac{(1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d} + \pi_d(1-\pi_d)\Delta_c^2}{n}, \tag{4.4.3}$$

where $\Delta_c = \mu_{a2d} - \mu_{a2o}$, with the inverse mills ratio calculated assuming $N(\mu_{r2}, \sigma_{22})$.

*Proof*: For the derivation of equations (4.4.2) and (4.4.3) refer to Appendices F and G. The proof of lemma 2 we delay until we have stated the information anchoring theorem in full.

Now, $E[V(\hat{\theta}_{MI,J2R})]$ in equation (4.4.2) is a rather complicated expression, but if we focus on terms of $o(\frac{1}{n})$ or larger, it simplifies to an expression quite similar to that which was derived for the CAR case. In fact, the expression is dominated by the first term in brackets, but this time the expression under J2R starts with a term in $\frac{\sigma_{22}}{n}$ rather than $\frac{2\sigma_{22}}{n}$, which we had as the first term of the analogous expression under the CAR assumption in equation (4.3.20).

Again, this of course makes sense because $n_d$ censored observations have been replaced with new event times of a similar magnitude to those on the reference arm (in terms of the hazard). Therefore, and in line with what might be expected, the variability in the difference between the arms at time $T_2$ is somewhat reduced due the hazard dilution effect. Equations (4.4.2) and (4.4.3) provide the building blocks for the main result concerning information anchoring.

**Theorem 1**: *For bivariate log normally distributed right censored data, the de-facto variance estimate,* $E[V(\hat{\theta}_{MI,J2R})]$*, following multiple imputation under J2R is information anchored.*

*Proof (sketch)*: We hypothesise that despite using the J2R approach for sensitivity analysis, the variance inflation following MI is the same as that under CAR. Therefore, we compare the

expression for the estimated variance under J2R in equation (4.4.2), $E[V(\hat{\theta}_{MI,J2R})]$, with the *predicted* variance under J2R, $E[\hat{V}_{anchored}]$, which we calculate using the other three terms in the equality in Lemma 2, which we recall relates the ratios for information anchoring to hold,

$$\frac{E[V(\hat{\theta}_{MI,J2R})]}{E[V(\hat{\theta}_{full,J2R})]} \approx \frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]} \implies E[V(\hat{\theta}_{MI,J2R})] \approx E[V(\hat{\theta}_{full,J2R})] \times \frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]}.$$

Therefore, using the expressions for the three terms on the right hand side, which we know from earlier calculations in this chapter, we can obtain the predicted *anchored* variance,

$$E[V_{anchored}] \approx E[V(\hat{\theta}_{full,J2R})] \times \frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]}. \tag{4.4.4}$$

Now, if we subtract the predicted term $E[\hat{V}_{anchored}]$ above from the expression for the newly derived expression for $E[\hat{V}(\hat{\theta}_{MI,J2R})]$ in equation (4.4.2), we will obtain an estimate of the difference, which, if information anchoring holds, should be rather small numerically.

It turns out that we obtain the following expression:

$$E[V(\hat{\theta}_{MI,J2R})] - E[V_{anchored}] \lessapprox 2\pi_d(1-\pi_d)\sqrt{\sigma_{22}}\lambda_{r2}\Delta_c +$$

$$\sigma_{22}\left[\frac{\rho^2}{2n} + \pi_d(1-\pi_d)\lambda_{r2}^2\right] + \pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) + \pi_d^2 VAR(w_{k,\lambda}) -$$

$$\sigma_{a2o}\left[\frac{\rho^2}{2}(1-\pi_d) + \frac{3}{2}\pi_d + \left[\frac{\pi_d(1-\pi_d)}{2}\right]\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right]\right] - \sigma_{a2d}\left[\frac{\rho^2\pi_d}{2}\right] - \Delta_c^2\left[\frac{\rho^2\pi_d}{2}\right], \tag{4.4.5}$$

where we only consider terms greater than or equal to $o\left(\frac{1}{n}\right)$, and assume both $\left(1-\frac{1}{n}\right) \approx 1$ and $C \approx 0.5$ for large $n$. Workings are shown in Appendix H.

The upper bound on the difference in equation (4.4.5) is dominated in absolute magnitude by the first two terms, and the negative ones in $\sigma_{a2o}$ and $\Delta_c^2$. Focussing only on these terms, we see that the difference depends on the number of patients on each arm ($n$), the censoring level ($\pi_d$),

the variance of the data at time 2 ($\sigma_{22}$), the variance of those observed ($\sigma_{a2o}$), the correlation between measurements at times 1 and 2 ($\rho^2$), the difference between the mean of those observed and those deviating on the active arm at time 2 ($\Delta_c$), and the inverse Mills ratio relating to the censoring point $\alpha$.

Furthermore, using the same argument as Cro (2016) in her PhD thesis, we can apply t-test power calculation arguments to provide an upper bound on $\Delta_c$, assuming, for example, 80% power and 5% significance:

$$\Delta_c \lessapprox (\mu_{a2} - \mu_{r2}) = \sqrt{\frac{15.68\sigma_{22}}{n}}.$$

The first term in equation (4.4.5) becomes,

$$\frac{2\sqrt{15.68}\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}}{\sqrt{n}},$$

which is approximately of order $o(1/n)$, and the final term in equation (4.4.5) is now,

$$\frac{15.68\sigma_{22}\rho^2\pi_d}{2n},$$

which is also of approximately $o(1/n)$.

Since $\pi_d \ll 0.5$ can be expected for most sensitivity analyses with realistic applications, the whole expression is of the order of approximately 10% of the total variance $\sigma_{22}$.

Therefore, we may conclude that the upper bound on the difference is relatively small in comparison to the absolute information anchored variance, and the principles of information anchoring have been approximately upheld following MI under J2R, confirming the proposition in Lemma 2 and Theorem 1 above.

In the following sections, we validate these results first by using simulated data, and then by applying the principles to the RITA-2 data.

## 4.5 Simulation study

We now present the results of a simulation study which uses the information anchoring results derived in this chapter. The summary statistics of the example data sets were taken from Cro *et al.* (2018). This helped in the code verification process, that is, to make sure the results were as expected under CAR, before moving to the more complex Jump to Reference scenario.

The simulation study has patients with times $T_1$ and $T_2$ generated from a bivariate normal distribution with means and covariances as follows:

$$\mu_{\text{reference}} = [2, 1.9],\ \mu_{\text{active}} = [2, \mu_{a2}],$$

$$\Sigma_{\text{reference}} = \Sigma_{\text{active}} = \begin{pmatrix} 0.4 & 0.2 \\ 0.2 & 0.6 \end{pmatrix},$$

with a sample size $n = n_r = n_a = 250$ in each arm. We imputed new event times for those censored at time point $\alpha$ using the standard MI methodology we presented earlier in this chapter using the Tobit model. We assume multivariate normality of the estimates from fitting the model to the observed data (with $K$=50 imputed data sets), and compared these results with those from the the theoretically calculated results derived in the previous sections of this chapter. Censoring was varied between 10% and 50% on the active arm, all data being observed on the reference arm, with 500 simulated data sets.

The results summarised in Table 4.5.1 show a considerable degree of alignment when we compare the *predicted* variance calculated using $V_{\text{anchored}}$ with the formula for the estimated variance following MI using J2R (column "Difference theory (A-B)"), similarly when we use simulated data (column "Difference simulation (C-D)). (In the table we have dropped the $\hat{\theta}$ from the expressions to ease readability).

The discrepancies increase as the censoring level increases as we move down the table (column "Difference simulation (C-D)"), from 0.00002 at 10% censoring to 0.0002 at 50% censoring, which are of the approximately of the order of mangnitude of the Monte Carlo simulation error

(0.00016). This was also the case for the analogous results with longitudinal data (Cro *et al.*, 2018). Therefore, we conclude that the simulation results are consistent with our expectations and our information anchoring arguments appear to hold.

In the next section, we investigate whether information anchoring principles hold in a real data setting.

| $n_a$ | Proportion of missingness ($\pi_d$) | Information anchored variance bound (A) | Information anchored variance - simulation (C) | $E[\hat{V}_{MI,J2R}]$ theory (B) | $E[\widehat{V_{MI,J2R}}]$ simulation (D) | Difference theory (A-B) | Difference simulation (C-D) |
|---|---|---|---|---|---|---|---|
| 250 | 0.1 | 0.0081 | 0.0080 | 0.0080 | 0.0080 | 0.00009 | 0.00002 |
| 250 | 0.2 | 0.0081 | 0.0080 | 0.0082 | 0.0080 | -0.00004 | 0.00005 |
| 250 | 0.3 | 0.0085 | 0.0082 | 0.0083 | 0.0081 | 0.00017 | 0.00012 |
| 250 | 0.4 | 0.0090 | 0.0085 | 0.0087 | 0.0084 | 0.00034 | 0.00017 |
| 250 | 0.5 | 0.0098 | 0.0091 | 0.0092 | 0.0089 | 0.00064 | 0.00020 |

Table 4.5.1: Difference between Rubin's Jump to Reference MI variance estimator and the information anchored variance estimate comparing theoretical bounds with simulated data for J2R (calculated in multiples of $\sigma_{22}$).
Column (A): The predicted variance following MI under the *de-facto* assumption of J2R using Rubin's rules, calculated based on information anchoring principles with *a priori* values (i.e. without using simulated data):

$$E[\hat{V}_{MI,J2R}] = E[\hat{V}_{\text{anchored}}] = \frac{E[\hat{V}_{MI,CAR}]}{E[\hat{V}_{full,CAR}]} \times E[\hat{V}_{full,J2R}],$$

utilising the theoretical bound for Rubin's variance estimate for $k = 1 \ldots K$ multiply imputed data sets: $\hat{V}_{.,MI} = \hat{W} + \left(1 + \frac{1}{K}\right)\hat{B}$, where $\hat{W} = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2$, is the within imputation variance $\hat{\sigma}_k$, and the between imputation variance is $\hat{B} = \frac{1}{(K-1)}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_{MI})^2$, with MI point estimator of $\theta$, $\hat{\theta}_M I$ defined as $\hat{\theta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$, for the $k$th estimate $\hat{\theta}_k$.
Column (B) is the estimate of Rubin's variance under the *de-facto* assumption of J2R, $E[\hat{V}_{MI,J2R}]$, calculated using *a priori* values (i.e. using without simulated data).
Column (C) applies the same calculation for information anchoring as defined in column (A) but uses simulated data (denoted by the wide hat):

$$E[\widehat{V_{MI,J2R}}] = E[\widehat{V_{\text{anchored}}}] = \frac{E[\widehat{V_{MI,CAR}}]}{E[\widehat{V_{full,CAR}}]} \times E[\widehat{V_{full,J2R}}].$$

Column (D) is the estimate of Rubin's variance under the *de-facto* assumption of J2R, $E[\widehat{V_{MI,J2R}}]$, calculated using simulated data.

Column (A-B) is the theoretical difference using the bounds calculated in this chapter ($E[\hat{V}_{MI,J2R}] - E[\hat{V}_{anchored}]$), whereas column (C-D) is the difference using simulated data ($E[\widehat{V_{MI,J2R}}] - E[\widehat{V_{anchored}}]$).
Using the uncensored simulated data sets, the Monte Carlo error is 0.000156.

### 4.5.1 Information anchoring for the RITA-2 data

We now return to the RITA-2 data to provide further evidence of the validity of the information anchoring principle.

The underlying assumptions used for the theoretical results on information anchoring are slightly different to the setting in the RITA-2 data. For example, the fixed censoring threshold $\alpha$ is not the same for all patients. This being the case, instead of using the RITA-2 data directly, we copy the statistical characteristics of the data as the basis for simulating log normally distributed patient times, censored at a time point $\alpha$.

To briefly reiterate the RITA-2 setting, the primary end-point is the difference in mean log time to death on the medical arm (the reference arm) compared to that for those receiving PTCA (the active arm). Patients are censored administratively at the end of the study, or earlier, should they require a non-random intervention (i.e. PTCA/CABG).

For the sensitivity analysis, we consider the potential effect of NRIs on the medical arm *only* (the reference arm); that is, those patients "jumping" to PTCA (here the active arm) following censoring — that is, as before in Chapter 3, they "Jump to PTCA arm". Again, as in Chapter 3, those administratively censored, or having a second surgical intervention on the active arm are considered CAR for this illustrative example.

In summary, our sensitivity analysis investigates the possible effects of an NRI on those originally randomised to the medical arm, under the *de-facto* assumption of "Jump to PTCA arm".

The original analysis for the RITA-2 trial was based on an intention to treat approach, providing us with the observed outcomes for patients followed up to the end of the study, including those having an NRI. This allows us to broadly compare results multiply imputed under "Jump to PTCA arm" with those actually observed (refer to the "observed" column A in Table 4.5.2), albeit under slightly different assumptions.

To validate the theoretical results, we generate bivariate normally distributed data according to the properties of the observed RITA-2 trial data, and choose $\alpha$ to result in a censoring level of 27% on the medical arm, as with the RITA-2 data. The variable $T_1$ we fix as the date of randomisation and $T_2$ is the censoring or event time (as appropriate). The "true" mean of the log time at $T_2$ on the medical arm, $\mu_{r2}$, is unknown because of the censoring:

$$\hat{\mu}_{\text{reference=medical}} = [0.94, \hat{\mu}_{r2}],\ \hat{\mu}_{\text{active=PTCA}} = [0.94, 1.75],$$

$$\Sigma_{\text{reference}} = \Sigma_{\text{active}} = \begin{pmatrix} 0.15 & -0.04 \\ -0.04 & 0.22 \end{pmatrix},$$

so that $\sigma_{a1} = \sigma_{r1} = \sigma_{11} = 0.15$, $\sigma_{a2} = \sigma_{r2} = \sigma_{22} = 0.22$, and $\sigma_{12} = -0.04$. For the medical arm, we also know the mean of the events at time 2, $\mu_{a2o} = 1.60$ with associated variability, $\sigma_{a2o} = 0.11$. Again, these summary statistics reflect those of the RITA-2 data.

Table 4.5.2 summarises the parameter estimates and standard errors, along with the primary endpoint for the difference in group means between the treatment arms. Under the *de-jure* assumption of CAR, there is a significant difference (-0.15 [-0.21, -0.10], p<0.001 [column [1], row 3]), whereas the difference is only marginally significant following MI using the Jump to Active (J2A) approach (here "Jump to PTCA arm") for those censored on the medical arm (0.06 [-0.004, 0.11], p = 0.07 [column [3], row 3]). This is approximately the same as the analogous results for the intention to treat endpoint from the original study (0.06 [0.001, 0.12], p = 0.05 [column [2], row 3]).

The information anchored variance comparison is shown in the lower part of Table 4.5.2. There is little difference between the *theoretically* predicted estimates using our calculated quantities (Difference - predicted, [A]-[B] = -0.00002), and those from performing multiple imputation under J2A (Difference - MI, [C]-[D] = 0.00005), which confirms our thinking regarding the relationship $E[V_{MI,J2R}] - E[V_{anchored}]$ from Theorem 1. Therefore, information anchoring appears to hold for this example data set.

| Treatment | N | De-jure estimand (CAR) [1] | Intention to treat observed[2] | *De-facto* estimand (J2A) multiple imputation ($K = 50$) [3] |
|---|---|---|---|---|
| Medical (reference) | 504 | $\mu_{a2o} = 1.60$ (0.34) | $\mu_{r2}$ =1.81 (0.42) | $\mu_{r2}$ =1.81 (0.42) |
| PTCA (active) | 514 | $\mu_{r2}$ =1.75 (0.47) | $\mu_{r2}$ =1.75 (0.47) | $\mu_{r2}$ =1.75 (0.46) |
| Difference in group means (t-test) | 1018 | -0.15 [-0.21, -0.10] p $< 0.001$ | 0.06 [0.001, 0.12] p=0.05 | 0.06 [-0.004, 0.11] p=0.07 |
| Rubin's variance estimator (calculated) | | $E[\hat{V}_{MI,CAR}] = 0.0042$ | - | $E[\hat{V}_{MI,J2R}] = 0.0043$ |
| Rubin's variance estimator (following MI) | | $E[\widehat{V_{MI,CAR}}] = 0.0042$ | - | $E[\widehat{V_{MI,J2R}}] = 0.0041$ |
| Variance under J2A | | | | |
| Information anchored $E[\hat{V}_{\text{anchored}}]$, predicted | [A] | - | - | 0.0042 |
| $E[\hat{V}_{MI,J2R}]$ calculated | [B] | - | - | 0.0043 |
| Information anchored $E[\widehat{V_{\text{anchored}}}]$, MI | [C] | - | - | 0.0042 |
| $E[\widehat{V_{MI,J2R}}]$ | [D] | | - | 0.0041 |
| Difference - predicted | [A]-[B] | - | - | -0.00002 |
| Difference - MI | [C]-[D] | - | - | 0.00005 |

Table 4.5.2: Top half of table: Mean and standard deviation (in brackets) for the medical and PTCA arms of the RITA-2 data set;
Bottom half of table: Comparisons of the variance under the *de-jure* and *de-facto* assumptions of CAR respectively "Jump to PTCA", following MI;
Information anchoring predicted [A], calculated theoretically [B], information anchoring predicted using simulated data [C], following MI under J2A using simulated data [D]; variance estimators expresses as a multiple of $\sigma_{22}$.

Definitions for the table (prior to normalisation with $\sigma_{22}$): Fully observed = $E[\hat{V}_{full,CAR}] = 0.00087$; under CAR (theoretically calculated) $E[\hat{V}_{MI,CAR}] = 0.0009$; under CAR (following MI) $E[\widehat{V_{MI,CAR}}] = 0.00092$; under J2A (fully observed, theoretical) $E[\hat{V}_{full,J2R}] = 0.00087$; under J2A, (fully observed, MI) $E[\widehat{V_{full,J2R}}] = 0.00087$; under J2A, (theoretical) $E[\hat{V}_{MI,J2R}] = 0.00094$; under CAR (following MI) $E[\widehat{V_{MI,J2R}}] = 0.00090$.

Information anchored variance theory - theory [A]: $\frac{\hat{V}_{MI,CAR}}{\hat{V}_{full,CAR}} \times \hat{V}_{full,J2R} = 1.07 \times 0.00087$; information anchored variance after MI [C]: $\frac{\widehat{V_{MI,CAR}}}{\widehat{V_{full,CAR}}} \times \widehat{V_{full,J2R}} = 1.05 \times 0.00087$.

## 4.6 Summary

In Chapter 3, we embarked on the investigation of information anchoring for reference based sensitivity analysis for time-to-event outcomes. It was demonstrated that, at least empirically, information anchoring principles hold.

In this chapter, we derived an expression that showed that the difference between the information anchored variance using Rubin's rules for the primary and sensitivity analyses is relatively small in magnitude, certainly compared to the variance of the outcome. This was then validated using both simulated and real data.

Taken together, Chapters 3 and 4 demonstrate that the information anchoring principle defined in the introductory remarks of Chapter 1 appears to hold both empirically, and more generally, albeit the latter under certain distributional assumptions. With these statements, we close the study of information anchoring for the reference based sensitivity analysis approaches.

In the next chapter we change direction somewhat, and focus on an application of the reference based sensitivity methods to *observational* data, again with a time-to-event outcome.

# Chapter 5

# Reference-based multiple imputation to investigate informative censoring: A *trial emulation* in COHERE

## 5.1 Preamble — sensitivity analysis born out of necessity

Prior to discussing the application of the methods to observational data, it is perhaps worth taking the time to explain the reasoning behind the change of direction from sensitivity analysis for RCTs in Chapters 2-4, to sensitivity analysis for observational data in this chapter.

The plan for the PhD always included a final part investigating the use of the methods for observational data. However, at the time the plan was conceived there did not seem to be a clear path to achieve this — of course, inverse probability methods and other methodological building blocks had been developed (Sterne *et al.*, 2005; Hernan *et al.*, 2006) — but the link between these methods and reference based sensitivity analysis had not been made (at least by the author). The work was put on the back burner (by the author) for some years.

In the meantime, the requirement for sensitivity analysis for observational data was confirmed and underscored (again, to the author) following an analysis of the incidence of tuberculosis in patients with HIV for data from a southern African cohort (Fenner *et al.*, 2017). This served to re-ignite the search for practical, yet statistically valid, methods for sensitivity analysis applica-

ble to observational data.

At the same time several publications by Hernan and colleagues focussing on so-called “trial emulation” techniques for handling time varying confounding and selection bias issues typical in observational cohort data sets were published. If a trial emulation method was being applied to observational data, then it seems logical to apply the same type of sensitivity analysis methods from RCTs to such a trial.

Driven by the availability of new data from COHERE, and the need to clarify the benefits of prophylaxis on the risk of Pneumocystis pneumonia (PCP), the project was started including a sensitivity analysis to investigate possible information censoring.

## 5.2 Introduction

PCP is an opportunistic disease contracted by individuals having a weakened immune system, and it remains one of the most frequent AIDS defining diagnoses in resource rich countries. HIV viral load can be managed using combinational Antiretroviral Treatment (cART), with additional PCP prophylactic treatments recommended for those with low CD4 lymphocyte counts thought to be at risk. In addition to increased pill burden, these additional medications can cause adverse effects. Prolonged usage potentially increasing the risk of antimicrobial resistance which should be avoided especially in this high risk population.

The COHERE data and motivation for the study were introduced in Chapter 1.11.3. Given the wealth of new data available in COHERE since the last of the studies focussing on PCP was carried out using COHERE data (2014), the goal of the project was to investigate whether PCP prophylaxis might be withheld in all patients on antiretroviral therapy with suppressed plasma HIV RNA (<400c/mL). We use observational data from COHERE to compare the risk of continuing versus stopping the usually required PCP prophylaxis.

Estimating such a treatment effect using observational data is made complicated due to the presence of time dependent confounders. For example, CD4 count is not only used as a biomarker for disease progression, but is also itself affected by patients taking their antiretroviral treatment. Such inherent "feedback loops" often make analyses estimating causal effects more complex.

To achieve our goal, the risk of primary PCP was estimated in patients on cART using an established causal inference approach in which observational data are used to *emulate* a hypothetical randomised trial (the *target* trial). We use Inverse Probability Weighting (IPW) to adjust for potential selection bias, but this still implicitly makes the assumption that the censoring was at random (CAR). Since this is an untestable assumption, in a further step we went on to apply the "Jump to Reference" reference based sensitivity analysis approach, introduced and explored in previous chapters, to investigate inferences when censoring is informative. Such sensitivity analyses are equally as important in an emulated trial using observational data, as for the RCT setting.

The set-up of the data records and implementation approach for the trial emulation follows the methods outlined in Danaei *et al.* (2013). This provided an excellent step-by-step blueprint of how to apply the trial emulation method. The recent publications by Caniglia *et al.* (2017), Lodi *et al.* (2017), and Garcia-Albeniz *et al.* (2017) were also invaluable in terms of providing

guidance for the definition of the target trial, along with the guidelines and recommendations for avoiding "self inflicted injuries" from Hernan *et al.* (2016) with such methods.

The next section briefly summarises the causal inference literature, the trial emulation approach and sensitivity analysis methods used to date.

## 5.3 Causal methods, *trial emulation* and the rationale for a different approach to sensitivity analysis

There is a vast literature tracking the introduction and methodology progress of causal inference approaches. The book by Hernan and Robins takes the reader through the motivation and implementation of such methods (Hernan and Robins, 2018). Also, the recent review by Newsome *et al.* (2017) provides a short overview of the main methodological approaches. In this section, we briefly review the literature, focussing primarily on the basics of the approach taken to emulate our hypothetical target trial.

A causal inference approach moves beyond describing "associations" derived from fitting a model to the observational data — it attempts to identify the underlying, ideally nonconfounded, risk factors for the outcome. This is a different concept to what can sometimes be a rather "scatter gun" approach to finding potential associations following model fitting. Causal inference methods endeavour to disentangle the cause and effect by using structured arguments regarding the potential risk factors, often supported by helpful "causal diagrams" tracking the effects and their directions (Robins *et al.*, 2000).

By way of introduction we provide a quote from Hernan and Robins which perfectly sets the scene for our trial emulation approach:

> "Ideally, questions about comparative effectiveness or safety would be answered using an appropriately designed and conducted experiment. When we cannot conduct a randomized experiment, we analyze observational data. Causal inference from large observational databases (big data) can be viewed as an attempt to emulate a randomized experiment - the target experiment or target trial - that would answer the question of interest" (Hernan and Robins, 2016).

Therefore, we fall back on analysing observation data since the "target" trial which would address the particular causal question is not always feasible, ethical or timely (Hernan and Robins, 2016). Another reason, and this has been particularly relevant in research associated with the study of HIV treatments, it is often not feasible to run a trial comparing *many* different treatment regimes simultaneoulsy. A typical example of this has been the definition of optimal points for starting ART treatment based on CD4 count, a biomarker for disease progression in HIV positive patients (Hernan *et al.*, 2006).

For an emulated trial, due to the nature of observational data, the target trial will often be one which focusses on *de-facto* estimands, that is, one with an "intention to treat" flavour, so that treatments are compared under the usual conditions in which they will be applied, rather than under the strict monitoring of, for example, deviations associated with an RCT (Hernan and Robins, 2016).

Using observational data in this way does however leads to some complications. To illustrate these we again consider examples from the HIV field. The main issue is exemplified when we have a time varying exposure, such as ART, and a time varying marker, such as CD4 count. In the past ART has been initiated when the CD4 count reached a certain level, since the individual was at higher risk for opportunistic diseases. Treatment with ART improves the health of the patient, increasing CD4 count. This creates what is essentially a "feedback loop" between treatment and the biomarker. Nowadays, reflecting newer data on the benefit of ART in asymptomatic individuals with high CD4 count, ART is indicated for all HIV-infected patients.

If we consider a time-to-event outcome, such as time to disease progression, then the usual statistical analysis approach would be to estimate the effect of the time varying treatment (ART) on survival by fitting, for example, a Cox Proportional Hazards model including time varying treatment and CD4 count as dependent variables. Robins showed this approach may be biased, irrespective of whether there is further adjustment for past covariate history (Robins, 1997), whenever

1. "there exists a time dependent covariate (CD4) that is both a risk factor for the outcome and also predicts subsequent treatment (ART), and,
2. past treatment history predicts the risk factor (CD4) . . . " (Hernan *et al.*, 2000).

When conditions 1 and 2 above apply, which is often the case in observational studies, then this

is known as "confounding by indication" (Robins *et al.*, 2000). The solution is to apply methods which adjust the model to take into account, or perhaps better said, *remove* any potential feedback loop(s).

Robins and colleagues defined three main methods to estimate causal effects involving a time varying treatment when there are also time varying confounders: the parametric g-computation algorithm estimator, g-estimation of structural nested models, and Inverse Probability of Treatment weighting (IPTW, here shorthanded to IPW) estimation of marginal structural models (MSM) (Robins, 1998a,b, 2000; Hernan *et al.*, 2000). Adjustment can also be performed using matching methods (e.g. using propensity scores, Rosenbaum and Rubin, 1984; Xu and Kalbfleisch, 2010), or doubly robust methods (as mentioned in Chapter 1). More recently, other methods such as targeted maximum likelihood estimation have been developed which are also becoming popular, perhaps because they (can) employ machine learning methods borrowed from other fields (Luque-Fernandez *et al.*, 2017; van der Laan and Rose, 2018).

In our application, we avoid the confounding by indication issue by censoring individuals when they change treatment. We then use a marginal structural modelling approach using inverse probability weighting to adjust for potential selection bias from censoring (details follow in section 5.7.2).

Hernan *et al.* recommend using MSM for a variety of reasons, not least of which is that they "resemble standard models", and are somewhat less complex to implement than either of the two g-type methods mentioned above (Hernan *et al.*, 2000). To apply the methods appropriately a number of conditions have to be taken account of:

- Exchangeability: For this assumption to hold we have to have measured a sufficient number of joint predictors of exposure (e.g. ART) and outcome (e.g. disease progression) so that, within each level of the predictor (e.g. a stratified CD4 count), associations between the exposure and outcome that are due to common causes will disappear. That is, and again using the HIV context, those patients within the same CD4 strata do not exhibit any significant differences after fitting a model with outcome and exposure (Cole and Hernan, 2008). Equivalently, exchangeability ensures the same conditions apply as if we had an RCT, so that randomisation at baseline ensures that patients assigned to treatment arms do not differ significantly (Toh and Hernan, 2008). Of course, it is difficult to determine if you have achieved this in practice. Exchangeability implies the assumption often known as *no unmeasured confounding*.

- Positivity: For this assumption to hold there have to be exposed and unexposed paticipants at each level of the confounders. For example, in the COHERE data, for positivity to hold we would have to ensure that there were indivuduals on and off PCP prophylaxis in all ethnic groups, if we wanted to adjust for all ethnic groups in the model appropriately (Cole and Hernan, 2008).

- Consistency: We paraphrase Hernan *et al.* again — this assumption is tantamount to making sure that the intervention is clearly defined (Hernan and Swanson, 2017). One would think that this would go without saying, but as our later example application shows, complying with this requirement proves more difficult in practice when using real observational data. This point is discussed again in section 5.4.1. Here, it suffices to mention that the consistency assumption boils down to answering two questions relating to the emulated trial: 1.) *what* treatments are we comparing? and 2.) *when* is "time zero" in our trial?

Our emulated trial aims to mimic the design of a randomised trial as closely as possible. This involves the iterative process of comparing the hypothetical "target trial" with the emulated one to ascertain if the observational data is sufficient to address the research question. This process requires repeated discussion since it is critical "to systematically articulate the tradeoffs that we were willing to accept" (Hernan and Robins, 2016), and enunciates these clearly to the study team. The tradeoffs made for our illustrative example are discussed in section 5.4.1, where we compare the target and emulated trials. Of course, there are still differences between a real trial and an emulated one, for example, blinding is not possible. However, we endeavour to minimise the differences, and be precise and open about the limitations of the approach when documenting it.

Notwithstanding the potential pitfalls, Hernan and Robins recommend a pragmatic approach — "we will rarely be able to emulate the ideal trial in which we are most interested ... a number of compromises will have to be made regarding eligibility criteria, strategies to be compared, etc", (Hernan and Robins, 2016).

Once we have emulated the target trial, our goal is to apply the "Jump to Reference" sensitivity analysis approach to investigate potential informative censoring. To date there seems to be few cases of sensitivity analyses to investigate informative censoring in such a setting. Danaei *et al* explain that:

> "IP weights can also be estimated to adjust for informative censoring due to loss to follow-up, which may arise in all types of analyses ... We examined the effect of censoring due to loss to follow-up [in the data] by using IP weights for censoring. As expected, the results were almost identical to those without IP weights for censoring ...", (Danaei *et al.*, 2013).

We interpret this statement to mean that for those lost to follow-up in the trial, the IP weights were altered in some way to investigate informative censoring.

Such "weight manipulation" methods also suffer from similar difficulties as the *delta* methods discussed in Chapters 1 and 2. Namely, that although the manipulation itself is straightforward, the size of the change in the weights, the "$\delta$", and its distribution is not easy for the trial team to dimension.

Lodi *et al.* use the parametric g-formula rather than trial emulation, and interestingly pursue something more akin to the "Extreme Hazard" methods defined in Chapter 2 for their sensitivity analysis:

> "Because a nonnegligible proportion of death events had unknown cause of death, as a sensitivity analysis we estimated the ... risk ratio of non-AIDS mortality *assuming that all deaths due to unknown causes were non-AIDS related*. This extreme case scenario is unrealistic in practice, but provides an illustration of how sensitive the analyses may be to assumptions regarding the missing data." (italics added), (Lodi *et al.*, 2017).

Aside from the methodological simplicity of the reference based MI methods, the apparent lack of adoption of sensitivity analysis methods in observational data settings provides a realistic rationale for the application presented here.

In the next section, we define and compare the hypothetical target trial and the emulated trial, before going on to describe the primary analysis model, inverse probability weighting and the sensitivity analysis in more detail.

## 5.4 Methods

### 5.4.1 Target trial

The aim of the analysis was to emulate a randomised control trial (RCT) using observational data, and the natural starting point for this approach is to first define the hypothetical RCT to investigate the hypothesis. The target trial is a two arm, open label trial comparing the time to PCP diagnosis for individuals who continue taking PCP prophylaxis with those who stop taking PCP prophylaxis. Up to the point of randomisation, all patients are assumed to be taking PCP prophylaxis according to existing NIH guidelines (NIH, 2018) (refer to Table 1.11.1 in Chapter 1). That is, they have a CD4 count of less than or equal to 200 cells/$\mu$L.

As secondary endpoint, we also examined the risk for all-cause mortality for those on an off PCP prophylaxis.

We first define in more detail the protocol of the hypothetical target trial for the effect of stopping PCP prophylaxis (Table 1 left hand side). Prophylaxis might be stopped if a patient no longer required prophylaxis according to NIH guidelines. Alternatively, prophylaxis might be started as a rescue medication.

The components of the estimand for the trial, as defined generically in section 1.2 on page 9, are also presented is Table 1: “eligibility” summarises the population targeted by the scientific question; “outcome” defines the endpoint for each patient; “protocol deviation” reviews the reasons for censoring (i.e. the intercurrent events) — for example, patients deviating from protocol are censored for the primary analysis, as well as those no longer adhering to the originally assigned treatment; finally, “statistical analysis” describes the population level summary which provides the treatment effect of interest.

Figure 5.4.1 summarises the enrolment, monitoring phase, “randomization” and primary endpoint for the hypothetical target trial.

| Component | Hypothetical target trial | Emulated trial using observational data |
|---|---|---|
| Aim | To compare the risk of PCP diagnosis between those patients *continuing* PCP prophylaxis and those *stopping* PCP prophylaxis | To compare the risk of PCP diagnosis between those patients *taking* PCP prophylaxis and those *not taking* PCP prophylaxis |
| Eligibility | Individuals on cART with a CD4 count below 200 cells/$\mu$L and current HIV RNA measurement $<$ 400 copies/mL (i.e. virally suppressed) and on PCP prophylaxis | Individuals on cART with a CD4 count below 200 cells/$\mu$L and current HIV RNA measurement $<$ 400 copies/mL (i.e. virally suppressed) and on **and** off PCP prophylaxis |
| Treatment strategies | 1. Continue taking PCP prophylaxis at baseline<br>2. Stopping PCP prophylaxis at baseline | 1. Taking PCP prophylaxis<br><br>2. Not taking PCP prophylaxis. |
| Treatment assignment | Patients are randomly assigned to either strategy | Patients are assigned to PCP prophylaxis if they are taking prophylaxis when the eligibility criteria are met, and to off PCP prophylaxis if they are not taking prophylaxis when the eligibility criteria are met. Baseline was defined to be the first time these criteria were met, and patients would be assigned to their respective arms accordingly. Emulated point of randomisation is therefore the time point at which eligibility criteria are met. Randomisation is emulated by adjustment for baseline covariates and using censoring weights. |
| Follow-up | Follow-up starts at treatment assignment and end at first PCP diagnosis, at death, at loss to follow-up (2 years with no contact), or 10 years after baseline, or on 31.3.15, whichever occurs first. | Same, plus censoring at discontinuation of the treatment strategy assigned at baseline |
| Protocol deviation | Any patient no longer fulfilling the eligibility criteria, stopping prophylaxis for those on prophylaxis, or re-starting prophylaxis for those no longer taking prophylaxis. These patients are censored at their time of deviation, and their time-to-event not considered in the primary analysis | Same |
| Outcome | Primary endpoint: Time from randomisation until a PCP diagnosis<br>Secondary endpoint: Time from randomisation until death (all-cause) | Same |
| Causal contrast | Per-protocol effect i.e. effect of taking PCP prophylaxis versus stopping PCP prophylaxis | Observational analogue of per-protocol effect |
| Statistical Analysis | Per protocol analysis comparing hazard ratio for continuing versus stopping prophylaxis, adjusted for baseline covariates | Per protocol analysis comparing hazard ratio for on versus off prophylaxis, adjusted for baseline covariates, with inverse probability weighting used to adjust for potential selection bias |



Table 5.4.1: Target trial and emulated trial using observational data from COHERE.

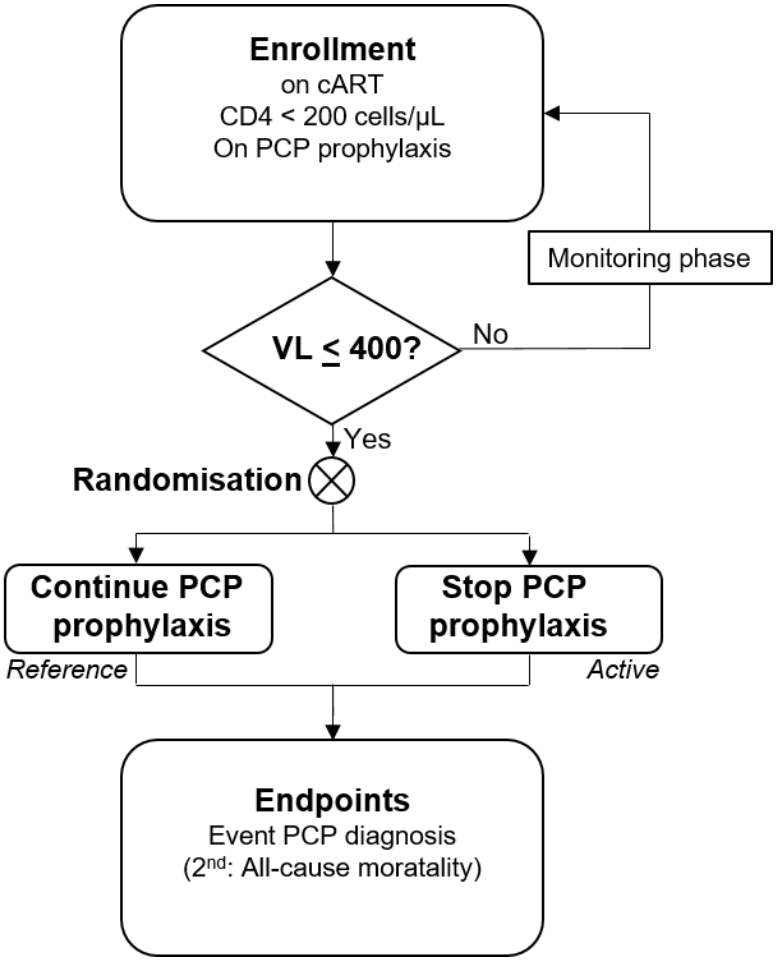


Figure 5.4.1: Hypothetical target trial: Enrolment, monitoring phase, "randomization" and primary endpoint.

## 5.5 Emulated trial using COHERE data

To emulate the target trial, we included data from the 2015 merger of the COHERE database from 23 of the cohorts, with information on patient characteristics (age, sex, geographical origin, and transmission category), use of ART (type of regimes and dates of start and discontinuation), CD4 cell counts and plasma HIV-RNA over time and their dates, AIDS-defining conditions and indicator variables for drop-out/death.

HIV infected individuals are eligible to enter the hypothetical study if they began follow-up in their cohort after 1st June 1998. Additionally, they must have started cART on or after this date, be 16 years or older, and have no history of previous PCP. cART was defined as any combination of 3 or more antiretrovirals of any type.

We selected patients in COHERE compliant with the same entry criteria as in the target trial. A total of 9,743 patients with approximately 49,000 follow-up visits were eligible for "pseudo-randomisation". We defined the emulated point of randomisation to be the first and possible subsequent time points at which the eligibility criteria were met (refer to Table 5.4.1 right hand side).

In the target hypothetical trial, we randomise patients already taking prophylaxis to either *continue* or *stop* taking prophylaxis. In contrast, in the emulated trial we choose patients fulfilling the eligibility criteria in terms of CD4 count and HIV RNA measurements, and then allocate them to an arm depending on whether, at the time they fulfil these eligibility criteria, they are on or off prophylaxis.

This assumes that there is no, or negligible, influence from the duration that PCP prophylaxis has been taken or not taken, prior to this point of randomisation. Discussions with clinical experts implied that PCP prophylaxis might have an effect up to one month after stopping treatment, so the above assumption seemed appropriate.

At this point we note that the hypothetical and emulated trials address slightly different treatment strategies. In the hypothetical target trial, we compare continuing taking prophylaxis against the risk of stopping prophylaxis, whereas in the emulated trial we compare the risk of taking prophylaxis versus not taking prophylaxis. This was a pragmatic approach, taken since there were few PCP events when considering the hypothetical target trial defined on the left hand side of Table 5.4.1. Opinion was divided as to whether this pragmatic approach was sufficient (the view of our clinicians), or whether the trials should be implemented exactly as on the left hand side of Table 5.4.1 (the "purist" approach advocated by some trialists).

To be completely compliant with the target trial, the patients would need to be eligible in terms of CD4 count and HIV RNA suppression *and* on prophylaxis, and then stop prophylaxis *at this visit*. The combination of these events occurring was so rare that the possible patient donor pool made the analysis less tractable. Instead, the more pragmatic approach was taken. As noted by Cain *et al.* this seems reasonable in a practical setting,

> "In order to emulate randomized experiments using observational data, we consider 'having data consistent with a regime in the observational data' analogous to 'following a regime in a randomized experiment with perfect adherence' ", (Cain *et al.*, 2010).

In addition, and in response to several reviewers, as a separate subgroup analysis we also emulated two additional trials. One in which all patients are off PCP prophylaxis at randomisation, and then *at this point in time* (or within the next quarter) some start prophylaxis. A second trial was also emulated which was exactly in line with the hypothetical target trial, so that all patients are on prophylaxis at randomisation, and at this point in time (or within 3 months) patients which stop prophylaxis are considered to be on the off prophylaxis arm.

In contrast to an RCT, our emulated trial does not randomise patients to each of the arms. Table 5.5.1 shows the baseline characteristics of the patients on and off PCP prophylaxis at the point at which they are randomised for the first time in an emulated trial. Although most of the values are comparable between the two groups, for example, those off prophylaxis have similar median baseline CD4 count (149 vs 133 cells/$\mu$L) and HIV RNA levels (50 vs 90 copies/ml), there are some differences between the two groups which could be clinically relevant for the analysis. Of note, there are more IDUs in the off prophylaxis group (28% vs 21%, $p < 0.001$). In addition, there is a large number of missing values for the *geographical origin* variable for those off prophylaxis. In an RCT, we would not expect to see such differences following randomisation, but for the emulated trial we have to adjust for the potential confounding effects in the analysis. Although we adjust for several variables that have been collected in the data, we cannot rule out unmeasured confounding, which may lead to biased results.

There are also potential issues concerned with this type of trial, irrespective of whether the RCT was actually performed as defined hypothetically, or emulated using the observational data. The presence of undiagnosed PCP at the time the trial is started would most probably be a potential risk in both the hypothetical target trial and the emulated trial. In addition, we cannot rule out potential behavioural changes associated with a patient knowing that he/she is on prophylaxis. Finally, certain physicians may be more or less cautious about prescribing prophylaxis perhaps depending on unrecorded characteristics which may influence the outcome. All the above points are common risks associated with any open label trial.

| | On PCP prophylaxis | | Off PCP prophylaxis | | p-value |
|---|---|---|---|---|---|
| | N / median | (%) / [IQR] | N / median | (%) / [IQR] | |
| N | 3150 | (32%) | 6593 | (68%) | |
| PCP diagnosis | 11 | (0.3%) | 28 | (0.4%) | 0.7 |
| Male | 2371 | (75%) | 4794 | (73%) | 0.008 |
| Transmission mode | | | | | $< 0.001$ |
| Heterosexual | 1200 | (38%) | 2157 | (33%) | |
| IDU | 666 | (21%) | 1863 | (28%) | |
| MSM | 961 | (31%) | 1835 | (28%) | |
| Other | 131 | (4%) | 287 | (4%) | |
| *Missing* | *192* | *(6%)* | *451* | *(7%)* | |
| Geographical origin | | | | | $< 0.001$ |
| Europe | 2470 | (78%) | 4555 | (69%) | |
| Africa | 341 | (11%) | 555 | (8%) | |
| Asia | 62 | (2%) | 84 | (1%) | |
| Latin America | 183 | (6%) | 272 | (4%) | |
| North Africa & Middle East | 46 | (2%) | 63 | (1%) | |
| *Missing* | *48* | *(2%)* | *1064* | *(16%)* | |
| Age (yrs) (median [IQR]) | 41 | [35, 48] | 42 | [36, 49] | 0.001 |
| CD4 (median [IQR]) | 133 | [84, 170] | 149 | [101, 180] | $< 0.001$ |
| HIV RNA (median [IQR]) | 90 | [50, 204] | 50 | [49, 199] | $< 0.001$ |
| Calender year | 2005 | [2001, 2009] | 2006 | [2002, 2010] | $< 0.001$ |
| Follow-up (yrs) (median [IQR]) | 1.0 | [0.4, 3.1] | 0.7 | [0.3, 2.1] | $< 0.001$ |
| Death | 573 | (18%) | 830 | (13%) | $< 0.001$ |
| Lost to follow-up | 542 | (17%) | 931 | (14%) | 0.2 |

Table 5.5.1: Characteristics for eligible COHERE patients on and off PCP prophylaxis at randomisation to the first emulated trial.

IQR Interquartile range; IDU Intravenous drug users; MSM Men having sex with men

## 5.6 Emulation of multiple trials

There were only 248 virally suppressed individuals starting follow-up in the first quarter of 1998, none of whom had a later PCP diagnosis. Clearly, using only these data to emulate an RCT would be of limited value.

In line with our blueprint analysis from Danaei *et al.* (2013), instead of emulating a single trial, we emulated multiple trials, each starting at consecutive quarters from the first quarter of 1998 until the end of the first quarter of 2015. Hernan and Robins make the following comment concerning the ideal inter-trial timecale,

> "if there is a fixed schedule for data collection at prespecified times ...., then we can emulate a new trial starting at each specified time", (Hernan and Robins, 2016).

In light of this, we chose to allow new trials to start every 3 months from 1.1.98 until 31.3.15.

Patients becoming eligible in a specific quarter are enrolled into the trial starting in that quarter, remaining in the trial according to the follow-up conditions, unless censored (see Table 5.4.1). A patient censored from a trial may become eligible again at a later date, and in this case the patient also participates in the later trial, starting as a new patient in the trial beginning in the respective quarter. For example, an individual having multiple non-contiguous periods of having CD4 count less than 200 cells/$\mu$L and being virally suppressed would be involved in multiple trials. At each visit that such a patient fulfils the eligibility criteria, the patient is treated anew as though they were a new participant in the respective trial.

This means there are multiple concurrent trials running during the follow-up period, each of which are monitored separately. This approach led to in total 69 emulated trials, involving on average approximately 60 patients on, and 120 patients off, prophylaxis, with each individual being involved on average in 1.25 trials.

In addition, and as pointed out by Hernan and colleagues, individuals may start/stop taking prophylaxis at or following a scheduled follow-up visit *during* a contiguous period of eligibility, leading to the issue of having a "multiplicity of regimes" (Hernan *et al.*, 2006). In such cases, we censor these individuals from their respective treatment regime at this time, and simultaneously reallocate them to the trial starting in the following quarter on the other arm. Of course, this again assumes a negligible washout period for those stopping prophylaxis.

In summary, patients may be involved in more than one trial throughout the follow-up period, but are involved in only one trial at any specific time point.

Two example participants are shown in Figure 5.6.1.

- Patient a.) is a 58 year old European male who was first eligible for an emulated trial in the 3rd quarter of 2003, and at this time he was not taking PCP prophylaxis. He was followed up for the next three quarters until the end of the first quarter of 2004. At this point he was diagnosed with PCP (the event of interest), and follow-up was stopped for the primary endpoint. This patient was admitted into the respective cohort at the beginning of 1998 and died shortly after the PCP diagnosis was made in the 1st quarter of 2004. Therefore, in terms of the secondary endpoint of all-cause mortality, this patient also experienced the event in the same quarter (1st quarter of 2004).

- Patient b.) in Figure 5.6.1 is a 30 year old European male who was first eligible for an emulated trial in the 2nd quarter of 2008 (trial number 43), and at this time was taking PCP prophylaxis. He was followed up for the two quarters until the end of 2008, at which point he stopped taking prophylaxis, and was "artificially" censored in this quarter. In the first quarter of 2009 he was no longer eligible since his CD4 count was temporarily above 200 cells/$\mu$L (2009 to 2009.25). Following this in the 2nd quarter of 2009, he was again eligible in terms of CD4 count and viral suppression, and consequently was assigned to the off prophylaxis arm in the trial beginning in the 2nd quarter of 2009 (trial number 46). Follow-up continued in this new trial for 2 quarters, and at this point he was censored since he was no longer eligible — his CD4 count rose to above 200 cells/$\mu$L. This patient was admitted into the respective cohort in the 2nd quarter of 2007, and was followed up until the end of the study period in the 1st quarter of 2015.

a.) Patient with a PCP diagnosis

| patient | Quarter start | Quarter end | Calendar yr. trial starts | Trial number | Treatment group (0 = On, 1 = Off) | PCP diagnosis | Censoring indicator |
|---|---|---|---|---|---|---|---|
| 1 | 2003.5 | 2003.75 | 2003 | 23 | 1 | FALSE | 0 |
| 1 | 2003.75 | 2004 | 2003 | 23 | 1 | FALSE | 0 |
| 1 | 2004 | 2004.25 | 2004 | 23 | 1 | TRUE | 0 |

| patient | Gender | mode of transmission | geographical origin | Age at baseline | sqrt(CD4) count at baseline | sqrt(CD4) count at in quarter | $\log_{10}$ RNA at baseline | $\log_{10}$ RNA in quarter |
|---|---|---|---|---|---|---|---|---|
| 1 | M | MSM | Europe | 58.48871 | 11.40175 | 11.40175 | 1.69897 | 1.69897 |
| 1 | M | MSM | Europe | 58.48871 | 11.40175 | 11.40175 | 1.69897 | 1.69897 |
| 1 | M | MSM | Europe | 58.48871 | 11.40175 | 11.40175 | 1.69897 | 1.69897 |

b.) Patient in multiple trials, assigned to different arms following censoring

| patient | Quarter start | Quarter end | Calendar yr. trial starts | Trial number | Treatment group (0 = On, 1 = Off) | PCP diagnosis | Censoring indicator |
|---|---|---|---|---|---|---|---|
| 2 | 2008.25 | 2008.5 | 2008 | 43 | 0 | FALSE | 0 |
| 2 | 2008.5 | 2008.75 | 2008 | 43 | 0 | FALSE | 1 |
| 2 | 2009.25 | 2009.5 | 2009 | 46 | 1 | FALSE | 0 |
| 2 | 2009.5 | 2009.75 | 2009 | 46 | 1 | FALSE | 1 |

| patient | Gender | mode of transmission | geographical origin | Age at baseline | sqrt(CD4) count at baseline | sqrt(CD4) count at in quarter | $\log_{10}$ RNA at baseline | $\log_{10}$ RNA in quarter |
|---|---|---|---|---|---|---|---|---|
| 2 | M | IDU | Europe | 30.90486 | 13.37909 | 13.37909 | 1.69897 | 1.69897 |
| 2 | M | IDU | Europe | 30.90486 | 13.37909 | 13 | 1.69897 | 1.69897 |
| 2 | M | IDU | Europe | 31.81109 | 11.95826 | 11.95826 | 1.69897 | 1.69897 |
| 2 | M | IDU | Europe | 31.81109 | 11.95826 | 13.45362 | 1.69897 | 1.69897 |

Figure 5.6.1: Patient examples.

## 5.7 Statistical methods

### 5.7.1 Analysis model: Estimating the observational analogue of the per-protocol effect

As analysis model, we fitted a pooled logistic regression model to the "expanded data", with one record per patient per quarter, to estimate the hazard ratio comparing the risk of those on and off PCP prophylaxis.

This is exactly the same approach Hernan and colleagues adopt in many of their publications (for example, Hernan *et al.*, 2000; Danaei *et al.*, 2013). They orginally chose this approach as their software (SAS) did not support per subject time varying weights (page 561 of Hernan *et al.* (2000)), and sandwich based standard errors (personal communication at causal inference course, summer 2017) for the Cox proportional hazards model at that time.

However, as mentioned in Chapter 3, a parametric modelling approach provides advantages over the Cox PH model when multiply imputing data in the context of a sensitivity analysis. Since, in this case, we have a function for the baseline hazard we were able to avoid using an additional model to estimate the baseline hazard using, for example, the Kaplan-Meier product limit estimate. The pooled logistic model provides a reasonable approximation to the Cox proportional hazards model when the risk of an event is small in any particular time window (quarter in our trial set-up) (Thompson, 1977; Greene, 2003; Efron, 1988; D'Agostino *et al.*, 1990). More details of the pooled logistic model, and rationale for its approximating the Cox PH model are found in Appendixes I and J.

At the start of each emulated trial, each patient's baseline characteristics were derived, and adjusted for, in the analysis model. These were then held constant for all the quarters they remain in the study.

Therefore, to model the baseline hazard, we included a term for the "time" within the trial, along with its square and cubic terms. "time" is a continuous variable, measured in quarters, from the first to the final quarter of the specific trial. (Of course, a spline could also be fitted to estimate the baseline hazard). We then included the treatment indicator, and adjusted for the following (non-time varying variables for each patient within each trial) baseline covariates:

- gender (reference female),
- probable mode of transmission (with categories heterosexual (reference), intravenous drug use (IDU), men having sex with men (MSM), and "other"),
- geographical region of origin (with categories Europe (reference), Africa, Asia, Latin America and North Africa and Middle East),
- $\log_{10}$ baseline HIV RNA level (and its square),
- square root of CD4 count (and CD4 count),
- age (and its square),
- the calendar year in which the trial was started (to capture guideline/clinical practice changes since 1998),

- and finally, the quarter in which the trial was started (and its square) (e.g. 2011.25 for the second quarter of 2011).

Since patients may be involved in multiple emulated trials we calculated robust sandwich errors to account for intra-patient correlation (for example, see Enders *et al.* (2018)). Where CD4 measurements were not available for a patient in a particular quarter, they were estimated using a general additive model with a restricted cubic spline fitted for "time on cART", adjusted for age, gender, geographical origin mode of transmission and RNA. This approach was inspired by Caniglia *et al.* who use a similar approach in their analysis (Caniglia *et al.*, 2017). Missing RNA measurements were estimated in an analogous manner. This method has the same potential drawbacks of other single imputation methods, but was considered preferable to using last observation carried forward, which can be conservative or anti-conservative depending on the situation (refer to Chapter 1). An alternative would be to multiply impute the time varying covariates, which could be implemented using the method outlined in the recent paper by Keogh and Morris (2018).

We pooled data from all the person-periods from all emulated trials and fitted a single model. Again, this is the approach used by Hernan *et al.* in many of their publications.

There is an argument for reversing the order of the calculation, whereby we would fit a model to each emulated trial and then pool afterwards. We would assume that the results would be similar. However, analogous arguments applied to multiple imputation techniques (e.g. Schomaker and Heumann, 2018)) would point towards the approach we take leading to more reliable variance estimates.

An identical modelling approach was taken for the all-cause mortality secondary endpoint.

All analyses were carried out with R version 3.2.4 (R Core Team, 2017), using the function *svyglm* in package "survey" to calculate robust sandwich errors from logistic models. Throughout we used a level of 0.05 as statistically significant.

### 5.7.2 Inverse probability weighting to account for covariate dependent censoring

As described in Table 1, patients are censored for a variety of reasons:

- Patients are administratively censored at the end of the trial when the event of interest, PCP diagnosis for the primary analysis, and all-cause mortality for the secondary analysis, has not occurred before the end of the study period (i.e. 1st quarter 2015).
- Patients are censored when they are lost to follow-up, that is, when they have not had a follow-up visit for at least 12 months, and the end of the study period has not yet been reached.
- In the case of the primary endpoint of PCP diagnosis, patients are censored when they die (any cause).
- Analogously, for the secondary endpoint of all-cause mortality, patients are not censored when they have been diagnosed with PCP.
- Patients are censored then they no longer fulfil the eligibility criteria for the emulated trial. For example, if their CD4 count rises above 200 cells/$\mu$L and/or the HIV RNA level is above 400 copies/mL (refer to Table 1 for eligibility criteria) they are also censored.
- Finally, there are special cases of censoring particular to the trial emulation approach we have followed. Patients are censored when they no longer follow their assigned trial arm. For example, patients changing from on to off PCP prophylaxis, or visa versa, are also censored.

To differentiate the last two types of censoring from the others, we refer to these as "artificial" censoring.

For the analysis model defined in the previous section, had we assumed censoring depends only on a patient's baseline covariates at the start of each trial, then we would not need any further adjustment for censored patients. However, since we suspect that censoring depends on time varying variables, for example, it is well established that lower CD4 count is associated with disease progression and mortality, subject based inverse probability weights per quarter were added to the model to adjust for the potential time varying selection bias *from censoring*.

By focussing specifically on individuals on cART, those adhering to the conditions required for PCP prophylaxis according the guidelines, and censoring patients when they deviate from their assigned treatment arm, we have "broken" the treatment-confounder time-based feedback loop. This means that we do not need to include additional inverse probability weights (IPW) to adjust for this type of confounding, which simplifies the approach somewhat. In the words of Hernan *et al.*:

> "Artificially censoring individuals when they deviate from one of the two regimes of interest [as we do here]... the treatment variable is effectively forced to be non time-varying - as soon as it varies, the person is censored there is no time-varying confounding and therefore no need for IPW or g-estimation to appropriately adjust for such confounding. However, the censoring itself may introduce time-dependent selection bias and IPW is therefore needed to adjust for such bias", Hernan *et al.* (2006).

Since all forms of censoring could potentially introduce time varying selection bias (Caniglia *et al.*, 2017), we do not differentiate between the different types in the calculation of the weights.

The following simple steps, paraphrased from Hernan *et al.* were used to account for potential selection bias using IP weighting (Hernan *et al.*, 2006).

1. We define two regimes of interest — on and off PCP prophylaxis in this case.
2. We artificially censor individuals when they stop following their assigned regime.
3. We estimate inverse probability weights to adjust for the possibility of informative censoring in the previous step.
4. We compare the survival of the uncensored individuals under each regime of interest in a a weighted analysis adjusted for baseline covariates.

We now illustrate the concept behind the inverse probability weights by using a simple example (copied verbatim) from Hernan *et al.* (2004): *"To adjust for selection bias due to non-administrative censoring, we use inverse probability weighting (IPW). The idea behind IPW is*

*to assign a weight to each selected subject so that she accounts in the analysis not only for herself but also for those with similar characteristics that were censored. The weight is the inverse of the probability of being uncensored. For example, if there are four untreated women, aged 40-45, with CD4 count $>$ 500 in our study, and three of them were lost to follow-up, then these three women do not contribute to the analysis (i.e. they receive zero weight) while the remaining woman receives a weight of four. In other words, the (estimated) conditional probability of remaining uncensored until the end of the study is 1/4 = 0.25, and therefore the (estimated) weight for the uncensored subject is 1/0.25 = 4. IPW creates a hypothetical population where the four subjects of the original population are replaced by four copies of the uncensored subject"*, thus creating a pseudo-population, where the representation is as desired.

For our study, we include a weight for each patient into the analysis model which is inversely proportional to the conditional probability of such a patient remaining uncensored until the end of the particular quarter. This is slightly more complex than the simple example quoted above, since in our analysis there are a number of covariates, some of which are continuous. We calculate the weights by fitting a logistic model with the censoring indicator as dependent variable, and independent variables the relevant covariates. Appendix K describes this procedure in more detail, with a patient example and code. If we then fit the model including these weights we have created a pseudo-population in which censoring has effectively been eliminated.

Therefore, to create a situation without censoring, we weighted each patient at each quarter by the inverse of having their observed history using stabilised inverse probability weights, which have a numerator and denominator part. To estimate the numerator weight, we calculate the inverse fitted probabilities from a logistic regression where the outcome was the censoring indicator at the particular quarter (column "censoring indicator" in Figure K.2.1), and the independent variables were the patient's baseline covariates: the square root of the CD4 count (and its square), $\log_{10}$ of the RNA (and its square), gender, mode of transmission, geographical origin, age (and its square), the number of the trial (and its square), and the calendar year in which the trial started.

To estimate the denominator weights, we calculate the inverse fitted probabilities from a logistic regression where, once again, the outcome is whether the patient was censored at the end of the quarter, but the covariates are now i.) the baseline covariates (as above) with in addition ii.) time updated (at the start of the quarter) values of $\sqrt{\text{CD4}}$ count and $\log_{10}$ HIV RNA level. Thus, for the first quarter that each patient was in a trial, their IP weight was 1 (consistent with

the assumption that in the first quarter, censoring is at random given the baseline covariates).

Stabilised weights were then constructed by dividing the numerator weights by the denominator weights. Finally, we truncated the stabilised weights at the 99th percentile to avoid including very large weights (Cole and Hernan, 2008). Again, more details and example R code are in Appendix K.

In summary, fitting the analysis model of the previous section, then including the IP weighting per subject and trial is the same as fitting this model to a pseudo-population in which censoring has been removed. This establishes the rationale for the analysis providing results which provide statistically valid inference, albeit with the assumption that there is no unmeasured confounding.

### 5.7.3 Sensitivity analysis

For the sake of our illustrative example, we hypothesise that those lost to follow-up in the off prophylaxis arm might be informatively censored. Therefore, it seems reasonable to investigate plausible departures from the censoring at random (CAR) assumption for this subgroup of patients. We assumed that censoring was at random (CAR) for all other censored patients. We also note at this point that the primary analysis using the IP weighting implicitly assumes CAR for all types of censoring.

The sensitivity analysis scenario was defined as follows: For the subgroup of patients on the off prophylaxis arm which were censored due to being lost to follow-up, we made the clinically plausible assumption that they started PCP prophylaxis at the time point at which the censoring occurred i.e. they "jumped" to the on prophylaxis arm in the respective emulated trial.

To model this post-censoring behaviour, we adopted the "Jump to Reference" sensitivity analysis approach for time-to-event data we came across in previous chapters. To briefly recap, under J2R the hazard for patients lost to follow-up for the off prophylaxis arm is constructed from the pre-censoring hazard for these patients, and the post-censoring hazard is assumed to be that from those on prophylaxis (refer to Figure 5.7.1). More details of the implementation including code can be found in Appendix L.

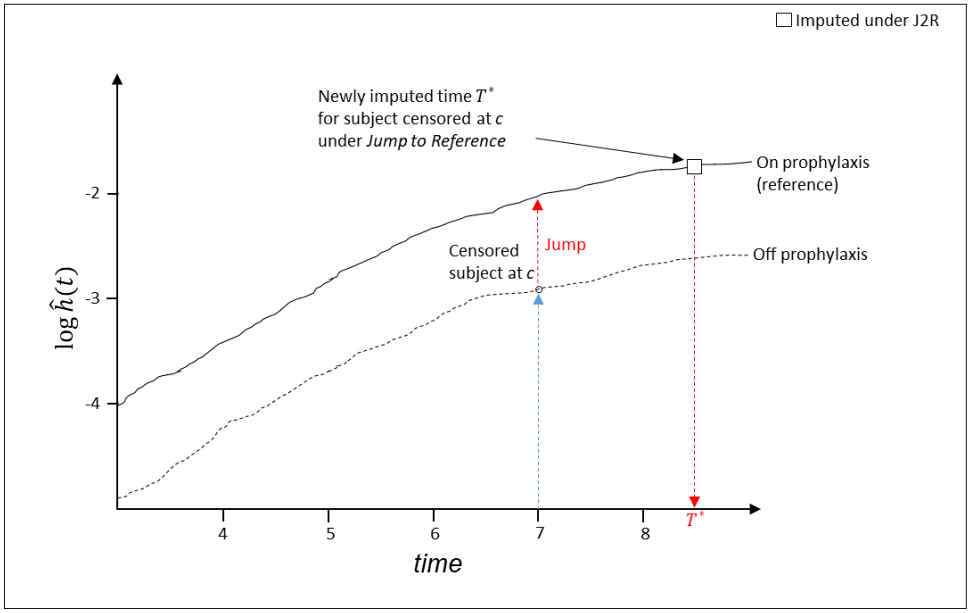


Figure 5.7.1: Schematic illustration of "Jump to Reference"

## 5.8 Results

### 5.8.1 Clinical endpoints

There were 9,743 patients complying with the conditions for the emulated trials with a total of 18,550 person years followed-up during 1998-2015 (median 0.8 yrs per patient per trial, IQR [0.3, 2.4]). The unadjusted incidence rate of PCP diagnosis was 1.5 (95% confidence interval [0.7, 2.7]) per 1000py on PCP prophylaxis compared to 2.8 [1.8, 4.0] off PCP prophylaxis.

Whereas for the secondary endpoint of all-cause mortality, the unadjusted incidence rate was 97.6 [90.6, 105.0] for those on PCP prophylaxis versus 84.3 [78.7, 90.1] off PCP prophylaxis per 1000py.

With PCP diagnosis as endpoint, and fitting the model using IP weights, the hazard ratio (HR) for those off versus on prophylaxis following adjustment for $\sqrt{CD4}$ and CD4, $\log_{10}$ RNA and $\log_{10}$ RNA$^2$, gender, age, age$^2$, transmission mode, geographical origin, calendar year at baseline, trial number and its square, and time, time$^2$ and time$^3$ was 1.24 (95% confidence interval [0.49, 3.15], p = 0.65, refer to the far right columns in Table 5.8.1 and Figure 5.8.1 for details).

For the secondary endpoint of all-cause mortality, in multivariable models being off PCP prophylaxis was associated with lower mortality (0.83 [0.75, 0.91], $p < 0.001$, refer to Table 5.8.2 and Figure 5.8.2 for details).

The results from all the fitted models are summarised in Table 5.8.3 and Figure 5.8.3.

| Variable | i.) Univariable | | ii.) Multivariable | | iii.) Multivariable (IP weighted) | |
|---|---|---|---|---|---|---|
| | Hazard Ratio | p-value | Hazard Ratio | p-value | Hazard Ratio | p-value |
| Prophylaxis | | | | | | |
| On | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| Off | 1.47 [0.70, 3.08] | 0.31 | 1.52 [0.70, 3.28] | 0.29 | 1.24 [0.49, 3.15] | 0.65 |
| Gender | | | | | | |
| Female | 1.0 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| Male | 0.70 [0.35, 1.38] | 0.31 | 0.63 [0.30, 1.36] | 0.24 | 0.68 [0.31, 1.47] | 0.32 |
| Transmission mode | | | | | | |
| Heterosexual | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| IDU | 2.17 [1.01, 4.64] | 0.05 | 2.46 [1.10, 5.32] | 0.03 | 2.46 [1.06, 5.67] | 0.04 |
| MSM | 0.79 [0.32, 1.97] | 0.62 | 1.03 [0.37, 2.85] | 0.96 | 1.03 [0.37, 2.87] | 0.96 |
| Other | 0.39 [0.05, 2.96] | 0.36 | 0.62 [0.08, 4.99] | 0.66 | 0.43 [0.05, 3.51] | 0.43 |
| Geographical origin | | | | | | |
| Europe | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| Africa | 0.41 [0.10, 1.71] | 0.22 | 0.59 [0.11, 3.09] | 0.53 | 0.51 [0.11, 2.71] | 0.43 |
| Asia | nE | | nE | | nE | |
| Latin America | 1.32 [0.39, 4.46] | 0.65 | 2.21 [0.64, 7.61] | 0.21 | 1.79 [0.51, 6.27] | 0.36 |
| North Africa & Middle East | nE | | nE | | nE | |
| Age at start of trial (yrs) | 0.99 [0.95, 1.02] | 0.44 | 0.89 [0.74, 1.07] | 0.23 | 0.90 [0.75, 1.08] | 0.24 |
| Age$^2$ at start of trial (yrs) | - | - | - | - | - | - |
| $\sqrt{\text{CD4}}$ | 0.85 [0.78, 0.93] | $< 0.001$ | 0.83 [0.62, 1.13] | 0.24 | 0.83 [0.60, 1.14] | 0.25 |
| CD4 | - | - | - | - | - | - |
| $\log_{10}$ RNA | 0.61 [0.39, 0.94] | 0.02 | 0.63 [0.36, 1.10] | 0.11 | 0.43 [0.24, 0.95] | 0.03 |
| $\log_{10}$ RNA$^2$ | - | - | 1.14 [0.82, 1.10] | 0.11 | 1.25 [0.87, 1.78] | 0.23 |
| Calender year at start of trial | 0.99 [0.92, 1.05] | 0.66 | 0.37 [0.14, 1.01] | 0.05 | 0.32 [0.11, 0.95] | 0.04 |
| Time | 0.37 [0.21, 0.62] | $< 0.001$ | 0.36 [0.22, 0.61] | $<$0.001 | 0.35 [0.21, 0.59] | $<$0.001 |
| Time$^2$ | 1.10 [1.03, 1.18] | 0.005 | 1.10 [1.03, 1.17] | 0.003 | 1.10 [1.03, 1.18] | 0.004 |
| Time$^3$ | 1.00 [1.00, 1.00] | 0.02 | 1.00 [1.00, 1.00] | 0.01 | 1.00 [1.00, 1.00] | 0.02 |
| Trial | 1.00 [1.00, 1.00] | 0.02 | 1.33 [1.03, 1.72] | 0.03 | 1.38 [1.05, 1.81] | 0.02 |
| Trial$^2$ | - | - | - | - | - | - |

Table 5.8.1: Estimates from fitting a pooled logistic regression model for the primary analysis considering the hazard ratio of being off versus on PCP prophylaxis; i.) univariate models, ii.) multivariable adjusted model and iii.) multivariable adjusted model with IP weighting; 95% confidence intervals are shown in brackets, with robust standard errors used to adjust for patient correlation.

nE not estimable; IP Inverse probability; IDU Intravenous drug users; MSM Men having sex with men

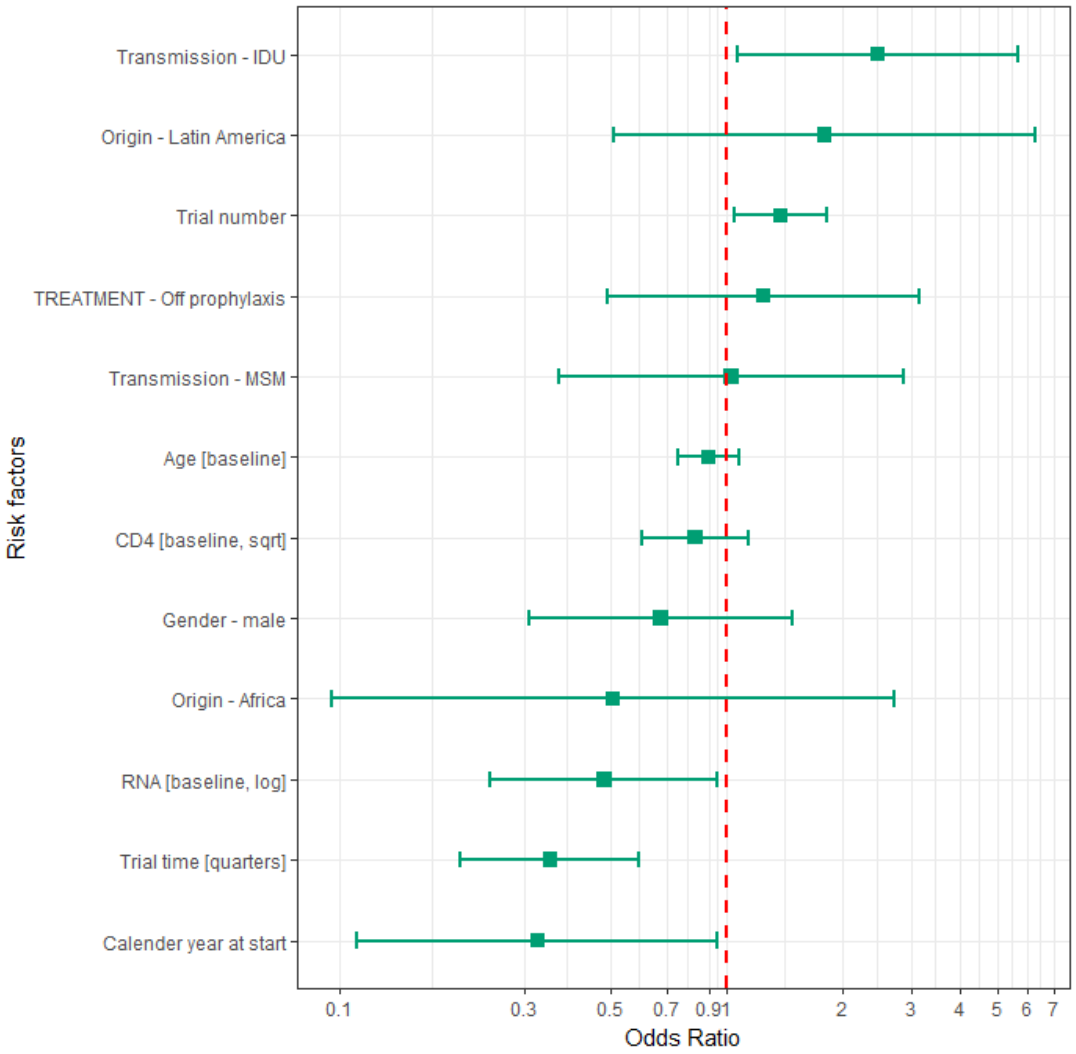


Figure 5.8.1: Adjusted hazard ratios (HR) for the PCP diagnosis primary endpoint including inverse probability weights; refer to Tables 5.8.1 and 5.8.3 for more details.

Estimates for CD4, $\log_{10}\text{RNA}^2$, $\text{age}^2$, $\text{trial}^2$, $\text{time}^2$, $\text{time}^3$ not shown.

| Variable | i.) Univariable | | ii.) Multivariable | | iii.) Multivariable (IP weighted) | |
|---|---|---|---|---|---|---|
| | Hazard Ratio | p-value | Hazard Ratio | p-value | Hazard Ratio | p-value |
| Prophylaxis | | | | | | |
| On | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| Off | 0.76 [0.69, 0.81] | <0.001 | 0.82 [0.74, 0.90] | <0.001 | 0.83 [0.75, 0.91] | <0.001 |
| Gender | | | | | | |
| Female | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| Male | 1.36 [1.14, 1.61] | 0.004 | 1.02 [0.86, 1.22] | 0.81 | 1.03 [0.85, 1.23] | 0.79 |
| Transmission mode | | | | | | |
| Heterosexual | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| IDU | 1.84 [1.55, 2.18] | <0.001 | 1.95 [1.63, 2.35] | <0.001 | 1.94 [1.60, 2.36] | <0.001 |
| MSM | 1.58 [1.32, 1.89] | <0.001 | 1.32 [1.09, 1.59] | 0.005 | 1.29 [1.05, 1.58] | 0.02 |
| Other | 1.36 [0.96, 1.94] | 0.09 | 1.35 [0.97, 1.88] | 0.08 | 1.34 [0.96, 1.88] | 0.09 |
| Geographical origin | | | | | | |
| Europe | 1.00 (reference) | | 1.00 (reference) | | 1.00 (reference) | |
| Africa | 0.37 [0.26, 0.51] | <0.001 | 0.69 [0.49, 0.99] | 0.04 | 0.64 [0.45, 0.91] | 0.01 |
| Asia | 0.47 [0.20, 1.11] | 0.09 | 0.66 [0.28, 1.59] | 0.36 | 0.62 [0.26, 1.48] | 0.29 |
| Latin America | 0.86 [0.60, 1.23] | 0.41 | 1.33 [0.97, 1.84] | 0.08 | 1.32 [0.94, 1.83] | 0.11 |
| North Africa & Middle East | 0.80 [0.48, 1.33] | 0.39 | 0.92 [0.57, 1.48] | 0.72 | 0.97 [0.61, 1.54] | 0.90 |
| Age at start of trial (yrs) | 1.03 [1.02, 1.04] | <0.001 | 1.06 [1.01, 1.10] | 0.01 | 1.05 [1.01, 1.10] | 0.02 |
| $\text{Age}^2$ at start of trial (yrs) | - | - | - | - | - | - |
| $\sqrt{\text{CD4}}$ | 0.97 [0.95, 0.99] | 0.005 | 0.89 [0.82, 0.98] | 0.01 | 0.89 [0.81, 0.98] | 0.02 |
| CD4 | - | - | - | - | - | - |
| $\log_{10}$ RNA | 1.17 [1.07, 1.27] | $< 0.001$ | 1.28 [1.08, 1.51] | 0.005 | 1.26 [1.06, 1.50] | 0.009 |
| $\log_{10}\text{RNA}^2$ | 0.96 [0.89, 1.02] | 0.18 | 0.93 [0.87, 0.99] | 0.03 | 0.92 [0.86, 0.99] | 0.02 |
| Calender year at start of trial | 0.93 [0.92, 0.94] | $< 0.001$ | 0.98 [0.82, 1.17] | 0.83 | 0.98 [0.82, 1.17] | 0.80 |
| Time | 0.92 [0.89, 0.95] | $< 0.001$ | 0.90 [0.88, 0.93] | <0.001 | 0.90 [0.87, 0.93] | <0.001 |
| $\text{Time}^2$ | 1.00 [1.00, 1.00] | 0.01 | 1.10 [1.00, 1.00] | 0.002 | 1.00 [1.00, 1.01] | 0.002 |
| $\text{Time}^3$ | - | - | 1.00 [1.00, 1.00] | 0.05 | 1.00 [1.00, 1.00] | 0.07 |
| Trial | 0.98 [0.98, 0.99] | <0.001 | 1.00 [0.96, 1.04] | 0.92 | 1.00 [0.96, 1.04] | 0.97 |
| $\text{Trial}^2$ | 1.00 [1.00, 1.00] | 0.05 | 1.00 [1.00, 1.00] | 0.01 | 1.00 [1.00, 1.00] | <0.02 |

Table 5.8.2: Estimates from fitting a pooled logistic regression model for the secondary analysis considering all-cause mortality as endpoint; hazard ratio of being off versus on PCP prophylaxis; i.) univariate models, ii.) multivariable adjusted model and iii.) multivariable adjusted model with IP weighting; 95% confidence intervals are shown in brackets, with robust standard errors used to adjust for patient correlation.

nE not estimable; IP Inverse probability; IDU Intravenous drug users; MSM Men having sex with men

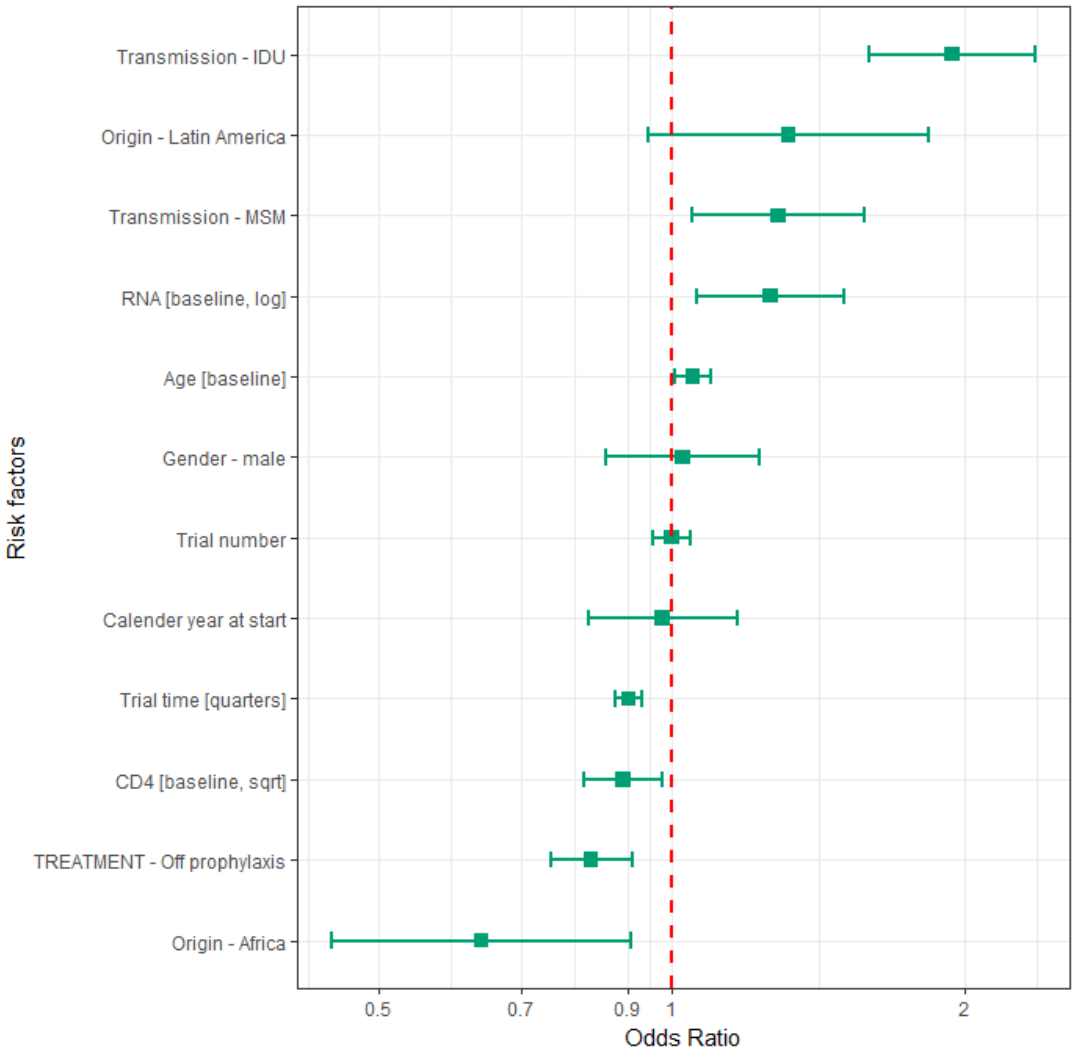


Figure 5.8.2: Adjusted hazard ratios (HR) for the all-cause mortality secondary endpoint including inverse probability weights; refer to Tables 5.8.1 and 5.8.3 for more details.

Estimates for CD4, $\log_{10}\text{RNA}^2$, $\text{age}^2$, $\text{trial}^2$, $\text{time}^2$, $\text{time}^3$ not shown.

| Endpoint | Model | Treatment estimate | 95% confidence interval | p-value | Number of events |
|---|---|---|---|---|---|
| PCP diagnosis | 1. Unadjusted / no IPW | 1.61 | [0.80, 3.24] | 0.19 | 39 (11 on / 28 off) |
| PCP diagnosis | 2. Unadjusted / IPW | 1.47 | [0.70, 3.08] | 0.31 | 39 |
| PCP diagnosis | 3. Adjusted / no IPW | 1.52 | [0.70, 3.28] | 0.29 | 39 |
| PCP diagnosis | 4. Adjusted / IPW | 1.24 | [0.49, 3.15] | 0.65 | 39 |
| All-cause mortality | 5. Unadjusted / no IPW | 0.75 | [0.68, 0.81] | $< 0.001$ | |
| All-cause mortality | 6. Unadjusted / IPW | 0.76 | [0.69, 0.83] | $< 0.001$ | 1581 (725 on, 856 off) |
| All-cause mortality | 7. Adjusted / no IPW | 0.82 | [0.74, 0.90] | $< 0.001$ | 1581 |
| All-cause mortality | 8. Adjusted / IPW | 0.83 | [0.75, 0.91] | $< 0.001$ | 1581 |
| All-cause mortality | 9. Multiple imputation under CAR Adjusted | 0.87 | [0.79, 0.95] | 0.002 | 1581+89 |
| All-cause mortality | 10. Sensitivity analysis under CNAR Adjusted | 0.89 | [0.81, 0.97] | 0.01 | 1581+290 |

Table 5.8.3: Results summary - hazard ratio estimates for each of the fitted models for endpoints (PCP diagnosis and all-cause mortality), model (unadjusted or adjusted, with and without using inverse probability weighting [IPW]), multiple imputation (under censoring at random [CAR], or censoring not at random [CNAR]); number of events on both arms in the far right column, average number of multiply imputed events indicated after the "+".

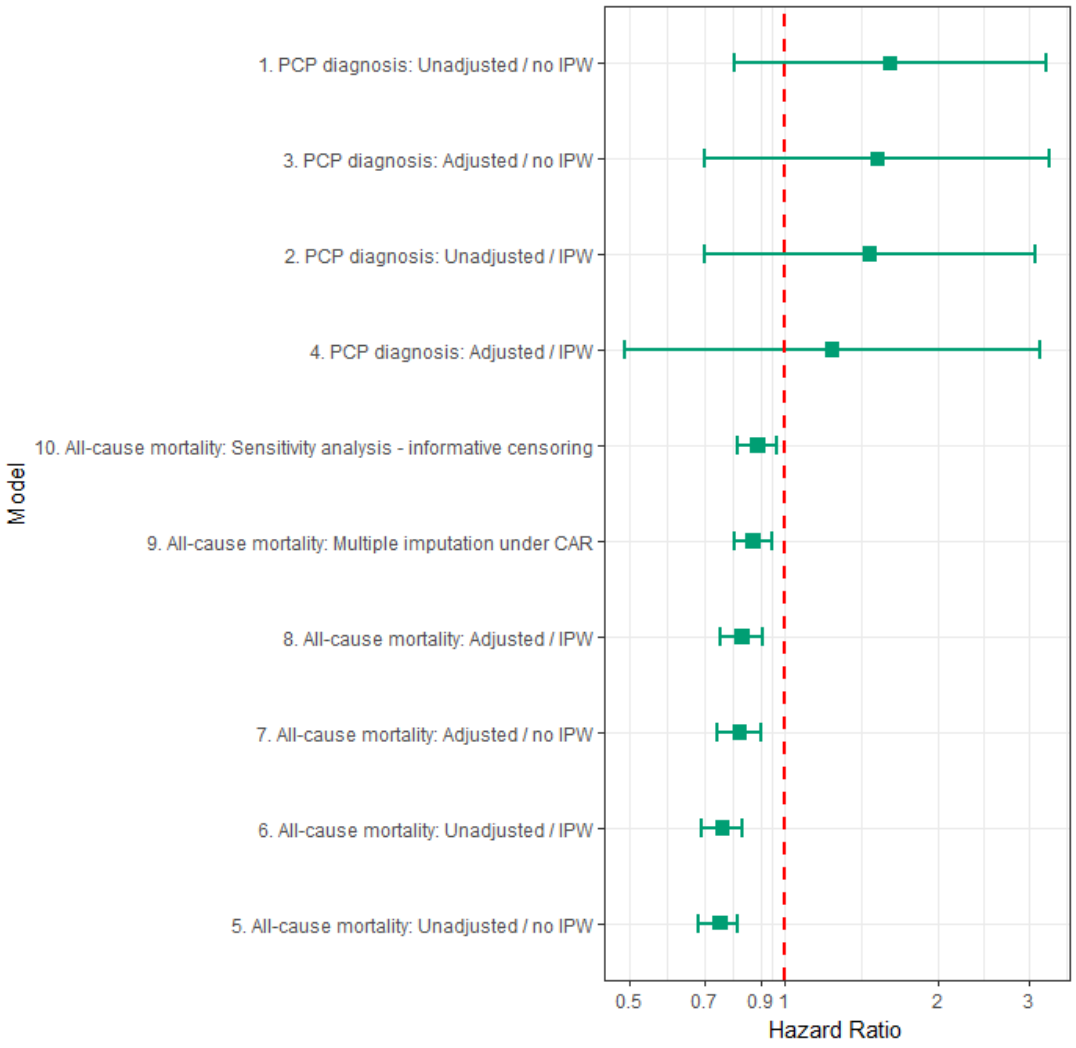


Figure 5.8.3: Hazard ratios (HR) for endpoints PCP diagnosis and all-cause mortality; HR $< 1$ indicates risk is lower off PCP prophylaxis compared to on prophylaxis; refer to Tables 5.8.1, 5.8.2 and 5.8.3 for more details.

Inverse probability weighted model adjusted for baseline $\sqrt{\text{CD4}}$, CD4, $\log_{10}$RNA, $\log_{10}\text{RNA}^2$, gender, age, age$^2$, transmission mode, geographical origin, calendar year at baseline, trial, trial$^2$, time, time$^2$, time$^3$.

### 5.8.2 Sensitivity analysis to investigate informative censoring

We illustrated the proposed sensitivity analysis approach using the analysis with all-cause mortality as endpoint. We compared the hazard ratio assuming censoring at random, with that for the sensitivity analysis in which we used the "Jump to Reference" (J2R) approach. We multiply imputed new events using J2R for the subgroup of 406 patients (6.9%) censored due to being lost to follow-up on the off prophylaxis arm (median time to censoring 0.75 yrs, IQR [0.50, 1.25]). All other censored patients were assumed to be censored at random.

In the previous section, the hazard ratio assuming CAR, estimated using the IP weighting method, was 0.83 with 95% confidence interval [0.75, 0.91], ($p < 0.001$, refer to the results from Model 8 in Table 5.8.3). Following multiple imputation assuming CAR, as expected we arrived at similar results with the hazard ratio being 0.87 [0.79, 0.95], p=0.002.

When we apply the post-censoring J2R behaviour to those not taking PCP prophylaxis that are lost to follow-up, the hazard ratio attenuates to 0.89 [0.81, 0.97] ($p = 0.01$). This is exactly what we might expect to happen since the multiply imputed patients now have event times consistent with the hazard of the "on prophylaxis" arm, creating more homogeneity between the patients in both arms. Once again, the dilution or mixing effect seen in previous chapters is also apparent in this example application.

Figure 5.8.4 shows the estimated cumulative hazard curves for patients on both arms. The solid blue line indicates the marginal cumulative hazard for patients taking prophylaxis assuming censoring was at random. The solid pink line shows the marginal cumulative hazard for patients not taking prophylaxis, again assuming CAR.

The dot-dashed pink line shows the marginal cumulative hazard of the patients not taking prophylaxis, including the event times which have been multiply imputed assuming post-censoring behaviour modelled as J2R for those lost to follow-up. Consistent with the decrease in the hazard ratio noted in the previous paragraph, the cumulative hazard lines now converge discernibly, with the 95% confidence interval (shaded pink) overlapping slightly with the cumulative hazard curve for those taking prophylaxis under the primary analysis assumption of CAR (blue solid line).

This example sensitivity analysis reveals that, despite a rather extreme scenario of informative censoring for those off prophylaxis but lost to follow-up, the results still broadly support the

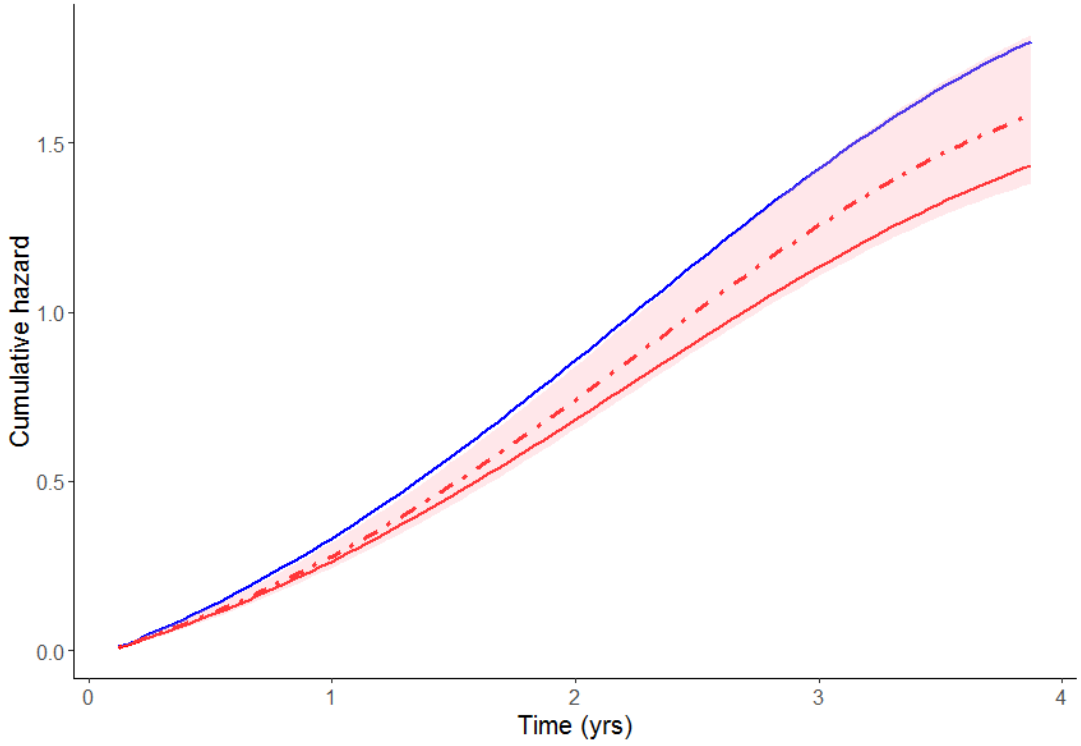


Figure 5.8.4: Comparison of those on (blue) and off (red) PCP prophylaxis under CAR (solid); Sensitivity analysis for scenario "Jump to PCP prophylaxis" shown in red, dot-dashed; 95% confidence interval for "Jump to PCP prophylaxis" is shaded pink.

outcome from the analysis of the secondary endpoint, that there is a significant difference between those on and off PCP prophylaxis in terms of all-cause mortality. As we expected, the effect has been attenuated due to dilution of the hazard on the off prophylaxis arm, but not to the extent that the findings have been overturned the p-value remains significant at the 5% level.

### 5.8.3 Subgroup analyses

As mentioned in section 5.5, we also carried out two subgroup analyses to investigate whether our more general definition for the emulated trial, with those on prophylaxis being compared to those off prophylaxis, provides different results to an emulated trial in which the target trial is followed more closely. We originally chose the more general definition for the emulated trial to reach an adequate number of PCP diagnosis events to power the study.

For the first subgroup analysis, Trial A, we estimated the treatment hazard ratio for the case in which all patients are eligible and *start on prophylaxis*. We then compare those continuing with prophylaxis with those that stop prophylaxis. We allowed patients to stop taking prophylaxis within three months either side of the point of eligibility, that is, when their CD4 $\leq$ 200 and RNA $\leq$400, and they were on prophylaxis.

For the second subgroup analysis, Trial B, we estimated the treatment hazard ratio for the case in which all patients are eligible and are *not taking prophylaxis*. We then compare those continuing to not take prophylaxis with those that start prophylaxis, again allowing a 3 month window either side of eligibility.

The results are shown in Table 5.8.4 and are broadly in line with those from the primary analysis results with PCP diagnosis as endpoint. Due to the more restricted trial definition, there were fewer events that in the main analysis (39 vs 24 or 33), leading to reduced power, and correspondingly wider confidence intervals.

| Endpoint | Model | Treatment estimate | 95% confidence interval | p-value | Number of events |
|---|---|---|---|---|---|
| Trial A | | | | | |
| PCP diagnosis | 1. Unadjusted / no IPW | 2.71 | [0.97, 7.54] | 0.06 | 24 (11 on / 13 off) |
| PCP diagnosis | 2. Unadjusted / IPW | 1.60 | [0.63, 4.06] | 0.32 | 24 |
| PCP diagnosis | 3. Adjusted / no IPW | 1.58 | [0.63, 3.98] | 0.33 | 24 |
| PCP diagnosis | 4. Adjusted / IPW | 1.99 | [0.61, 6.43] | 0.25 | 24 |
| Trial B | | | | | |
| PCP diagnosis | 1. Unadjusted / no IPW | 1.62 | [0.59, 4.46] | 0.35 | 33 (8 off / 25 on) |
| PCP diagnosis | 2. Unadjusted / IPW | 2.99 | [0.99, 9.04] | 0.05 | 33 |
| PCP diagnosis | 3. Adjusted / no IPW | 1.44 | [0.47, 4.46] | 0.52 | 33 |
| PCP diagnosis | 4. Adjusted / IPW | 2.82 | [0.80, 9.94] | 0.11 | 33 |

Table 5.8.4: Results summary - Hazard ratio estimates for each of the fitted models for the PCP diagnosis endpoint for Trial A (continue PCP prophylaxis vs stop ) and Trial B (No PCP prophylaxis vs starting)

## 5.9 Summary

HIV replication is a major risk factor for primary PcP. In virologically suppressed patients, irrespective of CD4 levels, the risk of PcP appears not to depend significantly on whether the individual is on or off prophylaxis. This suggests that primary PcP prophylaxis might be withheld in this patient group.

That being off prophylaxis was associated with lower all-cause mortality is intriguing, and probably indicates the presence of unmeasured confounding. For example, comorbidities were not collected systematically, and this might mean those allocated to a treatment arm in the emulated trial might be more or less sick than those in the other arm. Treatment non-adherence by the patient, or the physician not prescribing according to guidelines (for whatever reason), might lead to over or under re-reporting of those on and off prophylaxis. The presence of undiagnosed PCP at the time of a visit would also lead to systematic difference at the point of randomisation in our emulated trials.

Of course, the lower all-cause mortality in the off prophylaxis group might not be due to unmeasured confounding. It could also point towards a negative effect of long-term unnecessary exposure to antimicrobial drugs, most of them interfering with folate-metabolism (e.g. Bactrim toxicity). Alternatively, prolonged use of antibiotics has been shown to alter the microbiome, and this might have long term impact on patient health. COHERE does not document all co-medications, and it cannot be ruled out that this is adversely affecting this outcome. Furthermore, the nature of the all-cause mortality endpoint itself means that it is difficult to define an unambiguous causal argument between the presence or absence of prophylaxis and such a composite endpoint, which is itself dependent on the age profile and underlying condition of patients in the different cohorts. Finally, from a methodological point of view, the time horizon for the trial emulation approach was different between the primary endpoint (4 years) and all-cause mortality (all patient follow-up). The latter is not a realistic endpoint for a real trial, and therefore it seems questionable if it is appropriate for an emulated trial.

In terms of more general limitations associated with the COHERE data, we assume that taking prophylaxis is systematically recorded in all cohorts, and there is no under reporting in the database. For example, we assume that those defined as taking prophylaxis are on, and adhering to, treatment, and those recorded as not taking prophylaxis are indeed not taking prophylaxis, and that this has simply been forgotten on the information from the respective visit.

Notwithstanding these limitations, from a methodological perspective we demonstrated that reference-based imputation can be used to investigate possible informative censoring in an observational data setting, and can be combined with existing causal inference methods, which are often considered the gold standard for analyses with such data.

Whilst manipulation of the inverse probability weights is just as straightforward to implement as, for example, the Jump to Reference approach, the latter avoids discussion of the appropriate $\delta$-multiplier for the weights, and is also likely to be approximately information anchored, a desirable property for sensitivity analysis methods. We are of the opinion that this again recommends reference based methods in terms of both their *practicality* and *clinical plausibility*.

# Chapter 6

# Discussion

## 6.1 Sensitivity analysis for time-to-event data

Given the prominent role of sensitivity analysis in the analysis of clinical trials, not least exemplified by the proposed ICH E9 addendum published in 2018 (CHMP, 2018), it is important to provide methods which are not only easy to implement and use, but which are also clinically plausible and contextually relevant to the trial team and other stakeholders. However, an FDA mandated report by the US National Research Council in 2010 highlighted the lack of sensitivity analysis methods involving time-to-event data with just these properties.

Reference based methods have been developed and well received for continuous data. It seems natural to extend this approach to the time-to-event setting. Therefore, the overall aim of the PhD has been to propose, develop methods for, and critically evaluate, reference based sensitivity analysis approaches for time-to-event data.

In Chapter 2 we proposed a number of reference based approaches for survival data, and considered the consequences they have for the proportional hazards assumption, which is often used in survival analysis. In Chapter 3 we homed in on the most practically applicable of these methods — "Jump to Reference" — in the context of the RITA-2 trial, and introduced the concepts behind the principle of "information anchoring", alongside presenting simulation studies exploring the information anchoring properties of this approach. Chapter 4 builds on the empirical evidence from simulation results, presenting some theory proving that, to a good approx-

imation, information anchoring holds under a specific, analytically tractable working model. Finally, Chapter 5 applies the "Jump to Reference" method to a causal analysis of our HIV cohort data using an "emulated" trial method — so indicating the potentially wide applicability of the approach.

## 6.2 Reference based sensitivity analysis using multiple imputation

In Chapter 2, we showed that instead of the analyst specifying a (potentially large number) of sensitivity parameters, censored values can be imputed "by reference" to other groups of patients. For example, patients in the active arm may be imputed "by reference" to those in the control arm. We started by providing time-to-event analogues of the sensitivity analysis methods developed by Carpenter *et al.* (Jump to Reference, Last Mean Carried Forward/Hazard Carried Forward, Copy Increments in Reference, Copy Reference), and extended these with some new approaches (Immediate Event, Hazard Increases/Decreases to extremes, Hazard Tracks Back). The attraction of such Class-2 sensitivity analysis methods is that they are accessible, that is, they are both simple to understand and straightforward to implement. Taking such an approach avoids the alternative, as in Class-1 methods, in which we would have to explicitly model the event *and* censoring process, which is often rather complex to achieve in practice, even for experts in the field.

We demonstrated the applicability of these Class-2 methods in terms of both their practicality, that is, their ease of implementation and use, and their clinical plausibility, which we defined to be the ability to contextualise them to the trial team. The results from simulation studies and application to the GBC data suggested that "Jump to Reference" exhibited the most obvious utility in the time-to-event setting. The other methods either required the definition of some kind of sensitivity analysis parameter, which increases complexity since this then has to be defined and verified with the trial team, or in many settings are unlikely to provide significantly different results from imputing under censoring at random. This observation is not *per se* a drawback of the methods themselves in the sense that they generally are not suitable for investigating departures from CAR. Rather, it reflects our experience with the GBC data and similar studies. They may well be appropriate in other scenarios.

In light of the results in our chosen setting, we decided to focus on the "Jump to Reference" approach for the further analyses in later chapters. Reference based sensitivity analysis methods offer the natural advantages associated with pattern mixture modelling. It is possible, although we did not demonstrate this, and would be perhaps rather complex, to define a separate hazard for each *type* of censoring. For example, different assumptions for the hazard could be used for those dropping-out due to different adverse effects, or based on the type of non-random intervention in a pragmatic trial of the sort we encountered in the RITA-2 study. The inherent flexibility of these methods means that we can accommodate a wide range of contexts. Having said this, in any particular study, we need to seek the simplest approach that is sufficient (that is, the principle of parsimony continues to apply), avoiding the temptation these methods provide for spurious complexity.

Notwithstanding the flexibility of the approach, a potential limitation of the methods is that when there are low numbers of subjects in a trial and higher levels of censoring (i.e. over 40% on an arm), the basis for estimating the hazard in an arm becomes less certain. Similarly, when one (or both) arms of a trial has very few events then estimation of the hazard is made more difficult due to this lack of information. In such cases, our methods can of course still be applied but caution must be exercised when interpreting the results. Such situations would, in any case, generally raise concerns as to the appropriateness of using both the proportional hazards assumption, and perhaps applying survival analysis methods. Both of the above potential limitations are arguably unlikely in a well designed study.

In this thesis we use proportional hazards models, both for the primary analysis model and for imputing censored event times. If we suppose that proportional hazards holds for the primary analysis under CAR, then any sensitivity analysis making the CNAR assumption will blend hazards on one or more arms, and therefore will, at least strictly speaking, contravene the assumption of proportional hazards. However, we see this as a positive feature of our sensitivity analysis approach in the sense that proportional hazards cannot strictly hold for meaningful departures from CAR. The challenge in moving away from proportional hazards is not so much computational, as interpretational, as there is, by definition, no single number summarising the difference between the groups. Using the restricted mean survival time is an increasingly popular alternative, though this requires agreement on the "event horizon", and, as with proportional hazards models, also averages the treatment effect over the chosen time period.

We focussed on the "Jump to Reference" approach from Chapter 3 onwards, but there are a

number of ways this general approach can be used. For example, a reviewer of the manuscript in which we summarise the analysis of RITA-2 rightly pointed out that the reality of patient "Jumping to PTCA" is that, assuming the procedure is successful, the hazard arguably drops from the level for a medical patient to that of a patient *just following* surgery. In other words, "Jump to PTCA" might also be modelled as "Jump to PTCA immediately following successful surgery" where we re-set the clock to the hazard at time zero on the PTCA arm. This would be relatively straightforward to implement, and underlines the *flexibility* of pattern mixture models implemented using MI.

Another subtle point with reference based sensitivity analysis is that it might be argued that the use of the pre-censoring hazard to provide valid *de-jure* estimation, when our sensitivity analysis assumes censoring is not at random, might not be appropriate. Essentially, this again poses the question of whether using the pre-censoring hazard represents an appropriate departure point for the hazard of those who have just dropped out. As above, we argue that, as in most things, "context is everything". It is the responsibility of the stakeholders in a specific trial situation to consider the appropriate hazard, and its timing (e.g. a "pre-censoring" hazard could also be used), but once this definition is clear the trial team can at least be secure in the knowledge that the *method* itself is flexible enough to implement their choices.

## 6.3 Information anchored sensitivity analysis

There has been considerable discussion and numerous publications on the issue of congeniality with respect to multiply imputing data, and the degree of conservativeness of Rubin's variance estimator. In the context of the approach to sensitivity analysis proposed here, we argue this is a red herring, and circumvents the fundamental issue concerning reference based sensitivity analysis implemented using MI — namely, that with these Class-2 sensitivity analysis methods we have essentially entered, as it were, a "new world", in which we can no longer rely in the frequentist ("long run") variance properties of the variance estimator to provide us with a sensible estimator of the variance. Why is this so? Due to the incompatibility between the data generating mechanism and the assumptions we make for the post-censoring behaviour in the sensitivity analysis, these Class-2 methods could potentially inject information. We need to step beyond a classical frequentist view of variance estimation, and for this reason Cro *et al.* (2018) provided us with a new property, that of *information anchoring* to help us to navigate

this new world.

Information anchoring provides us with an equivalence property such that the statistical information is held constant across the primary and sensitivity analyses, and that it should certainly not be increased, otherwise, as explained by Cro *et al.* (2018): "an information positive sensitivity analysis is rarely justifiable, implying as it does that the more data are missing, the more certain we are about the treatment effect under the sensitivity analysis ...", and, "while information negative sensitivity analyses provides an incentive for minimising the missing data, there is no natural consensus about the appropriate loss of information". We seek to show that if we wish to follow the information anchoring principle, reference based sensitivity analysis implemented using multiple imputation provides statistically appropriate inference in the time-to-event setting.

Therefore, in Chapters 3 and 4 we presented the results of this investigation of the statistical properties of Rubin's variance estimator following multiple imputation under "Jump to Reference". We initially presented results from simulation studies and from the application to the RITA-2 data which provided empirical evidence that information anchoring holds. Given this encouragement, we went on to prove, under a specific distribution and modelling assumption, that indeed Rubin's variance estimator with reference based multiple imputation provides, to a good approximation, information anchored inference.

Whilst the model we adopted is relatively uncommon, we point to the increased use of the restricted mean survival time (RMST) to overcome situations in which the proportional hazards assumption is questionable, and note the clear parallels between the RMST and, as our theory does, using a t-test to determine treatment difference. Our theoretical results indicate that the principle of information anchoring has transferred seamlessly to the time-to-event setting, providing encouragement that more general results set out in Cro *et al.* (2018) could also transfer to other endpoints and analysis approaches. This would undoubtedly be an interesting and fruitful area of potential further study.

## 6.4 Observational data example

Previous analyses using the COHERE observational HIV data suggested that primary Pneumocystis Pneumonia (PCP) prophylaxis could be withdrawn in patients with CD4 counts of

100-200 cells/L if HIV-RNA is suppressed, suggesting HIV replication as major risk factor for PcP. We estimated the risk of primary PcP in COHERE patients on cART including time-updated CD4 counts, HIV-RNA and use of PcP prophylaxis. The primary endpoint was time to PCP diagnosis, with secondary endpoint all-cause mortality.

Causal inference methods are now established as the gold standard for analysing observational ("big") data. We emulated a hypothetical randomised trial using these established causal inference methods, using inverse probability (IP) weighting to adjust for potential censoring selection bias.

This type of analysis using IP weighting assumes that censoring is at random. As this is an empirically untestable assumption, it is important to explore the sensitivity of inferences to informative censoring. There have been few examples of sensitivity analysis in such settings — these usually involve manipulation of the IP weights — a method which has similar drawbacks to the "$\delta$-type" sensitivity analysis methods. Approaches that are based on the manipulation of the post-censoring hazard function provide an obvious alternative, but it can be difficult to verify their clinical plausibility. Our methods using reference-based multiple imputation (MI) methods provide a simple and more intuitive way for conducting sensitivity analysis.

We focussed on the all-cause mortality endpoint, and estimated the hazard ratio (HR) by fitting a pooled logistic model as analysis model including baseline characteristics, restricted cubic splines to capture CD4/RNA trajectories, and polynomial time effects for modelling the baseline hazard.

To assess the sensitivity of the conclusions to plausible departures from CAR for those patients not on prophylaxis, we used reference-based multiple imputation to construct a contextually plausible scenario in which those lost to follow-up immediately started prophylaxis at their time of censoring (i.e. they "jumped" to the reference arm).

The hazard ratio comparing patients on versus off prophylaxis for the all-cause mortality endpoint was 0.83 with 95% confidence interval [0.75, 0.91], ($p< 0.001$). When patients "off" prophylaxis are censored, and then "Jump to PCP prophylaxis", the HR attenuates to 0.89 ([0.81, 0.97], p=0.01). The sensitivity analysis reveals that, even under this relatively extreme scenario, the estimated HR is still broadly consistent with the primary analysis.

This final example demonstrated how our methods can also be used for observational "big" data when a trial is emulated to look like a randomized controlled trial. In a sense, our approach

exemplified how the two worlds of causal inference and missing data analysis using multiple imputation can be combined in a pragmatic but simple manner. A particular attraction of the approach here is the computational simplicity. Moving from analysing the emulated trial under CAR to J2R, only one line of code needs to be changed!

In our analysis of the COHERE observational data we defined our hypothetical target trial, then censored patients if they no longer conformed to their assigned treatment regimes, emulating a "per protocol", (also known as an "on treatment") type of analysis focussed on determining treatment *efficacy*. We estimated the risk of those continuing and stopping prophylaxis, using inverse probability weighting to adjust for potential confounding due to the time varying treatment (and covariates).

However, one drawback of the "per protocol" approach in a real RCT is that it introduces selection bias since the treatment effect is estimated assuming perfect adherence, which could be considered unrealistic in real-life clinical settings.

In many RCTs, such as the RITA-2 pragmatic trial encountered in Chapter 3, the focus is rather on estimate the effectiveness of the treatment in clinical practice, with the intention-to-treat comparison being the main analytical approach. This estimates the treatment effect based on the original group randomisation, irrespective of adherence. Whilst the ITT analysis is recognised as the pre-eminent approach for many RCTs, it is important to acknowledge its potential shortcomings. As argued succinctly in Hernan and Hernandez-Diaz (2012), an ITT analysis in a placebo-controlled trial "can underestimate the treatment effect, and [is] therefore non-conservative for both safety and non-inferiority trials", and in RCTs with an active comparator an ITT analysis "can overestimate a treatment's effect in the presence of differential adherence". Furthermore, even if these issues are not applicable, it can be argued that the RCT trial setting itself might be considered non-equivalent to a real clinical setting due to, for example, double blinding and the presence of increased patient monitoring (and accordingly increased adherence) amongst others.

Finally, the as treated analysis completes our triumvirate of approaches for RCTs. The "as treated" analysis is a halfway house of the per protocol and ITT analyses — patients are analysed according to the treatment they took rather than that assigned. This means that, in terms of the data used in the analysis, we essentially treat the data from a trial as if it were from an observational study. As such, and in common with the per protocol analysis, an "as treated" comparison is potentially confounded due to non-random selection of subjects into the assigned

groups. Again, in common with the per protocol analysis, IP weighting can be used to adjust for the potential confounding in an "as treated" comparison.

However, in using IP weighting to adjust our model we have assumed all potential confounders have been measured, and included the analysis. For example, in presenting the conclusions of the analyis with the all-cause mortality endpoint for the COHERE data in Chapter 5.9, we mention that the presence of unmeasured confounding (in this case comorbidities) could be the reason for the counter-intuitive results. Although not applicable in all settings, and not addressed in this thesis, instrumental variable (IV) methods have been developed to control for unmeasured confounding, and these provide another valuable and powerful tool for the analysis of observational data (Baiocchi *et al.*, 2014).

## 6.5 The "best" approach to sensitivity analysis

We have focussed on reference based methods, but there are numerous implementations using the "$\delta$-type" methods mentioned in Chapter 1. Recent examples include Zhao *et al.* which use so-called "Kaplan-Meier" and "Proportional Hazards" multiple imputation (Zhao *et al.*, 2014, 2016), while the study by Lipkovich *et al.* (2016) which compared MI strategies based on the the Cox model, piece-wise exponential and logistic models. As previously mentioned these methods, although straightforward to implement, can be rather tricky to use in terms of interpreting the $\delta$ sensitivity analysis parameter in clinical terms. Moreover, these "benchmarking" issues are accentuated if different $\delta$ parameters are used for different arms. There are further challenges if we wish to give a distribution for different $\delta$ parameters in each arm: their correlation then plays a key role in estimating the standard error of the treatment (Mason *et al.*, 2017a).

We have focussed on reference based multiple imputation methods applying Rubin's rules in the usual way. Both Lu *et al.* (2015) and Gao *et al.* (2017) state that using Rubin's rules in such a setting leads to overly conservative and over-estimation of the variance, both therefore use a bootstrapped variance estimator instead, despite this being more "computationally intensive" (Lu *et al.*). Liu and Peng (2016) make a similar point, again in the context of reference based imputation, stating that "a conventional approach [using MI and Rubin's combining rules] ... inflates the variance estimates, which results in an overly conservative test for the treatment effect". We have demonstrated that this is indeed the case — and make the point that this is

actually a good property in analyses — we think that the focus should be elsewhere. Namely, that Rubin's variance estimate displays the desirable property that, as we move from the primary to the sensitivity analysis, the information is anchored. By contrast, the bootstrap variance is information positive: it gets smaller as the level of censoring increases: it rewards researchers for losing data!

Rubin's rules undoubtedly provide the simplest panacea for calculating the variance following MI, and, as has been demonstrated, have the attractive properties we require.

## 6.6 Joint and shared parameter models

We introduced shared parameter models in Chapter 1, but at the time the PhD started we chose not to follow this modelling paradigm due to its interpretational and computational challenges, relative to other approaches. In the meantime, there has been considerable progress made in terms of both methods and off-the-shelf software implementations. Hickey *et al.* (2016) recently provided an overview of the current status of joint models for time-to-event and (multivariate) longitudinal data. This article also includes the software available for fitting such models (e.g. Rizopoulos, 2012; Crowther *et al.*, 2013), which could now be considered mainstream in terms of adoption. The complexity and software barriers having been removed it would seem natural to extend such models to include sensitivity analysis to investigate potentially informative censoring.

This step has yet to be included formally in the available software but there are numerous examples of how this might be achieved (e.g. Barrett and Su (2015), and the references therein). Hu *et al.* (2016) have developed an analogue to Full Conditional Specification methods in which iteratively and sequentially, they use a data augmentation approach to multiply impute respectively, longitudinal outcomes, event times and event types. This would avoid explicitly fitting a joint model.

Kim *et al.* (2017) propose a joint longitudinal survival model which combines many of the trickier aspects confronted in earlier chapters such as non-proportional hazards with those from Chapter 6 on observational data. They use a shared parameter model with cubic splines to model a time varying biomarker, combined with a flexibly defined cumulative hazard function allowing both censoring at random and informative censoring to be modelled.

As we have alluded to, such methods whilst providing flexibility to more closely model real clinical situations, are rather complex both to understand and implement. They are essentially an example of Class-1 sensitivity analyses in which the event and censoring processes are modelled simultaneously. We argue that just such a situation is ripe for the application of our reference based methods since they limit the *further* complexity required to carry out the sensitivity analysis. Reference based sensitivity analysis could be extended next to this area.

## 6.7 Software implementations and adoption

In terms of sensitivity analysis, the "mimix" package in Stata (Cro *et al.*, 2016) fully implements the reference based multiple imputation approach for longitudinal data with a continuous outcome outlined by Carpenter and Kenward (2012). For SAS users, the code from J. Roger in O'Kelly and Ratitch (2014) is available from the website www.missingdata.org.uk.

To date there is only one software package available in R to multiply impute time-to-event outcomes under informative censoring (Ruau *et al.* (2016)). This is surprising given the continued development of packages, for example in R, Stata and SAS, and the comparative simplicity of implementation. As shown in Chapter 5, inverse probability weighting implicitly assumes censoring at random and adjusts for non-informative censoring. This could conceivably be the reason behind this lack of further development — certainly for observational data.

The speed of adoption of such methods for time-to-event data will clearly be impaired by not having off-the-shelf software to perform multiple imputation process, and I would like to develop this.

## 6.8 Final remarks

Given the complexity of many clinical settings, it would be illusory to expect a single methodology to address all possible sensitivity analysis scenarios.

Our approach to sensitivity analysis extends reference based sensitivity analysis to the time-to-event setting. Reference based sensitivity analysis has found increasing application in settings where a non-trivial proportion of patients deviate from the protocol, thus the analysis cannot

proceed without making additional assumptions, which are not fully verifiable from the trial data. For example, building on earlier work by Little and Yau (1996) addressing intention-to-treat analyses using MI, Keene *et al.* (2014) show how to use controlled imputation for sensitivity analysis under a negative binomial distribution for recurrent events; Gao *et al.* (2017) use controlled imputation with a piecewise exponential model, whilst Tang (2018) propose an extension of control-based imputation to longitudinal binary and ordinal data. In the survival setting, Lu *et al.* (2015) compared two approaches to sensitivity analysis with controlled multiple imputation, whereas Lipkovich *et al.* (2016) propose an approach for a "tipping point" analysis for time-to-event data. Zhao *et al.* (2014) apply non-parametric multiple imputation which uses "nearest neighbour" algorithms to investigate potentially informative censoring, including a reference based approach. There are also numerous recent examples of applications of the method in trials (see, for example, Mallinckrodt *et al.*, 2013; Philipsen *et al.*, 2015; Jans *et al.*, 2015; Billings *et al.*, 2018; Atri *et al.*, 2018).

Reference based imputation has two advantages. Firstly, it avoids the user specifying numerous parameters describing the distribution of patient's post-withdrawal data. This difficulty has been widely acknowledged (Daniels and Hogan, 2008), and our methods solve this issue simply using the tools at hand. Secondly, when implemented using multiple imputation it is, to a good approximation, information anchored, holding the proportion of information lost due to missing data under the primary analysis constant across the sensitivity analyses. This property has been theoretically demonstrated in longitudinal data settings (Cro *et al.*, 2018), and we have proved that it also applies to time-to-event data under a specific model, and provided emirical evidence it applies more generally (Atkinson *et al.*, 2018).

The recent Addendum on estimands and sensitivity analysis in clinical trials will only strengthen the need for practical sensitivity analysis methods (CHMP, 2018). In conclusion, we believe reference-based sensitivity analysis via multiple imputation is a flexible, accessible and practical approach, as witnessed by its increasing use. We hope that, by showing how these ideas can be extended to survival data, practitioners will have confidence to use it in their own studies.

We started with a quote from Robert Burns to the effect that however much we plan (to avoid missing data) fate often intervenes (and missing data occur). However,

> "Doubt is not pleasant but certainty is absurd" — *Voltaire*,

> "The demand for certainty is one which is natural to man but is nonetheless an

intellectual vice" — *Bertrand Russell*,

"No great deed is done by falterers who ask for certainty" —*George Eliott*.

All of which we interpret to mean that it is important to acknowledge the presence of missing data, but nonetheless deal with such data in a systematic and well-founded manner.

Doubt is to be welcomed and drives innovation. This research shows that reference based sensitivity analysis is a well-founded, accessible and practical approach for time-to-event data.

# Appendices

# Appendix A

# German Breast Cancer Data set

## A.1 Exploratory Data Analysis

The primary end points of the original study were tumour recurrence, and death of the patient. We focus solely on the recurrence free survival time. This is defined as the time from mastectomy to the first occurrence of either locoregional or distant recurrence, contralateral tumor or secondary tumor.

The chemotherapy regime consisted of using the modified Bonnadonna CMF scheme consisting of 500 $mg/m^2$ cyclophosphamide, 40 $mg/m^2$ methotrexate and 600 $mg/m^2$ fluorouracil on days 1 and 8 of a 4-week treatment period. The hormonal treatment (HT) consisted of a daily dose of 3 $\times$ 10 mg tamoxifen orally administered over 2 years, starting after the third cycle of CMF (Schumacher *et al.*, 1994).

The study was defined as a 2 x 2 factorial design with the following treatment arms:

- 3 cycles of chemotherapy and no hormonal treatment.
- 3 cycles of chemotherapy and hormonal treatment.

- 6 cycles of chemotherapy and no hormonal treatment.
- 6 cycles of chemotherapy and hormonal treatment.

Patients were generally randomized following the completion of an initial 3 cycle phase of chemotherapy. It should also be noted that premenopausal patients admitted to the study after December 1986 were only randomised to the first 3 treatment regimes.

Due to patient preference in the non-randomised part of the trial, and a change in the protocol for premenopausal patients, only 40% of the 448 patients received hormonal treatment (Sauerbrei *et al.*, 1999). Of these, 38% had a recurrence of the disease, irrespective of the chemotherapy treatment. Of the patients randomized to hormonal treatment, two thirds received it for at least 1.5 years, 10% for less than 1 year, with the duration unknown for 15% of patients.

Patients were followed up regularly, with clinical examinations every 3 months during the first 2 years, every 3 months for the subsequent 3 years, and every 6 months in years 6 and 7.

Not all patients adhered to the schedule, with 63 patients having follow-up times longer than a year, and several patients missing information for more than 2 years. Therefore, the censored patients are a mixture of those surviving until the end of the study (i.e. administrative censoring), and those lost to follow-up during the study. Of the latter group, no additional information was available as to the reasons for dropping out. It may be assumed that these could be due to lack of tolerance to the 6 cycle chemotherapy treatment, lack of adherence to the daily hormonal treatment, other non-adherence to study protocol reasons, or the full recovery, or death, of the patient.

The data set contains the following variables for 448 patients:

- *oobs*: Integer patient identifier.
- *Hther*: Binary indicator variable for hormonal treatment HT (0 = no HT or 1 = HT).
- *THERC*: Indicator variable for chemotherapy taking values 1 or 2:
  - Reference = 3 cycles CT (*THERC* = 1)
  - Treatment = 6 cycles CT (*THERC* = 2)
- Patient characteristics with respect to prognostic factors:

  - *Alter*: Age in years.
  - *meno*: Indicator variable for the Menopausal state taking values 1 (pre) or 2 (post).
  - *tgroesse*: Tumour size in millimeters.
  - *grad*: Tumour grading - indicator variable taking values 1, 2 or 3.
  - *npos*: Number of involved nodes.
  - *nprog*: Progesterone Receptor, fmol.
  - *noest*: Oestrogen Receptor, fmol.

- End point Measurement
  - *rezfrei1*: Recurrence free time in days (RESDAT1 − MASDAT) i.e. time to event or administrative censoring.
  - *MASDAT*: Date of the patients mastectomy, and entrance to the study.
  - *RESDAT1*: Recurrent free survival date, later than MASDAT; either the event date, or administrative censoring time.

- Censoring and missingness indicators:
  - *zensrez1*: Binary indicator variable for censoring status (0 = censored or 1 = event occurred).
  - *foll*: Integer indicator variable taking values 4, 6 or 12. This is the theoretical number of follow-up times the patient has between the date of the mastectomy operation and their recurrence free time (*rezfrei1*, the event or censoring time), this time being known as the follow-up time. Depending on this follow-up time, *foll* takes the following values[1]:
    * less than 24 months: there are 3 follow-ups (*foll* = 3, does not occur in the data set);
    * greater than or equal to 24 months and less than 60 months: *foll* = 4;
    * greater than or equal to 60 months and less than 84 months: *foll* = 6;
    * otherwise foll = 12.
    * Note that this is the theoretical number of follow-up times, assuming that all patients attended all follow-up visits.

[1] Taken from the SAS-Code from M. Olschewski, sent by mail in May 2013.

- *miss*: This is the number of follow-up visits missed by a censored patient, calculated as the difference between the follow-up time (as defined above) and the patients censoring time. If this difference is at least 6 months, the variable *miss* takes the value 1, otherwise it is 0. There are 123 censored patients that have this variable set.

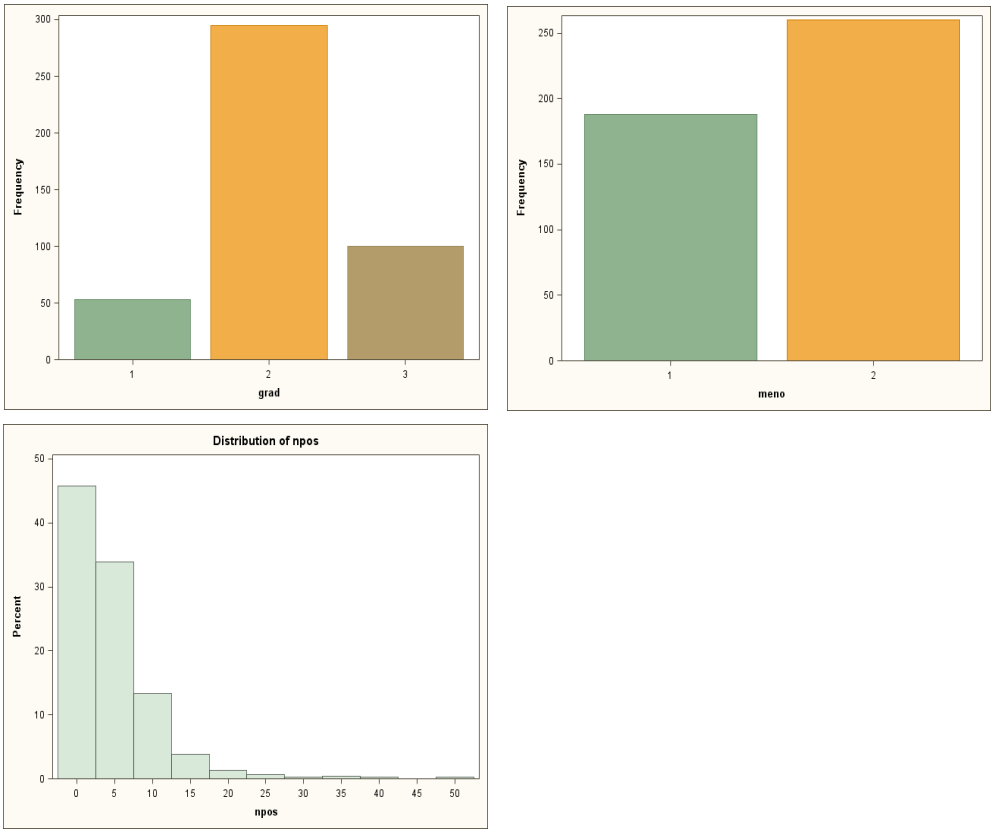


Figure A.1.1: Exploratory data analysis - factors *grad*, *meno*, *npos*

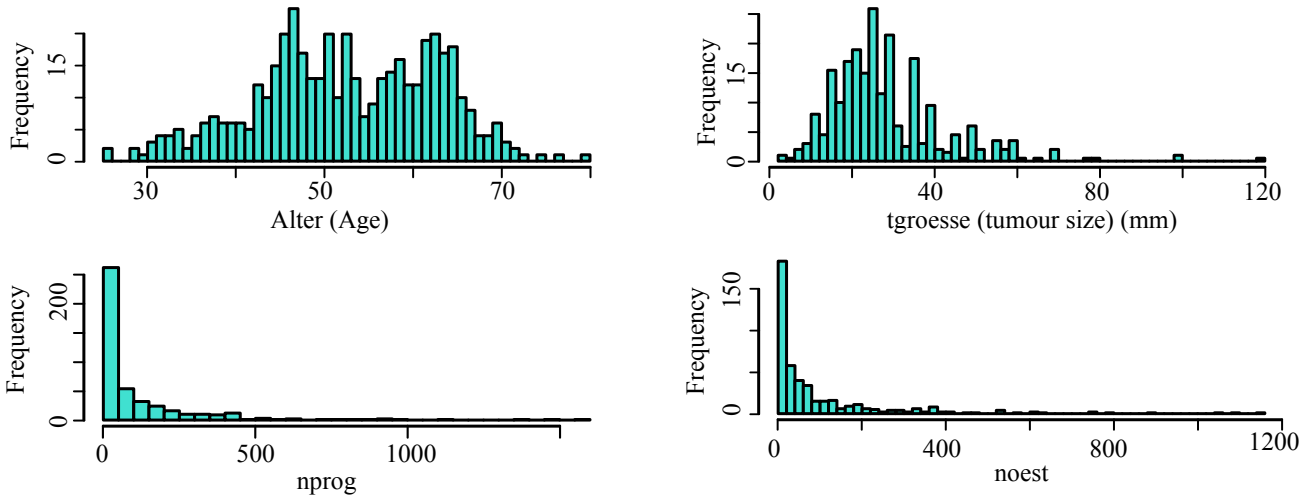


Figure A.1.2: Exploratory data analysis - continuous covariates (*Alter*, *tgroesse*, *nprog*, *noest*)

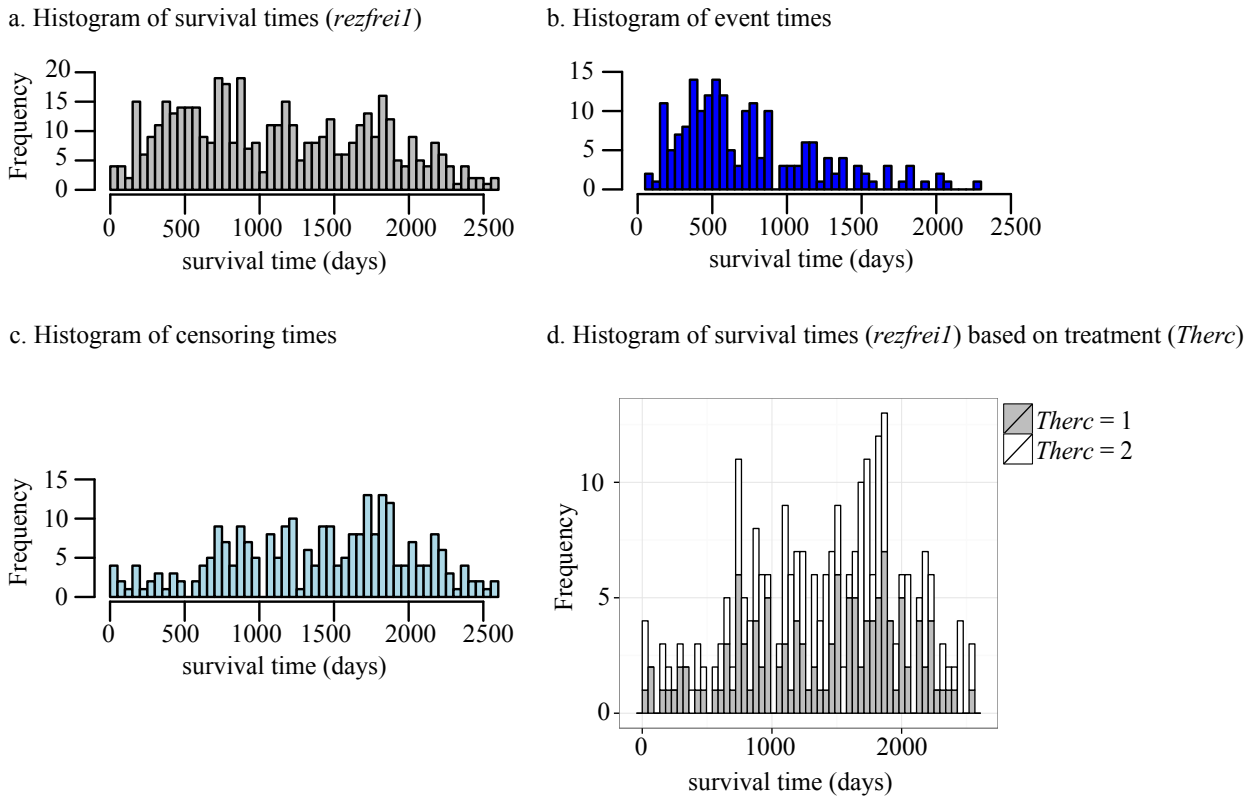


Figure A.1.3: Event and censoring profile for the data set;
a.) Top left panel: Histogram of time to the first of event, or censoring;
b.) top right panel: Histogram of survival times for patients experiencing an event (i.e. a recurrence of the disease);
c.) bottom left panel: Histogram of survival times for censored patients;
d.) bottom right panel: Histogram of the survival times (variable *rezfrei1*), based on chemotherapy treatment level;
*THERC*=1 is the lower treatment regime (3-cycle);
*THERC*=2 is the higher treatment regime (6-cycle).

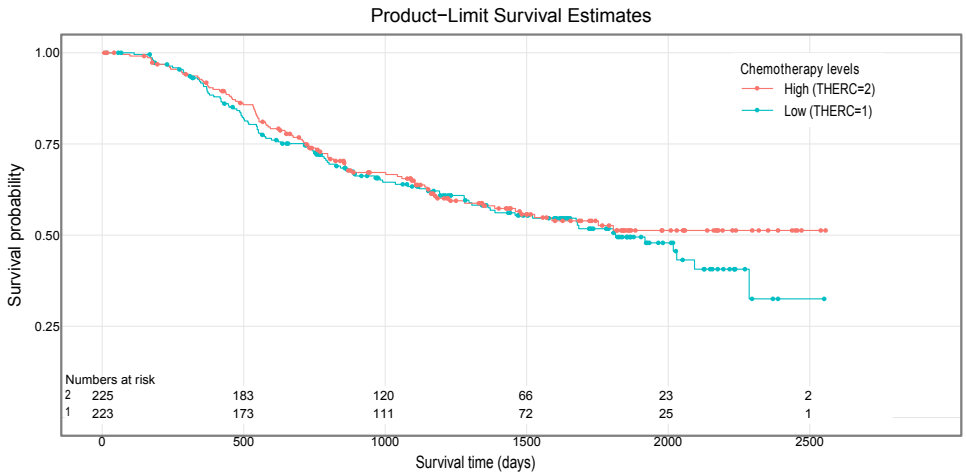


Figure A.1.4: Kaplan-Meier plot of the treatment effect of 3 cycles of chemotherapy (blue, *THERC* =1) versus 6 cycles of chemotherapy (orange, *THERC* = 2); dots marking censored times

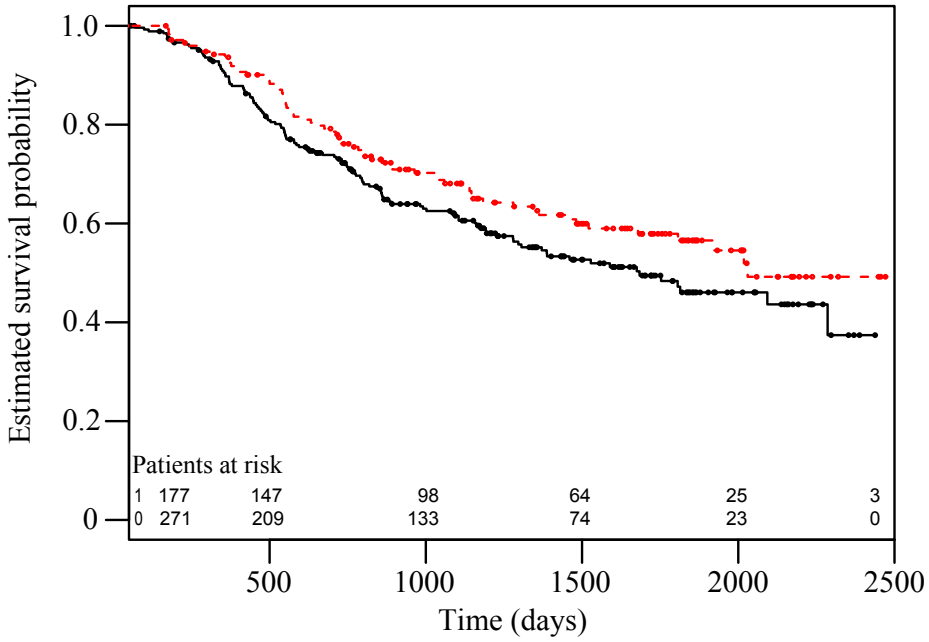


Figure A.1.5: Kaplan-Meier estimator of the survival function for the treatment effect without hormonal treatment (black solid line, *hther* = 0) versus with hormonal treatment (red dotted line, *hther* = 1); dots marking censored times.

# Appendix B

# Properties of the bivariate normal distribution

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be jointly normal random variables. The conditional expectation of $X$ given $Y$ satisfies:

$$E[X|Y] = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(Y - \mu_2),$$

a linear function in $Y$, where $\rho$ is the correlation coefficient of $X, Y$, $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y}} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$, with

$$Pr(X|y_i) \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(Y - \mu_2),\ (1-\rho^2)\sigma_1^2\right).$$

Making the linear relationship explicit we have the regression estimates:

$$\beta_0 = \mu_1 - \beta_1\mu_2,\ \ \beta_1 = \rho\frac{\sigma_1}{\sigma_2},\ \ \sigma_{X|Y} = (1-\rho^2)\sigma_1^2 = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2},$$

so that $x_i|y_i \sim N(\beta_0 + \beta_1 y, \sigma_{X|Y})$.

The above residual variance converted to the notation in the main body of the document be-

comes:
$\sigma_{2.1} = \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}$.

# Appendix C

# Adapted variance calculation for the truncated normal distribution

The variance formulae presented in section 4.3.2 led in practice to a slight under evaluation of the variance at the tail of the distribution — certainly with 10% censoring and potentially also up to 20% censoring levels.

In terms of implementation in code we preferred to calculate the variance using the method based on the chi-square density function defined by Barr and Sherrill (1999).

The variance is defined in this case as:

$$V(Z) = c(t)\left[\sqrt{\frac{\pi}{2}}(1 \mp C_3(t^2) - c(t)e^{-t^2}\right],$$

for $Z$ standard normal truncated at $t$, where $c(t) = 1/\left[\sqrt{2\pi(1-\Phi(t))}\right]$ with $\Phi(t)$ being the standard normal cumulative distribution function, and $C_3(t^2)$ the $\chi$-square density function with 3 degrees of freedom. The plus sign applying for $t < 0$ and the minus otherwise. For example, if we have 10% censoring with $\alpha = 3.2$ we evaluate the variance of the observed patients using the formula with a plus sign and $t = -3.2$. When the data are not standard normal then a normalising transformation should be applied prior to applying the formula.

# Appendix D

# Rubin's variance estimate under the de-jure estimate of CAR

Referring back to equation 4.3.13, for the active arm we decompose the summation into observed and deviating parts, substituting our new expressions for $\bar{Y}_{a2,k}$ and $\tilde{Y}_{aj2,k}$, assuming CAR.

$$(n_a - 1)\hat{\sigma}_a^2 = E\left(\sum_{j\in o}(Y_{aj2} - \hat{\mu}_{a2,k})^2\right) + E\left(\sum_{j\in d}(\hat{Y}_{aj2,k} - \hat{\mu}_{a2,k})^2\right)$$

For those observed we obtain the following expression:

$$E\left[\sum_{j\in o}(Y_{aj2} - \hat{\mu}_{a2,k})^2\right] =$$

$$E\left[\sum_{j\in o}((Y_{aj2} - \bar{Y}_{a2o}) - \frac{n_d}{n_a}u_k - \frac{n_d}{n_a}\left(\frac{r}{q} + b_k\right)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) - \frac{n_d}{n_a}\bar{\dot{\lambda}}\sqrt{\bar{\tilde{\sigma}}_{22,k}}\right.$$

$$\left. - \frac{n_d}{n_a}w_{k,\dot{\lambda}} - \frac{n_d}{n_a}\bar{\lambda}\sqrt{\bar{\tilde{\sigma}}_{22,k}} - \frac{n_d}{n_a}w_{k,\lambda} - \frac{n_d}{n_a}\bar{\epsilon}_k)^2\right].$$

We now derive the expectation of each of the squared expressions term by term.

For the first term, we can use the standard expressions for the truncated normal distribution:

$$E\left[\sum_{j\in o}(Y_{aj2}-\bar{Y}_{a2o})^2\right]=(n_o-1).$$
$$\sigma_{22}\left[1-\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}-\left(\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\right)^2\right].$$

In addition, we have the following terms:

$$\left(\frac{n_d}{n_a}\right)^2\sigma_{2.1}+n_o\left(\frac{n_d}{n_a}\right)^2\left[\frac{\sigma_{12}^2}{\sigma_{11}}+\frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]+\left(\frac{n_on_d}{n_a^2}\right)\sigma_{a2d},$$

with additional squared terms from the parts from the Mills Ratio expressions:

$$E\left[\sum_{j\in o}\bar{\dot{\lambda}}^2\left(\frac{n_d}{n_a}\right)^2\left(\sqrt{\bar{\tilde{\sigma}}_{22,k}}\right)^2\right]=n_o\dot{\lambda}^2\left(\frac{n_d}{n_a}\right)^2\sigma_{22},$$

$$E\left[\sum_{j\in o}\bar{\lambda}^2\left(\frac{n_d}{n_a}\right)^2\left(\sqrt{\bar{\tilde{\sigma}}_{22,k}}\right)^2\right]=n_o\lambda^2\left(\frac{n_d}{n_a}\right)^2\sigma_{22},$$

$$E\left[\sum_{j\in o}\left(\frac{n_d}{n_a}\right)^2w_{k,\dot{\lambda}}^2\right]=\left(\frac{n_d}{n_a}\right)^2.n_o.VAR(w_{k,\dot{\lambda}}),$$

$$E\left[\sum_{j\in o}\left(\frac{n_d}{n_a}\right)^2w_{k,\lambda}^2\right]=\left(\frac{n_d}{n_a}\right)^2.n_o.VAR(w_{k,\lambda}),$$

with $VAR(w_k)$ as defined as in the main body of the document.

If we consider a normal quadratic equation $(a+b)^2=a^2+b^2+2ab$, then analogously, the expressions above correspond to the $a^2$ and $b^2$ terms.

For the $2ab$ terms in the square for the observed patients, we have the definitions $E(u_k) = E(b_k) = E(w_{k,\cdot}) = E(\bar{\epsilon}_k) = 0$, ensuring that any expressions containing these terms disappear. In addition, under CAR we have both $E(\sum_{j \in o}(Y_{aj2} - \bar{Y}_{a2o})) = 0$ and due to randomisation $E(\bar{Y}_{a1d} - \bar{Y}_{a1o}) = 0$, so terms in these expressions both disappear.

We also have an additional term in $\sqrt{\sigma_{22}}$:

$$E\left[\sum_{j \in o} 2\left(\frac{n_d}{n_a}\right)^2 \bar{\dot{\lambda}}\, \bar{\lambda}(\sqrt{\tilde{\sigma}_{22,k}})^2\right]$$

$$= 2n_o\, \bar{\dot{\lambda}}\, \bar{\lambda}\left(\frac{n_d}{n_a}\right)^2 \sigma_{22}.$$

For the patients deviating, we again write out the full summation so that we can identify the terms:

$$E\left[\sum_{j\in d}(\tilde{Y}_{aj2,k}-\hat{\mu}_{a2,k})^2\right]=$$

$$E\left[\sum_{j\in d}\Big((\hat{Y}_{aj2,k}-\bar{Y}_{a2d,k})+\right.$$

$$\left.\frac{n_o}{n_a}\left(u_k+\left(\frac{r}{q}+b_k\right)(\bar{Y}_{a1d}-\bar{Y}_{a1o})+\sqrt{\hat{\sigma}_{22}}\bar{\dot{\lambda}}+\bar{w}_{k,\dot{\lambda}}+\sqrt{\hat{\sigma}_{22}}\bar{\lambda}+\bar{w}_{k,\lambda}+\bar{\epsilon}_k\right)\Big)^2\right],$$

using the reformulation from the main body of the document.

We now calculate this expression term by term:

$$E\left[\sum_{j\in d}(\hat{Y}_{aj2,k}-\bar{Y}_{a2,k})^2\right]=(n_d-1)\;\sigma_{22}\left[1-\frac{\phi(\frac{\alpha-\mu_{a2}}{\sigma_{22}})}{1-\Phi(\frac{\alpha-\mu_{a2}}{\sigma_{22}})}\left[\frac{\phi(\frac{\alpha-\mu_{a2}}{\sigma_{22}})}{1-\Phi(\frac{\alpha-\mu_{a2}}{\sigma_{22}})}-\left(\frac{\alpha-\mu_{a2}}{\sigma_{22}}\right)\right]\right]$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2 u_k^2\right]=\frac{n_o n_d}{n_a^2}\;\sigma_{2.1}$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2\left(\frac{r}{q}+b_k\right)^2(\bar{Y}_{a1d}-\bar{Y}_{a1o})^2\right]=n_d\;\left(\frac{n_o}{n_a}\right)^2\;\left[\frac{\sigma_{12}^2}{\sigma_{11}}+\frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2\;\bar{\dot{\lambda}}^2\left(\sqrt{\bar{\tilde{\sigma}}_{22,k}}\right)^2\right]=n_d\;\left(\frac{n_o}{n_a}\right)^2\;\dot{\lambda}^2\;\sigma_{22}$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2\;\bar{\lambda}^2\left(\sqrt{\bar{\tilde{\sigma}}_{22,k}}\right)^2\right]=n_d\;\left(\frac{n_o}{n_a}\right)^2\;\lambda^2\;\sigma_{22}$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2 w_{k,\dot{\lambda}}^2\right] = n_d\ \left(\frac{n_o}{n_a}\right)^2 VAR(w_{k,\dot{\lambda}})$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2 w_{k,\lambda}^2\right] = n_d\ \left(\frac{n_o}{n_a}\right)^2 VAR(w_{k,\lambda})$$

$$E\left[\sum_{j\in d}\left(\frac{n_o}{n_a}\right)^2 \bar{\epsilon}_k^2\right] = \left(\frac{n_o}{n_a}\right)^2 \sigma_{a2d}$$

For the $2ab$ term in the square for the deviating patients, we have the definitions $E(\bar{u}_k) = E(\bar{b}_k) = E(w_{k,\cdot}) = E(\bar{\epsilon}_k) = 0$, ensuring that any expressions containing these terms disappear.

We also have an additional term in $\sigma_{22}$:

$$E\left[\sum_{j\in d} 2\ \left(\frac{n_o}{n_a}\right)^2\ \bar{\dot{\lambda}}\ \bar{\lambda}\left(\sqrt{\bar{\tilde{\sigma}}_{22,k}}\right)^2\right] = 2\ n_d\ \left(\frac{n_o}{n_a}\right)^2\ \bar{\dot{\lambda}}\ \bar{\lambda}\ \sigma_{22}$$

.

Putting the observed and deviating parts together, we arrive at a revised estimate to plug into the pooled variance estimator for $E(\hat{W})$:

$$(n_a - 1)E(\hat{\sigma}_a^2) = (n_o - 1)\sigma_{22}$$

$$\left[1 - \left(\frac{\alpha - \mu_{a2}}{\sqrt{\sigma_{22}}}\right)\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)} - \left(\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\right)^2\right] +$$

$$\left(\frac{n_d}{n_a}\right)^2 \sigma_{2.1} + n_o \left(\frac{n_d}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o - 1)}\right]\left[\frac{1}{n_d} + \frac{1}{n_o}\right] + \left(\frac{n_d n_d}{n_a^2}\right)\sigma_{a2d} +$$

$$n_o\dot{\lambda}^2\left(\frac{n_d}{n_a}\right)^2\sigma_{22} + n_o\lambda^2\left(\frac{n_d}{n_a}\right)^2\sigma_{22} + n_d\left(\frac{n_o}{n_a}\right)^2\dot{\lambda}^2\,\sigma_{22} + n_d\left(\frac{n_o}{n_a}\right)^2\lambda^2\,\sigma_{22} +$$

$$\left(\frac{n_d}{n_a}\right)^2 n_o VAR(w_{k,\dot{\lambda}}) + \left(\frac{n_d}{n_a}\right)^2 n_o VAR(w_{k,\lambda}) + n_d\left(\frac{n_o}{n_a}\right)^2 VAR(w_{k,\dot{\lambda}}) + n_d\left(\frac{n_o}{n_a}\right)^2 VAR(w_{k,\lambda}) +$$

$$2\dot{\lambda}\,n_o\left(\frac{n_d}{n_a}\right)^2\left(\frac{\sigma_{12}}{\sigma_{11}}\right)\sqrt{\sigma_{22}}\,(\mu_{a1d} - \mu_{a1o}) + 2\lambda\,n_o\left(\frac{n_d}{n_a}\right)^2\left(\frac{\sigma_{12}}{\sigma_{11}}\right)\sqrt{\sigma_{22}}\,(\mu_{a1d} - \mu_{a1o}) +$$

$$2n_o\,\bar{\dot{\lambda}}\,\bar{\lambda}\left(\frac{n_d}{n_a}\right)^2\sigma_{22} + 2\,n_d\left(\frac{n_o}{n_a}\right)^2\bar{\dot{\lambda}}\,\bar{\lambda}\,\sigma_{22} +$$

$$(n_d - 1)\,\sigma_{22}\left[1 - \frac{\phi(\frac{\alpha-\mu}{\sigma})}{1 - \Phi(\frac{\alpha-\mu}{\sigma})}\left[\frac{\phi(\frac{\alpha-\mu}{\sigma})}{1 - \Phi(\frac{\alpha-\mu}{\sigma})} - \left(\frac{\alpha - \mu}{\sigma}\right)\right]\right] +$$

$$\left(\frac{n_o n_d}{n_a^2}\right)\,\sigma_{2.1} + \left(\frac{n_o}{n_a}\right)^2\sigma_{a2d} +$$

$$n_d \left(\frac{n_o}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]+$$

$$2\, n_d \left(\frac{n_o}{n_a}\right)^2 \left(\frac{\sigma_{12}}{\sigma_{11}}\right) (\mu_{a1d}-\mu_{a1o})\, \bar{\dot{\lambda}}\, \sqrt{\sigma_{22}} + 2\, n_d \left(\frac{n_o}{n_a}\right)^2 \left(\frac{\sigma_{12}}{\sigma_{11}}\right) (\mu_{a1d}-\mu_{a1o})\, \bar{\lambda}\, \sqrt{\sigma_{22}}.$$

Collecting similar terms we obtain,

$$(n_a-1)E(\hat{\sigma}_a^2) = (n_o-1)\,\sigma_{22}\left[1-\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)} - \left(\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\right)^2\right]+$$

$$(n_d-1)\,\sigma_{22}\left[1-\frac{\phi(\frac{\alpha-\mu}{\sigma})}{1-\Phi(\frac{\alpha-\mu}{\sigma})}\left[\frac{\phi(\frac{\alpha-\mu}{\sigma})}{1-\Phi(\frac{\alpha-\mu}{\sigma})}-\left(\frac{\alpha-\mu}{\sigma}\right)\right]\right]+$$

$$\left(\frac{n_d}{n_a}\right)^2 \sigma_{2.1} + n_o \left(\frac{n_d}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right] + \left(\frac{n_d n_o}{n_a^2}\right)\sigma_{a2d}+$$

$$\left(\frac{n_o\, n_d}{n_a}\right)\dot{\lambda}^2\sigma_{22} + \left(\frac{n_o\, n_d}{n_a}\right)\lambda^2\sigma_{22}+$$

$$\left(\frac{n_d n_o}{n_a}\right) VAR(w_{k,\dot{\lambda}}) + \left(\frac{n_d n_o}{n_a}\right) VAR(w_{k,\lambda})+$$

$$2\left(\frac{n_o n_d}{n_a}\right)\, \bar{\dot{\lambda}}\, \bar{\lambda}\, \sigma_{22}+$$

$$\left[\frac{n_d n_o}{n_a^2}\right]\sigma_{2.1} + \left(\frac{n_o}{n_a}\right)^2 \sigma_{a2d} + n_d \left(\frac{n_o}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]+$$

$$2\ n_d\ \left(\frac{n_o}{n_a}\right)^2\ \left(\frac{\sigma_{12}}{\sigma_{11}}\right)\ (\mu_{a1d}-\mu_{a1o})\ \bar{\dot{\lambda}}\ \sqrt{\sigma_{22}}+\ 2\ n_d\ \left(\frac{n_o}{n_a}\right)^2\ \left(\frac{\sigma_{12}}{\sigma_{11}}\right)\ (\mu_{a1d}-\mu_{a1o})\ \bar{\lambda}\ \sqrt{\sigma_{22}}.$$

Under CAR $\mu_{a1d}=\mu_{a1o}$, so this expression simplifies to

$$=(n_o-1)\ \sigma_{22}\left[1-\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}-\left(\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\right)^2\right]+$$

$$(n_d-1)\ \sigma_{22}\left[1-\frac{\phi(\frac{\alpha-\mu}{\sigma})}{1-\Phi(\frac{\alpha-\mu}{\sigma})}\left[\frac{\phi(\frac{\alpha-\mu}{\sigma})}{1-\Phi(\frac{\alpha-\mu}{\sigma})}-\left(\frac{\alpha-\mu}{\sigma}\right)\right]\right]+$$

$$\left(\frac{n_o\ n_d}{n_a}\right)\dot{\lambda}^2\sigma_{22}+\left(\frac{n_o\ n_d}{n_a}\right)\lambda^2\sigma_{22}+$$

$$\left(\frac{n_d n_o}{n_a}\right)VAR(w_{k,\dot{\lambda}})+\left(\frac{n_d n_o}{n_a}\right)VAR(w_{k,\lambda})+$$

$$2\left(\frac{n_o n_d}{n_a}\right)\ \bar{\dot{\lambda}}\ \bar{\lambda}\ \sigma_{22}+$$

$$\left(\frac{n_d}{n_a}\right)\sigma_{2.1}+n_d\ \left(\frac{n_o}{n_a}\right)^2\ \left[\frac{\sigma_{12}^2}{\sigma_{11}}+\frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]+$$

$$\left(\frac{n_o}{n_a}\right)\ \sigma_{a2d}+n_o\ \left(\frac{n_d}{n_a}\right)^2\ \left[\frac{\sigma_{12}^2}{\sigma_{11}}+\frac{2\sigma_{2.1}}{(n_o-1)}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]$$

Simplifying and collecting terms:

$$(n_a - 1)E(\hat{\sigma}_a^2) = (n_o - 1)\sigma_{a2o} + (n_d - 1)\sigma_{a2d} +$$

$$\left(\frac{n_o n_d}{n_a}\right)\sigma_{22}(\dot{\lambda} + \lambda)^2 +$$

$$\left(\frac{n_d n_o}{n_a}\right)VAR(w_{k,\dot{\lambda}}) + \left(\frac{n_d n_o}{n_a}\right)VAR(w_{k,\lambda}) +$$

$$\left(\frac{n_d}{n_a}\right)\sigma_{2.1} +$$

$$\left(\frac{n_o}{n_a}\right)\sigma_{a2d} +$$

$$\frac{n_d}{n_a}\sigma_{2.1} +$$

$$\left(\frac{n_d n_o}{n_a}\right)\left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o - 1)}\right]\left[\frac{1}{n_d} + \frac{1}{n_o}\right].$$

The final term in this expression is simplified to $\left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o - 1)}\right]$.

In a final step,

- we re-write terms such that $\pi_d = n_d/n_a$, $n_o/n_a = (1 - \pi_d)$,
- we approximate $(n_o - 1) \approx n_o$, $(n_d - 1) \approx n_d$,
- we let $n_r = n_a = n$, and finally,
- we divide by $(n_a - 1) = n_a = n$.

We obtain

$$E(\hat{\sigma}_a^2) = (1 - \pi_d)\sigma_{a2o} + \sigma_{a2d}\left(\pi_d + \tfrac{(1-\pi_d)}{n}\right) +$$

$$\pi_d(1 - \pi_d)\sigma_{22}\left(\dot{\lambda} + \lambda\right)^2 +$$

$$\pi_d(1 - \pi_d)\left(VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right) +$$

$$\tfrac{1}{n}\left[\tfrac{\sigma_{12}^2}{\sigma_{11}} + \tfrac{2\sigma_{2.1}}{(n_o-1)}\right] + \tfrac{\pi_d}{n}\sigma_{2.1}.$$

We now have an expression for $\hat{\sigma}_a^2$ which takes account of the within imputation variance, but has not yet included the between variance component $E(\hat{B})$.

For $E(\hat{B})$, we want to calculate the following expression:

$$E(\hat{B}) = E\left[\sum_{k=1}^{K}\left(\frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}(\bar{Y}_{a2o} + u_k + \left(\frac{r}{q} + b_k\right)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) + \right.\right.$$

$$\dot{\lambda}\sqrt{\bar{\tilde{\sigma}}_{22,k}} + w_{k,\dot{\lambda}} + \lambda\sqrt{\bar{\tilde{\sigma}}_{22,k}} + w_{k,\lambda} + \bar{\epsilon}_k) - \bar{Y}_{r2})$$

$$-(\frac{n_o}{n_a}\bar{Y}_{a2o} + \frac{n_d}{n_a}(\bar{Y}_{a2o} + \bar{u} + \left(\frac{r}{q} + \bar{b}\right)(\bar{Y}_{a1d} - \bar{Y}_{a1o}) +$$

$$\left.\left.\bar{\dot{\lambda}}\sqrt{\bar{\tilde{\sigma}}_{22,k}} + \bar{w}_{\dot{\lambda}} + \bar{\lambda}\sqrt{\bar{\tilde{\sigma}}_{22,k}} + \bar{w}_{\lambda} + \bar{\epsilon}) - \bar{Y}_{r2})\right)^2\right]$$

We note that terms in $\lambda\sqrt{\bar{\tilde{\sigma}}_{22,k}}$ and $\dot{\lambda}\sqrt{\bar{\tilde{\sigma}}_{22,k}}$ cancel out since they are constant when averaged over $k$. Similarly, the terms $w_{k,\dot{\lambda}}$ and $w_{k,\dot{\lambda}}$ cancel one another out since they are from the observed data and therefore don't vary over the $k$ imputations.

Now, we proceed term by term:

$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 u_k^2\right] = K\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_o}$

$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \left(\frac{r}{q} + b_k\right)^2 (\bar{Y}_{a1d} - \bar{Y}_{a1o})^2\right] = K\left(\frac{n_d}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o - 1)}\right]\left[\frac{1}{n_d} + \frac{1}{n_o}\right]$

$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{\epsilon}_k^2\right] = K\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$

$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{u}^2\right] = \left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_o}$

$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \left(\frac{r}{q} + \bar{b}\right)^2 (\bar{Y}_{a1d} - \bar{Y}_{a1o})^2\right] = K\left(\frac{n_d}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{(n_o - 1)}\frac{(K+1)}{K}\right]\left[\frac{1}{n_d} + \frac{1}{n_o}\right]$

$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{\epsilon}^2\right] = \left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$

$$E\left[\sum_{k=1}^{K} -2\left(\frac{n_d}{n_a}\right)^2 u_k\bar{u}\right] = -2\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_o}$$

$$E\left[\sum_{k=1}^{K} -2\left(\frac{n_d}{n_a}\right)^2 \left(\frac{r}{q}+b_k\right)\left(\frac{r}{q}+\bar{b}\right)(\bar{Y}_{a1d}-\bar{Y}_{a1o})^2\right] = -2K\left(\frac{n_d}{n_a}\right)^2 \left[\frac{\sigma_{12}^2}{\sigma_{11}}+\frac{2\sigma_{2.1}}{(n_o-1)}\frac{(K+1)}{K}\right]\left[\frac{1}{n_d}+\frac{1}{n_o}\right]$$

$$E\left[\sum_{k=1}^{K} -2\left(\frac{n_d}{n_a}\right)^2 \epsilon_k\bar{\epsilon}\right] = -2\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$$

We now focus on the additional terms from the Mills Ratio parts. We have the additional squared terms:

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 w_{k,\lambda}^2\right] = K.\left(\frac{n_d}{n_a}\right)^2 .VAR(w_{k,\lambda})$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{w}_{\lambda}^2\right] = \left(\frac{n_d}{n_a}\right)^2 .VAR(w_{k,\lambda})$$

Combining the latter two expressions we obtain:

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 w_{k,\lambda}\right] + E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{w}_{\lambda}^2\right] = (K+1).\left(\frac{n_d}{n_a}\right)^2 .VAR(w_{k,\lambda}).$$

For the $2ab$ terms in the square, we have to consider the variance terms in $w_{k,\lambda}$, but otherwise there are no other new terms since $E(w_{k,\lambda}) = 0$.

$$E\left[-2\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 w_{k,\lambda}\bar{w}_{\lambda}\right] = -2.\left(\frac{n_d}{n_a}\right)^2 .VAR(w_{k,\lambda})$$

Putting all this together we obtain the *additional* terms for $E(\hat{B})$ for the Mills Ratio part of the expression:

$$(K+1).\left(\frac{n_d}{n_a}\right)^2.VAR(w_{k,\lambda}) - 2.\left(\frac{n_d}{n_a}\right)^2.VAR(w_{k,\lambda}) = (K-1).\left(\frac{n_d}{n_a}\right)^2.VAR(w_{k,\lambda})$$

Simplifying the total expression for $E(\hat{B})$ we get:

$$E[\hat{B}] = \left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_o} + \left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$$

$$+ \left(\frac{n_d}{n_a}\right)^2 .VAR(w_{k,\lambda}) - \frac{1}{(K-1)}.\left(\frac{n_d}{n_a}\right)^2 \frac{2\sigma_{2.1}}{(n_o-1)}\left[\frac{1}{n_d} + \frac{1}{n_o}\right]$$

Asymptotically, as K $\to \infty$, the final term disappears. Furthermore, writing $\pi_d = \frac{n_d}{n_a}$,

$$E[\hat{B}] = \pi_d^2 \left(\frac{\sigma_{2.1}}{n_o} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda})\right).$$

This concludes the term by term derivation of $E[\hat{B}]$.

We now have the components for the within and between multiple imputation variance to plug into equation 4.3.17 in the main body of the document.

## Explanation of simplification step in equation 4.3.18

We claim that

$$\frac{\sigma_{22}}{n} \approx \frac{1}{n}\left[(1-\pi_d)\sigma_{a2o} + \sigma_{a2d}\left(\pi_d + \frac{(1-\pi_d)}{n}\right) + \pi_d(1-\pi_d)\sigma_{22}\left(\dot{\lambda}+\lambda\right)^2\right]. \tag{D.1}$$

We proceed by writing out the expression on the right hand side of D.1 in terms of its constituent parts, collecting terms, and then simplifying.

We use $\beta = (\alpha-\mu)/\sigma$ as shorthand, and assume $\frac{(1-\pi_d)}{n} \approx 0$ in the following,

$$\frac{1}{n}\left[(1-\pi_d)\sigma_{a2o} + \sigma_{a2d}\left(\pi_d + \frac{(1-\pi_d)}{n}\right) + \pi_d(1-\pi_d)\sigma_{22}\left(\dot{\lambda}+\lambda\right)^2\right] \approx$$

$$\frac{1}{n}\left[(1-\pi_d)\left[\sigma_{22}\left[1-\beta\frac{\phi(\beta)}{\Phi(\beta)} - \left(\frac{\phi(\beta)}{\Phi(\beta)}\right)^2\right]\right] +$$

$$\pi_d\left[\sigma_{22}\left[1-\frac{\phi(\beta)}{1-\Phi(\beta)}\left[\frac{\phi(\beta)}{1-\Phi(\beta)}-\beta\right]\right]\right] +$$

$$\pi_d(1-\pi_d)\sigma_{22}\left[\left(\frac{\phi(\beta)}{\Phi(\beta)} + \frac{\phi(\beta)}{(1-\Phi(\beta))}\right)^2\right]\Bigg]. \tag{D.2}$$

Noting that $(1-\pi_d) = \Phi(\beta)$, and taking out $\frac{\sigma_{22}}{n}$, we simplify equation D.2 as

$$\frac{\sigma_{22}}{n}\left[\Phi(\beta)\left[1-\beta\frac{\phi(\beta)}{\Phi(\beta)} - \left(\frac{\phi(\beta)}{\Phi(\beta)}\right)^2\right] +\right.$$

$$(1-\Phi(\beta))\left[1-\frac{\phi(\beta)}{1-\Phi(\beta)}\left[\frac{\phi(\beta)}{1-\Phi(\beta)}-\beta\right]\right] +$$

$$\Phi(\beta)(1-\Phi(\beta))\left[\left(\frac{\phi(\beta)}{\Phi(\beta)}+\frac{\phi(\beta)}{(1-\Phi(\beta))}\right)^2\right]\Bigg] =$$

$$\frac{\sigma_{22}}{n}\Bigg[\Phi(\beta)-\beta\phi(\beta)-\frac{\phi(\beta)^2}{\Phi(\beta)}+$$

$$(1-\Phi(\beta))-\frac{\phi(\beta)^2}{(1-\Phi(\beta))}+\beta\phi(\beta)+$$

$$\Phi(\beta)(1-\Phi(\beta))\phi(\beta)^2\left(\frac{1}{\Phi(\beta)(1-\Phi(\beta)}\right)^2\Bigg] =$$

$$\frac{\sigma_{22}}{n}\left[1-\frac{\phi(\beta)^2}{\Phi(\beta)}-\frac{\phi(\beta)^2}{(1-\Phi(\beta))}+\frac{\phi(\beta)^2}{\phi(\beta)(1-\Phi(\beta))}\right]=\frac{\sigma_{22}}{n}.$$

# Appendix E

# Proof of Lemma 1 regarding variance inflation under CAR

The ratio of the information in the complete data relative to the incomplete data as defined by Rubin's variance estimator, asymptotically as $K$ tends to infinity can be defined as,

$$\frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]} = \frac{n}{2\sigma_{22}} \times \left[ \frac{2\sigma_{22}}{n} + \pi_d(1-\pi_d)\left(VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right) \right.$$

$$\left. \frac{1}{n}\left[\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{2\sigma_{2.1}}{n_o}\right] + \frac{\pi_d}{n}\sigma_{2.1} + \pi_d^2\left(\frac{\sigma_{2.1}}{n_o} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda})\right)\right] =$$

$$1 + \frac{n}{2\sigma_{22}}\pi_d(1-\pi_d)\left(VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right) + \frac{\rho^2}{2} + \frac{(1-\rho^2)}{n_o} + \frac{\pi_d}{2}(1-\rho^2) +$$

$$\pi_d^2\left[\frac{(1-\rho^2)}{2(1-\pi_d)} + \frac{\sigma_{a2d}\pi_d}{2\sigma_{22}} + \frac{n}{2\sigma_{22}}VAR(w_{k,\lambda})\right] =$$

$$1 + \frac{\rho^2}{2} + \frac{(1-\rho^2)}{2}\left[\frac{2}{n_o} + \pi_d + \frac{\pi_d^2}{(1-\pi_d)}\right] +$$

$$\frac{\pi_d \sigma_{a2d}}{2\sigma_{22}} + \frac{n\pi_d}{2\sigma_{22}} \left[(1-\pi_d) VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right]$$

$$= 1 + \frac{\rho^2}{2} + \frac{(1-\rho^2)}{2} \left[\frac{2}{n_o} + \pi_d + \pi_d^2 + \pi_d^3 + \pi_d^4 + \dots\right] +$$

$$\frac{\pi_d \sigma_{a2d}}{2\sigma_{22}} + \frac{n\pi_d}{2\sigma_{22}} \left[(1-\pi_d) VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right], \tag{E.1}$$

where we have used the binomial expansion of $(1-\pi_d)^{-1}$ to replace the last term in the square bracket prior to the last equality. Since $\frac{\sigma_{a2d}}{\sigma_{22}} < 1$, and letting $\frac{2}{n_o} \approx \frac{1}{n_o}$, the expression in equation (E.1) is bounded above by

$$\frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]} \lesssim 1 + \frac{\rho^2}{2} + (1-\rho^2) \left[\frac{1}{n_o} + \pi_d + \pi_d^2 + \pi_d^3 + \pi_d^4 + \dots\right] +$$

$$\frac{\pi_d}{2} + \frac{n\pi_d}{\sigma_{22}} \left[(1-\pi_d) VAR(w_{k,\dot{\lambda}}) + VAR(w_{k,\lambda})\right]$$

$$= 1 + \frac{\rho^2}{2} + (1-\rho^2) \left[\frac{1}{n_o} + \pi_d + \pi_d^2 + \pi_d^3 + \pi_d^4 + \dots\right] +$$

$$\frac{\pi_d}{2} + \frac{n\pi_d}{\sigma_{22}} \left[(1-\pi_d) \frac{\dot{\lambda}^2 C \sigma_{22}}{N} + \frac{\lambda^2 C \sigma_{22}}{N}\right]$$

$$= 1 + \frac{\rho^2}{2} + (1-\rho^2) \left[\frac{1}{n_o} + \pi_d + \pi_d^2 + \pi_d^3 + \pi_d^4 + \dots\right] +$$

$$\frac{\pi_d}{2} + \pi_d C \left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right],$$

where as K tends to infinity, $VAR(w_{k,\lambda}) \approx \frac{\lambda^2 C}{N}$, with $VAR(w_{k,\dot{\lambda}})$ defined analogously, and letting $N = (n_a - 1) \approx n$, with $C = N - 1 - \left[\sqrt{2}\frac{\Gamma((N+1)/2)}{\Gamma(N/2)}\right]^2$.

# Appendix F

# Design based variance estimator when post-deviation data is observed for the de-facto estimand

We will be making use of the following results:

$$E(X^2) = Var(X) + (E(X))^2 \tag{F.1}$$

$$E(\bar{X}) = \mu \tag{F.2}$$

$$E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n} \tag{F.3}$$

We would like to estimate the expected design based variance when post censoring data is fully observed, behaving under the *de-facto* assumption of "Jump to Reference" (J2R) on the active arm (equation 4.4.3 in the main body of the document):

$$E[V_{full,J2R}] = \frac{\hat{\sigma}_{22,r}}{n_r} + \frac{\hat{\sigma}_{22,a}}{n_a} = \frac{\frac{1}{(n_r-1)}\sum_{j=1}^{n_r}(Y_{rj2} - \bar{Y}_{r2})^2}{n_r} + \frac{\frac{1}{(n_a-1)}\sum_{j=1}^{n_a}(Y_{aj2} - \frac{n_o}{n_a}\bar{Y}_{a2o} - \frac{n_d}{n_a}\bar{Y}_{a2d})^2}{n_a} \tag{F.4}$$

where, as for the CAR case,

$$E\left[\frac{\frac{1}{(n_r-1)}\sum_{j=1}^{n_r}(Y_{rj2} - \bar{Y}_{r2})^2}{n_r}\right] = \frac{\sigma_{r22}}{n_r}$$

which we calculate directly, since there is no censoring on the reference arm.

For the active arm, we expand the square on the right-hand side of equation F.4, and calculate the expectation of this expression term by term. Using equation F.1 after decomposing $Y_{aj2}$ for $j = 1, \ldots, n_a$ into its constituent parts of those observed and those deviating at time 2 on the active arm:

$$E\left[\sum_{j=1}^{n_a} Y_{aj2}^2\right] = n_o(\sigma_{a22} + \mu_{a2}) + n_d(\sigma_{d22} + \mu_{d2})$$

Taking the constant terms out of the summation, noting that we sum over $n_a$ terms, and using equation F.3 but with denominator only over the $n_o$ observed terms[1]:

$$E\left[\sum_{j=1}^{n_a}\left(\frac{n_o}{n_a}\right)^2 \bar{Y}_{a2o}^2\right] = n_a\left(\frac{n_o}{n_a}\right)^2\left(\mu_{a2}^2 + \frac{\sigma_{a22}}{n_o}\right).$$

In a similar vane, the expression for those deviating is,

$$E\left[\sum_{j=1}^{n_a}\left(\frac{n_d}{n_a}\right)^2 \bar{Y}_{a2d}^2\right] = n_a\left(\frac{n_d}{n_a}\right)^2\left(\mu_{d2}^2 + \frac{\sigma_{d22}}{n_d}\right).$$

For the mixed $2ab$ term in the square, we firstly take the constants outside the expectation and

[1] $\mu_{a2}$ and $\sigma_{a22}$ are the mean and variance of the terms on the active arm at time 2 for the *observed* patients only.

sum, splitting $\sum_{j=1}^{n_a} Y_{aj2} = \sum_{j=1}^{n_o} Y_{a2o} + \sum_{j=1}^{n_d} Y_{a2d}$,

$$E\left[\sum_{j=1}^{n_a} -2\left(\frac{n_o}{n_a}\right) Y_{aj2}\bar{Y}_{a2o}\right] = -2\left(\frac{n_o}{n_a}\right) E\left[\left(\sum_{j=1}^{n_o} Y_{a2o} + \sum_{j=1}^{n_d} Y_{a2d}\right) \times \sum_{j=1}^{n_o} \bar{Y}_{a2o}\right]$$

$$= -2\left(\frac{n_o}{n_a}\right)\left(n_o E\left[\sum_{j=1}^{n_o} \bar{Y}_{ajo}^2\right] + E\left[\sum_{j=1}^{n_a} Y_{a2d}\bar{Y}_{a2o})\right]\right)$$

The first term in the above expression we have already calculated using equation F.3, and for the second term we can take expectations separately by assuming the observed and deviating patients are independent. Using equation F.2, and noting the summation is over $n_a$, the expression above becomes,

$$-2\left(\frac{n_o^2}{n_a}\right)\left(\mu_{a2}^2 + \frac{\sigma_{a22}}{n_o}\right) - 2\left(\frac{n_o}{n_a}\right) \times n_d\mu_{d2}\mu_{a2}$$

Similarly, the term with the mean of those deviating:

$$E\left[\sum_{j=1}^{n_a} -2\left(\frac{n_d}{n_a}\right) Y_{aj2}\bar{Y}_{a2d}\right] = -2\left(\frac{n_d^2}{n_a}\right)\left(\mu_{a2}^2 + \frac{\sigma_{d22}}{n_d}\right) - 2\left(\frac{n_d n_o}{n_a}\right)\mu_{d2}\mu_{a2}$$

Finally, we take the expectations of each of the observed and deviating terms, noting the summation is over $n_a$, which cancels out one of the denominator terms in $n_a$:

$$E\left[\sum_{j=1}^{n_a} 2 \times \left(\frac{n_o n_d}{n_a^2}\right)\bar{Y}_{a2o}\bar{Y}_{a2d}\right] = 2\left(\frac{n_o n_d}{n_a}\right)\mu_{d2}\mu_{a2}$$

Putting all the terms together, cancelling where necessary, and noting that $n_a - n_o = n_d$ and $n_a - n_d = n_o$, we get the overall expression:

$$E\left[\frac{\hat{\sigma}_{22,a}^2}{n_a}\right] = E\left[\frac{\frac{1}{(n_a-1)}\sum_1^{n_a}(Y_{aj2} - \frac{n_o}{n_a}\bar{Y}_{a2} - \frac{n_d}{n_a}\bar{Y}_{d2})^2}{n_a}\right] =$$

$$\frac{\frac{1}{(n_a-1)}\left((n_a-1)\left(\frac{n_o}{n_a}\sigma_{a22}+\frac{n_d}{n_a}\sigma_{d22}\right)+\frac{n_o n_d}{n_a}(\mu_{a2}-\mu_{d2})^2\right)}{n_a}.$$

Simplifying, and assuming equal sized trial arms $n_r = n_a = n$, with $(n_a - 1) \approx n_a$ this expression becomes:

$$E\left[\frac{\hat{\sigma}^2_{22,a}}{n}\right] = \frac{\left(\frac{n_o}{n}\sigma_{a22}+\frac{n_d}{n}\sigma_{d22}\right)+\frac{n_o n_d}{n^2}(\mu_{a2}-\mu_{d2})^2}{n}$$

For censored data, we can use the standard results from the truncated normal distribution, so we substitute $\sigma_{a22} = \sigma_{a2o}$ and $\sigma_{d22} = \sigma_{a2d}$, and let $\Delta = (\mu_{a2} - \mu_{d2}) = (\mu_{a2o} - \mu_{a2d})$:

$$E\left[\frac{\hat{\sigma}^2_{22,a}}{n}\right] = \frac{\left(\frac{n_o}{n}\sigma_{a2o}+\frac{n_d}{n}\sigma_{a2d}\right)+\frac{n_o n_d}{n^2}\Delta_c^2}{n},$$

so that the full expression becomes:

$$E[V_{full,J2R}] = \frac{\sigma_{22}}{n} + \frac{\left(\frac{n_o}{n}\sigma_{a2o}+\frac{n_d}{n}\sigma_{a2d}\right)+\frac{n_o n_d}{n^2}\Delta_c^2}{n}.$$

# Appendix G

# Rubin's variance under the *de-facto* assumption of Jump to Reference (J2R)

We want to derive the expression for $E[V(\hat{\theta}_{MI,J2R})]$ which is equation 4.4.2 in the main body of the document.

We follow the same approach as for CAR. For the observed cases, we firstly write out the full summation to determine which new terms we have for the J2R case:

$$E\left[\sum_{j\in o}(Y_{aj2}-\hat{\mu}_{a2,k})^2\right]=$$

$$E\left[\sum_{j\in o}\left((Y_{aj2}-\bar{Y}_{a2o})+\frac{n_d}{n_a}(\bar{Y}_{a2o}-\bar{Y}_{r2})-\frac{n_d}{n_a}u_k-\frac{n_d}{n_a}\left(\frac{r}{q}+b_k\right)(\bar{Y}_{a1d}-\bar{Y}_{r1})-\right.\right.$$

$$\left.\left.\frac{n_d}{n_a}\bar{\lambda}_{r2}\sqrt{\bar{\tilde{\sigma}}_{22,k}}-\frac{n_d}{n_a}w_{k,\lambda_{r2}}-\frac{n_d}{n_a}\bar{\epsilon}_k\right)^2\right]$$

Deriving term by term, and simplifying we obtain something similar to the CAR case:

$$(n_a-1)E(\hat{\sigma}_a^2)=(n_o-1)\sigma_{22}\left[1-\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}-\left(\frac{\phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}{\Phi\left(\frac{\alpha-\mu_{a2}}{\sqrt{\sigma_{22}}}\right)}\right)^2\right]+$$

$$(n_d-1)\,\sigma_{22}\left[1-\frac{\phi(\frac{\alpha-\mu_{r22}}{\sigma_{22}})}{1-\Phi(\frac{\alpha-\mu_{r2}}{\sigma_{22}})}\left[\frac{\phi(\frac{\alpha-\mu_{r2}}{\sigma_{22}})}{1-\Phi(\frac{\alpha-\mu_{r2}}{\sigma_{22}})}-\left(\frac{\alpha-\mu_{r2}}{\sigma_{22}}\right)\right]\right]+$$

$$\left(\frac{n_d}{n_a}\right)^2\left[n_o(\mu_{r2}-\mu_{a2o})^2+\left(\sigma_{a2o}+\frac{n_o}{n_r}\sigma_{22}\right)\right]+\left(\frac{n_o}{n_a}\right)^2 n_d\left[(\mu_{a2o}-\mu_{r2})^2+\frac{\sigma_{a2o}}{n_o}+\frac{\sigma_{22}}{n_r}\right]+$$

$$n_o\left(\frac{n_d}{n_a}\right)^2\frac{\sigma_{2.1}}{n_r}+n_d\left(\frac{n_o}{n_a}\right)^2\frac{\sigma_{2.1}}{n_r}+$$

$$n_o\left(\frac{n_d}{n_a}\right)^2\left[\left(\frac{\sigma_{12}}{\sigma_{11}}\right)^2+\frac{2\sigma_{2.1}}{(n_r-1)\sigma_{11}}\right]\left[\sigma_{11}\left(\frac{1}{n_d}+\frac{1}{n_r}\right)+(\mu_{a1}-\mu_{r1})^2\right]+$$

$$n_d\left(\frac{n_o}{n_a}\right)^2\left[\left(\frac{\sigma_{12}}{\sigma_{11}}\right)^2+\frac{2\sigma_{2.1}}{(n_r-1)\sigma_{11}}\right]\left[\sigma_{11}\left(\frac{1}{n_d}+\frac{1}{n_r}\right)+(\mu_{a1}-\mu_{r1})^2\right]+$$

$$\left(\frac{n_o n_d}{n_a^2}\right)\sigma_{a2d}+\left(\frac{n_o}{n_a}\right)^2\sigma_{a2d}+$$

$$-2\left(\frac{n_d}{n_a}\right)^2 n_o\frac{\sigma_{12}}{\sigma_{11}}\left(\frac{\sigma_{12}}{n_r}+(\mu_{a1}-\mu_{r1})(\mu_{a2}-\mu_{r2})\right)-2n_d\left(\frac{n_o}{n_a}\right)^2\frac{\sigma_{12}}{\sigma_{11}}\left(\frac{\sigma_{12}}{n_r}+(\mu_{a1}-\mu_{r1})(\mu_{a2}-\mu_{r2})\right)+$$

$$n_o\lambda_{r2}^2\left(\frac{n_d}{n_a}\right)^2\sigma_{22}+n_d\lambda_{r2}^2\left(\frac{n_o}{n_a}\right)^2\sigma_{22}+$$

$$n_o\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda_{r2}})+n_d\left(\frac{n_o}{n_a}\right)^2 VAR(w_{k,\lambda_{r2}})+$$

$$2n_o\left(\frac{n_d}{n_a}\right)^2\left(\frac{\sigma_{12}}{\sigma_{11}}\right)\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{a1d}-\mu_{r1})+2n_d\left(\frac{n_o}{n_a}\right)^2\frac{\sigma_{12}}{\sigma_{11}}\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{a1d}-\mu_{r1})+$$

$$2n_o\left(\frac{n_d}{n_a}\right)^2\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{r2}-\mu_{a2o})+2n_d\left(\frac{n_o}{n_a}\right)^2\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{r2}-\mu_{a2o})$$

Collecting terms and simplifying again:

$$(n_a-1)E(\hat{\sigma}_a^2)=n_o\left(1-\frac{1}{n_a}\right)\sigma_{a2o}+n_d\left(1-\frac{1}{n_a}\right)\sigma_{a2d}+$$

$$\left(\frac{n_d n_o}{n_a}\right)VAR(w_{k,\lambda_{r2}})+$$

$$2\left(\frac{n_d n_o}{n_a}\right)\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{r2}-\mu_{a2o})+\left(\frac{n_o n_d}{n_a}\right)(\mu_{r2}-\mu_{a2o})^2+$$

$$\left(\frac{n_d n_o}{n_a}\right)\sigma_{22}\left(\lambda_{r2}^2+\frac{1}{n_r}\right)+$$

$$\frac{n_o^2}{n_a^2}\frac{\sigma_{12}^2}{\sigma_{11}}+$$

$\frac{n_o(3n_d+2n_a)}{n_a^3}\sigma_{2.1}$.

In the final term in this expression we have assumed $n_a = n_r$ and $n_r \approx (n_r - 1)$.

Furthermore, let $\pi_d = n_d/n_a$, $(1-\pi_d) = \frac{n_o}{n_a}$, $(n_o - 1) = n_o$, $(n_d - 1) = n_d$, $n_r = n_a = n$, and we also divide by $(n_a - 1) = n_a$, then let $\left(1 - \frac{1}{n_a}\right) \approx 1$:

$E(\hat{\sigma}_a^2) \approx (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d} +$

$\pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) +$

$2\pi_d(1-\pi_d)\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{r2} - \mu_{a2o}) + \pi_d(1-\pi_d)(\mu_{r2} - \mu_{a2o})^2 +$

$\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}^2 +$

$\frac{(1-\pi_d)^2}{n}\frac{\sigma_{12}^2}{\sigma_{11}} +$

$\frac{3\pi_d(1-\pi_d)^2}{n^2}\sigma_{2.1}$.

This expression has taken account of the within imputation variance, but we still need to add the between imputation variance $E(\hat{B})$.

The derivation for $E(\hat{B})$ proceeds as for the CAR case. Note that the $\lambda$ and $\bar{\lambda}$ terms cancel out since they are invariant over $k$.

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 u_k^2\right] = K\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_r}$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \left(\frac{r}{q}+b_k\right)^2 (\bar{Y}_{a1d}-\bar{Y}_{r1})^2\right] =$$

$$K\left(\frac{n_d}{n_a}\right)^2 \left[\left(\frac{\sigma_{12}}{\sigma_{11}}\right)^2 + \frac{2\sigma_{2.1}}{(n_r-1)\sigma_{11}}\right]\left[\sigma_{11}\left(\frac{1}{n_d}+\frac{1}{n_r}\right)+(\mu_{a1}-\mu_{r1})^2\right]$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{\epsilon}_k^2\right] = K\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{u}^2\right] = \left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_r}$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \left(\frac{r}{q}+\bar{b}\right)^2 (\bar{Y}_{a1d}-\bar{Y}_{r1})^2\right] =$$

$$K\left(\frac{n_d}{n_a}\right)^2 \left[\left(\frac{\sigma_{12}}{\sigma_{11}}\right)^2 + \frac{\sigma_{2.1}}{(n_r-1)\sigma_{11}}\frac{(K+1)}{K}\right]\left[\sigma_{11}\left(\frac{1}{n_d}+\frac{1}{n_r}\right)+(\mu_{a1}-\mu_{r1})^2\right]$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{\epsilon}^2\right] = \left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$$

$$E\left[\sum_{k=1}^{K} -2\left(\frac{n_d}{n_a}\right)^2 u_k\bar{u}\right] = -2\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{2.1}}{n_r}$$

$$E\left[\sum_{k=1}^{K} -2\left(\frac{n_d}{n_a}\right)^2 \left(\frac{r}{q}+b_k\right)\left(\frac{r}{q}+\bar{b}\right)(\bar{Y}_{a1d}-\bar{Y}_{r1})^2\right] =$$

$$-2K\left(\frac{n_d}{n_a}\right)^2 \left[\left(\frac{\sigma_{12}}{\sigma_{11}}\right)^2 + \frac{\sigma_{2.1}}{(n_r-1)\sigma_{11}}\frac{(K+1)}{K}\right]\left[\sigma_{11}\left(\frac{1}{n_d}+\frac{1}{n_r}\right)+(\mu_{a1}-\mu_{r1})^2\right]$$

$$E\left[\sum_{k=1}^{K} -2\left(\frac{n_d}{n_a}\right)^2 \epsilon_k\bar{\epsilon}\right] = -2\left(\frac{n_d}{n_a}\right)^2 \frac{\sigma_{a2d}}{n_d}$$

With the additional squared terms:

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 w_{k,\lambda}^2\right] = K\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda})$$

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{w}_{\lambda}^2\right] = \left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda})$$

Combining these terms:

$$E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 w_{k,\lambda}\right] + E\left[\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^2 \bar{w}_{\lambda}^2\right] = (K+1)\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda})$$

For the $2ab$ terms in the square, we have to also consider the variance terms in $w_{k,\lambda}$. However, we also note that since $E(w_{k,\lambda}) = 0$, other $2ab$ terms in the square disappear.

$$E\left[-2\sum_{k=1}^{K}\left(\frac{n_d}{n_a}\right)^3 w_{k,\lambda}\bar{w}_{\lambda}\right] = -2.\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda})$$

Putting all this together we obtain the *additional* Mills Ratio terms for $E(\hat{B})$:

$$(K+1)\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda}) - 2\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda}) = (K-1)\left(\frac{n_d}{n_a}\right)^2 VAR(w_{k,\lambda})$$

Simplifying the total expression we obtain,

$$E[\hat{B}] = \pi_d^2 \frac{\sigma_{2.1}}{n} + \pi_d^2 \frac{\sigma_{a2d}}{n_d} + \pi_d^2 VAR(w_{k,\lambda}) + \frac{\pi_d^2(1+\pi_d)}{n^2 \pi_d}\sigma_{2.1},$$

Writing $\pi_d = \frac{n_d}{n_a}$, letting $n_r \approx (n_r - 1)$ and assuming $n_r = n_a = n$:

$$E[\hat{B}] = \pi_d^2 \left( \frac{\sigma_{2.1}}{n} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) + \frac{(1+\pi_d)}{n^2 \pi_d}\sigma_{2.1} \right).$$

Altogether, we obtain an expression for the variance on the active arm:

$$E(\hat{\sigma}_a^2) = (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d} + \pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) +$$

$$2\pi_d(1-\pi_d)\lambda_{r2}\sqrt{\sigma_{22}}(\mu_{r2}-\mu_{a2o}) + \pi_d(1-\pi_d)(\mu_{r2}-\mu_{a2o})^2 +$$

$$\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}^2 + \frac{(1-\pi_d)^2}{n}\frac{\sigma_{12}^2}{\sigma_{11}} + \frac{3\pi_d(1-\pi_d)^2}{n^2}\sigma_{2.1} +$$

$$\left(1+\frac{1}{K}\right)\pi_d^2\left(\frac{\sigma_{2.1}}{n} + \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) + \frac{(1+\pi_d)}{n^2\pi_d}\sigma_{2.1}\right).$$

To obtain the variance of the treatment difference under J2R following MI, we just add the expression above to the variance for the reference arm $\frac{E[\hat{\sigma}_r^2]}{n} = \frac{\sigma_{22}}{n}$, and asymptotically assuming K $\to \infty$,

$$E[V(\hat{\theta}_{MI,J2R})] = \left[\frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d}\right] +$$

$$+\pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) + 2\pi_d(1-\pi_d)\lambda_{r2}\sqrt{\sigma_{22}}\Delta_c + \pi_d(1-\pi_d)\Delta_c^2 +$$

$$\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}^2 + \frac{(1-\pi_d)^2}{n}\rho^2\sigma_{22} + \frac{3\pi_d(1-\pi_d)^2}{n^2}\sigma_{22}(1-\rho^2) +$$

$$\pi_d^2\left(\frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) + \left[\frac{1}{n} + \frac{(1+\pi_d)}{n^2\pi_d}\right]\sigma_{22}(1-\rho^2)\right),$$

which is equation 4.4.2 in the main body of the document.

# Appendix H

# Proof for information anchoring property for Jump to Reference

In a first step, we calculate the predicted variance $E[V_{anchored}]$ using equation (4.3.21) from Lemma 1 and equation (4.4.3) from Lemma 2,

$$E[\hat{V}_{anchored}] \approx E[V(\hat{\theta}_{full,J2R})] \times \frac{E[V(\hat{\theta}_{MI,CAR})]}{E[V(\hat{\theta}_{full,CAR})]} =$$

$$\left[\frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d} + \Delta_c^2\pi_d(1-\pi_d)\right] \times$$

$$1 + \frac{\rho^2}{2} + (1-\rho^2)\left[\frac{1}{n_o} + \pi_d + \pi_d^2 + \pi_d^3 + \pi_d^4 + \dots\right] +$$

$$\frac{\pi_d}{2} + \pi_d C\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right],$$

with $\Delta_c = \mu_{a2o} - \mu_{a2d}$ as defined for the information anchoring CAR case.

Multiplying out term by term:

$$E[\hat{V}_{anchored}] = \left[\frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d} + \pi_d(1-\pi_d)\Delta_c^2\right] +$$

$$\frac{\rho^2\sigma_{22}}{2n} + \frac{\rho^2(1-\pi_d)\sigma_{a2o}}{2} + \frac{\pi_d\rho^2\sigma_{a2d}}{2} + \frac{\rho^2\pi_d(1-\pi_d)\Delta_c^2}{2} +$$

$$\frac{(1-\rho^2)\sigma_{22}}{nn_o} + \frac{(1-\rho^2)(1-\pi_d)\sigma_{a2o}}{n_o} + \frac{(1-\rho^2)\pi_d\sigma_{a2d}}{n_o} + \frac{(1-\rho^2)\pi_d(1-\pi_d)\Delta_c^2}{n_o} +$$

$$\frac{(1-\rho^2)\pi_d\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d\sigma_{a2o} + (1-\rho^2)\pi_d^2\sigma_{a2d} + (1-\rho^2)\pi_d^2(1-\pi_d)\Delta_c^2$$

$$\frac{(1-\rho^2)\pi_d^2\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d^2\sigma_{a2o} + (1-\rho^2)\pi_d^3\sigma_{a2d} + (1-\rho^2)\pi_d^3(1-\pi_d)\Delta_c^2$$

$$\frac{(1-\rho^2)\pi_d^3\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d^3\sigma_{a2o} + (1-\rho^2)\pi_d^4\sigma_{a2d} + (1-\rho^2)\pi_d^4(1-\pi_d)\Delta_c^2$$

$$\frac{(1-\rho^2)\pi_d^4\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d^4\sigma_{a2o} + \cdots +$$

$$\frac{\pi_d\sigma_{22}}{2n} + \frac{(1-\pi_d)\pi_d\sigma_{a2o}}{2} + \frac{\pi_d^2\sigma_{a2d}}{2} + \frac{\Delta_c^2\pi_d^2(1-\pi_d)}{2} +$$

$$\frac{\pi_d C\sigma_{22}}{n}\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right] +$$

$$\pi_d(1-\pi_d)\sigma_{a2o}C\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right] +$$

$$\pi_d^2 \sigma_{a2d} C \left[ (1-\pi_d)\dot{\lambda}^2 + \lambda^2 \right] +$$

$$\pi_d^2 (1-\pi_d) \Delta_c^2 C \left[ (1-\pi_d)\dot{\lambda}^2 + \lambda^2 \right].$$

To calculate the difference between the variance bound $E[V(\hat{\theta}_{MI,J2R})]$, and that predicted using information anchoring theory, $E[V_{anchored}]$, we write down the terms for both expressions:

$$E[V(\hat{\theta}_{MI,J2R})] - E[V_{anchored}] =$$

$$\left[ \frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d \sigma_{a2d} \right] +$$

$$+\pi_d(1-\pi_d) VAR(w_{k,\lambda_{r2}}) + 2\pi_d(1-\pi_d)\lambda_{r2}\sqrt{\sigma_{22}}\Delta_c + \pi_d(1-\pi_d)\Delta_c^2 +$$

$$\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}^2 + \frac{(1-\pi_d)^2}{n}\rho^2\sigma_{22} + \frac{3\pi_d(1-\pi_d)^2}{n^2}\sigma_{22}(1-\rho^2) +$$

$$\pi_d^2 \left( \frac{\sigma_{a2d}}{n_d} + VAR(w_{k,\lambda}) + \left[ \frac{1}{n} + \frac{(1+\pi_d)}{n^2 \pi_d} \right] \sigma_{22}(1-\rho^2) \right) -$$

$$\left( \left[ \frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d \sigma_{a2d} + \pi_d(1-\pi_d)\Delta_c^2 \right] + \right.$$

$$\frac{\rho^2 \sigma_{22}}{2n} + \frac{\rho^2(1-\pi_d)\sigma_{a2o}}{2} + \frac{\pi_d \rho^2 \sigma_{a2d}}{2} + \frac{\rho^2 \pi_d (1-\pi_d)\Delta_c^2}{2} +$$

$$\frac{(1-\rho^2)\sigma_{22}}{n n_o} + \frac{(1-\rho^2)(1-\pi_d)\sigma_{a2o}}{n_o} + \frac{(1-\rho^2)\pi_d \sigma_{a2d}}{n_o} + \frac{(1-\rho^2)\pi_d(1-\pi_d)\Delta_c^2}{n_o} +$$

$$\frac{(1-\rho^2)\pi_d\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d\sigma_{a2o} + (1-\rho^2)\pi_d^2\sigma_{a2d} + (1-\rho^2)\pi_d^2(1-\pi_d)\Delta_c^2$$

$$\frac{(1-\rho^2)\pi_d^2\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d^2\sigma_{a2o} + (1-\rho^2)\pi_d^3\sigma_{a2d} + (1-\rho^2)\pi_d^3(1-\pi_d)\Delta_c^2$$

$$\frac{(1-\rho^2)\pi_d^3\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d^3\sigma_{a2o} + (1-\rho^2)\pi_d^4\sigma_{a2d} + (1-\rho^2)\pi_d^4(1-\pi_d)\Delta_c^2$$

$$\frac{(1-\rho^2)\pi_d^4\sigma_{22}}{n} + (1-\rho^2)(1-\pi_d)\pi_d^4\sigma_{a2o} + \cdots +$$

$$\frac{\pi_d\sigma_{22}}{2n} + \frac{(1-\pi_d)\pi_d\sigma_{a2o}}{2} + \frac{\pi_d^2\sigma_{a2d}}{2} + \frac{\Delta_c^2\pi_d^2(1-\pi_d)}{2} +$$

$$\frac{\pi_d C\sigma_{22}}{n}\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right] +$$

$$\pi_d(1-\pi_d)\sigma_{a2o}C\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right] +$$

$$\pi_d^2\sigma_{a2d}C\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right] +$$

$$\pi_d^2(1-\pi_d)\Delta_c^2 C\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right]\Big).$$

We then subtract similar terms, and simplify, disregarding any terms of $o(1/n^3)$ or smaller:

$$E[V(\hat{\theta}_{MI,J2R})] - E[V_{anchored}] = \Big[\frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d}\Big] -$$

$$\Big[\frac{\sigma_{22}}{n} + (1-\pi_d)\sigma_{a2o} + \pi_d\sigma_{a2d}\Big] -$$

$$\sigma_{a2o}\Big[\frac{\rho^2(1-\pi_d)}{2} + \frac{(1-\rho^2)(1-\pi_d)}{n_o} + (1-\rho^2)(1-\pi_d)\pi_d + (1-\rho^2)(1-\pi_d)\pi_d^2 + \frac{(1-\pi_d)\pi_d}{2}\Big] -$$

$$\sigma_{a2d}\Big[\frac{\pi_d\rho^2}{2} + \frac{(1-\rho^2)\pi_d}{n_o} + (1-\rho^2)\pi_d^2 + \frac{\pi_d^2}{2}\Big] +$$

$$\pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) + \pi_d^2 VAR(w_{k,\lambda}) +$$

$$2\pi_d(1-\pi_d)\lambda_{r2}\sqrt{\sigma_{22}}\Delta_c + [\pi_d(1-\pi_d)\Delta_c^2 - \pi_d(1-\pi_d)\Delta_c^2] - \frac{\Delta_c^2\pi_d^2(1-\pi_d)}{2} - \pi_d^2(1-\pi_d)\Delta_c^2 C\Big[(1-\pi_d)\dot{\lambda}^2 + \lambda$$

$$-\frac{\rho^2\pi_d(1-\pi_d)\Delta_c^2}{2} - \frac{(1-\rho^2)\pi_d(1-\pi_d)\Delta_c^2}{n_o} - (1-\rho^2)\pi_d^2(1-\pi_d)\Delta_c^2 +$$

$$\pi_d(1-\pi_d)\sigma_{22}\lambda_{r2}^2 +$$

$$\frac{(1-\pi_d)^2}{n}\rho^2\sigma_{22} - \frac{\rho^2\sigma_{22}}{2n} - \frac{(1-\rho^2)\sigma_{22}}{nn_o} - \frac{(1-\rho^2)\pi_d\sigma_{22}}{n} - \frac{\pi_d\sigma_{22}}{2n} -$$

$$C\Big[\frac{\pi_d\sigma_{22}}{n} + \pi_d(1-\pi_d)\sigma_{a2o} + \pi_d^2\sigma_{a2d}\Big]\Big[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\Big].$$

The first two terms in the above expression cancel out. For the remaining terms we only consider terms of $o(\frac{1}{n})$ or larger and simplify accordingly,

$$E[V(\hat{\theta}_{MI,J2R})] - E[V_{anchored}] \lessapprox 2\pi_d(1-\pi_d)\sqrt{\sigma_{22}}\lambda_{r2}\Delta_c +$$

$$\sigma_{22}\left[\frac{\rho^2}{2n} + \pi_d(1-\pi_d)\lambda_{r2}^2\right] + \pi_d(1-\pi_d)VAR(w_{k,\lambda_{r2}}) + \pi_d^2 VAR(w_{k,\lambda}) -$$

$$\sigma_{a2o}\left[\frac{\rho^2}{2}(1-\pi_d) + \frac{3}{2}\pi_d + \left[\frac{\pi_d(1-\pi_d)}{2}\right]\left[(1-\pi_d)\dot{\lambda}^2 + \lambda^2\right]\right] - \sigma_{a2d}\left[\frac{\rho^2\pi_d}{2}\right] - \Delta_c^2\left[\frac{\rho^2\pi_d}{2}\right],$$

with $C \approx 0.5$ for large $n$.

In absolute terms this bound is dominated by the first two positive terms, and the negative terms in $\sigma_{a2o}$ and $\Delta_c^2$. Focussing just on these four terms, they approximately cancel one another out, with the remaining difference being of the order of 10% of $\sigma_{22}$.

This is the upper bound used in equation (4.4.5) in the main body of the document.

# Appendix I

# Survival function for the pooled logistic model

We start by discussing equivalency between the pooled logistic and Cox proportional hazards model from first principles. The approach is based on the early work of Green and Symons (1983), Efron (1988), and Thompson (1977), which has been recently reviewed in an article from Ngwa *et al.* (2016).

Let

$n_i$ = the number of patients at risk at the beginning of month $i$,
$s_i$ = the number of patients having the event of interest during month $i$,
$s_i^{'}$ =the number of patients censored during month $i$ ,

We assume that the number of events, $s_i$ is binomially distributed given $n_i$,

$$s_i|n_i \sim \text{Bin}(n_i, h_i) \text{ independently for } \ i = 1, 2, \ldots, N$$

with

$h_i$ the discrete hazard i.e. the probability that a patient experiences the event during the $i$th interval conditional on surviving until the beginning of the $i$th interval.

Further, we consider the $n_i$ to be fixed at their observed values and assume independence. Efron makes the point that although this may not be true in all cases, it is reasonable under "usual assumptions for censored data".

For the discretised data the survival function is defined as

$$S_i = \prod_{1 \leq j \leq i} (1 - h_j),$$

the probability that a patient survives during the first $i - 1$ intervals, with $S_1 = 1$ by definition. We can estimate the $h_i$ using logistic regression with parameter $\lambda_i$ defined as

$$\lambda_i = \log \left[ \frac{h_i}{(1 - h_i)} \right],$$

which can be rearranged to define the discrete hazard in terms of the parameters from the fitted logistic model

$$h_i = \frac{1}{1 + \exp(-\lambda_i)}$$

Analogously, if we introduce covariates $x_i$ (1 x $p$), the logistic model becomes $\lambda_i = x_i \alpha$ for unknown model parameters $\alpha$ ($p$ x 1), and we can determine the MLEs $\hat{\alpha}$, resulting in the MLE of the hazard:

$$\hat{h}_i = \frac{1}{1 + \exp(\alpha_0 - x_i \hat{\alpha})},$$

where $\alpha_0$ is the intercept term of the model and $\hat{\alpha}$ is the vector of estimates pertaining to the covariates.

Efron called this conditional logistic regression because of the conditionality of the definition of the original binomial distributions.

The probability of survival of the follow-up period $T$ is defined as (from Green and Symons):

$$S(T|\boldsymbol{X},\alpha) = \frac{\exp(-\alpha_0 - \boldsymbol{X}\hat{\alpha})}{1+\exp(-\alpha_0 - \boldsymbol{X}\hat{\alpha})}$$

For discrete time intervals, the survival function is defined as:

$$S_i = \prod_{1\leq j\leq i}[1-(1+\exp(\alpha_0 - x_i\hat{\alpha}))^{-1}],$$

We note that in all of these expressions the follow-up time is not explicitly defined for each individual since it is assumed to be the same for all individuals for discrete time follow-up data. This is of course not the case for the Cox model. Green and Symons explain that the Cox and logistic models may be viewed to be equivalent when the event of interest is relatively rare and follow-up short.

Formal arguments of equivalency between time dependent Cox and pooled logistic models are presented in the appendix of the paper by D'Agostino *et al.* (1990).

# Appendix J

# PCP risk models

We expanded the data set to have one record per patient per time slot (i.e. quarter) in which the patient was involved in a trial. Let $Y_{m+k,i}$ be the indicator for PCP diagnosis (or death) for subject $i$ at the end of period $k$ in trial $m$, $C_{m+k,i}$ (1: censored, 0 = uncensored) be the indicator for censoring at the end of period $k$ for subject $i$ in trial $m$, $A_{m,i}$ is the exposure arm assignment for patient $i$ in trial $m$, and $L_{m+k,i}$ are time fixed ($k = 0$) and time varying ($k = 1, \ldots, K_m$) covariates at the end of period $k$ for subject $i$ in trial $m$. In the following, over bars are used to denote histories up to and including the period defined by the period subscript $k$.

For the primary analysis, we fit an inverse probability weighted pooled logistic models to estimate the IPW adjusted hazard ratio for each trial $m$:

$$\text{logit}\left[Pr(Y_{m+k+1,i} = 1 | A_{m,i}, \bar{L}_{m+k,i}, \bar{Y}_{m+k,i} = \bar{0}, \bar{C}_{m+k+1,i} = \bar{0})\right] = \beta_{0,m+k,i} + \beta_1 A_{m,i} + \beta_2^T L_{m+k+1,i}, \tag{J.1}$$

where
$Pr(Y_{m+k+1,i})$ is the probability of PCP diagnosis for the $k$th time period which starts in the $m$th quarter after 1st quarter 1998 (when $m = 0$). $m$ is the baseline month of the trial i.e. an indicator for the emulated trial,
$\beta_{0,m+k,i}$ is a function for the time varying intercept (i.e. the function for the baseline hazard) for trial $m$ including terms for $time$, $time^2$ and $time^3$ within trial $m$,

$\beta_1$ is the estimated log hazard ratio for PCP prophylaxis averaged over the follow-up period, and,
$\beta_2$ is a vector of estimated log hazard ratios for the covariates.

The following code implements the analysis model in R:

```
amod1 <- svyglm(EVENT~
                # pcp diagnosis (or death for the secondary endpoint)
                  as.factor(armIND)
                # treatment 0=ON / 1= OFF PCP prophylaxis
                # continous variable for the quarter in which trial starts
                # with the squared and cubic terms to allow for a flexible
                # shape of the baseline hazard
                +trial_time+I(trial_time^2)+I(trial_time^3)
                +trial+I(trial^2)
                # the quarter in which trial starts = trial identifier
                # time fixed covariates and their squares
                +b_sCD4+I(b_sCD4^2)+b_log10RNA+I(b_log10RNA^2)+
                factor(gender)+
                factor(mode2)+ #  mode of transmission
                factor(origin)+ # geographical origin
                factor(cohort)+
                b_age +I(b_age^2) # baseline age
                +YRbase # calendar year in which this trial starts
                , family = quasibinomial()
                , design = svydesign(id = ~patient
                , weights = ~sw.trunc # truncated weights
                , data = dat))
```

# Appendix K

# Inverse probability weights

## K.1 Inverse Probability Weights

Let $C_{k,i}$ (1: censored, 0 = uncensored) be the indicator for censoring (all types) at the end of period $k$ (i.e. quarter) for subject $i$, $A_i$ is the exposure arm for patient $i$ in this trial, $V_i$ are time fixed (baseline) covariates for subject $i$ in this trial, and $L_{k,i}$ are the time varying covariates at the end of period $k$ for subject $i$ for this trial. Over bars are used to denote histories up to and including the period defined by the period subscript $k$. We drop the subscript for the trial $m$ to reduce notation complexity in the following.

The stabilised weights for all types of censoring are defined as:

$$\text{SW}_{k,i}^{C} = \prod_{k=1}^{k+1} \frac{Pr(C_{k,i} = 0 | \bar{C}_{k-1,i} = 0, A_i, V_i)}{Pr(C_{k,i} = 0 | \bar{C}_{k-1,i} = 0, A_i, \bar{L}_{k-1,i})},$$

where $\bar{L}_{0,i}$ are the baseline covariates for subject $i$. The denominator is, informally, the subject's probability of remaining uncensored through period $k$ given baseline and time varying confounders. When the outcome is also expected to have an effect on drop-out then this can also be added to the denominator of the model (not shown).

The probability of being uncensored through visit $k$ is estimated by fitting a pooled logistic model (see example code below):

$$\text{logit}\left[Pr(C_{k,i}=0|\bar{C}_{k-1,i}=0,A_i,\bar{L}_{k-1,i})\right]=\psi_0+\psi_1 A_i+\psi_2^T\bar{L}_{k-1},$$

where
$\psi_0$ is the intercept term,
$\psi_1$ estimates the odds ratio of being on treatment, and
$\psi_2$ is a vector of estimates for the covariate history up to time $k-1$.

The numerator being defined similarly, but without including time varying covariates:

$$\text{logit}\left[Pr(C_{k,i}=0|\bar{C}_{k-1,i}=0,A_i,V_i)\right]=\psi_0+\psi_1 A_i+\psi_2^T V_i,$$

The numerator stabilises the weights to reduce the variance of the estimates in the final model.

We note that using the IP weights in this way implicitly makes the assumption that censoring is at random, so the results should be equivalent to those from an analysis using MI under CAR.

The following code implements the IP weights in R:

```
dat$notcensor <- 1- dat$c # c = 1 is censored, 0 if not censored, or event

# denomator of IPWeights

mod <- glm(notcensor ~ as.factor(armIND) # treatment indicator
b_sCD4+I(b_sCD4^2)+s_CD4+I(s_CD4^2)+ # baseline and time varying covariates
b_log10RNA+I(b_log10RNA^2)+log10_RNA+I(log10_RNA^2)+factor(gender)+
factor(mode2)+factor(origin)+b_age+I(b_age^2)+YRbase
            ,family = binomial()
           , data = dat)

test$probC.d <- predict(mod, type = 'response')

# numerator of IPWeights; as above but without time varying covariates
mod <- glm(notcensor ~ as.factor(armIND)+b_sCD4+I(b_sCD4^2)+b_log10RNA
```

```
+I(b_log10RNA^2)+factor(gender)+factor(origin)+factor(mode)
+b_age+I(b_age^2)+YRbase
           ,family = binomial()
           , data = test)

dat$probC.n <- predict(mod, type = 'response')

# products
dat$C.numcum <- ave(dat$probC.n,dat$patient,
                    FUN=function(x) cumprod(x))
dat$C.dencum <- ave(dat$probC.d,dat$patient,
                    FUN=function(x) cumprod(x))
dat$swC <- dat$C.numcum/dat$C.dencum

summary(dat$swC);hist(dat$swC, col="lightblue", breaks=50)

#-------------------------------------------------------------------------
# Truncate weights at 1%, 99% for stability

trunc.cutoff <- quantile(dat$swC,0.99,na.rm=TRUE)
test$sw.trunc <- ifelse(test$swC<trunc.cutoff, test$swC,trunc.cutoff)
summary(test$sw.trunc);hist(test$sw.trunc, col="lightblue", breaks=50)
```

## K.2 Patient example

Patient 3 is a 34 year old European heterosexual male of European descent who was first eligible for an emulated trial in the 3rd quarter of 2000, and at this time he was taking PCP prophylaxis. The inverse probability weights are calculated based on the covariates in the figure, using the code above.

| patient | Quarter start | Quarter end | Calendar yr. trial starts | Trial number | Trial time | Treatment group (0 = On, 1 = Off) | PCP diagnosis | Censoring indicator | Gender | mode of transmission | geographical origin | Age at baseline | sqrt(CD4) count at baseline | sqrt(CD4) count at in quarter | $\log_{10}$ RNA at baseline | $\log_{10}$ RNA in quarter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2000.75 | 2001 | 2000 | 13 | 1 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 4.90 | 1.32 | 1.32 |
| 3 | 2001 | 2001.25 | 2000 | 13 | 2 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 4.90 | 1.32 | 1.32 |
| 3 | 2001.25 | 2001.5 | 2000 | 13 | 3 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 5.66 | 1.32 | 1.70 |
| 3 | 2001.5 | 2001.75 | 2000 | 13 | 4 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 5.66 | 1.32 | 1.70 |
| 3 | 2001.75 | 2002 | 2000 | 13 | 5 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 5.29 | 1.32 | 1.70 |
| 3 | 2002 | 2002.25 | 2000 | 13 | 6 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 5.29 | 1.32 | 1.70 |
| 3 | 2002.25 | 2002.5 | 2000 | 13 | 7 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |
| 3 | 2002.5 | 2002.75 | 2000 | 13 | 8 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |
| 3 | 2002.75 | 2003 | 2000 | 13 | 9 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |
| 3 | 2003 | 2003.25 | 2000 | 13 | 10 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |
| 3 | 2003.25 | 2003.5 | 2000 | 13 | 11 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |
| 3 | 2003.5 | 2003.75 | 2000 | 13 | 12 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |
| 3 | 2003.75 | 2004 | 2000 | 13 | 13 | 0 | FALSE | 0 | M | Heterosexual | Europe | 34.39 | 4.90 | 6.16 | 1.32 | 1.70 |

Figure K.2.1: Patient example with covariate data used in calculating the inverse probability weights.

# Appendix L

# Sensitivity analysis for the PCP study

## L.1 Multiple imputation under Censoring at Random

We describe multiple imputation (MI) in the context of a censored time to event outcome. MI most often assumes that data are censored at random (CAR), but also allows the investigation of contextually clinically plausible departures from CAR. The approach may be implemented using standard software, or as here, using computationally straightforward R code, with minimal changes required to implement sensitivity analysis scenarios.

We assume that either there are no missing baseline and time varying covariates, or that these have already been (multiply) imputed in some way.

If there are missing baseline and time varying covariates then we multiply impute these prior to moving on to the process described below. For the COHERE data we implemented multiple imputation by chained equations using the MICE package in R to impute baseline covariates (van Buuren and Groothuis-Oudshoorn, 2011). Although this has yet to be implemented, for missing time varying covariates we could potentially fit a mixed effects spline model to impute CD4 and HIV RNA counts based on their trajectory. Alternatively, the expectation-maximization with bootstrapping (EMB) algorithm might be used for multiply imputing the time varying covariates (Honaker *et al.*, 2011), as implemented in the "Amelia" package in R. These approaches assume the covariates are missing at random. This constitutes the first step of the MI process, prior to the imputation step for generating new event times for those censored.

We explore sensitivity of our primary analysis results by assuming that those patients censored on the off prophylaxis arm which have been lost to follow-up have the hazard of those on the on prophylaxis arm. This is implemented using the "Jump to Reference" (J2R) approach. We assume that all other censored patients are censored at random.

However, to start with we begin by performing the multiple imputation assuming *all* censored patients are censored at random. Of course, this repeats the IP weighting based primary analysis, but it serves as a useful cross-check since the results should be approximately the same. We then move onto the J2R approach in a second step.

To recap, in the following we consider all censored patients in the data set, and multiply impute new events times assuming censoring is at random (CAR). We now briefly describe the MI approach for generating "new" event times for censored individuals (according to chapter 8.1.3 of Carpenter and Kenward (2012)).

1. As the imputation model, we fit a model predicting survival time based on all covariates necessary for CAR, along with those not involved in the censoring mechanism but nonetheless predictive of survival.
2. Impute the censored survival times, creating e.g. $K = 50$ imputed data sets resulting in all patients having event times and no censoring.
3. Fit the analysis model to each of these data sets in turn.
4. Combine the results for inference using Rubin's rules.

We now expand each of these steps in more depth.

*Step 1: Imputation model*

For the imputation step 1 above, each subject $i$ censored at time $T_i$ we calculate

$$p_i = 1 - \hat{S}(T_i|\boldsymbol{A}, \boldsymbol{L}),$$

for indicator variable of the treatment $\boldsymbol{A}$ and covariates $\boldsymbol{L}$. We use this to draw a new value

$$u_i \sim \text{uniform}\left[p_i, 1\right],$$

which is the basis for calculating the new event time as the solution of

$$u_i = 1 - \hat{S}(t|\boldsymbol{A}, \boldsymbol{L}),$$

thus ensuring that this time is greater than the existing censoring time.

There are a number of ways of defining the imputation model for the survival function. In the light of our analysis model, we use an IP weighted, adjusted pooled logistic model to predict an event time:

$$\text{logit}\left[Pr(Y_{m+k+1,i} = 1|A_{m,i}, V_{m,i}, \bar{Y}_{m+k,i} = \bar{0}\right] = \beta_{0,m+k,i} + \beta_1 A_{m,i} + \beta_2^T V_{m,i},$$

where
$Pr(Y_{m+k+1,i})$ is the probability of PCP diagnosis for the $k$th trial which starts in the $m$th quarter after 1st quarter 1998 (when $m = 0$). $m$ is the baseline month of the trial i.e. an indicator for the emulated trial,
$V_{m,i}$ are the baseline covariates for subject $i$ in trial $m$. In the current implementation we have not used the time varying CD4/RNA to predict the survival function.
$\beta_{0,m+k,i}$ is a function for the time varying intercept for trial $m$ including terms for $time$, $time^2$ and $time^3$ within trial $m$,
$\beta_1$ is the estimated log hazard ratio for PCP prophylaxis averaged over the follow-up period, and,
$\beta_2$ is a vector of estimated log hazard ratios for the covariates.

The inverse probability weights being defined exactly as in Appendix K.

The following model has been fitted:

```
amod1 <- svyglm(mod=formula(EVENT~ # PCP diagnosis (or death)
             as.factor(armIND) # treatment
            +trial_time # time in the trial
```

```
+I(trial_time^2)
+I(trial_time^3)
+trial # quater in which the trial starts
+I(trial^2)
# baseline covariates for this trial
+b_sCD4
+I(b_sCD4^2)
+b_log10RNA
+I(b_log10RNA^2)
+factor(gender)
+factor(mode2)
+factor(origin)
+b_age
+I(b_age^2)
+YRbase
, family = quasibinomial()
, design = svydesign(id = ~patient
, weights = ~sw.trunc # truncated weights
, data = temp))
```

This model estimates the risk of the outcome quarter by quarter, conditional on the treatment and baseline covariates for each subject.

We employ the associated survival function:

$$\hat{S}(T|\boldsymbol{A},\boldsymbol{V}) = \prod_{j:t_j \leq T} [1 - (1 + \exp(-\boldsymbol{\beta}_0(t_j) - \beta_1 \boldsymbol{A}(t_j) - \boldsymbol{\beta}_2^T \boldsymbol{V}(t_j)))^{-1}],$$

for distinct time periods (i.e. quarters in our study) $j = 1, 2, \ldots, J$, for treatment indicator $\boldsymbol{A}$, and baseline covariates $\boldsymbol{V}$.

Given the censoring time $T_i$ for a specific patient, we can use the survival function to predict the survival probability $p_i = 1 - \hat{S}(T_i|\boldsymbol{A},\boldsymbol{V})$.

*Step 2: Generate $K$ multiply imputed data sets*

We need to generate some variability in the imputed data sets, so we assume the parameter estimates from fitting the imputation model to the observed data are multivariate normally distributed. We then sample these estimates multiple times generating slightly different survival functions each time. These are used as the basis for generating the new event times for those censored.

Formally, we approximate the Bayesian posterior distribution by drawing $K$ estimates for the parameters from the asymptotic normal sampling distribution, $\mathcal{N}(\beta, I(\beta)^{-1})$, where the expected information is estimated by the observed sampling information (i.e. $VAR(\hat{\beta})$). This results in, for example, $K = 50$ sets of parameter estimates $\hat{\beta}_k$, k = 1,. . . , $K$. The linear predictors are used to calculate new event times for those censored for the $k$th imputed data set by using the formula for the survival function.

Concretely, $u_i$ is drawn from a uniform distribution on $[p_i, 1]$. This ensures that the imputed survival time is greater than the existing censoring time. The new event time for the censored patient $j$ is generated by evaluating:

$$T_j^* = \frac{-\log(u_i)}{\prod_{j:t_j \leq T}[1 - (1 + \exp(-\boldsymbol{\beta}_0(t_j) - \beta_1 \boldsymbol{A}(t_j) - \boldsymbol{\beta}_2^T \boldsymbol{V}(t_j)))^{-1}]},$$

from the logistic model. Due to the discrete time intervals, we have a step function for the survival probability by quarter, so we can find a new event time (i.e. quarter) by using a reverse look up, rather than solving a continuous function for time (refer to Figure L.1.1).

We repeat the process for each censored patient in data set $k$, and then for each of the 50 imputed data sets in turn. For each completed data set $k$ we re-calculate the stabilised IP weights, since the original weights were were calculated for the observed data only.

The analysis model is then fitted to each of the now complete imputed data sets.

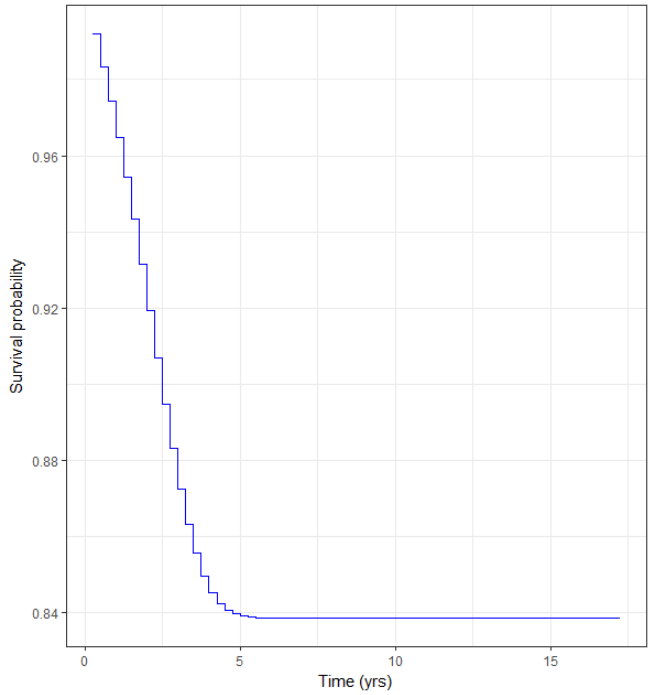


Figure L.1.1: Example survival function

*Step 3: Fit the analysis model*

The primary endpoint of our study is the hazard ratio for the effect of PCP prophylaxis conditional on the baseline covariates. We fit the adjusted pooled logistic regression model as analysis model. In a first step, we assume all censoring is at random, so we fit the analysis model to each of the multiply imputed data sets in turn and average the resulting estimated using Rubin's rules.

For the sensitivity analysis, we take a slightly different approach at this stage. We focus on the subgroup of patients not taking prophylaxis which were lost to follow-up and use the J2R approach to impute new times. That is, we use multiple imputation under J2R for this subgroup, but assume CAR for all other censored patients. Once we have multiply imputed $K = 50$ fully observed data sets assuming J2R for the subgroup, we then fit the primary analysis model, not forgetting to *recalculate* the IP weights to ensure the appropriate adjustment for other censored patients (i.e. those assumed to be censored is at random).

We then use Rubin's rules to combine point and variance estimates as usual.

## L.2 Sensitivity analysis using "Jump to Reference" approach

To illustrate the methodology, we focus on the secondary endpoint of all-cause mortality for the sensitivity analysis.

We modify the hazard in the imputation model for censored patients not taking prophylaxis which were lost to follow-up. This is implemented in Step 2 above by identifying these patients and altering their linear predictor so that the indicator function for treatment is set to "on prophylaxis" instead of "off prophylaxis" (denoted be (****) in the pseudo-code below).

## L.3 Algorithm

The complete algorithm in pseudo-code is as follows:

```
Fit the imputation model to the data set incorporating the
stabilised IP weights, resulting in estimates := beta
and covariance matrix := sigma.

Sample K sets of estimates from MVN(beta, sigma)

for each of k imputed data sets
{
Create the linear predictor for data set k_j := lp_k at times j=1,....,J

for each censored patient i
{
Calculate the hazard h_ijk := 1/(1+exp(-lp_ik)) at each time j=1, .., J
**** <for sensitivity analysis: manipulate the lp at this stage>
Calculate the survival function S_t := cumul. product(1-h_ijk) for j=1,..,J
Calculate S = the survival probability for patient i censored at T_j
Calculate p = 1-S
Calculate U = uniform[p, 1]
Find time interval in which the survival probability S_t is
```

```
U := new event time

if (new event time) > study period (e.g. 4yrs) then assume
administratively censored
else (new event time) counts as an event
} end of loop for each censored patient

Re-calculate the stabilised IP weights for completed data set k

Fit the substantive model including stabilised IP weights to data set k

} end of loop for imputing K imputed data sets


Combine the K estimates using Rubin's rules
```

The following R code is used as the basis for the imputation process.

```
## FUNCTIONS

pred.S<-function(pat, k, t) # calculates the linear predictor
{
  tempo<-unname(sum(c(imp.coef[k,1]
                      , ifelse(pat$armIND==0
                      , 0
                      , imp.coef[k,2]) # CAR version
#**** SENSITIVITY ANALYSIS CHANGE
                     #, 0) # J2R
                      , imp.coef[k,3]*t
                      , imp.coef[k,4]*t^2
                      , imp.coef[k,5]*t^3
                      , imp.coef[k,6]*pat$trial
                      , imp.coef[k,7]*pat$trial^2
                      , imp.coef[k,8]*pat$b_sCD4
                      , imp.coef[k,9]*pat$b_sCD4^2
                      , imp.coef[k,10]*pat$b_log10RNA
                      , imp.coef[k,11]*pat$b_log10RNA^2
                      , ifelse(pat$gender=="F", 0, imp.coef[k,12])
                      , ifelse(pat$mode2=="Heterosexual", 0,
                               ifelse(pat$mode2=="IDU", imp.coef[k,13],
                                      ifelse(pat$mode2=="MSM"
                                      , imp.coef[k,14], imp.coef[k,15])))
                      , ifelse(pat$origin=="Europe", 0,
                               ifelse(pat$origin=="Africa", imp.coef[k,16],
                               ifelse(pat$origin=="Asia", imp.coef[k,17],
                               ifelse(pat$origin=="Latin America",
                                      imp.coef[k,18], imp.coef[k,19]))))
                      , imp.coef[k,20]*pat$b_age
                      , imp.coef[k,21]*pat$b_age^2
                      , imp.coef[k,22]*pat$YRbase
  )))
```

```
  return(tempo)
}


### Step 2: Multiple Imputation of new event times for censored individuals
# We assume that all baseline and time varying covariates are
# fully observed or have been imputed.

# the imputation model has been fit to the censored data set
# defined as "amod1"

set.seed(12353)
K<-50

est.log<-summary(amod1)$coefficients[,1]

# sample from the MVN matrix of the estimates
nparam=length(names(est.log)) # number of parameters in imputation model

# Assume MVN and sample the coefficients from the imputation model
betahats <- est.log

# need the full variance-covariance matrix
Sigma<-unname(summary(amod1)$cov.scaled)

imp.coef<-mvrnorm(n = K
                  , mu=c(est.log)
                  , Sigma=as.matrix(Sigma))


newtime<-list()
R<-list()

for (i in 1:length(temp$patient))
      {
        pat<-temp[i,]
```

```
        if(temp$imp.ind[i]==1) # identify patients to be imputed
        {
          # define the X matrix for this censored subject
          # NOTE: we only consider baseline covariates in the imputation mod
          # calculate S based on the lp, varying t
          t<-seq(1, max_study)

          # need to re-calculate the IPW model excluding time
          # since we want the whole of S(t) for all t
          # see function at top of code "pred.S"
          lp<-sapply(1:max_study, function(i) pred.S(pat, k, t[i]))

          # Estimate of the hazard
          hall<-1/(1+exp(-lp))
          # Calculate survival probability censoring time
          Sall<-cumprod(1-hall)

          # sort S to find the interval using findInterval
          d<-data.frame(t, Sall)

# print the survival function
# pl<-ggplot() +
#  geom_step(data=d, mapping=aes(x=t/4, y=Sall), color="lightblue")+
#  labs(x = "Time (yrs)")+
#  labs(y="Survival probability")
# p1
# SEE FIGURE L.1.1

          # calculate S at the censoring time
          S<-Sall[pat$trial_time]

          # Generate Uniform[p.i,1] variables only for the censored patients
          # This ensures that each imputed survival time is greater than the
          # at which the unit was censored.
          p<-1-S
```

```
            U<-runif(1, p, 1)

            # invert S to find T*
            # calculate the whole S curve for all t for this censored subject
            # and then do a reverse check
            # check the new probability is within the range of S
            minS<-min(Sall)
            # sort data frame to use findInterval
            d<-d[with(d, order(-t)), ]

            if((1-U)<=minS)
            {
              newtime[[i]]=max_study
              #Time > longest study time - assume CAR & maximum study time
              R[[i]]=0
            }else{
                  newtime[[i]]<-max(d[findInterval((1-U), d$S),1]-1, 0)
                  tempo<-newtime[[i]]
                  R[[i]]=1

            }
} # end of i loop
```

# Bibliography


Akacha, M., F. B., Ohlssen, D., Rosenkranz, G. and Schmidli, H. (2017) Estimands and their role in clinical trials. *Statistics in Biopharmaceutical Research*, **9(3)**, 268–271.

Allison, P. (2002) *Missing data*. Thousand Oaks: Sage.

Atkinson, A., Cro, S., Carpenter, J. and Kenward, M. (2018) Reference based sensitivity analysis for time-to-event data. *submitted to Pharmaceutical Statistics*, **April**.

Atri, A., Frölich, L., Ballard, C., *et al.* (2018) Effect of idalopirdine as adjunct to cholinesterase inhibitors on change in cognition in patients with alzheimer disease: Three randomized clinical trials. *JAMA*, **319**(2), 130–142.

Baiocchi, M., Cheng, J. and Small, D. S. (2014) Tutorial in biostatistics: Intrumental variable methods for causal inference. *Stat. Med.*, **33(13)**, 2297–2340.

Bang, H. and Robins, J. M. (2005) Double robust estimation in missing data and causal inference problems. *Biometrics*, **61**, 962–973.

Barr, D. R. and Sherrill, E. T. (1999) Mean and variance of truncated normal distribution. *The American Statistician*, **53(4)**, 357–361.

Barrett, J. and Su, L. (2015) Dynamic predictions using flexible joint models and time-to-event data. *Statistics in Medicine*, **36**, 1447–1460.

Bell, M., Fiero, M., Horton, N. J. and Hsu, C.-H. (2014) Handling missing data in rcts; a review of the top medical journals. *BMC Medical Research Methodology*, **14**, 118.

Bender, R., Augustin, T. and Bletter, M. (2005) Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, **24**, 1713–1723.

Beunckens, C., Molenberghs, G. and Kenward, M. (2005) Tutorial: Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, **2**, 379–386.

Beunckens, C., Molenberghs, G., Kenward, M. and Mallinckrodt, C. (2008) A latent-class mixture model fro incomplete longitudinal gaussian data. *Biometrics*, **64**, 96–105.

Billings, L. K., Doshi, A., Gouet, D., Oviedo, A., Rodbard, H. W., Tentolouris, N., Grøn, R., Halladin, N. and Jodar, E. (2018) Efficacy and safety of ideglira versus basal-bolus insulin therapy in patients with type 2 diabetes uncontrolled on metformin and basal insulin; dual vii randomized clinical trial. *Diabetes Care*.

Bradshaw, P., Ibrahim, J. and Gammon, M. (2010) A bayesian proportional hazards regression model with non-ignorably missing time-varying covariates. *Statistics in Medicine*, **29**, 3017–3029.

Brinkhof, M., Spycher, B., Yiannoutsos, C., Weigel, R., Wood, R., Messou, E., A., A. B., Egger, M. and Sterne, J. (2010) Adjusting mortality for loss to follow-up: analysis of five art programmes in sub-saharan africa. *PLoS ONE*, **5**, American Journal of Epidemiology.

van Buuren, S. (2012) *Flexible imputation of missing data*. Boca Raton, USA: CRC Press.

Cain, L., Robins, J., Lanoy, E., Logan, R., Costagliola, D. and Hernan, M. (2010) When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. *International Journal of Biostatitics*, **6(2)**, Article 18.

Caniglia, E. *et al.* (2017) Comparison of dynamic monitoring stategies based on cd4 counts in virally suppressed, HIV-positive individuals on combination antiretroviral therapy in high-income countries: a prospective, observational study. *Lancet HIV*, **4(6)**, 251–259.

Carpenter, J. and Kenward, M. (2012) *Multiple Imputation and its Applications*. New Jersey: Wiley.

Carpenter, J., Kenward, M., Evans, S. and White, I. (2003) Letter to the editor: Last observation carried forward and last observation analysis. *Stastics in Medicine*, **23**, 3241–3244.

Carpenter, J., Kenward, M. G. and White, I. R. (2007) Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res*, **16**, 259.

Carpenter, J., Roger, J. and Kenward, M. (2013) Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharm Stat*, **23(6)**, 1352–71.

Carpenter, J. R., Roger, J. H., Cro, S. and Kenward, M. G. (2014) Response to Comments by Seaman et al on Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation, Journal of Biopharmaceutical Statistics, 23, 1352-1371. *J Biopharm Stat*, **24**, 1363–9.

CHMP (2010) Committee for medicinal products for human use guidelines on missing data in confirmatory clinical trials. *European Medicines Agency*, **download from http://www.ema.europa.eu on 15th January 2014**.

CHMP (2018) Committee for human medicinal products, ich e9 (r1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. *EMA/CHMP/ICH/436221/2017*.

Cole, S. R. and Hernan, M. A. (2008) Contructing inverse probability weights for marginal structural models. *Am. Journal of Epid.*, **168(6)**, 656–664.

Cox, D. R. (1972) Regression model and life-tables. *Journal of the Royal Statistical Society, Series B*, **34 (2)**, 187–220.

Cro, S. (2016) *Relevant, accessible sensitivity analysis for longitudinal clinical trials with dropout*. Ph.D. thesis, London School of Hygiene & Tropical Medicine.

Cro, S., Morris, T., Kenward, M. and Carpenter, J. (2016) Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. *The Stata Journal*, **16(2)**, 443–463.

Cro, S., J., J. C. and Kenward, M. (2018) Information anchored sensitivity analysis: Theory and application. *accepted for Journal of the RSS Series A*.

Crowther, M., Abrams, K. and Lambert, P. (2013) Joint modeling of longitudinal and survival data. *Stata Journal*, **13(1)**, 165–184.

D'Agostino, R. B., Lee, M.-L., Belanger, A. J., Cupples, L. A., Anderson, K. and Kannel, W. B. (1990) Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study. *Stat. Med.*, **9**, 1501–1515.

Danaei, G., Rodriguez, L. A. G., Cantero, O. F., Logan, R. and Hernan, M. A. (2013) Observational data for comparative effectiveness research: am eumulation of randomised trials to estimate the effect of statins on primary prevention of coronary heart disease. *Stat Methods Med Res*, **22 (1)**, 70–96.

Daniel, R. and Kenward, M. (2012) A method for increasing the robustness of multiple imputation. *Computation Statistics and Data Analysis*, **56**, 1624–1643.

Daniels, M. and Hogan, J. (2008) *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Baton Rouge: Chapman and Hall.

EACS (2018) European aids clinical society guidelines. Online; accessed 28.06.18.

Eekhout, A., de Boer, M. R., Twisk, J. W. R., de Vet, H. and Heymans, M. W. (2012) Missing data: A systematic reviewof how they are reported and handled. *Epidemiology*, **23 (5)**.

Efron, B. (1988) Logistic regression, survival analysis, and the kaplan-meier curve. *J. Am. Stat. Soc.*, **84 (402)**, 414–25.

Emoto, S. and Matthews, P. (1990) A weibull model for informative censoring. *The Annals of Statistics*, **18**, 1556–1577.

Enders, D., Engel, S., Linder, R. and Pigeot, I. (2018) Robust versus consistent variance estimators in marginal structural models. *Statistics in Medicine*, **DOI: 10.1002/sim.7823**.

Fenner, L., Atkinson, A., Boulle, A., Fox, M., Prozesky, H., Zürcher, K., Balliff, M., Zwahlen, M., Davies, M.-A., Egger, M. and the International epidemiologic Database to Evaluate AIDS in Southern Africa (IeDEA-SA) (2017) HIV viral load as an independent risk factor for tuberculosis in south africa: collaborative analysis of cohort studies. *Journal of the International AIDS Society*, **20:21327**.

Fiero, M. H., Huang, S., Oren, E. and Bell (2016) Statistical analysis and handling of missing data in cluster randomised trials: a systematic review. *Trials*, **17**, 72.

Furrer, H. *et al.* (2015) HIV replication is a major predictor of primary and recurrent pneumocystis pneumonia - implications for prophylaxis recommendations. *European AIDS Clinical Society (EACS) conference*, **Poster PS5/2**.

Gao, F., G. L., Zeng, D., Xu, L., Lin, B., Diao, G., Golm, G., Heyse, J. and Ibrahim, G. (2017) Control-based imputation for sensitivity analyses in informative censoring for recurrent event data. *Pharm. Stat.*, **16**, 424–432.

Garcia-Albeniz, X., Hsu, J. and Hernan, M. (2017) The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur. J. Epidemiol.*, **32**, 495–500.

Gilbert, P., Shepherd, B. and Hudgens, M. (2013) Sensitivity analysis of per-protocol time-to-event treatment efficacy in randomized clinical trials. *Journal of the American Statistical Association*, **108(503)**.

Grambsch, P. M. and Therneau, T. M. (1994) Proportional hazards tests and diagnosis based on weighted residuals. *Biometrika*, **Vol. 81, No. 3**, 515–526.

Green, M. S. and Symons, M. J. (1983) A comparison of the logistic risk function and the proportional hazards model in prospective epidemiological studies. *J. Chron. Dis.*, **36 (10)**, 715–23.

Greene, W. H. (2003) Econometric analysis. *5th ed.*, **Prentice Hall**.

Harel, O., Mitchell, E., Perkins, N., Cole, S., Tchetgen-Tchetgen, E., Sun, B.-L. and Schisterman, E. (2018) Multiple imputation for incomplete data in epidemiologic studies. *American Journal of Epidemiology*, **187(3)**, 576–591.

Heitjan, D. F. (2017) Commentary on "Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE" Trial by Mason *et al*. *Clinical Trials*, **14**, 368–369.

Henderson, R. A., Pocock, S. J., Clayton, T. C., Knight, R., Fox, K. A., Julian, D. G. and Chamberlain, D. A. (1997) Coronary angioplasty versus medical therapy for angina: the second randomised intervention treatment of angina (rita-2) trial. *Lancet*, **350**, 461–8.

Henderson, R. A., Pocock, S. J., Clayton, T. C., Knight, R., Fox, K. A., Julian, D. G. and Chamberlain, D. A. (2003) Seven-year outcome in the rita-2 trial: Coronary angioplasty versus medical therapy. *J. Am. Coll. Cardiol.*, **42(7)**, 1162–70.

Hernan, M. and Hernandez-Diaz, S. (2012) Beyond the intention to treat in comparative effectiveness trials. *Clin. Trials*, **9(1)**, 48–55.

Hernan, M. and Robins, J. (2018) *Causal Inference*. forthcoming: Chapman and Halls/CRC.

Hernan, M. and Swanson, S. (2017) Estimation of causal effects in observational studies. *Causal Inference course*, **Erasmus summer school program 2017**, Day 2.

Hernan, M., Hernandez-Diaz, S. and Robins, J. (2004) A structural approach to selection bias. *Epidemiology*, **15(5)**, 615–625.

Hernan, M., Sauer, B., Hernandez-Diaz, S., Platt, R. and Shrier, I. (2016) Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational studies. *Journal of Clinical Epidemiology*, **79**, 70–75.

Hernan, M. A. and Robins, J. M. (2016) Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.*, **183(8)**, 758–764.

Hernan, M. A., Brumback, B. and M.Robins, J. (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, **11(5)**, 561–570.

Hernan, M. A., Lanoy, E., Costagliola, C. and Robins, J. M. (2006) Comparison of dynamic treatment regimes via inverse probability weighting. *Basic and clinical pharmacology and toxocology*, **98**, 237–242.

Herring, A., Ibrahim, J. and Lipsitz, S. (2004) Non-ignorably missing covariate data in survival analysis: a case study of an international breast cancer study group trial. *Applied Statistics*, **53**, 293–310.

Hickey, G., Philipson, P., Jorgensen, A. and Kolamunnage-Dona, R. (2016) Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, **16:117**.

Hogan, J. and Laird, N. (1997a) Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**, 259–272.

Hogan, J. W. and Laird, N. M. (1997b) Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**, 259–272.

Honaker, J., King, G. and Blackwell, M. (2011) Amelia ii: A program for missing data. *Journal of Statistical Software*, **45 (7)**.

Hu, B., Li, L. and Greene, T. (2016) Joint multiple imputation for longitudinal outcomes and clinical events which truncate longitudinal follow-up. *Statistics in Medicine*, **25(17)**, 2991–3006.

Huang, X. and Wolfe, R. (2002) A frailty model for informative censoring. *Biometrics*, **58**, 510–520.

Hughes, R., Sterne, J. and Tilling, K. (2014) Comparison of imputation variance estimators. *Stat. Meth. Med Res.*, **Epub ahead of print**, PMID: 24682265.

Ibrahim, J., Chen, M. and Sinha, D. (2001) *Bayesian Survival Analysis*. New York: Springer.

Ibrahim, J., Chu, H. and Chen, M.-H. (2012) Missing data in clinical studies: Issues and methods. *Journal of Clinical Oncology*, **30(26)**, 3297–3303.

Jackson, D., White, A., Seaman, S., Evans, H., Baisley, K. and Carpenter, J. (2014) Relaxing the indendependent censoring assumption in the cox proportional hazards model using multiple imputation. *Statistics in Medicine*, **33**, 4681–4694.

Jakobsen, J., Gluud, C. and Winkel, P. (2017) When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide. *BMC Medical Research Methodology*, **17(1)**, 162.

Jans, T., Jacob, C., Warnke, A., Zwanzger, U., Gro-Lesch, S., Matthies, S., Borel, P., Hennighausen, K., Haack-Dees, B., Rsler, M., Retz, W., von Gontard, A., Hnig, S., Sobanski, E., Alm, B., Poustka, L., Hohmann, S., Colla, M., Gentschow, L., Jaite, C., Kappel, V., Becker, K., Holtmann, M., Freitag, C., Graf, E., Ihorst, G. and Philipsen, A. (2015) Does intensive multimodal treatment for maternal ADHD improve the efficacy of parent training for children with adhd? a randomized controlled multicenter trial. *Journal of Child Psychology and Psychiatry*, **56**(12), 1298–1313.

Keene, O. N., Roger, J. H., Hartley, F. H. and Kenward, M. G. (2014) Missing data sensitivity analysis for recurrent event data using congtrolled imputation. *Pharm. Statistics*, **13**, 258–264.

Kenney, J. F. and Keeping, E. S. (1951) *The distribution of the standard deviation, in section 7.8 of Mathematical Statistics, Part 2*. Princeton, NJ: D. Van Norstrand.

Keogh, R. H. and Morris, T. P. (2018) Multiple imputation in cox regression when there are time-varying effects of covariates. *Statistics in Medicine*, pp. 1–18.

Kim, J. (2004) Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, **32(2)**, 766–783.

Kim, S., Zeng, D. and Taylor, M. (2017) Joint partially linear model for longitudinal data with informative drop-out. *Biometrics*, **73(1)**, 72–82.

van der Laan, M. and Rose, S. (2018) *Targeted Learning in Data Science*. Cham, Switzerland: Springer International Publishing.

Lambert, P. C. and Royston, P. (2009) Further development of flexible parametric models for survival analysis. *The Stata Journal*, **9**, 265–290.

LaVange, L. and Permutt, T. (2016) A regulatory perspective on missing data in the aftermath of the nrc report. *Statistics in Medicine*, **35**, 2853–2864.

Leacy, F., Floyd, S., Yates, T. and White, I. (2017) Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to tuberculosis/HIV prevalence survey with incomplete HIV-status data. *Am. J. Epidemiol.*, **185(4)**, 304–315.

Letué, F. (2008) A semi-parametric shock model for a pair of event related dependent censored failure times. *Journal of Statistical Planning and Inteference*, **138**, 3869–3884.

Li, Q. and Su, L. (2018) Accomodating informative dropout and death: a joint modelling approach for longitudinal and semicompeting risks data. *RSS Applied Statistics (Series C)*, **67(1)**, 145–163.

Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **72**, 13–22.

Lipkovich, I., Ratitch, B. and O'Kelly, M. (2016) Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints. *Pharmaceutical Statistics*, **15**, 216–229.

Little, R. and Yau, L. (1996) Intent-to-treat analysis for longitudinal studies with dropouts. *Biometrics*, **52**, 471–483.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data, 2nd Edition*. New Jersey: Wiley.

Liu, G. and Peng, L. (2016) On analysis of longitudinal clinical trials with missing data using reference-based imputation. *Journal of Biopharmaceutical Statistics*, **26:5**, 924–936.

Lodi, S. *et al.* (2017) Effect of immediate initiation of antiretroviral treatment in HIV-positive individuals aged 50 years or older. *J. AIDS*, **76(3)**, 311–318.

Lu, K., Li, D. and Koch, G. (2015) Comparison between two controlled multiple imputation methods for sensitivity analyses of time-to-event data with possibly informative censoring. *Stat. Biopharm. Res.*, **7(3)**, 199–213.

Luque-Fernandez, M., Schomaker, M. and Ratchet, B. (2017) Targeted maximum likelohood estimation for a binary treatment: A tutorial. *Statistic in Medicine*, **37(16)**, 2530–2546.

Mallinckrodt, C., J. Watkin, G. M. and R.Carroll (2004) Choice of the primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics*, **3**, 161–169.

Mallinckrodt, C., Molenberghs, G. and Rathmann, S. (2017) Choosing estimands in clinical trials with missing data. *Pharmaceutical Statistics*, **16**, 29–36.

Mallinckrodt, C. H., Lin, Q. and Molenberghs, G. (2013) A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharm Stat.*, **12(1)**, 1–6.

Mason, A., Gomes, M., Grieve, R., Ulug, P., Powell, J. and Carpenter, J. (2017a) Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the improve trial. *Clinical Trials*, **14**, 357–367.

Mason, A., Gomes, M., Grieve, R., Ulug, P., Powell, J. and Carpenter, J. (2017b) Rejoinder to commentary on 'Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE Trial'. *Clinical Trials*, **14**, 372–373.

Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input. *Statistical Sciences*, **9(4)**, 538–573.

Mocroft, A., Reiss, P., Kirk, O., Mussini, C., Giardi, E., Morlat, P., Wit, S. D., K., K. D., Ghosn, J., Bucher, H., Lundgren, K., Chene, G., Miro, J. and Furrer, H. (2010) Is it safe to discontinue primary pneumocystis jiroveci penumonia prophylaxis in patients with vorologically suppressed HIV infection and a cd4 cell count $<$ 200 cells/$\mu$l? *CID*, **51**, 611–619.

Molenberghs, G. and Kenward, M. (2007) *Missing Data in Clinical Studies*. New Jersey: Wiley.

Mussini, C., Pezzoti, P., Govini, A., Borghi, V., Antinori, A., d'Arminio Monforte, A., Luca, A. D., Mongiardo, N., Cerri, M., Chiodo, F., Concia, E., Bonazzi, L., Moroni, M., Ortona, L., Esposito, R., Cossarizza, A. and for the changes in Opportunistic Prophylaxis (CIOP) study, B. D. R. (2000) Discontinuation of primary prophylacis for pneumocystis carinii pneumonia and tocoplasmic encephalitis in human immunodeficiency virus type i-infected patients: The changes in opportunistic prophylaxis study. *Journal of Infectious Diseases*, **181**, 1635–1642.

Muthen, B., Asparouhov, T., Hunter, A. and Leuchter, A. (2011) Growth modeling with non-ignorable dropout: Alternative analyses of the star*d antidepressant trial. *Psychol Methods*, **16(1)**, 1733.

Nelson, W. (1972) Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945–965.

Newsome, S., Keogh, R. and Daniel, R. (2017) Estimating long-term treatment effects in observational data: A comparison of the performance of different methods under real-world uncertainty. *Statistics in Medicine*, **37(15)**, 2367–2390.

Ng'andu, N. H. (1997) An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Statistics in Medicine*, **16(6)**, 611–26.

Ngwa, J. S., Cabral, H. J., Cheng, D. M., Pencina, M. J., Gagnon, D. R., LaValley, M. P. and l. A. Cupples (2016) A comparison of time dependent cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the framingham heart study. *BMC Med. Res. Meth.*, **16:148**.

Nielson, S. (2003) Proper and improper multiple imputation. *International Statistical Review*, **71**, 593–627.

NIH (2018) Guidelines for the prevention and treatment of opportunistic infections in HIV-infected adults and adolescents. Online; accessed 28.06.18.

NRC (2010) *National Research Council report: The prevention and Treatment of Missing Data in Clinical Trials*. Washington DC; The National Academic Press: Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education.

O'Kelly, M. and Ratitch, B. (2014) *Clinical trials with missing data: A guide for practitioners*. New Jersey: Wiley.

Perkins, N., Cole, S., Harel, O., Tchetgen-Tchetgen, E., Sun, B., Mitchell, E. and Schisterman, E. (2018) Principled approaches to missing data in epidemiological studies. *American Journal of Epidemiology*, **187(3)**, 568–575.

Philipsen, A., Jans, T., Graf, E. *et al.* (2015) Effects of group psychotherapy, individual counseling, methylphenidate, and placebo in the treatment of adult attention-deficit/hyperactivity disorder: A randomized clinical trial. *JAMA Psychiatry*, **72**(12), 1199–1210.

Powney, M., Williamson, P., J., J. K. and Kolamunnage-Dona, R. (2014) A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*, **15:237**.

Proust-Lima, C., Sene, M., Taylor, J. and Jacqmin-Gadda, H. (2014) Joint latent class models for longitudinal and time-to-event data: A review. *Stat. Methods Med. Res.*, **23(1)**, 74–90.

Qiros, J. D., Miro, J., Pena, J. M., Podzamczer, D., Alberdi, J. C., Martinez, E., Cosin, J., Claramonte, X., Gonzalez, J., Domingo, P., Casado, J. L. and Ribera, E. (2001) A randomized trial of the discontinuation of primary and secondary prophylaxis against pneumocystic carinii penumonia after highly active antiretroviral therapy in patients with HIV infection. *NEJM*, **344(3)**, 159–167.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rezvan, P. H., Lee, K. J. and Simpson, J. A. (2015) The rise of multiple imputation: a review of the reporting and implementation of the methods in medical research. *BMC Medical Research Methodology*, **15**, 30.

Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton, USA: CRC Press.

Robins, J. (1997) Causal inference from complex longitudinal data. *In: Berkane M, ed. Latent variable modelling and applications to causality: Lecture notes in statistics 120*, **New York: Springer-Verlag**, 69–117.

Robins, J. (1998a) Correction of non-compliance in equivalence trials. *Stat Med*, **17**, 269–302.

Robins, J. (1998b) Marginal structural models. *1997 Proceedings of the Section on Bayesian Statistical Science*, **Alexandria, Virginia: American Statistical Society**, 1–10.

Robins, J. (2000) Marginal structutal models *versus* structural nested models as tools for causal inference. *In: Halloran E. Berry D, eds. Statistical Models in Epidemiology: The Environment and Clinical Trials*, **New York: Springer-Verlag**, 95–134.

Robins, J. and Wang, N. (2000) Inference for imputation estimators. *Biometrika*, **87(1)**, 112–124.

Robins, J., Rotnitzky, A. and Zhao, L. (1995) Analysis of semiparametric regression models for repeated outcomes with missing data. *American Statistical Association*, **90(429)**, 106–121.

Robins, J., Hernan, M. and Brumback, B. (2000) Marginal structal models and causal inference in epidemiology. *Epidemiology*, **11(5)**, 550–560.

Rosenbaum, P. and Rubin, S. (1984) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rotnitzky, A., Farall, A., Bergesion, A. and Scharfstein, D. (2002) Analysis of failure time data in the presence of competing censoring mechanisms. *Journal of the Royal Statistical Society*, **Series B, 69**, 307–327.

Royston, P. and Parmar, M. (2013) Survival analysis - coping with non-proportional hazards in randomized trials. *presentation from http://www.methodologyhubs.mrc.ac.uk accessed on 31.8.13.*

Royston, P. and Parmar, M. K. B. (2011) The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med.*, **30;30(19)**, 2409–21.

Ruau, D., Burkoff, N., Bartlett, J., Jackson, D., Jones, E., Law, M. and Metcalfe, P. (2016) *InformativeCensoring: Multiple Imputation for Informative Censoring*. R package version 0.3.4.

Rubin, D. (1976) Inference and missing data. *Biometrika*, **63(3)**, 581–592.

Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91 (434)**, 473–489.

Sauerbrei, W. and Royston, P. (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J. R. Statist. Soc. A*, **162 (1)**, 71–94.

Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., Schumacher, M. and the German Breast Cancer Study (1999) Modelling the effects of standard prognostic factors in node-positive breast cancer. *British Journal of Cancer*, **79 (11/12)**, 1752–1760.

Scharfstein, D. and Robins, J. (2002) Estimation of the failure time distribution in the presence of information censoring. *Biometrika*, **89**, 617–634.

Scharfstein, D., Rotnitzky, A. and Robins, J. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94 (448)**, 1096–1120.

Scharfstein, D., Robins, J., W., W. E. and A.Rotnitzky (2001) Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics*, **57**, 404–413.

Scharfstein, D., McDermott, A., Diaz, I., Carone, M., Lunardon, N. and Turkoz, I. (2018) Global sensitivity analysis for repeated measures studies with informative drop-out: A semi-parametric approach. *Biometrics*, **74**, 207–219.

Schmoor, C., Olschweski, M. and Schumacher, M. (1996) Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Satist. Med.*, **15**, 263–271.

Schoenfeld, D. A. (1982) Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239–241.

Schomaker, M. and Heumann, C. (2018) Bootstrap inference when using multiple imputation. *Statistics in Medicine*, **37(14)**, 2252–2266.

Schumacher, M., Bastert, G., Bojar, H., Huebner, K., Olschweski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Newman, R. L. and F., H. F. R. H. (1994) Randomized 2x2 trial outcome evaluating hormone treatment and the duration of chemotherapy in node-positive breast cancer patients. *J. Clin. Oncology*, **12**, 2086–2093.

Seaman, S., White, I. and Leacy, F. (2014) Comment on "analysis of longitudinal trials with protocol deviations: A framework for relevant, accessible, assumptions, and inference via multiple imputation". *Journal of Biopharmaceutical Statistics*, **24**, 1358–1362.

Shardell, M., Scharfstein, D., Viahov, D. and Galai, N. (2008) Inference for cumulative incidence functions with informatively coarsened discrete event-time data. *Statistics in Medicine*, **27(28)**, 5861–5879.

Siannis, F. (2004) Applications of a parametric model for informative censoring. *Biometrics*, **60**, 704–714.

Siannis, F. (2011) Sensitivity analysis for multiple right censoring: Investigating mortality in psoriatic arthritis. *Statistics in Medicine*, **30**, 356–367.

Siannis, F., Copas, J. and Lu, G. (2005) Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, **6**, 77–91.

Sterne, J., Hernan, M., Ledergerber, B., Tilling, K., Weber, R., Sendi, P., Rickenbach, M., Robins, J., Egger, M. and the Swiss HIV Cohort Study (2005) Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet*, **366**, 378–84.

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A. and Carpenter, J. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **339**, 157–160.

Sun, B., Perkins, N. and and, S. C. (2018) Inverse-probaibility-weighted estimation for monotone and nonmonotone missing data. *American Journal of Epidemiology*, **187(3)**, 585–591.

Taffé, P., May, M. *et al.* (2008) A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort. *Statistics in Medicine*, **27 (23)**, 4835–4853.

Tang, Y. (2018) Controlled pattern imputation for sensitivity analysis of longitudinal binary and ordinal outcomes with nonignorable dropout. *Statistics in Medicine*, **10.1002/sim.7583**, 1–15.

Thiebaut, R., Jacqmin-Gadda, H., Babiker, A., Commenges, D. and Collaboration, T. C. (2005) Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of cd4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statist. Med.*, **24**, 6582.

Thompson, W. A. (1977) On the treatment of grouped observations in life studies. *Biometrics*, **33**, 463–470.

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24–36.

Toh, S. and Hernan, M. (2008) Causal inference from longitudinal studies with baseline randomization. *International Journal of Biostatitics*, **4(1)**, Article 22.

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J. and White, I. R. (2018) On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statistics in Medicine*, **37**, 2338–2353.

Tsiatis, A., Davidian, M. and Cao, W. (2011) Improved doubly robust estimation when data are monotonely coarsenes, with application to longitudinal studies with dropout. *Biometrics*, **67**, 536–545.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, **45**(3), 1–67.

White, I. and Carlin, J. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **29**, 2920–2931.

White, I. R. and Royston, P. (2009) Imputing missing covariate values for the cox model. *Statistics in Medicine*, **28**, 1982–1998.

White, I. R., Royston, P. and Wood, A. M. (2011) Multiple imputation using chained equations: Issues and guidelines for practice. *Statistics in Medicine*, **30**, 377–399.

Wood, A. M., White, I. R. and Thompson, S. (2004) Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clin Trials*, **2**, 368.

Xu, Z. and Kalbfleisch, J. (2010) Propensity score matching in randomized clinical trials. *Biometrics*, **66(3)**, 813–23.

Zhao, Y., Herring, A. H., Zhou, H., Mirza, A. W. and Koch, G. G. (2014) A multiple imputation method for sensitivity analysis of time-to-event data with possibly informative censoring. *J. Biopharm. Stat.*, **24(2)**, 229–253.

Zhao, Y., Saville, B., Zhou, H. and Koch, G. (2016) Sensitivity analysis for missing outcomes in time-to-event data with covariate adjustment. *Journal of Biopharmaceutical Statistics*, **26(2)**, 269–279.