Statistical modelling of competing risks with incomplete data: with applications to allogeneic stem cell transplantation

Bonneville, E.F.

Statistical modelling of competing risks with incomplete data

With applications to allogeneic stem cell transplantation

Edouard Francis Bonneville

Cover design: Edouard Francis Bonneville & Ridderprint
Printed by: Ridderprint | www.ridderprint.nl
ISBN: 978-94-6522-314-8

Statistical modelling of competing risks with incomplete data

With applications to allogeneic stem cell transplantation

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 2 juli 2025
klokke 11:30 uur

door

Edouard Francis Bonneville
geboren te Madrid, Spanje
in 1995

Promotor:	Prof. dr. H. Putter
Co-promotor:	Dr. L. C. de Wreede
Leden promotiecommissie:	Prof. dr. J. Wallinga
	Prof. dr. J. Beyersmann *Universität Ulm, DE*
	Prof. dr. S. van Buuren *Universiteit Utrecht, NL*
	Dr. S. Iacobelli *Sapienza Università di Roma, IT*

En memoria de mis abuelos,
Ramiro y Pierre

Table of contents

1 General introduction **1**
- 1.1 Competing risks . . . 3
- 1.2 Multiple imputation of missing covariate data . . . 9
- 1.3 Joint models and informative dropout . . . 12
- 1.4 Thesis outline and contributions . . . 14

2 Handling missing covariate data in clinical studies in haematology **17**
- 2.1 Introduction . . . 19
- 2.2 Assumptions underlying methods for handling missing data . . . 19
- 2.3 Common approaches and associated pitfalls . . . 20
- 2.4 Systematic review . . . 24
- 2.5 Illustrative example . . . 28
- 2.6 Discussion and recommendations . . . 31

3 Multiple imputation for cause-specific Cox models: Assessing methods for estimation and prediction **35**
- 3.1 Introduction . . . 37
- 3.2 Motivating example . . . 39
- 3.3 Cause-specific competing risks analysis . . . 39
- 3.4 Methods . . . 41
- 3.5 Simulation study . . . 45
- 3.6 Illustrative analysis . . . 55
- 3.7 Discussion . . . 59

Appendix A: Imputation model derivations . . . 62

4 Impact of comorbidities and body mass index on the outcomes of allogeneic hematopoietic cell transplantation in myelofibrosis: A study on behalf of the Chronic Malignancies Working Party of EBMT **67**

Appendix A: Methods . . . 73

5 Multiple imputation of missing covariates when using the Fine–Gray model **77**
5.1 Introduction . . . 79
5.2 Notation . . . 80
5.3 MI approaches with a Fine–Gray substantive model . . . 82
5.4 Simulation study . . . 91
5.5 Applied data example . . . 105
5.6 Discussion . . . 108
Appendix A: Imputed censoring times, and resulting cumulative subdistribution hazards . . . 110
Appendix B: Additional simulations . . . 113
Appendix C: Data dictionary . . . 115

6 Why you should avoid using multiple Fine–Gray models: insights from (attempts at) simulating proportional subdistribution hazards data **117**
6.1 Introduction . . . 119
6.2 Competing risks and the Fine–Gray model . . . 120
6.3 Data-generating mechanisms . . . 122
6.4 Discussion . . . 132

7 Joint models quantify associations between immune cell kinetics and allo-immunological events after allogeneic stem cell transplantation and subsequent donor lymphocyte infusion **135**
7.1 Introduction . . . 137
7.2 Methods . . . 138
7.3 Results . . . 144
7.4 Discussion . . . 153
Appendix A: Statistical supplement . . . 157

8 Conclusions **163**

Bibliography **171**

Nederlandse samenvatting **193**

Summary **197**

List of publications **199**

Acknowledgements **201**

Curriculum vitae **203**

Chapter 1

General introduction

A variety of biomedical applications involve studying the time between a relevant starting point and the occurrence of a particular event. For instance, time until death is regularly of interest to researchers, be it from birth, diagnosis of a given disease, or the start of a treatment for that disease (van Houwelingen and Putter, 2012). One context in which time to death is particularly relevant is that of allogeneic haematopoietic stem cell transplantation (alloSCT), which is a type of treatment primarily given to patients with haematological malignancies such as acute myeloid leukaemia (AML) (Horowitz *et al.*, 1990). Specifically, we may be interested in both how long patients diagnosed with AML survive on average after an alloSCT, and how individual patient characteristics, such as age at alloSCT or the presence of particular genetic mutations, influence the rate of death over time. The latter question in particular can be sharpened in order to investigate the rate of death due to different causes (e.g. mortality related to treatment or due to an infection post-alloSCT), assuming these can be properly ascertained. Additionally, we may be interested in the rate of different clinical events occurring prior to death, such as disease relapse.

In order to research these types of questions, clinical data needs to be collected from patients receiving an alloSCT. This is typically done via clinical registries, which coordinate the collection of both baseline (i.e. patient- and treatment-related characteristics at time of alloSCT) and 'follow-up' information (e.g. the occurrence and timing of relevant clinical events post-alloSCT) for individual patients in transplantation centres, possibly across different countries and over extended periods of time (Horowitz, 2008). The complexities intrinsic to the collection of such data, such as changes in data collection forms over time and across centres, mean that *missing* or *incomplete* data are an unavoidable feature of registry data.

Different kinds of incomplete data can emerge when collecting data for a single patient who has undergone an alloSCT. The first, which is a defining characteristic of *survival* or *event history* data, is that of *censoring* (Aalen *et al.*, 2008). For an example study interested in time to disease relapse from alloSCT, some patients may still be alive and relapse-free by the end of study, or may have dropped out of the study at an earlier point in time without having relapsed. In both cases, the individuals are said to be *right-censored*. That is, we only know that they were alive and relapse-free until they dropped out or the study ended, but not whether or when they would have relapsed had the study period been longer or if they had remained under follow-up. Furthermore, patients may die prior to having relapsed, for example due to treatment-related toxicity. Since death precludes the occurrence of relapse (or any other event), death here is considered a *competing risk*. These and related points were concisely summarised over 60 years ago by Chiang (1961):

> When the period of observation is ended, there will usually remain a number of individuals on whom the mortality data in a typical study will be incomplete. Of first importance among these are the persons still alive at the close of the study. Secondly, if the investigation is concerned with mortality from a specific cause, the necessary information is incomplete and unavailable for patients who died from other causes. In addition, there will usually be a third group of patients who were 'lost' to the study because of follow-up failure. These three groups present a number of statistical problems in the estimation of the expectation of life and survival rates.

Another kind of incomplete data is *missing covariate data*, where relevant baseline information is only partially observed. For example, some variables may be collected more systematically later in time after their clinical importance has been established (e.g. age of the stem cell donor), or some measurements may be expensive and/or time-consuming to collect (e.g. high-resolution genetic typing). Furthermore, for more complex research questions, one may also choose to collect *longitudinal* or repeated measurements data, such as immune cell counts at regular visit times. These measurements may be missing intermittently (e.g. patients not attending some scheduled visits), or missing permanently from the moment a patient dies or drops out from a study prematurely. For both missing covariates and longitudinal data, the assumptions we make about *why* the data are missing are crucial in determining what kind of methodology we can use to draw valid conclusions, and the potential consequences of using naive approaches.

Lau and Lesko (2018) reported on the lack of research regarding the intersection of incomplete data and competing risks. They stated that while the awareness and use of methods that appropriately account for competing risks has grown over time, the potential impact of missing data in this setting has received comparatively little attention. In particular, the paucity of methodological research for dealing with

missing data in competing risks settings suggests that researchers are at risk of a) naively excluding individuals with missing data; b) using advanced methods such as *multiple imputation*, but failing to optimally account for the competing risks outcomes at the imputation stage. Consequently, depending on the extent and nature of a given missing data problem, the results of such analyses may be biased and/or make poor use of the observed information in the data. The development of methodological theory and associated user-friendly software implementations, the design and execution of simulation studies, and real data applications are all crucial in understanding when and how different statistical methods can be used for dealing with missing data in competing risks settings.

The present thesis is therefore about statistical methodology for dealing with incomplete data in observational studies with competing risks outcomes. Primarily, the focus is on the use of multiple imputation for handling missing covariate data in the context of major competing risks regression models. The use of *shared-parameter joint models* for dealing with informative drop-out is also addressed. The introduced methods are applied to multiple datasets of patients that have undergone an alloSCT.

The remainder of the introduction is structured as follows. In Section 1.1, we introduce notation for competing risks data and prominent regression models. In Section 1.2, we present basic concepts for missing covariate data and multiple imputation. In Section 1.3, we briefly introduce shared-parameter joint models. Finally, the outline and contributions of the thesis is given in Section 1.4.

1.1 Competing risks

In competing risks settings, individuals experience only one of K mutually exclusive events. We let $\tilde{D} \in \{1, ..., K\}$ be the variable indicating which of these events occurred at time $\tilde{T}$. In practice, $\tilde{T}$ is subject to right-censoring, and is therefore only partially observed. Denoting that right-censoring time as C, the observable data are realisations (t_i, d_i) for individual i of $T = \min(C, \tilde{T})$ and $D = I(\tilde{T} \leq C)\tilde{D}$, where $I(\cdot)$ is the indicator function and $D = 0$ indicates a right-censored observation.

Suppose interest initially lies in studying the time until any of these K events. For example, we may seek to estimate (for a given population) the cumulative probability $P(\tilde{T} \leq t)$ of failing from any cause by a certain timepoint t, or its complement: the event-free survival (EFS) probability $S(t) = P(\tilde{T} > t)$. Without any right-censoring, the latter quantity can be estimated simply by the empirical proportion of individuals still event-free by t.

In the presence of right-censoring, it is necessary to work with the so-called all-cause

hazard function, given by

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq \tilde{T} < t + \Delta t \mid \tilde{T} \geq t)}{\Delta t},$$

assuming the distribution of $\tilde{T}$ is continuous. It is the instantaneous *rate* at which individuals move from an initial event-free state to a second state representing failure from any cause. Importantly, the hazard function has a one-to-one correspondence with the EFS function, given by

$$S(t) = \exp\left\{ - \int_0^t h(u)\, \mathrm{d}u \right\},$$

and further visualised in Figure 1.1. Therefore, by modelling the all-cause hazard we can also estimate the EFS function. This is useful since by assuming that $\tilde{T}$ and C are stochastically independent of each other, the hazard function remains 'undisturbed' by any right-censoring (Beyersmann *et al.*, 2012). That is, the hazard for those censored at a given timepoint is equal to the hazard for those that remain under follow-up. This in turn makes it estimable from observed (i.e. censored) data. Further discussion on the subtleties regarding the random, non-informative, and independent censoring assumptions can be found in Kalbfleisch and Prentice (2011).

The most well-known estimator of the EFS function is the *Kaplan–Meier* or *product-limit* estimator (Kaplan and Meier, 1958). For a set of N ordered event times $0 < t_1 < t_2 < \cdots < t_N$, the estimator is given by

$$\hat{S}(t) = \prod_{j:\ t_j \leq t} \left\{ 1 - \frac{d_j}{n_j} \right\}, \tag{1.1}$$

where d_j and n_j respectively denote the number of events (of any type) and number of individuals at-risk at time t_j (Putter *et al.*, 2007).

In order to investigate the properties of specific events occurring in a competing risks setting, we can make use of the *cause-specific hazards*, defined for the k^{th} event as

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq \tilde{T} < t + \Delta t, \tilde{D} = k \mid \tilde{T} \geq t)}{\Delta t}.$$

These are the instantaneous rates of moving from an event-free state to a state representing failure from the k^{th} specific event, after which one can no longer fail of other causes. Furthermore, the cause-specific hazards sum up to the all-cause hazard function, and therefore fully define the EFS function defined earlier,

$$S(t) = \exp\left\{ - \sum_{k=1}^{K} \int_0^t h_k(u)\, \mathrm{d}u \right\} = \exp\left\{ - \sum_{k=1}^{K} H_k(t) \right\},$$

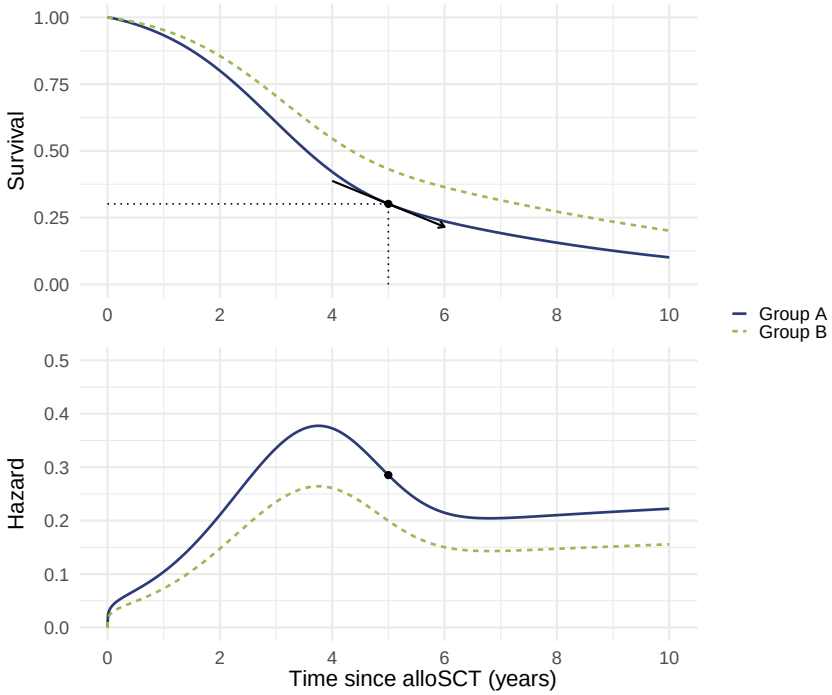

Figure 1.1: Example (all-cause) hazard and survival functions for two subgroups from a given population. The hazard is based on two aspects of the survival function at a particular timepoint: its value (i.e. the proportion of individuals still alive), and its slope (i.e. how rapidly the curve is decreasing, expressed as a rate per unit time). Mathematically, $h(t) = \{-\,\mathrm{d}S(t)/\,\mathrm{d}t\} \times S(t)^{-1}$.

where $H_k(t)$ is known as cause-specific cumulative hazard function for the k^{th} event. The cumulative probability of the k^{th} event occurring, known as cause-specific *cumulative incidence function* is defined as

$$F_k(t) = P(\tilde{T} \leq t, \tilde{D} = k) = \int_0^t h_k(u) S(u-) \, \mathrm{d}u, \tag{1.2}$$

where $S(u-)$ is the event-free survival probability just prior to u. As depicted in Figure 1.2, the cumulative incidence summarises two key aspects of the competing risks process: surviving event-free until just prior to t, and then experiencing the k^{th} specific event in the next instant.

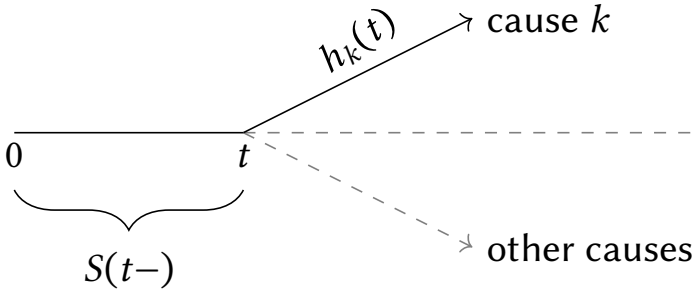

Figure 1.2: Summary of the competing risks process, adapted from Figure 1 in Geskus (2024).

The Aalen–Johansen estimator is the generalisation of the Kaplan–Meier method, allowing to estimate cumulative incidence functions non-parametrically (Aalen and Johansen, 1978). It has been extensively discussed in the literature that naively using the complement of the Kaplan–Meier estimator in the presence of competing risks (i.e. by treating competing events as censored) leads to overestimation of Equation 1.2, since it instead estimates

$$\int_0^t h_k(u) S_k(u-) \, \mathrm{d}u,$$

where $S_k(t) = \exp\{-H_k(t)\}$ (Putter *et al.*, 2007).

1.1.1 Competing risks regression models

Suppose we are interested in understanding the impact of individual-specific characteristics on the occurrence of one of multiple competing events (the 'primary' event of interest). In the case of a single categorical covariate of interest (e.g. presence or absence of a genetic mutation), a straightforward approach would be to use the Aalen–Johansen estimator in subsets defined by the categorical covariate. In order

to assess the 'impact' of this covariate, one may opt to then test the null hypothesis of equality of the resulting strata-specific cumulative incidence functions, usually by means of Gray's test (Gray, 1988).

Two important remarks can be made about the above strategy. The first is that its extension to multiple covariates is suboptimal: groups defined by combinations of categorical covariates will become increasingly smaller with the number of covariates, and continuous covariates need to be discretised. The second is that rejection of the aforementioned null hypothesis is only an indicator that being in a particular stratum relative to the others increases or decreases the cumulative probability of a specific event occurring, in the presence of competing risks. It does *not* however in isolation explain whether or to what extent this increase or decrease is attributable an existing association between the covariate and the competing risks.

For a more complete understanding of the competing risks process as a function of multiple covariates, one will usually need to specify regression models for the cause-specific hazards, which are identifiable and estimable from observed data (Prentice *et al.*, 1978). The most popular approach for modelling the cause-specific hazards is using the semi-parametric Cox proportional hazards model (Cox, 1972), given for the k^{th} event by

$$h_k(t \mid \mathbf{Z}) = h_{k0}(t) \exp(\boldsymbol{\beta}_k^\top \mathbf{Z}),$$

where $h_{k0}(t)$ is the cause-specific baseline hazard, $\mathbf{Z}$ is a p-dimensional vector of covariates, and vector $\boldsymbol{\beta}_k$ quantifies the impact of $\mathbf{Z}$ on the cause-specific hazard. Inference on the cause-specific, time-constant *hazard ratios* $\exp(\boldsymbol{\beta}_k)$ is based on maximising a *partial likelihood* which treats causes other than k as censored. The baseline hazard can be estimated non-parametrically by means of the Breslow estimator (Breslow, 1972).

It is important to note that a covariate which increases the cause-specific hazard of a particular event will not necessarily also increase the cumulative incidence of that same event, because the cumulative incidence function depends on the cause-specific hazards of *all* events. Depending on the direction and magnitude of the covariate effect on competing cause-specific hazards, the same covariate may even have opposing effects on the cause-specific hazard and cumulative incidence function of a particular event. As a result, regression models have been proposed that model the cumulative incidence function(s) *directly*, often through the so-called *subdistribution hazard*, given for a cause k by

$$\begin{aligned}\lambda_k(t) &= \lim_{\Delta t \downarrow 0} \frac{P\{t \leq \tilde{T} < t + \Delta t, \tilde{D} = k \mid \tilde{T} \geq t \cup (\tilde{T} \leq t \cap \tilde{D} \neq k)\}}{\Delta t}, \\ &= \frac{\mathrm{d}F_k(t)}{\mathrm{d}t} \times \{1 - F_k(t)\}^{-1}.\end{aligned}$$

The most prominent example of a model for the subdistribution hazard is the Fine–Gray model (Fine and Gray, 1999), given for a cause k by

$$\lambda_k(t \mid \mathbf{Z}) = \lambda_{k0}(t) \exp(\boldsymbol{\gamma}_k^\top \mathbf{Z}),$$

where $\lambda_{k0}(t)$ is the subdistribution baseline hazard, and $\boldsymbol{\gamma_k}$ is the vector of log subdistribution hazard ratios. It can also be expressed as a transformation model for the cumulative incidence function, as

$$F_k(t \mid \mathbf{Z}) = 1 - \exp\left\{ - \exp(\boldsymbol{\gamma}_k^\top \mathbf{Z}) \int_0^t \lambda_{k0}(u) \, \mathrm{d}u \right\},$$

using the complementary log-log link function.

Estimation for the Fine–Gray model is based on analogous partial-likelihood principles as in the cause-specific Cox model, except with a modified risk-set definition. Namely, as part of the estimation, individuals failing from competing events are kept in the risk-set indefinitely. In the presence of random right-censoring, individuals failing from competing events need to have their contribution to the partial likelihood re-weighted, since their competing event failure informatively censors their potential censoring time (i.e. how long they would have been in the risk-set for had they not experienced the competing event) (Geskus, 2011). Equivalently, potential censoring times for those failing from competing events can be multiply imputed (Ruan and Gray, 2008).

While there are many other existing modelling approaches in the presence of competing risks Geskus (2024), the cause-specific Cox and Fine–Gray models have arguably received the most attention in both the methodological and applied literature. Much of this attention has gone towards clarifying the different interpretation of the parameters from both models, and related estimands (see e.g. Andersen and Keiding, 2012). Specifically, the Fine–Gray model exclusively targets what historically has been called crude or 'mixed' probabilities (i.e. in presence of competing events), while cause-specific Cox models can also be used, under certain assumptions, to target so-called net or 'pure' probabilities (i.e. under hypothetical elimination of the competing events). These different potential quantities of interest were clearly summarised by Cornfield (1957):

> The estimate obtained is of a mixed probability. It provides an answer to the question: In a cohort subject for some future number of years to both the pure risk of developing the disease and to the pure risk of dying from some other cause what proportion will develop the disease in question?
>
> If we are interested in isolating effects, however, and wish to study, say, changes in the pure risk of developing a disease, without regard to changes in other causes of death, such a proportionate frequency may be misleading. Thus, if such a calculation tells us that the probability of developing

> cancer is higher now than it was in the past, this may be either because the pure risk of developing cancer has increased, or because the chance of dying of other causes has decreased.

More contemporary work has focused on providing a formal causal framework for these and related estimands, and providing the conditions under which they may be estimable from observed data (Martinussen and Stensrud, 2023; Young *et al.*, 2020)

Chapter 6 from this thesis provides a data-generating perspective on the Fine–Gray model, and argues in the favour of using cause-specific hazard models instead when multiple competing risks are of scientific interest.

1.2 Multiple imputation of missing covariate data

1.2.1 Missing data concepts

As discussed in the preceding section, right-censored survival times are a form of incomplete data. More specifically, they are a form of *coarsened* outcome data (Heitjan and Rubin, 1991), in that they are not 'fully' missing: we do at least know that an individual is event-free by their censoring time. When data are missing on an individual's (baseline) covariates, we are usually less fortunate, as these will tend to be fully missing.

The problem of missing covariate data in biomedical research remains pervasive to this day, and methods for dealing with missing covariate data have been extensively discussed in the methodological literature (Carpenter *et al.*, 2023; Enders, 2022; van Buuren, 2018). The core worry is that naively excluding individuals with missing data in at least one of multiple relevant covariates can result in estimates (e.g. of a survival probability) that are biased and/or inefficient (Sterne *et al.*, 2009).

The first step in assessing the extent of the missing data issue for a given application is to investigate the *missing data pattern*, and outline plausible assumptions for the *missing data mechanism*. As described in Enders (2022), 'patterns describe where the holes are in the data, whereas mechanisms describe why the values are missing'. Specifically, the missing data pattern tells us what variables have missing data and in what combination(s), while the missing data mechanism describes which variables contribute to the probability (and eventually in what functional form) of one of more these combinations occurring.

Rubin and colleagues (Little and Rubin, 1987; Rubin, 1976) developed a now ubiquitous framework for classifying missing data mechanisms. Using similar notation to Little and Rubin (2019), let $\mathbf{X} = (x_{ij})$ denote the hypothetically complete data matrix, where x_{ij} is the realised value of variable X_j for individual i. Additionally, define the

observation indicator matrix $\mathbf{M} = (m_{ij})$, where $m_{ij} = 1$ if x_{ij} is observed, and 0 if it is missing. The data can therefore be partitioned into observed and missing components as $\mathbf{X} = \{\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{mis}}\}$.

The general form of the mechanism (or missing data model, parametrised by ψ) is given by $f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}_i \,|\, \mathbf{x}_i; \psi)$, assuming that $(\mathbf{m}_i, \mathbf{x}_i)$, the i^{th} rows of $\mathbf{M}$ and $\mathbf{X}$, are independent and identically distributed across i. In other words, $\mathbf{M}$ is treated as a random matrix, where the probability of a *realised* missingness pattern $\mathbf{m}_i$ is allowed to depend on both observed ($\mathbf{x}_i^{\text{obs}}$) and unobserved ($\mathbf{x}_i^{\text{mis}}$) *realised* information from the same individual. This subtlety in Little and Rubin's definitions was clarified in work by Seaman *et al.* (2013), where broader definitions were also introduced (e.g. for missingness processes that hold across multiple data samples).

The missing data are said to be *missing completely at random* (MCAR, Marini *et al.*, 1980) when the probability of any particular unit being missing is independent of both observed and unobserved information, that is

$$f(\mathbf{m}_i \,|\, \mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}}; \psi) = f(\mathbf{m}_i \,|\, \psi) \text{ for all } \mathbf{x}_i.$$

MCAR missingness is often described as non-systematic or 'accidental', occurring for example as a result of measurement apparatus (e.g. heart rate monitor) malfunctioning or other administrative reasons.

Under *missing at random* (MAR), denoted as

$$f(\mathbf{m}_i \,|\, \mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}}; \psi) = f(\mathbf{m}_i \,|\, \mathbf{x}_i^{\text{obs}}; \psi) \text{ for all } \mathbf{x}_i^{\text{mis}},$$

the missingness depends only on observed information. For example, comorbidities may be monitored more closely in older patients, and therefore availability of information on the presence or absence of particular comorbidities may depend on (fully observed) patient age. However, if availability of comorbidity information was contingent on the underlying (partially missing) presence or absence of particular comorbidities, the mechanism is said to be *missing not at random* (MNAR). That is, the missingness mechanism depends at least partially on unobserved values. When missing data are multivariate, assumptions are formulated by relating the probability of any missing data pattern occurring (e.g. comorbidity information and patient age simultaneously missing) to the observed and unobserved values of that same pattern.

Furthermore, the missing data mechanism is said to be *ignorable* when a) the missing data are MAR; b) the parameters ψ of the missing data model, and the parameters of the analysis model θ (i.e. the one of scientific interest for the hypothetically complete data), are a priori distinct. Crucially, it implies that we do not need to model the missing data mechanism explicitly (hence ignorable), and that inference on θ is possible using only observed data. See also Little (2021) for an up-to-date discussion on missing data assumptions.

The above taxonomy does not in isolation fully inform what method(s) we can use to handle the missing values, or whether it is possible to obtain an unbiased estimate of some target estimand. Usually, one needs to also consider issues pertaining the structure of the outcome model itself, and whether or not the probability of missingness depends on the outcome variable(s) (Carpenter and Smuk, 2021). Related to the latter point, there has been a movement towards using graphical causal models to establish whether or not a particular estimand can be 'recovered' in the presence of missing data i.e. estimated consistently using only the observed data and no additional external information (Lee *et al.*, 2023).

1.2.2 Multiple imputation

Given that missing data usually mask values that would have been relevant or useful for analysis had they been observed, it makes sense to consider approaches which try to fill in or *impute* the unobserved values (Little and Rubin, 2019). For example, one may choose to replace all missing values in a given variable by the average or mode of the observed values from that same variable, and then proceed to analyse the data as if they were complete. This is an example of *single imputation*, which is usually problematic for inference purposes since it does not take into account the extra uncertainty due to the missing values. Additionally, the mean or mode may not be the most plausible guess for a given individual with an unobserved value given their other (observed) characteristics.

Multiple imputation (MI) instead repeatedly draws imputed values from a predictive distribution of the missing values given the observed ones, giving rise to multiple 'complete' datasets (Rubin, 1987). These can be analysed separately, and then the results can be combined in a way which reflects the additional uncertainty induced by the missing values. As described by Little (2024), approaches to MI will differ mainly in how this predictive distribution is derived. For example, one could specify a parametric joint model (e.g. multivariate normal, see Carpenter *et al.*, 2023) for the variables in the dataset, and impute the missing data based on the implied conditional distributions.

The more widely used approach to creating multiply imputed datasets is *multivariate imputation using chained equations* (MICE, van Buuren *et al.*, 1999). This approach is also known as *fully conditional specification* (FCS), since it involves specifying a univariate imputation model for each variable with missing data, fully conditional on the remaining variables in the dataset (or at least those relevant for the scientific question at hand). Under the ignorable missingness assumption, the conditional distribution of a variable with missing data given the remaining variables is the *same* for both the observed and unobserved components of the variable. Furthermore, the imputation models are iteratively 'chained' together until convergence: imputed values for each variable with missing data are drawn conditional on the most recently

imputed values for the other variables. The procedure is comprehensively outlined in the text by van Buuren (2018).

The flexibility of MICE lies in its variable-by-variable sampling, where each imputation model can be tailored to a particular variable type (e.g. ordered categorical). However, there is no guarantee that these univariate conditional distributions are actually mutually *compatible* with each other in the sense that together they form a coherent joint distribution. Nevertheless, as long as each conditional model fits the data well, this theoretical limitation may not have a large impact on pooled (i.e. after combining the results of the imputed datasets) inference (Liu *et al.*, 2014; Zhu and Raghunathan, 2015).

A related issue that has received more attention is that of compatibility between the imputation model(s) and the analysis model of interest, which is sometimes used interchangeably with the broader but closely related concept of *congeniality* (Meng, 1994). For example, for models with multi-level or longitudinal outcomes, or those that include non-linear terms such as interactions, the correct conditional distributions are usually difficult to specify directly (Du *et al.*, 2022; Erler *et al.*, 2016). This is also the case for regression models in the survival setting (Antunes *et al.*, 2021; Bartlett *et al.*, 2015; Beesley *et al.*, 2016; Haensch *et al.*, 2022; White and Royston, 2009).

Chapter 3 and Chapter 5 of this thesis directly tackle this compatibility issue respectively for cause-specific Cox and Fine–Gray models, while Chapter 2 provides a review of current practices in missing (covariate) data handling in clinical haematology. Additionally, the methodology in Chapter 3 is applied in Chapter 4.

1.3 Joint models and informative dropout

The focus of the preceding section was on handling missing data in baseline covariates. However, time-varying longitudinal information can also be collected, and may also have missing values. Suppose $\mathbf{y}_i^{\text{obs}} = \{y_{ij} = y_i(t_{ij}),\ j = 1, \dots, n_i\}$ is a vector of observed longitudinal measurements for individual i at timepoints t_{ij}. Here, we assume $\mathbf{y}_i^{\text{obs}}$ is a subset of an *intended* collection of measurements $\mathbf{y}_i$. After defining the vector of (random) observations indicators $\mathbf{m}_i$, the definitions of missingness mechanisms are analogous to those described in Section 1.2.1—see also Molenberghs and Fitzmaurice (2008) for further details.

Intermittent missingness is often assumed to be MAR, which allows the use of methodology based on the 'ignorable' likelihood (e.g. mixed models). That is, the marginal observed data density is used, and terms pertaining to the missingness process can be ignored (Rizopoulos, 2012). However, when missingness is due to *dropout* (i.e. $\mathbf{y}_i^{\text{obs}}$ are longitudinal measurements collected before an event or censoring time), the MAR assumption may not be as plausible. Dropout is said to be *non-random* (also referred

to as *informative*, even if the overarching concept is simply non-ignorability—see 17.2.3 of Molenberghs and Fitzmaurice, 2008), when conditioning on $\mathbf{y}_i^{\text{obs}}$ (or other fully observed baseline covariates) does not remove the dependence between the unobserved longitudinal measurements and the missingness process (Papageorgiou and Rizopoulos, 2021). This is for example the case when latent (unobserved) characteristics of the longitudinal measurements, such as the individual-specific true rate of increase or decrease at any timepoint, is related to the dropout probability. Analyses that ignore the dropout process in this context are subject to bias (Little, 1995).

An increasingly used approach to model the dependency between one or more longitudinal marker(s) and a dropout process are *shared-parameter joint models* (Rizopoulos, 2012). The first component of a basic joint model (for a single longitudinal marker and a univariate dropout process) is a longitudinal submodel, given by

$$y_{ij} = m_i(t_{ij}) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are independent random error terms, and $m_i(t)$ represents the true underlying (unobserved) value of the marker at time t. A mixed-effects structure is usually assumed for the latter, as

$$m_i(t) = \boldsymbol{\beta}^\top \mathbf{x}_i(t) + \boldsymbol{b}_i^\top \mathbf{z}_i(t),$$

where $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ are respectively (possibly time-dependent) design vectors for fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{b}_i \sim \mathcal{N}(0, D)$.

The second component is a survival submodel for the time to dropout (or other terminating event, such as death), and is usually assumed to follow a proportional hazards structure, given by

$$h_i(t \mid \mathcal{M}_i(t), \mathbf{w}_i) = h_0(t) \exp\{\boldsymbol{\gamma}^\top \mathbf{w}_i + \alpha m_i(t)\}, \tag{1.3}$$

where $\mathcal{M}_i(t) = \{m_i(s),\ 0 \leq s < t\}$ represents the history of the true marker value up to time t, and $\mathbf{w}_i$ is a vector of baseline covariates. The association between the true 'current value' of the longitudinal marker and the terminating event is quantified by α, and the baseline hazard $h_0(t)$ is often assumed to follow a flexible parametric form (e.g. using a B-spline basis).

Estimation of a joint model is based on a joint likelihood function which assumes that both longitudinal and dropout processes are independent, conditional on *shared* random effects $\boldsymbol{b}_i$. This latent structure means that joint models can provide unbiased estimates of the trajectories of the longitudinal marker when the dropout mechanisms depends on $\boldsymbol{b}_i$, which is an example of a MNAR mechanism. Furthermore, the joint model is suited for estimating the association between an *endogenous* or internal longitudinal marker and a time-to-event outcome, which is usually underestimated by a time-varying Cox model since it treats these as exogenous (Kalbfleisch and Prentice, 2011; Rizopoulos, 2012; Sweeting and Thompson, 2011).

Many extensions to the basic joint model have been proposed. For example, the longitudinal submodel can be extended to accommodate multiple longitudinal markers of mixed type (i.e. binary, count, and others), usually linked together via their respective random effects (Hickey *et al.*, 2016). Furthermore, Papageorgiou *et al.* (2019) outline various extensions for the association structure between the longitudinal marker(s) and the time-to-event outcome. In addition to the current value parametrisation in Equation 1.3, one could also consider modelling the current slope $\mathrm{d}m_i(t)/\,\mathrm{d}t$, or more complex transformations such as (weighted) cumulative effects.

The time-to-event submodel itself has also been extended in order to account for multiple (sequences of) events. Notably, Hickey *et al.* (2018) provide an overview of joint models that consider time to multiple competing events. In Chapter 7 of this thesis, we use the competing risks joint model described in Rizopoulos (2012) (based on cause-specific Cox proportional hazards) to model the trajectories of immune cell counts after T-cell depleted alloSCT.

1.4 Thesis outline and contributions

The current thesis aims to develop, assess, and apply statistical methodology for dealing with missing data in the context of alloSCT studies focusing on competing risks outcomes. In what follows, we outline the contributions of the individual chapters comprising this thesis.

Chapter 2 is a review aimed at haematological audiences concerning methodology for handling missing covariates. The first goal of this chapter is to provide a gentle introduction to major methods for handling missing covariate data, including multiple imputation. The second goal is to provide a contemporary overview, by means of a systematic review, of how missing covariate data are being handled across major haematological journals. While we know from other review articles in related applied fields that multiple imputation is being increasingly used (Hayati Rezvan *et al.*, 2015), we did not know whether this is also the case for alloSCT studies, or if the methodology is being applied and reported adequately. This chapter is based on the following publication:

> **Bonneville, E. F.**, Schetelig, J., Putter, H., de Wreede, L. C. (2023) Handling missing covariate data in clinical studies in haematology. *Best Practice & Research Clinical Haematology*, 36, 101477. DOI: 10.1016/j.beha.2023.101477.

Chapter 3 focuses on the use of multiple imputation when the analysis model of interest is one or more cause-specific Cox proportional hazards models. The first goal of this work is to formally derive approximately compatible imputation models for various types of missing baseline covariates (e.g. continuous, binary, etc.). The second

goal is to compare these approximately compatible imputation models to a previously proposed substantive-model-compatible approach (Bartlett and Taylor, 2016) as part of large simulation study. A novel aspect of this simulation study is the focus on the impact of missing covariate data on the estimation of individual-specific cumulative incidence functions. Furthermore, the methodology is applied on a dataset with long-term follow-up after an alloSCT for patients with myelodysplastic syndromes or secondary acute myeloid leukaemia (Schetelig *et al.*, 2019). This chapter is based on:

> **Bonneville, E. F.**, Resche-Rigon, M., Schetelig, J., Putter, H., de Wreede, L. C. (2022) Multiple imputation for cause-specific Cox models: Assessing methods for estimation and prediction. *Statistical Methods in Medical Research*, 31, 1860–1880. DOI: 10.1177/09622802221102623.

Chapter 4 represents another application of the methodology in Chapter 3, using a dataset of patients with myelofibrosis who have undergone an alloSCT. The goal of this applied work is primarily is assess the impact of partially observed comorbidities and body mass index (BMI) on the cause-specific hazard of non-relapse mortality. The imputation procedure in this chapter involved careful consideration of several non-standard issues such as the imputation of derived variables, the use of auxiliary variables, and imputation when several substantive models are of interest. This chapter is based on:

> Polverelli, N.[†], **Bonneville, E. F.**[†], de Wreede, L. C., Koster, L., Kröger, N. M., Schroeder, T., Peffault de Latour, R., Passweg, J., Sockel, K., Broers, A. E. C., Clark, A., Dreger, P., Blaise, D., Yakoub-Agha, I., Petersen, S. L., Finke, J., Chevallier, P., Helbig, G., Rabitsch, W., Sammassimo, S., Arcaini, L., Russo, D., Drozd-Sokolowska, J., Raj, K., Robin, M., Battipaglia, G., Czerw, T., Hernández-Boluda, J. C., McLornan, D. P. (2024) Impact of comorbidities and body mass index on the outcomes of allogeneic hematopoietic cell transplantation in myelofibrosis: A study on behalf of the Chronic Malignancies Working Party of EBMT. *American Journal of Hematology*, 99, 993—996. DOI: https://doi.org/10.1002/ajh.27262
>
> [†]These authors contributed equally to this work and share first authorship

In Chapter 5, novel methodology is developed for imputing missing covariate values compatibly with a Fine–Gray analysis model for a competing event of interest, without needing to specify a model for the competing event(s). As mentioned by Lau and Lesko (2018), there is a paucity of research regarding the use of multiple imputation for missing covariates in competing risks settings, and even more so in the context of subdistribution hazard models. The simulation study in this work is, to the best of our knowledge, the first systematic assessment of how substantive model misspecification at the covariate imputation stage affects inference in the competing risks setting. The methodology is also assessed using the dataset from Chapter 4. This chapter is based on the following preprint (currently under review in *Statistics in Medicine*):

> **Bonneville, E. F.**, Beyersmann, J., Keogh, R. H., Bartlett, J. W., Morris, T. P., Polverelli, N., de Wreede, L. C., Putter, H. (2024) Multiple imputation of missing covariates when using the Fine–Gray model. *arXiv preprint* arXiv:2405.16602. DOI: 10.48550/arXiv.2405.16602.

Chapter 6 provides an overview of data-generating mechanisms where a Fine–Gray model correctly holds for at least one competing event. In the context of the present thesis, this work mainly relates to the methodology developed in Chapter 5. That is, it outlines the possible assumptions that could be made for the competing event(s) at the imputation stage, should one want to develop substantive-model-compatible imputation methodology based on the complete data likelihood (in contrast to the methodology from Chapter 5, which is tailored for only one competing event). In isolation however, this article provides simulation-based arguments against specifying Fine–Gray models for more than one competing event, and suggests using cause-specific hazard models instead. This chapter is based on the following publication:

> **Bonneville, E. F.**, de Wreede, L. C. and Putter, H. (2024) Why you should avoid using multiple Fine–Gray models: Insights from (attempts at) simulating proportional subdistribution hazards data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae056. DOI: 10.1093/jrsssa/qnae056.

In Chapter 7, the immune cell kinetics of a cohort of patients with acute leukaemia during the 6 months after T-cell depleted alloSCT are analysed using competing risks joint models. Prior to this work, joint models were rarely used in the alloSCT context for modelling T-cell trajectories (to the best of our knowledge, only in Salzmann-Manrique *et al.*, 2018), and did not consider competing risks in the time-to-event submodel or post-baseline interventions such as a donor lymphocyte infusion. This chapter is based on:

> Koster, E. A. S.[†], **Bonneville, E. F.**[†], von dem Borne, P. A., van Balen, P., Marijt, E. W. A., Tjon, J. M. L., Snijders, T. J. F., van Lammeren, D., Veelken, H., Putter, H., Falkenburg, J. H. F., Halkes, C. J. M., de Wreede, L. C. (2023) Joint models quantify associations between immune cell kinetics and allo-immunological events after allogeneic stem cell transplantation and subsequent donor lymphocyte infusion. *Frontiers in Immunology*, 14. DOI: 10.3389/fimmu.2023.1208814.
>
> [†]These authors contributed equally to this work and share first authorship

Finally, the main conclusions from the different chapters are brought together in Chapter 8. Software code to reproduce the analyses of each chapter are openly available on https://github.com/survival-lumc.

Chapter 2

Handling missing covariate data in clinical studies in haematology

Chapter based on: **Bonneville, E. F.**, Schetelig, J., Putter, H., et al. (2023) Handling missing covariate data in clinical studies in haematology. *Best Practice & Research Clinical Haematology*, 36, 101477. DOI: 10.1016/j.beha.2023.101477

Abstract

Missing data are frequently encountered across studies in clinical haematology. Failure to handle these missing values in an appropriate manner can complicate the interpretation of a study's findings, as estimates presented may be biased and/or imprecise. In the present work, we first provide an overview of current methods for handling missing covariate data, along with their advantages and disadvantages. Furthermore, a systematic review is presented, exploring both contemporary reporting of missing values in major haematological journals, and the methods used for handling them. A principle finding was that the method of handling missing data was explicitly specified in a minority of articles (in 76 out of 195 articles reporting missing values, 39%). Among these, complete case analysis and the missing indicator method were the most common approaches to dealing with missing values, with more complex methods such as multiple imputation being extremely rare (in 7 out of 195 articles). An example analysis (with associated code) is also provided using haematopoietic stem cell transplant data, illustrating the different approaches to handling missing values. We conclude with various recommendations regarding the reporting and handling of missing values for future studies in clinical haematology.

2.1 Introduction

Missing data are widely encountered in clinical research across both covariate and outcome information. For example, a patient known to have relapsed from a particular malignant disease may be lacking information on the precise timing of the relapse (missing outcome), but may also have incomplete data on important prognostic factors that may be time-consuming or expensive to collect, or were not considered relevant at the moment of diagnosis or start of treatment (missing covariate).

While missing values generally represent a nuisance to the researcher, failure to adequately account for them may lead to results that are potentially biased and/or imprecise (Sterne *et al.*, 2009). Insufficient reporting of missing values may also complicate the interpretation of a study's results. Both the reporting of missing values and the methods used for handling them have been the subject of multiple reviews across a multitude of study types and outcome models (Bell *et al.*, 2014; Burton and Altman, 2004; Carroll *et al.*, 2020; Sullivan *et al.*, 2017).

On the subject of clinical studies in haematology, there has to the best of our knowledge been no explicit investigation into contemporary reporting and handling of missing data. Similarly, field-specific guidelines seem to be restricted to a section from Delgado *et al.* (2014), which are limited in scope and arguably outdated. The present work seeks to fill this knowledge gap, focusing mainly on missing covariate data. We will first introduce the assumptions underlying missing data, as well as the common methods used to handle them and their associated pitfalls. We follow this up with a systematic review of missing data related practices across major haematological journals, and an applied data example. We then conclude with recommendations for further practice.

2.2 Assumptions underlying methods for handling missing data

Any discussion concerning missing values in the context of a particular study should always begin with one simple question: why are the values missing? This enquiry into the possible causes of the missing values is in fact an assessment of the missing data mechanism. Little and Rubin (2019) formalised the concept, assuming that every value a priori had some probability of becoming missing, which could depend on either observed or unobserved information.

Suppose that in the context of a retrospective study, performance status (measured for example by means of the Karnofsky Performance Score, KPS) is missing for a portion of the patient cohort. These values are missing completely at random (MCAR) when all patients are equally likely to have a missing value. That is, no one subgroup defined by the data (e.g. older patients) has any comparatively higher likelihood of

having a missing KPS value. The missingness is independent from both observed and unobserved information.

In the case where observed information fully explains whether the data are missing or not, we refer to the mechanism as missing at random (MAR). For example, younger patients may have their KPS score recorded less routinely than older patients. In other words, whether or not the KPS score is observed can be explained using the observed age information. If the missing data are related to unobserved information, we refer to them as missing not at random (MNAR). An example of this would be if patients with a higher KPS score would also have a higher likelihood of having a corresponding missing value, e.g., because fitter—but not necessarily younger—patients are monitored less intensively. The missingness in this example is therefore related to the unobserved KPS values or patient fitness not measured by any other variable.

Given that MNAR data by definition depends on unobserved information, it is impossible to distinguish from MAR data based on data alone. Since the most common missing data handling methods will rely on data being MAR, it is critical to discuss how plausible the assumption is in a given context. Such a discussion should take place not only between statisticians and researchers, but also together with data managers, and other persons who possess intimate knowledge of the data collection process.

2.3 Common approaches and associated pitfalls

In what follows, we outline common approaches for handling missing values in the context of a regression model—referred to as the analysis or substantive model. We assume that the outcome of the model is observed for all individuals, and that one or more of the predictors used have missing values.

2.3.1 Complete case analysis

The simplest solution to deal with missing covariate data is to perform a complete case analysis (CCA). This excludes patients from an analysis as soon as they have an unknown value on at least one of the predictor variables. Therefore, in a multivariable analysis model with missingness spanning multiple predictors, the potential loss in statistical precision can be substantial—as reflected by wide confidence intervals. While a CCA is valid under the assumption of MCAR, it is also unbiased in various MAR and even MNAR situations (Hughes *et al.*, 2019). In turn, when missingness depends on the outcome, or observed variables explaining the missingness are not accounted for (e.g. by adjusting for them in a multivariable model), CCA will generally be biased. That is, the estimate(s) obtained using CCA will differ with respect to those that would have been obtained had the data been complete. Note also that CCA

is the (often implicit) default strategy for handling missing values in most software packages.

2.3.2 Missing indicator method

A straightforward way to exploit all available data is to use the missing indicator method (MIM). For categorical variables, this simply involves creating an additional level for the missing values. For continuous variables, the missing values are first replaced by a constant (commonly zero, or the observed average), and thereafter a binary variable indicating whether the value was missing in the first place or not is added to the model. Two major advantages of this method are that the patients with missing information are monitored in univariable analyses (e.g. checking whether those with missing data have much higher or lower overall survival compared to those with observed information), and that it retains all information for multivariable modelling. The MIM has historically been criticised, as it has been shown to potentially be biased, even in cases with data that are MCAR (Donders *et al.*, 2006). Nevertheless, it may prove to be a pragmatic solution under particular assumptions and mild confounding (Blake *et al.*, 2020).

2.3.3 Single imputation

One may also choose to replace all missing values in a variable (for example by the average among observed values) and proceed to analyse the data as if they were complete. This type of single imputation is generally discouraged as it does not take into account any of the uncertainty regarding the missing values (Sterne *et al.*, 2009). Data are analysed as if they were complete (i.e. failing to capture that each missing value could have taken a range of different values, other than for example the observed average), leading to confidence intervals that are too narrow. A less well known instance of single imputation occurs in categorical variables, where the analyst may choose to combine those individuals with missing values together with another factor level, usually the most frequent one. This may in fact also occur at a pre-processing stage, as may be the case for the calculation of the disease risk index (DRI), used in registry studies about patients undergoing allogeneic stem cell transplantation (alloSCT) (Armand *et al.*, 2014). In different implementations of the DRI, patients with myelodysplastic syndromes or acute myeloid leukemia lacking cytogenetic risk classification (a component of the DRI) are assumed to belong to (i.e. singly imputed as) the most commonly observed category, intermediate cytogenetic risk (Armand *et al.*, 2012; Saccardi *et al.*, 2023; Snowden *et al.*, 2020). In the original implementation of the DRI, this was justified on the basis of outcomes being similar between those unavailable and intermediate cytogenetics. Note also that the MIM is not a form of

single imputation: it assumes (in the case of a categorical variable), that missing values belong to a separate category, rather than to any one of the existing categories.

2.3.4 Multiple imputation

A more complex strategy to handle missing values is multiple imputation (MI). MI leverages the observed data to repeatedly replace the missing values by plausible (model-based) values. Doing so results in multiple 'complete' datasets that can separately be analysed, and whose results can be combined in a way that properly accounts for the uncertainty induced by the missing values (using so-called Rubin's rules). MI is one of multiple missing data methods that operate under the MAR assumption, and are capable of producing more precise estimates compared to CCA (White and Carlin, 2010).

There are different variants of MI, of which the best known is multivariate imputation by chained equations (MICE) (van Buuren *et al.*, 1999), implemented for example in the 'mice' R package (van Buuren and Groothuis-Oudshoorn, 2011). For each variable with missing data, an imputation model needs to be specified and fitted using the observed part of the variable, and can thereafter be used to generate imputed values. This approach allows to flexibly specify different models for different variable types, such as a logistic regression if the variable to be imputed is binary. Once these models are specified, there are two main settings to fix: the number of imputed datasets, and the number of iterations (also known as cycles). To generate a single imputed dataset, MICE will start with an initial guess for each missing value by choosing randomly from the observed values. Naturally, these initial guesses may be far from values we could have expected given a patient's other variables: for example, initial guesses for KPS scores among younger patients could be low due to the random choice, when we might have plausibly expected them to be high. In order to move in the direction of more plausible values given the observed data, we 'chain' the imputation models together: the most recently imputed values are used as predictors in the imputation model for the next variable to be imputed. One pass across all variables with missing data represents a single cycle or iteration. It is necessary to continue these cycles (i.e. imputed values at the end of the first cycle are used as starting values for the second cycle, and so forth) until 'convergence' is reached, i.e. the sequence of imputed values generated has stabilised. The imputed values at the end of the final cycle form a single imputed dataset. Independent (i.e. with different starting values) runs of these cycles give rise to multiple imputed datasets.

In studies where missing data are abundant, it is crucial to pick a sufficiently large number of imputed datasets. When the proportion of missing data is high, the imputation models are fitted on a smaller portion of data, which will lead to a larger variability in the imputed values. Intuitively, we need to make sure that conclusions made on the basis of for example 10 imputed datasets, do not substantially differ from

those that would have been made under another set of 10 imputed datasets (i.e. had we repeated the entire multiple imputation procedure). A rule of thumb is to use as many imputed datasets as the percentage of incomplete observations (e.g. 40 imputed datasets if a CCA discards 40% of observations), although recent research suggests that the number of imputed datasets should be even larger (Mertens *et al.*, 2020; von Hippel, 2020). Generally, at least 10-20 iterations are recommended, and convergence should be assessed using visual diagnostic tools, such as those described in section 6.4.2 of the text by van Buuren (van Buuren, 2018).

Another key consideration in MI is the issue of compatibility between analysis and imputation models, that is, that the imputation models should make assumptions that are consistent with those made by the proposed analysis model. Practically speaking, this means that all variables in the analysis model, including the outcome, should also be included as predictors in the imputation model. It also means that special terms, such as interactions, should be adequately handled. A version of MICE which naturally ensures compatibility between the analysis and imputation models in such situations is substantive-model-compatible fully conditional specification (SMC-FCS) (Bartlett *et al.*, 2015), implemented in the 'smcfcs' R package (Bartlett *et al.*, 2022).

A closely related approach to SMC-FCS is fully Bayesian imputation, implemented in the 'JointAI' R package (Erler *et al.*, 2021). This performs analysis and imputation simultaneously, again ensuring compatibility between the imputation and analysis models. An overview of these and related methods are found in the work of Carpenter and Smuk (Carpenter and Smuk, 2021).

Research concerning the theory and performance of different MI approaches extends to the field of survival analysis, with particular focus on the widely-used Cox proportional hazards model. When the analysis model of interest is a Cox model, White and Royston (2009) suggested that the imputation model for a partially observed variable should include as predictors the remaining covariates from the analysis model, the event indicator (indicating whether an patient experienced an event or was censored) and the Nelson–Aalen estimate of the cumulative hazard. This approach is expected to work well in settings with a low cumulative incidence, and when the true effects of the covariates are small. When this does not hold, results are expected to be biased: the imputation and analysis model are only approximately compatible. That is, this imputation model produces imputed values which are not completely consistent with data where the key assumption made by a Cox model (multiplicative covariate effects on the hazard) is assumed to hold.

The SMC-FCS approach, which addresses this compatibility issue directly, has shown superior performance (when the analysis model is well specified) compared to the standard MICE approach described in the previous paragraph across multiple simulation studies, including settings with time-varying effects of covariates (Keogh and Morris, 2018), excess hazard models (Antunes *et al.*, 2021), and competing risks (Bartlett and

Taylor, 2016; Bonneville *et al.*, 2022). While to our knowledge there has been no systematic evaluation of the fully Bayesian approach in the context of the Cox model, the expectation is that it should perform at least as well as the SMC-FCS approach. This is because it not only ensures that the imputations are consistent with the analysis model, but also that the different imputation models (when there are multiple variables to be imputed) are consistent with each other. The latter point may be of concern when there are complex non-linear relationships between covariates.

2.4 Systematic review

2.4.1 Search strategy and data extraction

We performed a systematic review to obtain a broad picture of current practices in missing data reporting and handling across research articles published in major haematological journals. We used the Ovid platform to search the MEDLINE and Embase databases for articles written in English in 2021. We excluded articles that did not contain new data analyses (such as review articles), as well as letters to the editor, since their brevity would preclude full reporting on the issues we are interested in. Meta-analyses, methodological publications, and articles that were co-authored by authors of the present work were also excluded. A total of 16 journals were selected, based on two primary criteria: a 5-year Journal Impact Factor (2021) larger than 3 (data obtained via Journal Citation Reports Science Edition, Clarivate Analytics 2018), and a journal scope focused on clinical research in haematological malignancies. The 5-year Journal Impact Factor criteria was used in order to target articles with (on average) better quality of methodology and reporting, and a larger readership.

The search terms in Ovid format are reported in Table 2.1. After narrowing down by journals and year of publication, the remaining criteria focused on the malignant disease group, and the type of analysis model used. A wide selection of malignant diseases was included: acute and chronic myeloid leukemia, acute and chronic lymphocytic leukemia, myelodysplastic syndromes (MDS), and non-Hodgkin lymphomas (NHL) . We searched for articles that used a multivariable Cox proportional hazards model as part of their statistical analysis. Models were further classified into standard Cox, Fine–Gray (competing risks), frailty, multi-state and relative survival. The standard Cox category included standard outcomes such as overall or progression-free survival, but also cause-specific Cox models for outcomes such as relapse incidence and non-relapse mortality (since these are applied by treating competing risks as censored). Search terms for both analysis model and malignant disease group were based on detection of relevant strings in either title, abstract or keywords.

For each included article, the information extraction spanned three main areas: 1) exclusion of patients at 'population selection' phase based on missing data, 2) presence

Table 2.1: Search terms used in Ovid format, as entered into MEDLINE and Embase.

1.	(0006-4971 or 2352-3026 or 1756-8722 or 2044-5385 or 0887-6924 or 0390-6078 or 0361-8609 or 2473-9529 or 0007-1048 or 2666-6367 or 0268-3369 or 2040-6207 or 0278-0232 or 0939-5555 or 1545-5009 or 1042-8194).is.
2.	(cox or HR or aHR or (hazard adj1 ratio) or hazard or (proportional adj1 hazards) or multivaria*).ti,ab,kf.
3.	((acute adj1 myeloid adj1 leukemia) or (myelodysplastic adj1 syndrome*) or (chronic adj1 lymphocytic adj1 leukemia) or non-Hodgkin* or (chronic adj1 myeloid adj1 leukemia) or (acute adj1 lymphocytic adj1 leukemia) or (multiple adj1 myeloma) or leukem* or leukaem*).ti,ab,kf.
4.	1 and 2 and 3
5.	limit 4 to ('conference review' or 'review')
6.	limit 4 to (article or article in press or journal article)
7.	6 not 5
8.	limit 7 to english language
9.	limit 8 to yr='2021 - 2021'
10.	remove duplicates from 9

and explicit reporting of missing data in baseline covariates used in one or more of the reported analysis models, and 3) explicit reporting of methods used to handle the missing data. The first part of the extraction is based on the findings in Carroll *et al.* (2020), namely that the population is occasionally filtered on the basis of information/variable availability, leaving little or no missing data at the analysis phase. Two possible examples of this are, a) retrospective analysis of data from multiple trials, and a particular trial being excluded because information on a covariate of interest was not collected; b) in a retrospective study, including only those with sequencing data at a point in time (thereby implicitly excluding those with unavailable sequencing data).

We checked whether there was any missing data in any of the covariates making up the multivariable model, whether it be reported in the descriptive 'Table 1' (hereafter referred to as the 'descriptives table'), in a figure or in the main text. If there were multiple multivariable Cox models reported in the article, we recorded whether in at least one of them there was a covariate with missing data. Note that sometimes the missing values are only implicitly reported, as is for example the case when the numbers per level of categorical variables are reported, and these fail to add up to the total.

In terms of missing data handling, we paid particular attention to the use of CCA. Specifically, when missing values are reported, authors can be explicit about use of complete cases in two main ways: specifying the number of subjects used when reporting the multivariable model, or including a sentence in-text explicitly mentioning the use of CCA (the sentence 'missing values were not imputed' is also appropriate). Otherwise, the use of CCA was considered implicit—which can be problematic since the reader is unaware of the extent of the power loss. Likewise, we also recorded

whether other methods were used, such as: MIM, single imputation, MI or other. We also recorded the software used for the analysis. The full extraction sheet is available in the online supplement.

An initial investigation was carried out by EB, LdW and HP using 10 randomly selected papers. This was done in order to assess the consistency of data extraction, sharpen the data extraction checklist and agree on how to extract information when answers were ambiguous. Data extraction was then carried out by EB.

2.4.2 Results

A total of 398 research articles were identified after eliminating duplicate records obtained via MEDLINE (n = 86) and Embase (n = 391). From those, 99 were excluded due to either co-authors of the present manuscript being involved as co-author on the publication (n = 8), meta-analyses (n = 6), no Cox models reported (n = 48) and absence of multivariable Cox model (n = 36). One article was excluded as the supplementary material (which described the predictors used in the multivariable model) was not available on the publisher's website. A total of 299 articles were therefore included in the review. The journals where these articles most frequently featured in were Bone Marrow Transplantation (n = 46), Transplantation and Cellular Therapy (n = 40), Blood Advances (n = 35), Annals of Hematology (n = 33), and Leukemia and Lymphoma (n = 28).

At population selection, 80 articles (27%) explicitly reported having excluded observations on the basis of missing information. At this stage of the extraction, the focus was not yet solely on covariates that would make part of the multivariable model(s) in a particular article. These 80 articles could thus for example comprise exclusions based on missing outcome data. Given the retrospective nature of many studies, many of these exclusions were based on lack of cytogenetic information or no minimal residual disease (MRD) assessments, as was the case (exclusion of patients without cytogenetic information) for example in Hansen *et al.* (2021). It is important to note that while this approach can seem natural, it could come at the cost of ending up with a slightly different population than the one originally targeted. A possible sanity test would be to compare the univariable outcomes (e.g. Kaplan–Meier based overall survival) between those excluded and those remaining.

The vast majority of articles (287 out of 299, 96%) included at least one standard Cox model. The Fine–Gray model for competing risks was used in 69 articles (23%), and frailty models were employed in 14 articles. Multi-state models (n = 2) and relative survival approaches (n = 3) were rarely employed. Furthermore, the software used for the analyses was R (n = 144), SPSS (n = 94), SAS (n = 55), Stata (n = 35), unknown/ not specified (n = 40), and other (n = 28). These numbers do not add up to the total as 86 articles used two or more of these software packages in combination.

A total of 195 articles (65%) reported missing data in at least one of the covariates that formed part of one or more of the presented multivariable models. In most cases, these were explicitly reported in the descriptives table (n = 124), in both the descriptives table and in text (n = 24) and in the main text only (n = 19). In some instances, these were reported as part of a figure/flowchart (n = 3). The missing values (for at least one of the variables) were implicit in 20 articles, deducted based on subcategories in the descriptives table not adding up to totals.

In 39% (n = 76) of the 195 articles reporting missing values, the method (or at least one of the methods, if multiple were used) for handling missing data was explicitly specified. The most common methods for dealing with missing values among these were CCA (n = 34), MIM (n = 29), MI (n = 6) and single imputation (n = 5). One article used both MIM and CCA together, while another used both MI and CCA. Han *et al.* (2021) provide a clear example of explicit method reporting in-text: they mention the use of MIM for a particular mutation status indicator, and the use of CCA on the remaining variables with missing values. Furthermore, in that same study, treatment data was not available for all patients: they proceeded to compare survival outcomes of patients with and without available treatment data, and checked that there were no differences. Occasionally, variables were explicitly excluded from the multivariable model if their proportion of missing values were deemed too large. This was the case in Sharma *et al.* (2021), where multiple variables (such as KPS and cytomegalovirus serostatus) were excluded from analyses on the basis on having a proportion of missing values larger than 35%.

In 67% (n = 131) of cases, all or part of reported missing data was handled implicitly, which was assumed to correspond to implicit CCA. Indeed, this meant that one or more predictors included in the multivariable model(s) had missing values, and that results were reported as if the data were complete: with no explicit mention of running a CCA in text, or no information provided on the reduction in sample size. Inoue *et al.* (2021) provide an example of both explicit and implicit reporting: the MIM was used for the HCT-CI comorbidity index, however the handling method for the other incomplete adjustment variables (such as performance score) was not stated.

From the few articles using MI, three explicitly mention using MICE, with the remaining simply referring to their approach to handling the missing values as general 'multiple imputation'. Information on the details of the imputation procedure was lacking across all these articles, with only three mentioning the number of imputed datasets (10, 15 and 30 imputed datasets used), and none of the articles outlining the contents of the imputation models or the number of cycles (also not in the supplementary materials).

2.5 Illustrative example

We make use of data published by Schetelig *et al.* (2019), describing long-term outcomes of patients with myelodysplastic syndromes (MDS) and secondary acute myeloid leukemia (sAML) following an alloSCT to demonstrate selected options how to deal with missing data. The dataset contained both outcome information (timing of relapse or non-relapse mortality, if either occurred) and variables measured at baseline for 6434 patients registered with the EBMT, transplanted between 2000 and 2012. Several of these recorded baseline predictors presented a substantial amount of missing data: IPSS-R cytogenetic classification (62.2%), HCT-CI comorbidity index (59.9%), donor age (49.5%), KPS (32.8%), and cytomegalovirus (CMV) status in both donor and patient (17.8%). For full description of the variables, we refer to Table 2.2. In our previous publication, we illustrated several approaches for dealing with these missing values in competing risks analyses (Bonneville *et al.*, 2022).

Table 2.2: Data dictionary with predictor variables and their descriptions, levels and proportion missing data for the illustrative example, adapted from Bonneville *et al.* (2022). The 'Summary' column reports median and interquartile range for continuous variables, as well as counts and proportion per level of categorical variables. Abbreviations: CMV = cytomegalovirus, CR = complete remission, IPSS-R = International Prognostic Scoring System, V. = very, interm. = intermediate, HLA = Human leukocyte antigen, HCT-CI = Hematopoietic stem cell transplantation-comorbidity index, M = male, F = female, MDS = myelodysplastic syndromes, sAML = secondary acute myeloid leukemia, w/= with, w/o = without.

Variable	Description	Levels	% Missing	Summary
Age (Donor)	Donor age at alloHCT (years)		49.49	42.1 (31.2, 52.5)
Age (Patient)	Patient age at alloHCT (years)		0	56 (46.9, 61.9)
CMV Patient/Donor	CMV status in patient and donor	Patient -/Donor -	17.8	1439 (27%)
		Patient -/Donor +		544 (10%)
		Patient +/Donor -		1281 (24%)
		Patient +/Donor +		2024 (38%)
Comorbidity score	HCT-CI score	Low risk (0)	59.93	1322 (51%)
		Interm. risk (1 – 2)		657 (25%)
		High risk (≥ 3)		599 (23%)
Cytogenetics	Cytogenetics categories used for IPSS-R	V. good/good/interm.	62.23	1784 (73%)
		Poor		287 (12%)
		V. poor		359 (15%)
HLA match patient/donor	HLA match between patient and donor	HLA-identical sibling	0	2666 (41%)
		Other		3767 (59%)

Karnofsky	Karnofsky performance status	≥ 90	32.8	3130 (72%)
		80		898 (21%)
		≤ 70		295 (7%)
MDS class	MDS groups based on subclassification at alloSCT	MDS w/o excess blasts	0	1355 (21%)
		MDS w/ excess blasts		2716 (42%)
		sAML		2362 (37%)
Patient/Donor sex match	Sex match patient and donor	M/M	1.68	2545 (40%)
		M/F		1196 (19%)
		F/M		1474 (23%)
		F/F		1110 (18%)
Stage	Stage at alloHCT	CR	3.17	2119 (34%)
		no CR		2156 (35%)
		Untreated		1954 (31%)

In the current work, using this dataset we present a multivariable Cox model for relapse-free survival (RFS), using the same covariates as in the cause-specific Cox models in Bonneville *et al.* (2022): the baseline variables with missing values described above, together with the completely observed variables patient age, stage at alloSCT, patient-donor human leukocyte antigen (HLA) match, patient-donor sex match and MDS classification. We compared various methods for dealing with the missing values: CCA, MIM, MICE, SMC-FCS, and fully Bayesian imputation using the 'JointAI' R package. The MI analyses were performed using 100 imputed datasets with 15 iterations, and the analysis using 'JointAI' used 2000 iterations following 200 adaptation iterations. Hazard ratios obtained using each method are presented in Figure 2.1 along with their corresponding 95% confidence interval. Full code to reproduce the analysis is available at https://github.com/survival-lumc/ReviewHaemaMissing.

Regarding results, we first note that the loss in efficiency when using a CCA is apparent in this application, with confidence intervals that are considerably larger than in any of the other methods. Indeed, the CCA made use of only 17.5% of the available observations (5309 patients were omitted from the analysis). CCA also presents point estimates that differ considerably compared to the remaining methods, as is the case for example with patient-donor sex match categories, the CMV status in patient and donor, and cytogenetic risk classification. Such differences can raise skepticism as to the validity of the MAR assumption made: assuming the specified model is 'correct' and contains all variables explaining the missingness mechanism (covariate dependent missingness), the point estimates obtained with CCA and MI should generally be in alignment. However, note that when missingness depends on the observed outcome (often unlikely in survival data), MI should theoretically outperform CCA, while the reverse holds true in the MNAR case where the missingness depends on the variable itself (Carpenter and Smuk, 2021). Of course in reality, missingness will often be a mixture of MAR and MNAR, and results should therefore be interpreted with care.

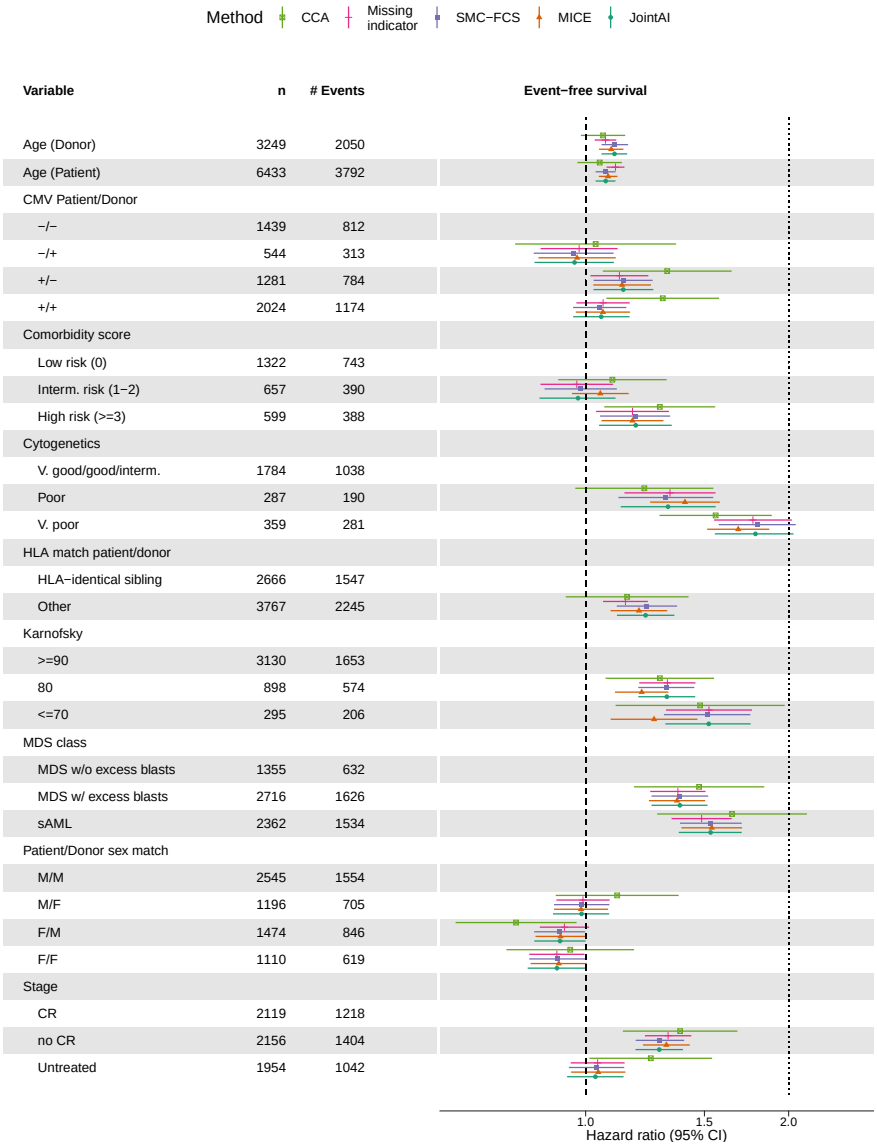

Figure 2.1: Point estimates and associated 95% confidence intervals for the Cox model for relapse-free survival, according to missing data handling method. Variables and their descriptions can be found in Table 2.2. Per level of factor and for continuous variables, we show the observed counts (n) and the number of events (# Events, which is the sum of relapse and non-relapse mortality events) in the full dataset.

Moreover, CCA point estimates will be more uncertain compared to their MI based counterparts, and one should therefore not expect to them to be equal (White *et al.*, 2022). One could indeed make the case that in the present example, given that the point estimates from the imputation methods fall within the confidence intervals of the CCA estimates (that is, within two standard errors), we should not be concerned.

Concerning the imputation methods, SMC-FCS and JointAI were in consistent agreement across all coefficients. The MICE approach differed noticeably from the previous two approaches for both the KPS and cytogenetic risk coefficients, which can perhaps be attributed to its theoretical limitations in the survival context: it does not ensure full compatibility between analysis and imputation model, in contrast to the other two MI approaches. The MIM was also consistently in agreement with estimates obtained with both SMC-FCS and JointAI. Given that the assumptions in Blake *et al.* (2020) concern missing data in adjustment variables (i.e. not in the outcome or in the exposure/treatment variable), it is difficult to reason about the validity of the MIM in the current context where the multivariable model is presented as prognostic.

2.6 Discussion and recommendations

Our systematic review demonstrated that missing data feature prominently across studies from major journals in the field of clinical haematology. Studies are often observational in nature, sometimes making use of large registries where whether or not a variable is recorded can depend on a multitude of reasons (e.g. variables collected only from a certain date, in particular centers). While presence of missing values was seemingly consistently reported (usually in the descriptives table), the method used for dealing with them was not. CCA was the dominant approach to handling missing values, and the associated loss of efficiency was generally poorly documented. Importantly, we also note that articles with seemingly complete data (particularly in observational studies) may have for simplicity chosen to filter out the missing values at a pre-processing step, without reporting it in the main manuscript—implying that the true prevalence of missing data is likely higher than reported in our review.

Various articles have attempted to raise the bar when it comes to reporting and handling of missing values more generally across observational studies. A convenient summary of guidelines found in Sterne *et al.* (2009), Sterne *et al.* (2016), and Vandenbroucke *et al.* (2007), can be found in Table 1 from Carroll *et al.* (2020), and should be considered alongside further guidelines given by Lee *et al.* (2021). The aforementioned guidelines put findings concerning missing values reporting in our systematic review into perspective: while the reporting may have been consistent across articles, it was by no means thorough. Paz *et al.* (2021) were perhaps the only standout article from the corpus in this respect. In their supplemental data, they provided a plot showing the missing data patterns (i.e. not only frequencies of missing values per variable,

but also frequencies for two or more variables being simultaneously missing), and presented a table showing the distribution of variables among complete cases versus cases with at least one missing value.

The review also showed that while multiple imputation remains an active topic of methodological research, its application in clinical haematological studies appears to be very limited. This does seem to contrast with findings regarding the increasing use of MI in the *Lancet* and *New England Journal of Medicine* over a period of 5 years (Hayati Rezvan *et al.*, 2015).

The systematic review had several strengths. The articles included covered the main haematological malignancies, as well as a (impact factor based) large range of journals. In order to obtain as representative a corpus as possible, we made sure to keep our initial criteria rather broad: filtering by journals, and by words (in abstract, title or keywords) directly referring to particular malignancies or multivariable Cox models. Importantly, we made sure not to include words such as imputation, missing or incomplete data in this initial search, as that could severely bias the sample towards articles which were exceptionally diligent with their missing data reporting.

One of the limitations of the present work is that the articles spanned a single year, and thereby may not be fully representative of trends in the literature. Furthermore, the extraction was limited in scope: we did not record many factors that may be of interest to stratify the results by, such as study type or goals (e.g. prediction), as was thoroughly done for example in Carroll *et al.* (2020). We effectively accepted the trade-off of having a larger sample of articles, in exchange for limited granularity on the information extraction.

As mentioned previously, extensive guidelines concerning the reporting and handling of missing values already exist, and are clearly not being adhered to. Among the few papers using multiple imputation, very little is reported except the number of imputed datasets: whether it be the number of iterations, the contents of the imputation model, or the discussion on the plausibility of the MAR assumption. We argue that at bare minimum, authors must (in the main article) a) state whether there were missing values, and if so, across which variables and in what frequency; b) make explicit whether and how their initial study population choice has been influenced by missing values; c) explicitly state the method used for handling the missing values. In the case of a CCA, one should then clearly report the sample size per analysis model. Given tight word count limitations for main text of many of the journals, authors should be encouraged to make use of supplemental materials to report additional information concerning the missing values. This may include discussion concerning the reasons why values were missing, comparison of outcomes between those with and without missing values, and full details behind the imputation procedure if an imputation method was used.

The choice of method for handling missing values in covariates will inevitably be highly context-dependent. For example, while the MIM is generally discouraged for observational studies (Groenwold *et al.*, 2012), its use has been approved for randomised trials (Sullivan *et al.*, 2018; White and Thompson, 2005). For observational studies, the flowchart shown in Figure 3 in Lee *et al.* (2021) provides a solid guide to motivating the choice of method for handling missing data. We would add that, if the MAR assumption is deemed plausible, results from a MI procedure (ideally SMC-FCS or fully Bayesian imputation, at least in the survival analysis context) should always be presented alongside the results of a CCA. Note that in cases where there are relatively few missing values, the imputation step could be skipped entirely: a CCA will likely be efficient enough, and will be unbiased under (covariate-dependent) MAR. In other cases, particularly when (auxiliary) variables related to either the variables with missing values or the mechanism are at the researcher's disposal, MI should be considered in order to mitigate a potentially important loss of statistical power. Nevertheless, the proportion of incomplete cases should not be the main criterion in deciding which method to use: when the imputation model is well specified and data are MAR, MI can yield unbiased results even with a large proportion of incomplete cases (Madley-Dowd *et al.*, 2019). Conversely, when either condition is not met (well specified imputation model and MAR), bias in the results will likely increase with the proportion of incomplete cases. Additionally, we caution that there is a lack of research concerning the use of MI for missing covariates in the presence of multiple outcomes. This is clearly relevant to articles in clinical haematology, which often present models for outcomes such as overall survival, relapse of a disease, and occurrence of graft-versus-host disease (GvHD) as part of a single study.

In conclusion, missing values are a prominent issue across studies in clinical haematology, and increased attention should be given in particular to the methods used to handle them, and the assumptions they make. In particular, we hope to stimulate a more open discussion about missingness mechanisms (and their implications on the validity of analyses), which can encourage better and more complete data collection in future studies. Nevertheless, researchers should remain prepared to discuss how the missing values in their particular context could affect the results of their study, both in terms of bias and statistical power.

Supplementary materials

The systematic review extraction sheet and associated reference list are available in the online supplement at https://doi.org/10.1016/j.beha.2023.101477.

Chapter 3

Multiple imputation for cause-specific Cox models: Assessing methods for estimation and prediction

Chapter based on: **Bonneville, E. F.**, Resche-Rigon, M., Schetelig, J., et al. (2022) Multiple imputation for cause-specific Cox models: Assessing methods for estimation and prediction. *Statistical Methods in Medical Research*, 31, 1860–1880. DOI: 10.1177/09622802221102623.

Abstract

In studies analysing competing time-to-event outcomes, interest often lies in both estimating the effects of baseline covariates on the cause-specific hazards, and predicting cumulative incidence functions. When missing values occur in these baseline covariates, they may be discarded as part of a complete case analysis (CCA) or multiply imputed. In the latter case, the imputations may be performed either compatibly with a substantive model pre-specified as a cause-specific Cox model (SMC-FCS), or approximately so (MICE). In a large simulation study, we assessed the performance of these three different methods in terms of estimating cause-specific regression coefficients and predicting cumulative incidence functions. Concerning regression coefficients, results provide further support for use of SMC-FCS over MICE, particularly when covariate effects are large and the baseline hazards of the competing events are substantially different. CCA also shows adequate performance in settings where missingness is not outcome-dependent. With regard to cumulative incidence prediction, SMC-FCS and MICE performed more similarly, as also evidenced in the illustrative analysis of competing outcomes following a hematopoietic stem cell transplantation. The findings are discussed alongside recommendations for practising statisticians.

3.1 Introduction

Missing covariate data are of perennial concern in observational studies in medicine (Carroll *et al.*, 2020). The backbone of such studies are clinical registries, which collect patient data potentially spanning many countries and centres over long periods of time. These and other data management complexities can lead to various patterns of (possibly informative) missingness. Furthermore, these registries are often set up for multiple purposes leading to multiple studies where different potentially exclusive survival outcomes could be considered. Consequently, *competing risks* outcomes are frequently investigated. This refers to a setting in which individuals can only experience one of several mutually exclusive events.

In studies considering competing risks outcomes, interest can lie in both the probabilities of events occurring over time and the effect of covariates on the different competing events. Appropriate handling of missing data is then of central concern in view of avoiding potential bias and/or loss of power when estimating these quantities, as could be expected when using simple methods such as complete-case analysis (CCA) (White and Carlin, 2010).

A more principled approach to handling missing covariate data is to use multiple imputation (MI), where a set of complete datasets is generated using samples based on an imputation model to fill in the missing values (Murray, 2018). A substantive model is then run on each of these datasets, before combining the estimates using rules that adequately reflect the uncertainty in the imputation procedure (Rubin, 1987). The imputation model and the substantive model should ideally be compatible, that is, deriving from a joint model under which both models are conditionals. If data are missing across multiple covariates, the fully conditional specification approach can be used (van Buuren *et al.*, 2006). This involves specifying an imputation model for each variable with missing values, fully conditional on the other variables, including the outcome. The procedure is better known under its more popular name 'multivariate imputation by chained equations' (MICE) (van Buuren *et al.*, 1999).

In time-to-event analysis, a popular choice of substantive model is the Cox proportional hazards model. White and Royston (2009) showed that when using MICE in the context of a Cox model (in absence of competing events), for each covariate with missing data, the corresponding imputation model should include the remaining covariates, the event indicator, and the cumulative baseline hazard. To implement this model, the cumulative baseline hazard can be approximated by the marginal Nelson–Aalen estimate of the cumulative hazard. Moreover, depending on the type of covariate, the imputation model is simplified with a Taylor approximation for the non-linear terms from the Cox likelihood. In view of this approximate compatibility between the substantive and imputation model, Bartlett *et al.* (2015) proposed a variant of MICE called 'substantive-model-compatible fully conditional specification' (SMC-FCS). The

approach ensures full compatibility between the imputation model and the substantive model by imputing missing covariate values in a rejection sampling procedure.

In competing risks settings, where the analysis model of interest is often a *cause-specific* Cox proportional hazards model, there has been little research addressing the appropriate use of MI when imputing missing covariate data (Lau and Lesko, 2018). The most prominent work is that of Bartlett and Taylor (2016), where the SMC-FCS approach was extended for cause-specific Cox models. In a simulation study as part of their work, Bartlett and Taylor compared SMC-FCS to an approximate MICE procedure proposed by Resche-Rigon *et al.* (2012). The proposal was an extension of the work of White and Royston for cause-specific Cox models. Simulation results suggested using SMC-FCS generally leads to estimates with little bias and nominal coverage (Bartlett and Taylor, 2016). In contrast, the approximate MICE approach was often biased, with some mitigation using interaction terms in the imputation model.

Importantly, we remark that the algebraic motivation behind the approximate MICE approach is currently unpublished. Moreover, the work of Bartlett and Taylor is to our knowledge the only empirical comparison of this approximate MICE approach with the SMC-FCS approach. Thus, questions regarding performance of both methods in a wider range of situations still remain. In addition, the question of how both the approaches perform with regard to predicted cumulative incidence functions is hitherto unexplored.

The aim of the present research is thus threefold. First, we aim to formally extend the work of White and Royston for cause-specific Cox models. Specifically, we will derive the approximately compatible imputation models for continuous, binary and multi-level categorical missing covariates. This extension was originally initiated by one of the authors of the current manuscript and shared as part of an oral presentation (Resche-Rigon *et al.*, 2012). Second, we aim to replicate and extend the simulations of Bartlett and Taylor; additionally manipulating the shape of the competing baseline hazards and the strength of missingness mechanisms, among other extensions. Third, we will explore how biases in cause-specific Cox models affect predicted cumulative incidence functions for patterns of reference covariate values. Simulation results will be interpreted alongside an illustrative analysis using a dataset from the field of allogeneic hematopoietic stem cell transplantation (alloHCT).

In Section 3.2, we present the motivating dataset, and in Section 3.3 we introduce notation for cause-specific competing risks analysis. In Section 3.4, the algebraic motivation behind the imputation model for a cause-specific Cox analysis model is shown. The simulation study is presented in Section 3.5, followed by an illustrative analysis in Section 3.6. Findings are discussed alongside recommendations for practice in Section 3.7.

3.2 Motivating example

Schetelig *et al.* (2019) assessed long-term outcomes of patients with myelodysplastic syndromes (MDS) or secondary acute myeloid leukemia (sAML) after an alloHCT. MDS is characterised by the production of deficient clonal blood cells in the bone marrow and can rapidly progress to more severe sAML (Adès *et al.*, 2014). AlloHCT is the only treatment that can offer long-term remission of the disease. Therefore, alloHCT is recommended for disease stages at high risk of transformation into AML or death from other complications. However, this procedure is associated with a high risk of adverse outcomes, either due to relapse of MDS or sAML, or due to side effects of the (pre-)treatment. This leads to the competing risks outcomes relapse and non-relapse mortality.

The dataset contains 6434 patients transplanted between 2000-2012, and registered with the EBMT. Several possible predictors measured at time of transplantation have a substantial amount of missing values. Some examples of variables with missing values are cytogenetic classification (62.2% missing), comorbidity index (59.9% missing) and the Karnofsky performance score (32.8% missing). A cause-specific model for relapse with the aforementioned three variables as predictors, performed on complete cases only, makes use of a mere 20% of the full dataset. The immediate lack of efficiency here prompted an investigation as to the performance of MI for such examples.

3.3 Cause-specific competing risks analysis

In a competing risks setting, we assume that individuals can 'fail' from only one of K distinct events. We denote that failure time as $\tilde{T}$, and the competing event indicator as $\tilde{D} \in \{1, ..., K\}$. In practice, individuals are subject to some right-censoring time C, which is assumed to be independent of $\tilde{T}$ and $\tilde{D}$, possibly given covariates. We thus only observe realisations (t_i, d_i) of $T = \min(C, \tilde{T})$ and $D = I(\tilde{T} \leq C)\tilde{D}$, where $D = 0$ indicates a right-censored observation.

If we view competing risks as a multi-state process, with a single (event-free) initial state and K absorbing states, interest often lies in the cause-specific hazard, defined for a single event k as

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq \tilde{T} < t + \Delta t, \tilde{D} = k \mid \tilde{T} \geq t)}{\Delta t}.$$

This hazard function can be interpreted as the instantaneous force of transition, or intensity, of moving between the initial state and state k (Beyersmann *et al.*, 2012;

Putter *et al.*, 2007). A model can then be specified, conditional on a covariate vector $\mathbf{Z}$. A Cox model is the common choice, defined for a failure cause k as

$$h_k(t \mid \mathbf{Z}) = h_{k0}(t) \exp(\boldsymbol{\beta}_k^\top \mathbf{Z}),$$

where $h_{k0}(t)$ is the cause-specific baseline hazard, and $\boldsymbol{\beta}_k$ represents the effects of covariates $\mathbf{Z}$ on the cause-specific hazard. We note that in what follows, we use 'effect' to refer to the impact of a covariate in a multivariable model where there may be non-negligible additional confounding, and this should hence not be interpreted as a fully causal quantity. Furthermore, the K hazard functions define the failure-free survival probability:

$$S(t \mid \mathbf{Z}) = \exp\left(-\sum_{k=1}^{K} \int_0^t h_k(u \mid \mathbf{Z}) \, \mathrm{d}u \right) = \exp\left(-\sum_{k=1}^{K} H_k(t \mid \mathbf{Z}) \right),$$

where $H_k(t \mid \mathbf{Z}) = \int_0^t h_k(u \mid \mathbf{Z}) \, \mathrm{d}u$ is the cause-specific cumulative hazard for cause k. Assuming conditional non-informative censoring, the likelihood contribution of an individual with observations $(t_i, d_i, \mathbf{z}_i)$ is then

$$p(t_i, d_i \mid \mathbf{z}_i) = S(t_i \mid \mathbf{z}_i) \prod_{k=1}^{K} \left[h_k(t_i \mid \mathbf{z}_i) \right]^{I(d_i = k)}, \tag{3.1}$$

where $I(\cdot)$ is the indicator function. The covariate effects $\boldsymbol{\beta}_k$ on the cause-specific hazard can then be estimated by optimising the partial likelihood (Cox, 1975). This follows from the observation that the above expression factorises into separate factors for each cause k, which each corresponding to a standard Cox likelihood function where events from all other causes are treated as censored observations (Prentice *et al.*, 1978).

3.3.1 Cumulative incidence functions

Beyond assessing covariates, cause-specific hazards can also be used to estimate the so-called cumulative incidence functions, defined as

$$P(\tilde{T} \leq t, \tilde{D} = k) = \int_0^t h_k(u) S(u-) \, \mathrm{d}u, \;\; k = 1, ..., K, \tag{3.2}$$

where $S(u-)$ is the failure-free survival probability just prior to u (Andersen *et al.*, 2002). This cumulative incidence function, or transition probability, is the probability of experiencing event k before or at time t. It is also known as the absolute, or crude risk. It can be computed either non-parametrically, or semi-parametrically if Cox models are specified for the $h_k(u)$. In the latter case, the cumulative hazards derived

from the Breslow estimator of the cumulative cause-specific baseline hazards are used as ingredients for estimating the cumulative incidence for cause k.

This implies that we do not need to model the cumulative incidence function *directly* in order to obtain these predicted probabilities, as is done when using the Fine–Gray model (Fine and Gray, 1999). This is helpful given that in observational studies, interest is seldom in prediction alone: predictions are often presented after first reporting and interpreting model coefficients. The cause-specific hazards framework provides a more natural scale on which to interpret covariate effects, and allows to obtain predicted patient-specific cumulative incidence functions for all causes.

3.4 Methods

In this section, we provide a framework for using MICE and SMC-FCS for both estimation of cause-specific regression coefficients, and cumulative incidence functions. Throughout, we assume that data are missing according to a missing (completely) at random mechanism, hereafter abbreviated as M(C)AR.

3.4.1 Fully conditional approach (MICE)

We introduce X as a single, partially observed covariate, and Z as a fully observed covariate. We note that Z could also represent a vector of complete covariates. Appropriate use of MICE for cause-specific competing risks analysis requires the specification of an *imputation model* $p(X \mid T, D, Z)$, from which a number of imputed datasets are generated. Detailed derivations for $p(X \mid T, D, Z)$ are provided in Appendix A, which we summarise in the present subsection.

To begin with, we note that by Bayes' Theorem,

$$\log p(X \mid T, D, Z) = \log p(T, D \mid X, Z) + \log p(X \mid Z) + c, \tag{3.3}$$

where c is a constant term that does not depend on X. For $p(T, D \mid X, Z)$, a cause-specific Cox proportional hazards model for each failure cause k is specified as $h_k(t \mid X, Z) = h_{k0}(t)\exp(\beta_k X + \gamma_k Z)$. In case of binary or continuous X and Z, β_k and γ_k are scalars; for categorical X or Z with two or more levels, β_k and γ_k are vectors and X and Z represent dummy codings for the levels of the covariates. To impute from the fully conditional distribution in Equation 3.3, we also need to specify a model for the missing data, $p(X \mid Z)$. This model will generally vary depending on the covariate type of X.

3.4.1.1 Binary X

If X is binary, we could assume $\text{logit}\, P(X = 1 \mid Z) = \zeta_0 + \zeta_1 Z$. If Z is categorical with $J \geq 2$ levels (without loss of generality assuming that Z takes values in $1, \ldots, J$), we can write

$$\begin{aligned} \text{logit}\, P(X = 1 \mid T, D, Z) = \alpha_0 &+ \sum_{k=1}^{K} \alpha_k I(D = k) + \sum_{k=1}^{K} \alpha_{K+k} H_{k0}(T) \\ &+ \sum_{j=1}^{J-1} \alpha_{2K+j} I(Z = j) \\ &+ \sum_{j=1}^{J-1} \sum_{k=1}^{K} \alpha_{(j+1)K+(J-1)+k} I(Z = j) H_{k0}(T), \end{aligned} \tag{3.4}$$

which implies that for categorical Z we can impute missing X values using a logistic regression with D (as a factor variable), the cumulative baseline hazards for all causes of failure, Z (as a factor variable), and the complete interactions between the cumulative baseline hazards and Z. For continuous Z, results are no longer exact. Using a first-order Taylor approximation for the $\exp(\gamma_k Z)$ term, we can write

$$\begin{aligned} \text{logit}\, P(X = 1 \mid T, D, Z) \approx \alpha_0 &+ \sum_{k=1}^{K} \alpha_k I(D = k) + \sum_{k=1}^{K} \alpha_{K+k} H_{k0}(T) \\ &+ \sum_{k=1}^{K} \alpha_{2K+k} H_{k0}(T) Z + \alpha_{3K+1} Z, \end{aligned} \tag{3.5}$$

which is valid if $\text{Var}(\gamma_k Z)$ is small. This approximate imputation model thus uses D, Z, all $H_{k0}(T)$ and the interactions between all $H_{k0}(T)$ and Z as predictors in a logistic regression. Note that the α parameters used above and in the next subsections represent the imputation model coefficients, and are themselves functions of other (substantive and missing data model) parameters. Therefore, these will vary depending on the covariate types of X and Z, and the parametrisation of the substantive model (i.e. whether each cause-specific model has the same predictors, and their functional forms).

3.4.1.2 Nominal categorical X

If X is a categorical covariate with $J \geq 2$ levels and $j = \{0, ..., J-1\}$, we can specify different imputation models depending on whether X is ordered or not. In the unordered (nominal) case, we can specify a multinomial logistic regression for $p(X \mid Z)$,

yielding

$$\log \frac{P(X=j\,|\,T,D,Z)}{P(X=0\,|\,T,D,Z)} \approx \alpha_{j,0} + \sum_{k=1}^{K} \alpha_{j,k} I(D=k) + \sum_{k=1}^{K} \alpha_{j,K+k} H_{k0}(T) + \sum_{k=1}^{K} \alpha_{j,2K+k} H_{k0}(T) Z + \alpha_{j,3K+1} Z.$$

This comes as a result of generalising $\text{logit}\, P(X=1|Z) = \zeta_0 + \zeta_1 Z$ to $\log \frac{P(X=j|Z)}{P(X=0|Z)} = \zeta_0 + \zeta_j Z$, and holds for continuous Z as in Equation 3.5. For categorical or no Z, where for the former $I(Z=j)$ should be used as in Equation 3.4, the expression for the fully conditional distribution is exact as in the binary case. The predictors to be included in the imputation model are exactly the same as for binary X.

3.4.1.3 Ordered categorical X

For ordered categorical X, a proportional odds model could be assumed as $\text{logit}\, P(X \leq j\,|\,Z) = \zeta_j + \zeta_Z Z$. This however implies that the fully conditional distribution requires specifying $p(T, D\,|\,X \leq j, Z)$, which does not have a standard proportional hazards density. Instead, it has a *weighted sum* of proportional hazards densities. Thus, the expression for $P(X \leq j\,|\,T, D, Z)$ does not extend from the binary case in any simple form. Nevertheless, a proportional odds model including D, Z and all $H_{k0}(T)$ could still be used to impute the missing X values, though the properties of such a model are not currently well known. We refer the reader to the book written by McCullagh and Nelder (1989) for a detailed description of both the multinomial logistic regression and proportional odds models.

3.4.1.4 Continuous X

If X is a continuous covariate, we could assume it to be normal conditional on Z (possibly after transformation), as $X\,|\,Z \sim \mathcal{N}(\zeta_0 + \zeta_1 Z, \sigma^2)$. The implied expression for $p(X\,|\,T, D, Z)$ is not normal due to the $\exp(\beta_k X + \gamma_k Z)$ term, and so a bivariate Taylor approximation is used around the sample means $\bar{X}$ and $\bar{Z}$. To the first degree, the approximate fully conditional distribution is expressed as

$$X\,|\,T, D, Z \sim \mathcal{N}(\alpha_0 + \alpha_1 Z + \sum_{k=1}^{K} \alpha_{k+1} I(D=k) + \sum_{k=1}^{K} \alpha_{K+k+1} H_{k0}(T), \sigma^2).$$

This suggests a model for imputing continuous X should be a linear regression with D, Z and all $H_{k0}(T)$ again as predictors. With a quadratic approximation for $\exp(\beta_k X + \gamma_k Z)$, the accuracy of the above model can be improved by additionally including

the interactions between all $H_{k0}(T)$ and Z. The approximations are valid under the assumption of small $\mathrm{Var}(\beta_k X + \gamma_k Z)$.

We note that the above models, like in the simple time-to-event settings, cannot be implemented without a working estimate of $H_{k0}(T)$, whose true values we will assume are unknown. For the competing risks setting, we can use the marginal Nelson–Aalen estimate of the cumulative cause-specific hazard (which requires treating all events other than k as censored) as an approximation for $H_{k0}(T)$. As explained by White and Royston (2009), this approximation becomes poorer with larger true covariate effects. We may then expect the estimated covariate effects after the imputation procedure to be biased.

3.4.2 Substantive-model-compatible approach

We refer the reader to the work of Bartlett *et al.* (2015) for a detailed introduction of the SMC-FCS method, and to the work of Bartlett and Taylor (2016) for its specific extension to cause-specific Cox proportional hazards models. Briefly, the SMC-FCS method (in the current setting) is based on application of Bayes' theorem,

$$p(X \mid T, D, Z) \propto p(T, D \mid X, Z) p(X \mid Z), \tag{3.6}$$

which was already introduced on the logarithmic scale in Equation 3.3. The parameters associated with both $p(T, D \mid X, Z)$ and $p(X \mid Z)$ are omitted for readability. In essence, the procedure involves choosing $p(X \mid Z)$ as a proposal density and using rejection sampling to draw possible values for missing X from a density proportional to $p(T, D \mid X, Z) p(X \mid Z)$. This is under the assumption that $p(X \mid Z)$ is simple to sample from, as is the case if we specify a model for it, e.g. a linear regression of X conditional on Z. The imputation model is then compatible with the substantive model in the sense that a joint distribution exists which contains both the substantive model and the imputation model as its conditional distributions. If multiple covariates have missing data, it is still possible to specify mutually incompatible models for $p(X \mid Z)$, but each fully conditional distribution will be compatible with the substantive model.

In contrast to MICE, the SMC-FCS approach does not require any approximations—neither for the non-linear terms, nor for the cumulative baseline hazard. Of course, the cumulative baseline hazard still needs to be evaluated in order to draw from Equation 3.6. In order to do so, the Breslow estimate is used, and is updated at each iteration of the imputation procedure conditional on the most recent draws from the posterior distribution of the regression coefficients.

3.4.3 Regression coefficients

Both the MICE and SMC-FCS procedures result in $m = 1, ..., M$ imputed datasets. In each of these datasets, the cause-specific Cox model for one or more of the K causes of failure is fitted. Let θ denote a cause-specific regression coefficient of interest, and let $\hat{\theta}_m$ and $\widehat{\mathrm{Var}}(\hat{\theta}_m)$ respectively denote the estimate and associated variance of this coefficient in the m^{th} imputed dataset. We can combine these M estimates using Rubin's rules, with estimator

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m.$$

The associated variance estimator is

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{Var}}(\hat{\theta}_m) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta})^2,$$

which combines estimates of within and between imputation variance (Rubin, 1987). The estimate of the standard error is then readily obtained as $\widehat{\mathrm{SE}}(\hat{\theta}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}$.

3.4.4 Predicted probabilities

To obtain the predicted cumulative incidence functions for an individual with fully observed covariates after an MI procedure, there are at least two possible options. The first is to pool the regression coefficients and baseline hazards separately, and use those to produce a single predicted curve. The second approach is to use the substantive models fitted in each imputed dataset to create *imputation-specific* predictions, and then pool those (possibly after transformation) using Rubin's rules. The articles by Wood *et al.* (2015) and Mertens *et al.* (2020) recommend the second approach, which is the one we employ in the present paper.

3.5 Simulation study

We designed a simulation study with the aim of comparing the performance of CCA, MICE and SMC-FCS in the presence of missing baseline covariate values for cause-specific Cox proportional hazards models with two competing events. We assessed performance with respect to estimated regression coefficients, and predicted cumulative incidence functions.

3.5.1 Data-generating mechanisms

We generated datasets containing $n = 2000$ individuals, with one record each containing both predictor and outcome information.

3.5.1.1 Covariates

Two covariates X and Z were generated in each dataset. We varied the covariate type of X as either continuous or binary, and Z was fixed as continuous. When both covariates were continuous, they were generated from a bivariate standard normal distribution $X, Z \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with means $\boldsymbol{\mu} = \{0, 0\}$, variances $\text{diag}(\Sigma) = \{1, 1\}$ and correlation $\rho = 0.5$.

When X was binary, we assumed $X \sim \text{Bern}(0.5)$ and $Z \sim N(0, 1)$, with a *point-biserial* correlation between the two variables of $\rho = 0.5$. We can generate observations in this way by first generating X' and Z from a bivariate standard normal distribution with correlation $\rho' \approx 0.63$, and then dichotomising X' at 0 (the value of the standard normal quantile function for a probability of 0.5) to produce X. We refer the reader to the work of Demirtas and Hedeker (2016) for a description of this well-established procedure.

3.5.1.2 Competing event times

We based our simulation of event times on the motivating alloHCT example described in Section 3.2, focusing on the two competing events relapse (REL) and non-relapse mortality (NRM) over a 10-year follow-up period. To generate the failure times for the competing events, we made use of latent failure times, denoted T_1 and T_2 for REL and NRM respectively (Beyersmann *et al.*, 2009).

Typically in alloHCT studies, patients are at very high risk of both REL and NRM in the initial period after alloHCT, with this risk gradually decreasing thereafter as they survive longer. For this reason, generating failure times from a distribution with a decreasing hazard function is appropriate. The Weibull distribution, with probability density function $f(t) = \kappa\lambda t^{\kappa-1}\exp(-\lambda t^{\kappa})$, with shape $\kappa > 0$ and rate $\lambda > 0$, accommodates decreasing hazards for $\kappa < 1$. This is the parametrisation used in the text by Klein and Moeschberger (2003).

We thus generated both latent failure times from independent Weibull distributions, assuming cause-specific proportional hazards conditional on X and Z. We furthermore generated independent censoring times from an Exponential distribution. In

summary:

$$\begin{aligned}\tilde{T}_1 &\sim \text{Weibull}(\kappa_1, \lambda_1 = \lambda_{10} e^{\beta_1 X + \gamma_1 Z}),\\ \tilde{T}_2 &\sim \text{Weibull}(\kappa_2, \lambda_2 = \lambda_{20} e^{\beta_2 X + \gamma_2 Z}),\\ C &\sim \text{Exp}(\lambda_C),\end{aligned}$$

where λ_C is the censoring rate, and λ_{10} and λ_{20} are the baseline hazard rates for REL and NRM respectively. We then defined $\tilde{T} = \min(\tilde{T}_1, \tilde{T}_2)$, with an associated factor variable $\tilde{D}$, where $\tilde{D} = 1$ if REL occured first, and $\tilde{D} = 2$ otherwise. The generated observed (event or censoring) time was then defined as $T = \min(C, \tilde{T})$, with corresponding indicator $D = I(\tilde{T} \leq C)\tilde{D}$.

We used estimates from cause-specific marginal accelerated failure time (AFT) models on the motivating dataset to fix the parameters values of the baseline shape and hazard rates for the latent failure times. Weibull AFT models for both causes of failure led to fixing $\kappa_1 = 0.58$, $\lambda_{10} = 0.19$, $\kappa_2 = 0.53$, and $\lambda_{20} = 0.21$. An exponential AFT model for the censoring distribution motivated setting $\lambda_C = 0.14$. Since the baseline hazards for both competing events were estimated to be very similar, we decide to also vary $\{\kappa_1, \lambda_{10}\} = \{1.5, 0.04\}$, such that REL had a steadily increasing hazard. Both these 'similar' and 'different' baseline hazard configurations lead to comparable marginal 10-year cumulative incidences of both events, in the 35–45% range. Regarding cause-specific regression coefficients, we varied $\beta_1 = \{0, 0.5, 1\}$, and fixed $\gamma_1 = 1$, $\beta_2 = 0.5$ and $\gamma_2 = 0.5$.

3.5.1.3 Missing data mechanisms

Z was conserved as a complete covariate, and missingness was induced in X. Let R_X indicate whether elements of X were missing ($R_X = 0$) or observed ($R_X = 1$). We varied the proportion of missing values as either 'low' with 10% missing, or 'high' with 50%. We defined four separate missingness mechanisms:

1. Missing completely at random (MCAR), defined as $P(R_X = 0) = 0.5$ or $P(R_X = 0) = 0.1$.
2. Missing at random conditional on Z (MAR), which was defined as $\text{logit}\, P(R_X = 0 \mid Z) = \eta_0 + \eta_1 Z$.
3. Outcome-dependent missing at random (MAR-T), which was defined as $\text{logit}\, P(R_X = 0 \mid T_{\text{stand}}) = \eta_0 + \eta_1 T_{\text{stand}}$. T_{stand} is $\log T$, standardised to have zero mean and unit variance. Note that T was the observed (event or censoring) time; if missingness depended on the true event time, this would lead to a missing not at random mechanism.

4. Missing not at random conditional on X (MNAR), which was defined as $\text{logit}\, P(R_X = 0 \mid X) = \eta_0 + \eta_1 X$.

For mechanisms 2–4, η_1 represented the strength and direction of the missingness mechanism. For example, if $\eta_1 < 0$ in the MAR mechanism, observations with smaller values of the Z had a larger probability (increasing with more extreme η_1) of the corresponding X being missing. In the present study, we varied $\eta_1 = \{-1, -2\}$, representing 'weak' and 'strong' mechanisms respectively. In this context, the MAR-T mechanism could reflect a measurement that is only collected if a subject survives long enough into a study and is in follow-up, as may be the case with a genetic test. Although this kind of measurement is collected or only available at a later point in time, it can still be considered as baseline information and does *not* constitute conditioning on the future.

The value of η_0 was chosen (in each simulated dataset) such that the average missingness probability was equal to either 0.5 or 0.1. This was done via standard root-solving for a fixed value of η_1.

3.5.1.4 Design

The simulation study is chosen to follow a partially factorial design, where the parameters outlined above are varied systematically. A full factorial design would result in 4 (missingness mechanisms) × 2 (mechanism strengths) × 2 (proportions missing data) × 2 (covariate types for X) × 2 (baseline hazard parametrisations) × 2 (effects magnitudes of X on cause-specific hazard of REL) = 128 scenarios. However, the strength of the missingness mechanism cannot be varied for MCAR settings by definition, leaving 112 scenarios in total.

3.5.2 Estimands

The analysis models of interest are the cause-specific Cox proportional hazards models for REL and NRM, $h_k(t \mid X, Z) = h_{k0}(t) \exp(\beta_k X + \gamma_k Z)$ for $k = \{1, 2\}$. We then have two main sets of estimands of interest:

- $\theta_{\text{regr}} = \{\beta_1, \gamma_1, \beta_2, \gamma_2\}$, which are the data-generating regression coefficients from both cause-specific Cox models.
- θ_{pred}, which is a vector containing the REL and NRM probabilities (cumulative incidences) for a set of reference patients at 6 months, 5 years and 10 years after baseline.

These reference patients were defined by all combinations of $Z_{\text{ref}} = \{-1, 0, 1\}$ with $X_{\text{ref}} = \{-1, 0, 1\}$ for continuous X, and $X_{\text{ref}} = \{0, 1\}$ for binary X. Since the data-generating coefficients for both competing events had a positive effect on the cause-specific hazards, one could for example refer to $\{X_{\text{ref}}, Z_{\text{ref}}\} = \{1, 1\}$ as a 'high risk' individual, and 'low risk' for $\{-1, -1\}$.

3.5.3 Methods

Five missing data methods were compared in each simulation scenario:

- CC - an analysis run on a dataset after listwise deletion.
- CH_1 - MI with imputation model predictors including Z, the event indicator solely for event one i.e. $I(D = 1)$, and the cumulative hazard for REL $\hat{H}_1(T)$ (at the end of follow-up for each individual), based on the Nelson–Aalen estimator, as an approximation of the cumulative baseline hazard $H_{10}(T)$.
- CH_{12} - MI with imputation model predictors including Z, the event indicator D as a three level factor variable, and the cumulative hazards for both events $\hat{H}_1(T)$ and $\hat{H}_2(T)$; outlined in Section 3.4.1.
- $CH_{12,\text{Int}}$ - identical to the CH_{12}, with the addition of the interactions $\hat{H}_1(T) \times Z$ and $\hat{H}_2(T) \times Z$; outlined in Section 3.4.1.
- smcfcs - the approach outlined in Section 3.4.2}, using Z as sole predictor in the $X \mid Z$ model (default setting).

The CH_1 method corresponds to the 'FCS survival' method explored in the simulation study by Bartlett and Taylor, where failures other than cause one are treated as censored and the cumulative hazard of cause two is omitted from the imputation model. It corresponds to a direct application of the White and Royston (2009) results to the cause-specific Cox model for cause one, which may present itself as intuitive when interest lies in a single failure cause.

Additionally, the model was also fitted on the complete dataset prior to any missingness being induced in X. For the imputation methods, the number of imputed datasets was varied as $m = \{5, 10, 25, 50\}$. We set $\max(m) = 50$ since no substantial reduction in empirical standard errors was observed over trial runs with $m = 100$. We also note that for $m \neq 50$, the imputations were not re-run independently. Results were instead pooled across the first 5, 10 or 25 imputed datasets from the original 50.

When X was continuous, the imputation model was a linear regression. For binary X, the imputation model was a logistic regression. We note that since there was only one partially observed covariate, chained equations were not needed. Nevertheless, we

still refer to methods CH_1, CH_{12} and $CH_{12,\text{Int}}$ under the general umbrella term 'MICE' methods when reporting the results.

3.5.4 Performance measures

For θ_{regr}, we recorded the point estimates, empirical and estimated standard errors, absolute bias and coverage probabilities. As our primary measure of interest was bias, we based the number of simulation replications per scenario n_{sim} on a desired Monte Carlo standard error (MCSE) of bias. As per Morris *et al.* (2019) this is defined as $\text{MCSE(Bias)} = \sqrt{\text{Var}(\hat{\theta}_{\text{regr}})/n_{\text{sim}}}$. We assumed that $\text{SD}(\hat{\theta}_{\text{regr}}) \leq 0.125$ (largest empirical standard error to be expected with binary X, based on small trial run), and we deemed a $\text{MCSE(Bias)} \leq 0.01$ to be appropriate. We thus required $n_{\text{sim}} = 0.125^2/0.01^2 \approx 156$ replications per scenario, which we rounded up to $n_{\text{sim}} = 160$. We thus generated 160 independent datasets per simulation scenario.

For $\hat{\theta}_{\text{pred}}$, we recorded the point estimates, empirical standard errors, absolute bias, coverage probabilities and root mean square error (RMSE). We focus primarily on reporting bias and RMSE. Based on trial runs, we assumed $\text{SD}(\hat{\theta}_{\text{pred}}) \leq 0.05$, which for 160 replications would result in a $\text{MCSE(Bias)} \leq 0.05/\sqrt{160} \approx 0.004$. We thus proceeded with the same number of simulated datasets.

3.5.5 Software

All analyses were performed using R version 3.6.2 (R Core Team, 2020). The substantive model compatible imputation was performed using the smcfcs package version 1.4.1 (Bartlett *et al.*, 2022) and MICE was performed using the mice package version 3.8.0 (van Buuren and Groothuis-Oudshoorn, 2011). The cause-specific Cox models were run and subsequent predicted cumulative incidences were obtained using the mstate package version 0.2.12 (de Wreede *et al.*, 2011).

3.5.6 Results

We focus primarily on β_1 (the regression coefficient for X in the cause-specific REL model) and the 5-year probabilities of REL and NRM. For the imputation methods, we present results only with $m = 50$. Full results are reported in the supplementary materials, linked at the end of the present text.

3.5.6.1 Regression coefficients

Figure 3.1 summarises the results with regard to bias in the estimation of β_1 with a MAR mechanism induced on continuous X. The plot is a variant of a nested-loop plot, where each colour-cluster of points represents a scenario defined by the step functions at the bottom of the plot (Rücker and Schwarzer, 2014). For example, the left-most bin in the plot corresponds to a scenario with data-generating $\beta_1 = 0.5$, 10% missing data, similar hazard shapes and a weak missingness mechanism. For readability, the CH_1 method and the analysis ran on the full dataset prior to inducing missing data are omitted from the figure.

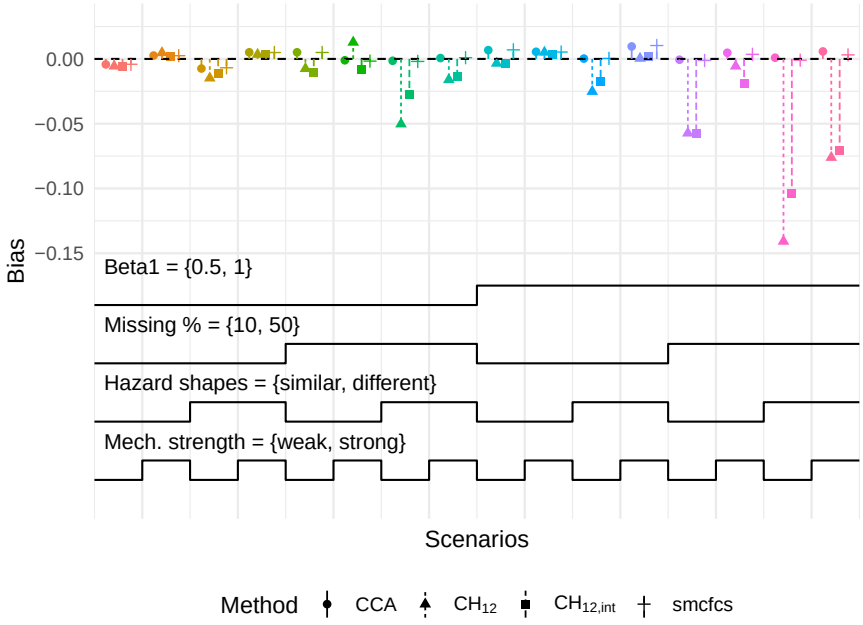

Figure 3.1: Bias in β_1 for MAR mechanism with continuous X. Each cluster of points corresponds to a scenario defined by the step functions at the bottom of the plot. Each step represents a level of a factor being varied, and is read from left-to-right (e.g. for Hazard shapes, first step is 'similar' while the second is 'different'). Monte Carlo standard errors of bias for all scenarios were below 0.008. Mech. = missingness mechanism.

First, we note that in the 16 scenarios depicted, both CC and smcfcs showed little to no bias in the estimation of β_1. For CC, no bias was expected given that this was a case of covariate-dependent MAR, and results for smcfcs were in line with the simulations

of Bartlett and Taylor (2016). Second, the MICE methods showed varying amounts of bias depending on the scenario. With increasing true covariate effects, and higher proportion of missing values, the bias was larger. This was to be expected in light of the approximations employed in Section 3.4.1, which are valid for small covariate effects. Moreover, the magnitude of the bias was also larger when the baseline hazard shapes were different. Last, adding the interaction terms in the imputation model did not significantly reduce bias, except when the missingness mechanism was weak, and the baseline hazard shapes were different.

For contrast, we also present the results for β_1 with a MAR-T mechanism in Figure 3.2, again with continuous X. In this case, CC was consistently biased, as is expected when missingness is dependent on the outcome. Particularly for a high proportion of missing values, the bias in both MICE methods was even more severe than that of CC, reaching close to 20% (relatively). Conversely, smcfcs was consistently unbiased across the depicted scenarios.

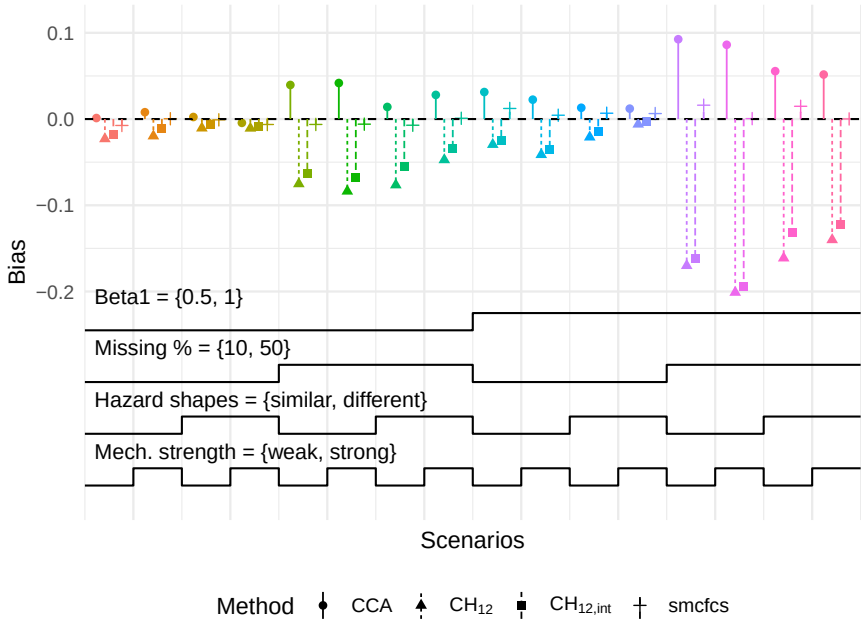

Figure 3.2: Bias in β_1 for MAR-T mechanism with continuous X. Monte Carlo standard errors of bias for all scenarios were below 0.008. Refer to Figure 3.1 for a description on how to read this type of plot. Mech.: missingness mechanism; MAR-T: outcome-dependent missing at random.

We also briefly summarise some of the more general findings across the simulations

reported in the supplementary material. First, efficiency gains (in the form of smaller estimated standard errors) were mainly observed for γ_1 and γ_2. Second, the CH_1 method yielded the largest biases, and lowest coverage probabilities of all methods. This was unsurprising, as CH_1 corresponded to imputing X as if competing outcomes were considered as censoring. Third, the findings with MCAR missingness were largely analogous to those of the MAR reported above; and in presence of MNAR, all imputation methods (including smcfcs) showed appreciable bias. Last, in scenarios with binary X, the overall bias in the MICE methods was lower with respect to scenarios with continuous X. This could be attributed to the different terms that are being approximated in the imputation models. In addition to the cumulative baseline hazards, only $\exp(\gamma_k Z)$ is being approximated in the case of binary X, whereas in the continuous case a fuller $\exp(\beta_k X + \gamma_k Z)$ is being approximated.

In terms of RMSE, which summarises both bias and variance, the differences in performance between the methods in M(C)AR scenarios was smaller, aside from when missingness was high and the baseline hazard shapes were different (see for example Figure 2.1.2 of supplementary material on regression coefficients).

3.5.6.2 Predicted probabilities

Concerning predicted probabilities, we focus on the estimation of 5-year REL and NRM probabilities for a 'low-risk' individual, i.e. $\{X, Z\} = \{-1, -1\}$ with continuous X. Figure 3.3 summarises the RMSE of these probabilities under a MAR mechanism where 50% of values are missing.

We point the reader to the y-axis of the plot, where results are now on the probability scale. The largest RMSE reported in the plot was just under 2.5%, with most RMSE values for the imputation methods being under 1.5%, with little to no difference between them. In these scenarios, the imputation methods outperform CC, but with the finest of margins. This is part of a general finding across the simulations: the predicted probabilities when using the imputation methods overall had very little bias, and little reduction in variability was observed beyond $m = 25$ imputations. We note that since all methods were similarly biased under M(C)AR (as seen for example in Figure 1.2.1 of supplementary material on predictions), the RMSE for CC is expected to be a factor of $\sqrt{2}$ larger than for the imputation methods when missingness was 'high', given that CC used half as much data.

We propose various explanations for this behaviour. First, we note that the prediction results for $\{X, Z\} = \{0, 0\}$ (with X continuous or binary) can be taken as a proxy for how precisely the cause-specific baseline hazards are estimated. For all non-MNAR scenarios, little to no bias was found in the predicted probabilities for these reference patients. This may be additionally linked to the fact that X and Z are centered and

normal, which could imply that $H_{k0}(T)$ is adequately approximated by the Nelson–Aalen estimator. Second, regarding regression coefficients, bias was primarily observed in β_1 and β_2, with the former showing more extreme bias when data-generating $\beta_1 = 1$. Estimates of γ_1 and γ_2 however generally only exhibited biases of up to 5% in the MAR scenarios, and slightly higher for CC in MAR-T scenarios. Well-estimated cause-specific baseline hazards in tandem with close to unbiased estimates of γ_k could then explain the small bias in the predictions, since bias in the linear predictor as a whole ($\beta_k X + \gamma_k Z$) only reached 10% in the most extreme cases, and was mostly below the 5% mark otherwise.

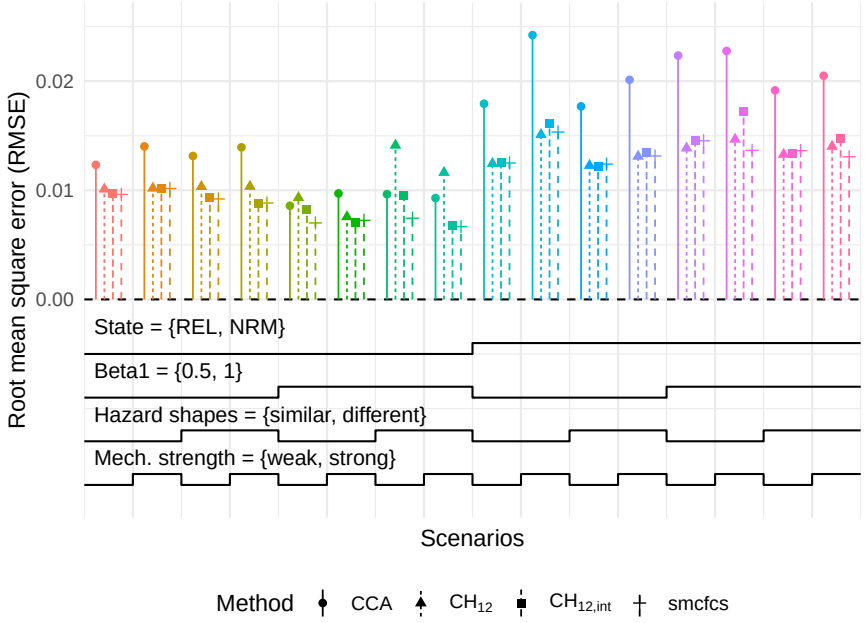

Figure 3.3: RMSE of 5-year REL and NRM probabilities with $\{X, Z\} = \{-1, -1\}$ for MAR with 50% missing values. Monte Carlo standard errors of RMSE for all scenarios were below 0.002. Refer to Figure 1 for a description on how to read this type of plot. Mech.: missingness mechanism; RMSE: root mean square error; REL: relapse; MAR: missing at random; NRM: non-relapse mortality.

3.5.6.3 Additional simulations

In supplementary material I (available online), we performed two additional simulation studies. The first investigated the use of the Breslow estimates of the cumulative baseline hazards in the imputation model, updated at each iteration of the imputation procedure. Consistent with earlier results in the standard survival setting, MICE using intra-iteration updates of the Breslow estimates performed no better than using the marginal cumulative hazards in the imputation model (White and Royston, 2009). The second study assessed the performance of the MI methods in the presence of $K = 3$ competing events. In this setting, SMC-FCS remained unbiased, while the MICE methods including additional interaction terms performed slightly better than those without.

3.6 Illustrative analysis

We used the motivating alloHCT dataset introduced in Section 3.2 to illustrate the methods described in the simulation study. Cause-specific Cox proportional hazards models were fitted for both REL and NRM, conditional on a set of baseline predictors chosen on the basis of substantive clinical knowledge. An overview of these predictors, including their names, descriptions and proportion of missing values, can be found in Table 2.2 from the previous chapter. The same predictors were used in the models for REL and NRM.

We used the CC, CH_{12} and smcfcs methods to handle the missing baseline covariate data, which we assumed to be missing at random. Given that $CH_{12,\text{Int}}$ did not show much improvement over CH_{12} in the simulation study, we decided to use the more parsimonious latter. Therefore, the imputation model for a partially observed covariate using CH_{12} contained as predictors the remaining fully and partially observed covariates from the substantive model, and the marginal cumulative hazards for both events. For smcfcs, the imputation model similarly contained the remaining fully and partially observed covariates from the substantive model, which is the default setting. Continuous covariates were imputed using linear regression, binary covariates using logistic regression, ordered categorical using proportional odds regression and nominal categorical using multinomial logistic regression. Since missingness spanned multiple covariates, chained equations were required.

To motivate the choice of m for CH_{12} and smcfcs, we used von Hippel's quadratic rule based on the fraction of missing information (FMI) rather than the proportion of complete cases (Madley-Dowd *et al.*, 2019; von Hippel, 2020). We first ran a set of $m = 20$ imputations, with $n_{\text{iter}} = 20$ iterations. After pooling, the coefficient with largest FMI was that of donor age in the model for NRM, with a value of approximately 0.49. Based on an 95% upper-bound for this FMI, and for a desired coefficient of

variation (CV) of 0.05, we would require approximately $m = 84$ imputed datasets. We rounded this upwards, and performed our final analysis with $m = 100$. We conserved $n_{\text{iter}} = 20$ as convergence was generally observed from 10 iterations onwards.

Figure 3.4 summarises the exponentiated point estimates (hazard ratios, HR) and associated 95% confidence intervals (CI) from the cause-specific model for REL. The CIs for CH_{12} and smcfcs are based on the pooled standard errors and the t-distribution. First, we observed a clear gain in efficiency across all coefficients for both imputation methods relative to CC. Second, there was general agreement between the estimates obtained from both CH_{12} and smcfcs; a finding which was also reported in the illustrative analysis in the work by Bartlett and Taylor (2016). Third, we did note some differences between CC and the imputation methods for certain variables, such as remission status or Karnofsky score. The most surprising case of this was with the MDS class of the patient, which was completely observed. In the model for REL, the HR for the sAML category estimated with CC is just above three, whereas the imputation methods estimate it much closer to two. This also raises the point that for categorical variables, differences in methods can be seen on the category level rather than on the variable level as a whole—as also evidenced by the estimated HRs for the cytogenetics variable. Results for the cause-specific NRM model are summarised in Figure 3.5.

Furthermore, we computed the predicted 5-year cumulative incidences of REL and NRM for a set of three reference patients. These corresponded to the three MDS classes, all with the median patient and donor ages at transplant, and with reference levels for the remaining categorical covariates. Table 3.1 summarises the point estimates, and corresponding 95% CIs. For the imputation methods, the variances of the predicted probabilities were obtained with the Aalen estimator (de Wreede *et al.*, 2010). Subsequently, the 95% CIs were constructed after transformation on the complementary log-log scale, as described in the work of Morisot *et al.* (2015). For comparability, the CIs for CC were also constructed on the complementary log-log scale. In line with the results from the estimated regression coefficients, both imputation methods yielded quasi identical results. By contrast, CC yielded cumulative incidences that were generally lower by approximately 3 to 7 percentage points, with CIs that were up to twice as wide.

Such differences between the MI methods and CC do question the validity of the M(C)AR assumption made. In the EBMT registry, many missing values can be considered MCAR, for reasons relating to data management. Variables such as comorbidity score, cytogenetic classification and donor age became more frequently collected over time as their clinical relevance grew clearer. Missingness may also be related to the transplant center, i.e. particular measurements not being recorded in certain clinics. In the current analysis, both calendar date and transplant center (categorical, large number of levels) were not included in the imputation model for simplicity. An option would have been to include them as auxiliary variables (added as predictor to $X \mid Z$, but

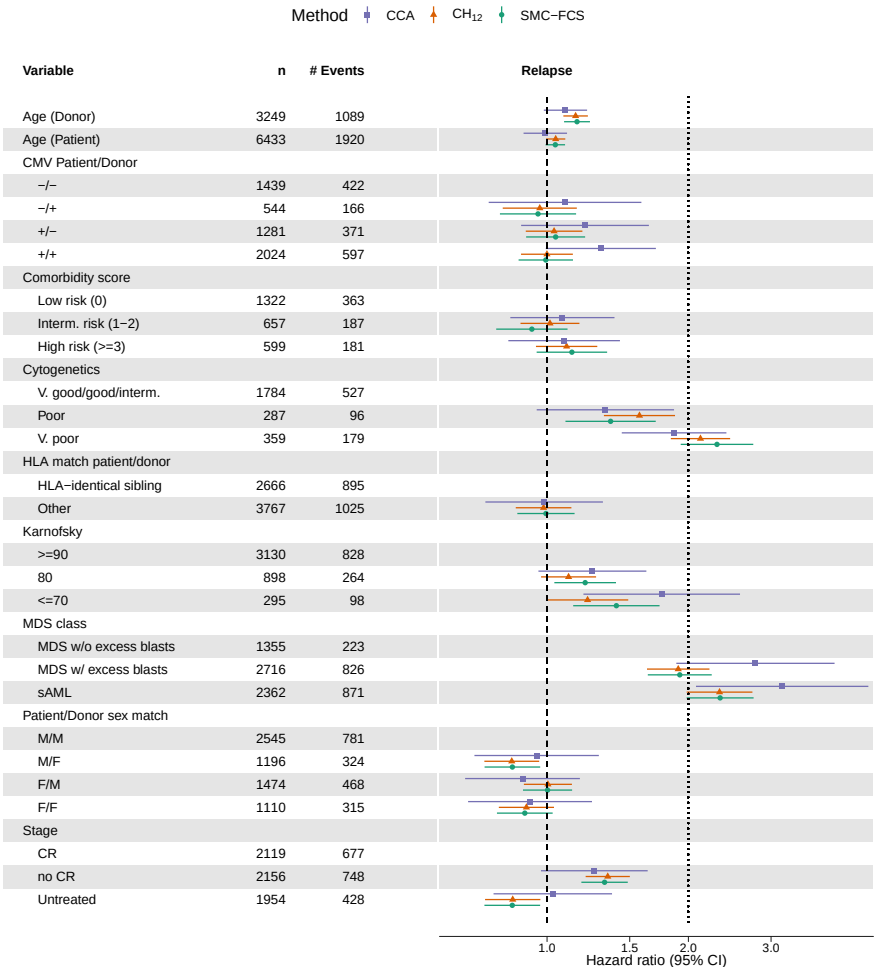

Figure 3.4: Forest plot with point estimates and 95% confidence interval for the cause-specific Cox model for Relapse. On the x-axis are the hazard ratios, which is plotted on the log scale where the confidence intervals are symmetric. Variables and their descriptions can be found in the data dictionary. Per level of factor and for continuous variables, we show the observed counts (n) and the number of relapse events (# Events) in the full data set.

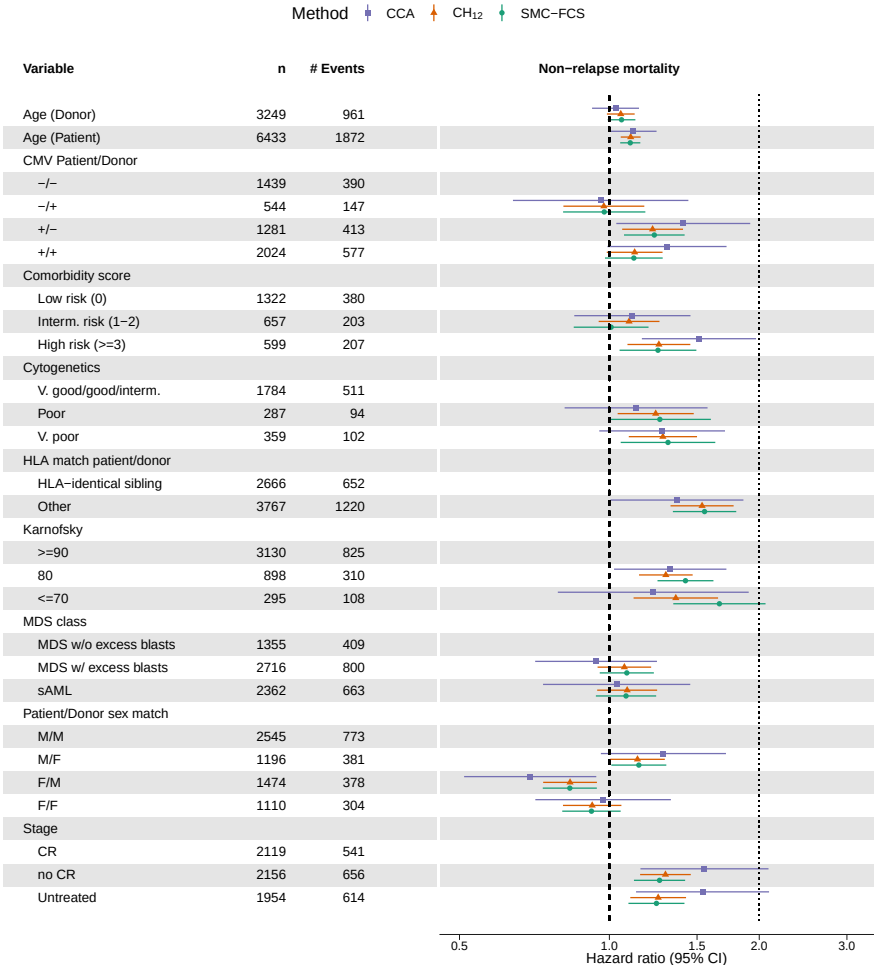

Figure 3.5: Forest plot with point estimates and 95% confidence interval for the cause-specific Cox model for non-relapse mortality (NRM). On the x-axis are the hazard ratios, which is plotted on the log scale where the confidence intervals are symmetric. Variables and their descriptions can be found in the data dictionary. Per level of factor and for continuous variables, we show the observed counts (n) and the number of NRM events (# Events) in the full data set.

Table 3.1: Predicted cumulative incidence (%) of both REL and NRM at 5-years for three reference patients with different MDS classes, reference levels for categorical covariates and sample median values for continuous covariates. The 95% confidence intervals were constructed based on a complementary log-log transformation.

MDS class	CC	CH_{12}	smcfcs
REL			
MDS without excess blasts	10.7 [6.5; 17.2]	17.2 [14.1; 20.8]	17.1 [13.9; 20.9]
MDS with excess blasts	26.9 [19.4; 36.4]	29.7 [25.8; 34.2]	29.7 [25.6; 34.4]
sAML	29.7 [21.4; 40.2]	34.9 [30.7; 39.6]	34.9 [30.4; 39.8]
NRM			
MDS without excess blasts	15.1 [9.7; 23]	18.1 [15.1; 21.7]	17.8 [14.7; 21.5]
MDS with excess blasts	13.1 [9; 18.8]	17.8 [15.2; 20.8]	17.7 [14.9; 20.9]
sAML	14.0 [9.5; 20.4]	17.4 [14.9; 20.2]	17.0 [14.4; 20]

not to substantive model), however the use of auxiliary variables was not a focus of this manuscript, and both MICE and SMC-FCS make different assumptions with respect to inclusion of these variables in the imputation model. Specifically, SMC-FCS would assume independence of center and outcome given the covariates in the substantive model—an assumption which likely does not hold in the registry (Snowden *et al.*, 2020).

3.7 Discussion

In this paper, we assessed the performance of currently implemented MI methods, MICE and SMC-FCS, that deal with missing baseline covariate data when the analysis model of interest is a cause-specific Cox proportional hazards model. For the MICE approach, we provided motivation for the imputation models to be used for continuous, binary, multi-level nominal and ordered categorical covariates with missing values. This is an extension to the work of White and Royston (2009) on Cox proportional hazards models for standard survival outcomes.

We covered a wide range of scenarios in our simulation study, also investigating parameters commonly not addressed in simulation studies for this or similar problems, such as the shape of the baseline hazard and strength of association in the missingness model. Our results confirm the findings of the earlier work of Bartlett and Taylor (2016). Namely, in terms of estimating regression coefficients, SMC-FCS categorically outperforms MICE across all investigated non-MNAR scenarios. Adding the $\hat{H}_k(T) \times Z$ interactions in the imputation model improves performance somewhat, but substantial bias remains. When using MICE, bias grows more extreme as both covariate effects

and the proportion of missing values increase, and seems also affected by the shape of the baseline hazards and the strength of the missingness mechanism. Interestingly, in scenarios where missingness was outcome-dependent, the MICE approach produced biases even larger than those with CCA, which is expected to be biased in these scenarios. Although this is clearly concerning, we do acknowledge that given the longitudinal nature of survival data, a missingness mechanism that depends on the observed event time may be rare.

To our knowledge, our work is the first systematic assessment of the performance of MI for missing covariates with regard to prediction of cumulative incidences. In this respect, the imputation methods performed comparably, which may be attributed to a solid estimation of both the baseline hazards and of the regression coefficients from the complete covariates. The low biases found are consistent with those reported in the work by Mertens et al. on multiple imputation and prediction in the context of logistic regression (Mertens *et al.*, 2020). Furthermore, empirical standard errors did not become smaller beyond around $m = 50$ imputed datasets. If interest lies in reducing the variability of individual predictions between replications of an MI procedure, or replications of a particular study, a choice of m in the order of hundreds will likely be required, as suggested by the same work by Mertens and colleagues. We also emphasise that since we are predicting for reference patients (for which we have *true* data-generating probabilities over time), the assesment of the estimated probabilities is not hindered by any optimism that we would need to correct for, using for example a cross-validation procedure.

There are various limitations to the present work. First, we remark that the explored scenarios are naturally limited as a result of the vast possible parameter space for simulation studies in the field of missing data. For example, missingness was only induced in a single variable. Naturally, more realistic data will be subject to missingness across multiple variables, among which could be interactions in the substantive model. Second, the imputation of covariates with more complex distributions (conditional on other variables) fell outside of the scope of this work. There is a clear need for research and guidance on how to properly impute such variables, particularly for continuous measurements which are heavily skewed (Lee and Carlin, 2017). This may in turn prevent unnecessary categorisation of these variables, and thus further loss of power. Last, we note that in the illustrative analysis, various multi-level nominal and ordinal categorical variables were multiply imputed. These covariate types were not investigated in the simulation study, but are pertinent for further research. Avenues for further exploration could include issues like category inbalance, and comparisons between imputing with proportional odds, multinomial logistic and even a latent normal model (Falcaro *et al.*, 2015; Quartagno and Carpenter, 2019).

Furthermore, a noteworthy difference between the MICE and SMC-FCS approaches in the present context lies in the treatment of cumulative cause-specific baseline hazards functions $H_{k0}(T)$. While the SMC-FCS approach updates $\hat{H}_{k0}(T)$ at each

iteration of the imputation procedure using the Breslow estimate, the MICE approach approximates $H_{k0}(T)$ once using the Nelson–Aalen estimate and keeps them fixed throughout the imputation procedure. Updating $\hat{H}_{k0}(T)$ iteratively with MICE was investigated in the single event setting by White and Royston (2009), with simulations failing to justify its use over inclusion of the Nelson–Aalen estimates in the imputation model. The additional simulation study reported in supplementary material I of the present work appears to show that these earlier results do extend to the competing risk setting. This in turn suggests that the differences in performance between MICE and SMC-FCS could almost entirely be attributed to the functional form of the imputation model, rather than to any error in estimating $H_{k0}(T)$.

For practising statisticians, our work in combination with that of Bartlett and Taylor (2016) shows that SMC-FCS should be the current standard when applying MI in the cause-specific competing risks setting. Although in many controlled situations differences between MICE and SMC-FCS may be small (as in our alloHCT example), the latter seems to be the safest choice given the inherent lack of knowledge regarding the true underlying missingness mechanism. Naturally, SMC-FCS can still be biased, and so the researcher is encouraged to think meticulously about the assumptions underlying their data.

We also recommend that a CCA still be a starting point before performing MI, as it will be unbiased when M(C)AR and covariate-dependent MNAR hold. When biases occur, they may not be as extreme as expected, particularly when the proportion of incomplete cases is low. However, in applications where the proportion of incomplete cases is very high and the M(C)AR assumption is deemed plausible, efficiency gains can be substantial when using MI. This was particularly the case in our alloHCT example, where smaller standard errors were observed with the MI methods for both regression coefficients and predicted cumulative incidence.

The present findings add to a broader literature concerning missing covariates in the context of Cox models (Keogh and Morris, 2018; Marshall *et al.*, 2010; Shah *et al.*, 2014). Studies investigating methods for dealing with missing covariates for a substantive Fine–Gray model remain scarce. For the Fine–Gray model, multiple imputation has predominantly been assessed in the context of missing or interval censored outcomes (Bakoyannis *et al.*, 2010; Delord and Génin, 2016) We conclude by remarking that likelihood-based and fully Bayesian approaches have also not yet been explored or implemented in the context of competing risks, despite already showing promise in other applications (Erler *et al.*, 2016).

Supplementary materials

There are two supplements to the present manuscript. The first, supplementary material I, is available online at https://doi.org/10.1177/09622802221102623. It presents two additional simulation studies, the non-parametric cumulative incidence curves from the alloHCT data and additional simulation results referred to in-text. Supplementary material II is an online supplement, hosted at https://github.com/survival-lumc/CauseSpecCovarMI. It contains full code, simulation data and results, in addition to a synthetic version of illustrative analysis dataset.

Appendix A: Imputation model derivations

Without loss of generality, we assume that X and Z are scalars. The following derivations are valid under the MAR assumption, but also apply if the missing data are MCAR. Letting R_X indicate whether elements of X are missing ($R_X = 0$) or observed ($R_X = 1$), MAR implies $R_X \perp X \,|\, \{T, D, Z\}$ while for MCAR $R_X \perp X$.

We can express the log conditional density of the (right-censored) competing risks outcomes given the covariate data as

$$\log p(T, D \,|\, X, Z) = \log S(T \,|\, X, Z) + \sum_{k=1}^{K} I(D = k) \log h_k(T \,|\, X, Z).$$

Using $\log S(T \,|\, X, Z) = -\sum_{k=1}^{K} H_k(T \,|\, X, Z)$, and assuming a cause-specific Cox proportional hazards model for each failure cause k as $h_k(t \,|\, X, Z) = h_{k0}(t) \exp(\beta_k X + \gamma_k Z)$, we can write

$$\log p(T, D \,|\, X, Z) = \sum_{k=1}^{K} \Big\{ I(D = k) \big[\log h_{k0}(T) + (\beta_k X + \gamma_k Z) \big] - H_{k0}(T) \exp(\beta_k X + \gamma_k Z) \Big\}.$$

We can then plug-in the above expression into Equation 3.3 describing the fully conditional density of X, yielding

$$\begin{aligned} \log p(X \,|\, T, D, Z) = \log p(X \,|\, Z) &+ \sum_{k=1}^{K} I(D = k)(\beta_k X + \gamma_k Z) \\ &- \sum_{k=1}^{K} H_{k0}(T) \exp(\beta_k X + \gamma_k Z) + c, \end{aligned}$$

where c may depend on T, D or Z, but not on X. We note that if Z is a categorical variable with more than two levels, γ_k represents a vector of coefficients for the dummy codes of Z.

Binary X

Suppose X is a binary covariate, depending on Z through a logistic regression model $\text{logit}\, P(X = 1 \mid Z) = \zeta_0 + \zeta_1 Z$. Given this missing data model, the objective now is to derive an expression for $\text{logit}\, P(X = 1 \mid T, D, Z)$. In general we have that

$$\begin{aligned}\text{logit}\, P(X = 1 \mid T, D, Z) &= \log p(T, D \mid X = 1, Z) - \log p(T, D \mid X = 0, Z) \\ &\quad + \text{logit}\, P(X = 1 \mid Z) \\ &= \zeta_0 + \zeta_1 Z + \sum_{k=1}^{K} I(D = k)\beta_k - \sum_{k=1}^{K} H_{k0}(T) \exp(\gamma_k Z)(e^{\beta_k} - 1).\end{aligned} \tag{3.7}$$

If Z is categorical with $J \geq 2$ levels $(0, \ldots, J-1)$, the above expression extends to

$$\begin{aligned}\text{logit}\, P(X = 1 \mid T, D, Z) = \alpha_0 &+ \sum_{k=1}^{K} \alpha_k I(D = k) + \sum_{k=1}^{K} \alpha_{K+k} H_{k0}(T) \\ &+ \sum_{j=1}^{J-1} \alpha_{2K+j} I(Z = j) + \sum_{j=1}^{J-1} \sum_{k=1}^{K} \alpha_{(j+1)K+(J-1)+k} I(Z = j) H_{k0}(T),\end{aligned}$$

which suggests that for categorical Z we can impute missing X values using a logistic regression with D (as a factor variable), the cumulative baseline hazards, evaluated at T, for all causes of failure as covariates, Z (as a factor variable), and the complete interactions between the cumulative baseline hazards at T and Z. The number of parameters, including intercept, equals $JK + J + K$. For example, if there are $K = 2$ competing risks and Z has $J = 3$ categories, there are 11 logistic regression parameters to be estimated (intercept included).

If Z is continuous, there are no exact results due to the $\exp(\gamma_k Z)$ terms. Analogously to White and Royston (2009), we use a first order Taylor series around $\bar{Z}$ to approximate it as $\exp(\gamma_k Z) \approx \exp(\gamma_k \bar{Z})[1 + \gamma_k (Z - \bar{Z})]$, which is valid if $\text{Var}(\gamma_k Z)$ is small. We can thus write

$$\begin{aligned}\text{logit}\, P(X = 1 \mid T, D, Z) &\approx \zeta_0 + \zeta_1 Z + \sum_{k=1}^{K} I(D = k)\beta_k \\ &\quad - \sum_{k=1}^{K} H_{k0}(T)(e^{\beta_k} - 1) \exp(\gamma_k \bar{Z})(1 - \gamma_k \bar{Z}) - \sum_{k=1}^{K} H_{k0}(T) Z \gamma_k (e^{\beta_k} - 1) \exp(\gamma_k \bar{Z}) \\ &= \alpha_0 + \sum_{k=1}^{K} \alpha_k I(D = k) + \sum_{k=1}^{K} \alpha_{K+k} H_{k0}(T) + \sum_{k=1}^{K} \alpha_{2K+k} H_{k0}(T) Z + \alpha_{3K+1} Z.\end{aligned}$$

This suggests an approximate imputation model for binary X and continuous Z including the following predictors: Z, D (as a factor) and all $H_{k0}(T)$. Adding an interaction between all $H_{k0}(T)$ and Z improves the accuracy of this approximation.

Nominal categorical X

Suppose X is categorical with $J > 2$ levels. If we assume the variable to be unordered, we can specify a polytomous logistic regression (also known as multinomial regression) for $p(X \mid Z)$. The model can be expressed in log odds form as

$$\log \frac{P(X = j \mid Z)}{P(X = 0 \mid Z)} = \zeta_{j0} + \zeta_{j1} Z.$$

By coding the reference category as $X = 0$, analogously to Equation 3.7 we obtain

$$\log \frac{P(X = j \mid T, D, Z)}{P(X = 0 \mid T, D, Z)} = \zeta_{j0} + \zeta_{j1} Z + \sum_{k=1}^{K} I(D = k)\beta_{jk} - \sum_{k=1}^{K} H_{k0}(T) \exp(\gamma_k Z)(e^{\beta_{jk}} - 1).$$

From this it becomes clear that all derivations and approximations of $\operatorname{logit} P(X = 1 \mid T, D, Z)$ for binary X continue to hold for $\log \frac{P(X=j \mid T,D,Z)}{P(X=0 \mid T,D,Z)}$, replacing ζ_0 and ζ_1 by ζ_{j0} and ζ_{j1}, and α_k's by α_{jk}'s. This implies that the imputation model for an unordered categorical X should contain the same predictors as for a binary X. The above expression is again exact, given no Z, and also for categorical Z, provided the full interactions between the levels of Z and the cumulative baseline hazards are used; the expression for categorical Z with more than 2 levels extends in the same way.

Ordered categorical X

We consider X as categorical with $J > 2$ ordered levels. A proportional odds model can then be specified for $p(X \mid Z)$, which can be expressed by

$$\log \frac{P(X \leq j \mid Z)}{1 - P(X \leq j \mid Z)} = \zeta_j + \zeta_Z Z,$$

or simply $\operatorname{logit} P(X \leq j \mid Z) = \zeta_j + \zeta_Z Z$ for $j = \{1, ..., J - 1\}$.

To motivate the fully conditional imputation model for X, we need an expression for $\operatorname{logit} P(X \leq j \mid T, D, Z)$. This involves specifying $p(T, D \mid X \leq j, Z)$, which no longer has a proportional hazards density, but instead a *weighted sum* of proportional hazards densities. The imputation model for an ordered categorical X thus does not have a simple extension from binary case. Nonetheless, including the cumulative cause-specific baseline hazards, D as a factor variable and the remaining covariates in the imputation model is still reasonable as an ad hoc solution.

Continuous X

In the case of a continuous X, we specify an exposure model $X \mid Z \sim \mathcal{N}(\zeta_0 + \zeta_1 Z, \sigma^2)$. We can write

$$\log p(X \mid T, D, Z) = \sum_{k=1}^{K} I(D = k)\beta_k X - \sum_{k=1}^{K} H_{k0}(T)\exp(\beta_k X + \gamma_k Z) - \frac{X^2 - 2X(\zeta_0 + \zeta_1 Z)}{2\sigma^2} + c, \tag{3.8}$$

where the terms that do not depend on X from the normal density are subsumed into c. Note that in what follows, the constant c is used to present anything that is not a function of X, and as such may not be equal from line to line. With the same reasoning as in the previous section, a bivariate Taylor approximation is used for the $\exp(\beta_k X + \gamma_k Z)$ around the sample means $\bar{X}$ and $\bar{Z}$. By taking $y = \beta_k(X - \bar{X}) + \gamma_k(Z - \bar{Z})$, we can write the quadratic approximation as

$$\exp(\beta_k X + \gamma_k Z) \approx \exp(\beta_k \bar{X} + \gamma_k \bar{Z})\left[1 + y + \frac{1}{2}y^2\right].$$

Using first the linear component of the above approximation in Equation 3.8 yields

$$\begin{aligned}\log p(X \mid T, D, Z) \approx & \sum_{k=1}^{K} I(D = k)\beta_k X - \sum_{k=1}^{K} H_{k0}(T)\beta_k X \exp(\beta_k \bar{X} + \gamma_k \bar{Z}) \\ & - \frac{X^2 - 2X\zeta_0 - 2X\zeta_1 Z}{2\sigma^2} + c, \\ = & \frac{2\sigma^2 \sum_{k=1}^{K} I(D = k)\beta_k X - 2\sigma^2 \sum_{k=1}^{K} H_{k0}(T)\beta_k X \exp(\beta_k \bar{X} + \gamma_k \bar{Z})}{2\sigma^2} \\ & - \frac{X^2 - 2X\zeta_0 - 2X\zeta_1 Z}{2\sigma^2} + c,\end{aligned} \tag{3.9}$$

and hence

$$\log p(X \mid T, D, Z) \approx - \frac{X^2 - 2X\Big[\overbrace{\zeta_0}^{\alpha_0} + \overbrace{\zeta_1}^{\alpha_1} Z + \sum_{k=1}^{K} I(D = k)\overbrace{\beta_k \sigma^2}^{\alpha_{k+1}} + \sum_{k=1}^{K} H_{k0}(T)\overbrace{(-\beta_k)\sigma^2 \exp(\beta_k \bar{X} + \gamma_k \bar{Z})}^{\alpha_{K+k+1}}\Big]}{2\sigma^2} + c.$$

Thus, approximately

$$X \mid T, D, Z \sim \mathcal{N}(\alpha_0 + \alpha_1 Z + \sum_{k=1}^{K} \alpha_{k+1} I(D = k) + \sum_{k=1}^{K} \alpha_{K+k+1} H_{k0}(T), \sigma^2).$$

Based on a linear approximation for $\exp(\beta_k X + \gamma_k Z)$, the suggested imputation model for missing X is a linear regression containing Z, D (as a factor variable) and all $H_{k0}(T)$ as covariates. This approximation is valid for small $\text{Var}(\beta_k X + \gamma_k Z)$.

For a more precise imputation model, we can revisit Equation 3.9 and add the remaining terms from the quadratic part of the approximation (that do not depend on X). After setting $w = \exp(\beta_k \bar{X} + \gamma_k \bar{Z})$, adding the quadratic terms yields

$$\log p(X \mid T, D, Z) \approx \sum_{k=1}^{K} I(D = k)\beta_k X - \frac{X^2 - 2X\zeta_0 - 2X\zeta_1 Z}{2\sigma^2} - \sum_{k=1}^{K} H_{k0}(T) w \left[\beta_k X + \frac{1}{2}\beta_k^2 (X - \bar{X})^2 + \beta_k \gamma_k X (Z - \bar{Z})\right] + c.$$

After adjusting by $2\sigma^2$ and adding $-\sum_{k=1}^{K} H_{k0}(T) w(\beta_k^2 \bar{X}^2/2)$ to c, we can write

$$\log p(X \mid T, D, Z) \approx -\frac{X^2\left[1 + \sigma^2 \sum_{k=1}^{K} \beta_k^2 H_{k0}(T)\right]}{2\sigma^2} + 2X \times \left\{ \frac{\zeta_0 + \zeta_1 Z + \sigma^2 \sum_{k=1}^{K} \beta_k \left[I(D = k) - H_{k0}(T) w(1 - \beta_k \bar{X})\right]}{2\sigma^2} \right\} - 2X \times \left\{ \frac{\sigma^2 \sum_{k=1}^{K} H_{k0}(T) w \beta_k \gamma_k (Z - \bar{Z})}{2\sigma^2} \right\} + c.$$

Thus, the conditional distribution of X, given (T, D, Z) is approximately normal with mean

$$\frac{\zeta_0 + \zeta_1 Z + \sigma^2 \sum_{k=1}^{K} \beta_k \left[I(D = k) - H_{k0}(T) w(1 - \beta_k \bar{X})\right] - \sigma^2 \sum_{k=1}^{K} H_{k0}(T) w \beta_k \gamma_k (Z - \bar{Z})}{1 + \sigma^2 \sum_{k=1}^{K} \beta_k^2 H_{k0}(T)},$$

and variance

$$\frac{\sigma^2}{1 + \sigma^2 \sum_{k=1}^{K} \beta_k^2 H_{k0}(T)}.$$

Based on a quadratic approximation for $\exp(\beta_k X + \gamma_k Z)$, the suggested imputation model for missing X is a linear regression containing Z, D (as a factor variable), all $H_{k0}(T)$ and all $H_{k0}(T)Z$ interactions as covariates. As explained by White and Royston (2009), this is only valid by ignoring terms in β_k^2 and for small $\sigma^2 \sum_{k=1}^{K} \beta_k^2 H_{k0}(T)$. We also note that the variance of $X \mid T, D, Z$ is non-constant in time.

Chapter 4

Impact of comorbidities and body mass index on the outcomes of allogeneic hematopoietic cell transplantation in myelofibrosis: A study on behalf of the Chronic Malignancies Working Party of EBMT

Chapter based on: Polverelli, N., **Bonneville, E. F.**, de Wreede, L. C., et al. (2024) Impact of comorbidities and body mass index on the outcomes of allogeneic hematopoietic cell transplantation in myelofibrosis: A study on behalf of the Chronic Malignancies Working Party of EBMT. *American Journal of Hematology*, 99, 993—996. DOI: 10.1002/ajh.27262.

The process of selection of feasibility for transplant in myelofibrosis (MF) is determined by several factors such as age, disease stage, comorbidities, performance status, and donor availability (Acosta-Medina *et al.*, 2023). The Myelofibrosis Transplant Scoring System (MTSS) has emerged as a valuable tool for selecting suitable candidates for transplant in MF. By incorporating patient-, transplant-, and donor-specific variables, the MTSS has proven its effectiveness in stratifying patients at varying risks of non-relapse mortality (NRM) and overall survival (OS). More recently, a CIBMTR/EBMT score has been identified as an effective tool for MF transplant candidates' prognostication. However, it should be noted that the prognostic ability of both scores may be reduced by a lack of information on the presence of comorbidities and body mass index (BMI) prior to the transplant, which was not available in these analyses (Kröger *et al.*, 2024).

In order to assess the role of comorbidities and BMI in MF patients undergoing transplantation the Chronic Malignancies Working Party (CWMP) of the EBMT performed a retrospective study with the aim to provide more comprehensive and reliable data on the impact of these factors on transplant outcomes and to identify potential areas for improvement in current MF transplantation protocols. The policy of such study is consistent with that previously published (Polverelli *et al.*, 2021). Inclusion and exclusion criteria, definitions, and methodology are available in Appendix A.

Overall, 4086 patients were included in the final analysis. Patients' characteristics are available in the online supplementary materials. Out of 3157 patients with fully reported comorbidity data, 1701 patients (54%) had at least one comorbidity, with pulmonary conditions being the most prevalent (12.7% moderate and 6.8% severe), as documented also in other transplant scenarios (Polverelli *et al.*, 2020). Other comorbidities present in more than 5% of cases were cardiac disorders (8.6%), diabetes (5.7%) and prior-solid tumor (5.4%). An overview of all comorbidities is available in Figure 4.1.

Concerning the HCT-CI, 1701 (54%) patients had a low (0), 762 (24%) intermediate (1, 2) and 694 (22%) high-risk (=3) score, respectively. Table S2 reports the clinical characteristics stratified according to different HCT-CI classes (see also Table 5.1 from Chapter 5 for the unstratified characteristics). As expected, higher risk class did correlate with increased use of RIC regimens (high risk with 70% vs. 69% and 62% in intermediate, and low risk, respectively), decreased KPS (KPS <80 in 11% vs. 8.1 and 5%). Moreover, the proportion of the splenectomised patients was higher for the high-risk HCT-CI category (14% vs. 9% and 6% in high, intermediate, and low HCT-CI categories, respectively), leading to a lower prevalence of massive splenomegaly (=15 cm) (17% vs. 22% and 28%). Compared to previous cohorts in which the HCT-CI had been developed and subsequently validated (Sorror *et al.*, 2005), the prevalence of comorbidities was higher in our study. Overall, these differences underscore a significant shift in the characteristics of the transplant population over time, as transplantation is increasingly considered in older patients with comorbidities.

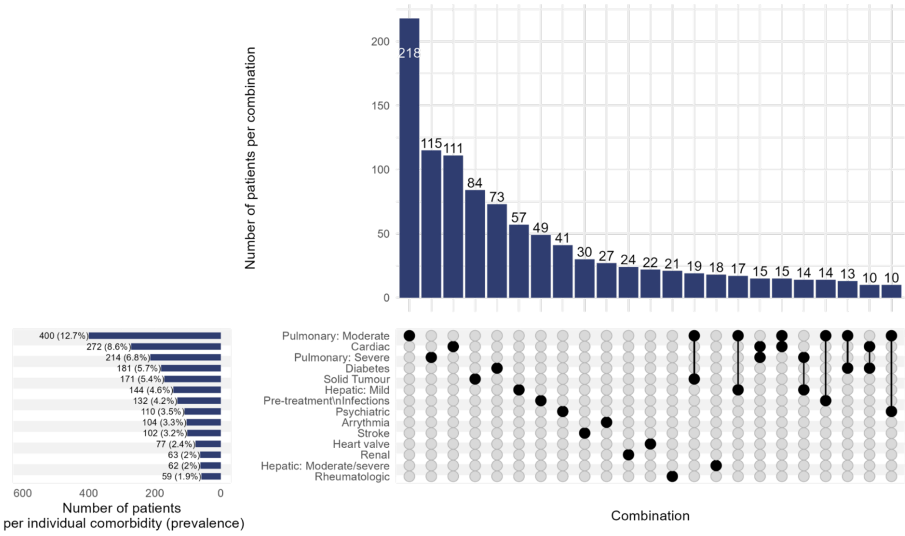

Figure 4.1: UpSet plot visualising the prevalences of individual comorbidities, together with most commonly occurring combinations of comorbidities in the cohort. The bar chart in the bottom-left hand corner of the plot shows the prevalences of individual comorbidities, while main bar chart shows the number of patients with a specific combination of comorbidities (defined by the black dots under it). For example, 400 (12.7%) patients had a moderate pulmonary comorbidity (either as only comorbidity, or in combination with other comorbidities), while 218 patients had a moderate pulmonary comorbidity as only comorbidity. 10 patients had both cardiac comorbidity and diabetes, but no other comorbidities. Combinations of comorbidities occurring in fewer than 10 patients are not shown. As a result, both IBD (n = 22, 0.7%) and Peptic Ulcer (n = 37, 1.17%) do not feature in this plot.

By univariable analysis, both NRM and OS were statistically associated with HCT-CI risk categories. The 5-year expected NRM was 27% (25%–30%), 33% (29%–36%), and 36% (32%–40%) in low, intermediate, and high-risk HCT-CI groups ($p < .001$), respectively. The 5-year estimated OS was 58% (55%–61%), 52% (47%–56%), and 46% (42%–51%) for the low, intermediate, and high HCT-CI scores, respectively ($p < .001$) (Figure S2). No statistical differences were observed in relapse incidence ($p = .22$), and incidence of grade 2-4 acute GvHD ($p = .056$) or chronic GvHD ($p = .46$) depending on the HCT-CI. Table S3 details the causes of NRM.

After adjusting for other variables well known to be associated with NRM and OS in MF, high-risk HCT-CI was strongly associated with both NRM (HR 1.32, 95% CI 1.12–1.55, $p < .001$) and OS (HR 1.27, 95% CI 1.11–1.46, $p < .001$), relative to patients with a low-risk HCT-CI (score of 0) (Figure S3). Also, splenectomy status did not appear to affect NRM in the context of high HCT-CI class ($p = .95$). Therefore, the presence of comorbidities continues to play a negative prognostic role on allo-HCT outcomes and should be integrated into the selection process for MF patients undergoing transplantation along with the existing MTSS and CIBMTR/EBMT tools.

A total of 2679 patients had information on BMI at time of transplant: 50 patients were classified as underweight (1.9%), 1318 as normal weight (49.2%), 964 as overweight (36%), and 347 as grade 1 to 3 obese (13%). Median BMI was 24.9 (range 12.1–46.1). The high prevalence of overweight and obese individuals suggested that patients with robust nutritional reserves were more often considered suitable for transplantation, while cachectic or sarcopenic patients may have had their transplant deferred due to a general tendency among physicians to avoid transplantation in such conditions, generally associated with worse transplant course. As compared to under-normal weight patients (1368, 51.1%), overweight/obese patients were more frequently males (69% vs. 57%), and had been more frequently exposed to ruxolitinib (40% vs. 34%). Continuous BMI was weakly correlated with all other comorbidities. Aside from a correlation of 0.11 with diabetes, correlations with any other comorbidity did not exceed ±0.07 (Table S4).

Despite differences in comorbidities and patient characteristics between different BMI classes, on univariable analysis, no significant differences were found across the BMI groups in terms of NRM ($p = .5$), OS ($p = .3$), grade II-IV acute GvHD ($p = .73$), or chronic GvHD ($p = .6$). By contrast, a modest difference was found regarding relapse incidence ($p = .031$). Furthermore, within a multivariable model that accounted for other variables known to correlate with NRM and OS in MF, including weight loss before allo-HCT, BMI was determined to have no significant impact on either NRM ($p = .59$) or OS ($p = .41$). Figure 4.2 shows the hazard ratio plots of BMI (relative to a reference BMI of 21.75) on OS and NRM respectively, and highlight the very limited impact of BMI on OS and NRM. Likelihood ratio tests also confirm the lack of non-linear effects on both OS ($p = .25$) and NRM ($p = .33$). These findings contrast with the original HCT-CI data, which identified a BMI >35 as a risk factor in both NRM and

OS after allo-HCT (Sorror *et al.*, 2005). In this context, it seems evident that overweight and obese MF patients should not be excluded from a potentially curative life-saving procedure. In MF, overweight or obesity can be associated with milder disease activity, resulting in better nutritional status and suggesting a greater likelihood of improved survival. On the other hand, even patients with lower BMIs can derive benefits from a transplant procedure.

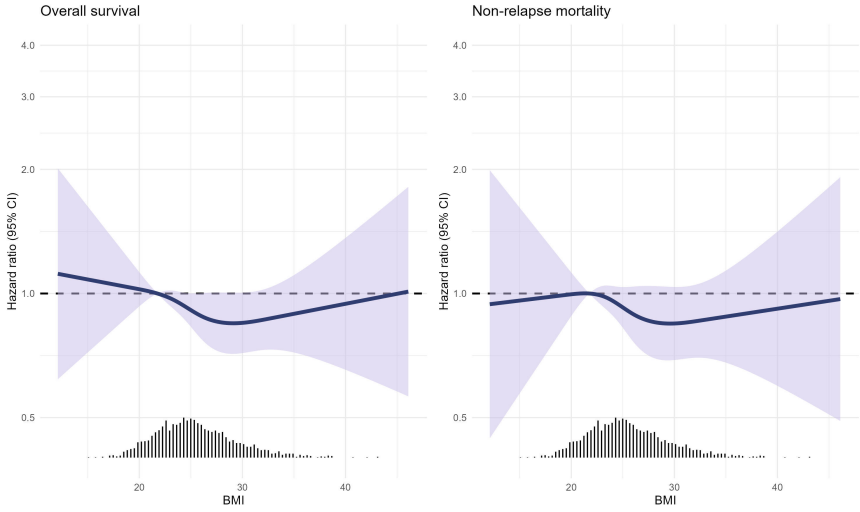

Figure 4.2: Non-linear effect of BMI on overall survival (left) and non-relapse mortality (right) as part of multivariable model adjusting for variables known to be associated with both outcomes in patients with MF. Models are based on n = 3982 patients with complete outcome information on OS and NRM, where covariates were multiply imputed using MICE (100 imputed datasets). Displayed are therefore the pooled coefficients. Hazard ratios should be interpreted relative to a reference BMI of 21.75 (mid-point of the 'normal' BMI category). The plots also show the (marginal) distribution of observed BMI values.

Importantly, evidence of weight loss >10% within 6 months prior to allo-HCT was significantly associated with higher risk of NRM (HR 1.19, 95% CI 1.01–1.39, p = .042), and a trend toward shortened OS (HR 1.19, 95% CI 0.96–1.46, p = 0.108). Therefore, it seems to be vital to consider transplantation not when the disease is already symptomatic with ongoing weight loss. In this case, the optimisation of nutritional status where possible should be considered.

The major strengths of this study rely on the novelty of this information in the largest sample of MF transplant patients, with comprehensive assessment of comorbidities

and BMI and significant follow-up. These numbers permit a comprehensive identification of factors associated with allo-HCT outcomes. Nevertheless, it's important to acknowledge some limitations in this study. First, despite adjusting for weight-loss prior to allo-HCT, the conclusions regarding the effect of BMI on mortality were likely also affected by selection effects that could not be modeled. In particular, patients with BMI >35 (i.e., for whom one would expect a clear negative impact on mortality) in this study may have been selected for allo-HCT based on other more favorable disease characteristics or lack of other comorbidities. Second, the study lacks detailed information regarding specific treatments administered for the management of comorbidities and weight loss before transplantation, which could potentially impact allo-HCT results; additionally, the definition of comorbidities lacks an understanding of their functional impact, and there is a current suggestion to include concomitant frailty assessment in cancer patients (Polverelli *et al.*, 2020). Third, the amount of missing data for both main variables and important adjustment factors was substantial. We chose not to exclude patients on the basis of unavailable comorbidity and BMI information (which may have rendered the cohort less representative), and instead made use of the observed information in the data in order to multiply impute the missing values, thereby potentially enhancing the robustness of the study results. However, the authors believe that despite these limitations, the significance of the topic, which pertains to an ever-growing number of MF patients over the years, outweighs these concerns.

In conclusion, this study, for the first time in a robust fashion, highlights the prognostic significance of HCT-CI in MF patients undergoing allo-HCT. Additionally, it suggests that BMI at the time of transplantation has a limited impact on transplant outcomes in this patient population. These findings enhance our understanding of risk factors and can guide clinical decision-making for MF patients considering allo-HCT. Nevertheless, future research should aim to validate these findings and explore the possibility of integrating comorbidity assessment alongside existing scoring systems and splenomegaly evaluation (Polverelli *et al.*, 2021, 2023) to develop effective tools for selecting MF patients as candidates for transplantation.

Supplementary materials

Supplementary tables and figures are available online at https://doi.org/10.1002/ajh.27262.

Appendix A: Methods

Inclusion criteria:

- All primary and secondary MF patients submitted to allo-HCT in chronic phase from any donor (except syngeneic or matched-related), conditioning regimen and stem cell source.
- Age greater or equal to 18 years-old.
- First transplant.
- Transplant performed between 2009 and 2019.

Exclusion criteria:

- Accelerated or Blastic phase MF.

Definitions

The comorbidities encountered at allo-HCT and their severity cut-offs were categorised according to the original study by Sorror *et al.* (2005). BMI at allo-HCT was measured by dividing a person's weight in kilograms by the square of his or her height in meters. According to WHO criteria, the following categories were defined: underweight (BMI <18.5), normal weight (18.5–24.9), overweight (25–29.9), class I obesity (30–34.9), class 2 obesity (35–39.9), and class 3 obesity ($\geq$ 40).

Statistical analysis

Pre-transplant patient characteristics were reported using the median and interquartile range (IQR) for continuous variables, and frequencies and proportions for categorical variables. Baseline was defined as time of first allo-HCT. Primary endpoints were overall survival (OS) and cumulative incidence of non-relapse mortality (NRM). OS was defined as time from allo-HCT to death, and NRM was defined as death without evidence of relapse or progression. Secondary endpoints included the cumulative incidence of relapse (including progression), as well as the cumulative incidences of acute (grades II-IV) and chronic Graft-versus-host disease (aGvHD and cGvHD, respectively). Median follow-up was determined using the reverse Kaplan–Meier method. OS was estimated using the Kaplan–Meier product limit estimation method, and differences in subgroups until 60 months were assessed by means of the Log-Rank test. Cumulative incidences of relapse and NRM were analysed together in a competing risks framework. Competing risks analyses were also separately applied to estimate the cumulative incidences of aGvHD and cGvHD, where death and second allo-HCT were considered as competing events. Subgroup differences in competing risks analyses were assessed using Gray's test.

HCT-CI was calculated using information for individual comorbidities reported by centers. The impact of HCT-CI (as low 0 risk, intermediate 1–2 risk or high ≥ 3 risk) on OS was assessed using a multivariable Cox proportional hazards model, and its impact of NRM was assessed using a cause-specific Cox model. Both models adjusted for the following variables: patient sex, patient age (per decade), donor type (identical sibling donor vs. other), Karnofsky performance score (KPS; ≥ 90, 80 and < 80), patient/donor cytomegalovirus match (patient/donor negative or other), ruxolitinib treatment prior to allo-HCT (given or not), conditioning (standard or reduced), and the individual components of the Dynamic international prognostic scoring system (Passamonti *et al.*, 2010) risk score other than patient age. These DIPSS components included continuous white blood cell count $x10^9$/L, hemoglobin g/L, peripheral blood blasts (per 5%), as well as weight loss (more than 10% in 6 months prior to allo-HCT) and presence of night sweats. Continuous components of the DIPSS were modelled linearly, after testing for non-linear effects using likelihood ratio tests (comparing the model with restricted cubic splines with one using linear effects).

A separate multivariable model was also fit (only for outcome NRM), evaluating the impact of individual comorbidities comprising the HCT-CI, instead of the HCT-CI itself. This was done in order to compare the implied weights to those assigned by the original HCT-CI. Furthermore, a subset analysis was considered to explore HCT-CI's prognostic value alongside MTSS-specific parameters. However, due to a small sample (around 100 patients) with complete MTSS data, this investigation was not feasible.

The impact of BMI on OS and NRM was also investigated by means of multivariable Cox models. We allowed for a potential flexible non-linear effect of BMI by using restricted cubic splines with two internal knots, placed at the 33rd and 67th percentile of BMI values. The reference BMI value was set to 21.75, the center point of the normal BMI category. This model adjusted for the same variables as those described above, together with a (continuous) partial HCT-CI score which omits the contribution of BMI. Importantly, we adjust for characteristics pertaining to disease risk, among which weight loss prior to allo-HCT (which is related to both BMI at allo-HCT and the risk of death). This is in view of trying to obtain a 'direct' effect of BMI on NRM, while mitigating potential bias from not adjusting for illness-related weight loss (which could make being overweight/obese appear protective against mortality—see Lennon *et al.*, 2016).

Missing values in covariates, which were assumed to be missing at random, were multiply imputed using multivariate imputations by chained equations (MICE) (van Buuren *et al.*, 1999), with 100 imputed datasets and 20 iterations. This approach efficiently makes use of the observed information in the data, and can offer estimates that are more precise and potentially less biased (Bonneville *et al.*, 2023). The imputation model for each partially observed covariate included remaining covariates in the multivariable models, the competing event indicator as a categorical variable, and the estimated marginal cause-specific cumulative hazards of both relapse and NRM

at an individual's event or censoring time (Bonneville *et al.*, 2022). Each comorbidity variable was imputed separately (rather than imputing the HCT-CI itself), and BMI was imputed as a continuous variable. The year of allo-HCT, and the interval between diagnosis and allo-HCT were used as continuous auxiliary variables in the imputation procedure. Default imputation methods were used: predictive mean matching for continuous variables, logistic regression for binary variables, proportional odds for ordered categorical, and multinomial regression for underordered multi-level categorical variables. Prior to imputing the missing covariates, missing relapse and aGvHD times (for patients with known event, but missing event time) were singly imputed by sampling the observed times from patients for whom the times were observed.

Outcomes in the analysis for aGvHD and death before aGvHD were artificially censored at 4 months after allo-HCT, while remaining outcomes were censored at 60 months after allo-HCT. All statistical tests were two-sided, at a significance level of 0.05. All analyses were performed in R version 4.2.1, using 'survival', 'cmprsk', 'prodlim', 'rms', and 'mice' packages.

Chapter 5

Multiple imputation of missing covariates when using the Fine–Gray model

Chapter based on: **Bonneville, E. F.**, Beyersmann, J., Keogh, R. H., et al. (2024) Multiple imputation of missing covariates when using the Fine–Gray model. arXiv:2405.16602. arXiv. DOI: 10.48550/arXiv.2405.16602.

Abstract

The Fine–Gray model for the subdistribution hazard is commonly used for estimating associations between covariates and competing risks outcomes. When there are missing values in the covariates included in a given model, researchers may wish to multiply impute them. Assuming interest lies in estimating the risk of only one of the competing events, this paper develops a substantive-model-compatible multiple imputation approach that exploits the parallels between the Fine–Gray model and the standard (single-event) Cox model. In the presence of right-censoring, this involves first imputing the potential censoring times for those failing from competing events, and thereafter imputing the missing covariates by leveraging methodology previously developed for the Cox model in the setting without competing risks. In a simulation study, we compared the proposed approach to alternative methods, such as imputing compatibly with cause-specific Cox models. The proposed method performed well (in terms of estimation of both subdistribution log hazard ratios and cumulative incidences) when data were generated assuming proportional subdistribution hazards, and performed satisfactorily when this assumption was not satisfied. The gain in efficiency compared to a complete-case analysis was demonstrated in both the simulation study and in an applied data example on competing outcomes following an allogeneic stem cell transplantation. For individual-specific cumulative incidence estimation, assuming proportionality on the correct scale at the analysis phase appears to be more important than correctly specifying the imputation procedure used to impute the missing covariates.

5.1 Introduction

The presence of missing covariate data continues to be pervasive across biomedical research. Among the many existing approaches for dealing with missing covariate data, multiple imputation (MI) methods in particular have become increasingly popular in practice (Carpenter and Smuk, 2021). Compared to a complete-case analysis (CCA), MI can provide inferences that are both less biased and more efficient, under certain missingness mechanisms and given that the imputation models are appropriately specified (Sterne *et al.*, 2009).

The most common approach to MI is to specify and fit univariate regression models for partially observed covariates, from which imputations are then generated. Ideally, each one of these imputation models should be compatible with the substantive model of interest. That is, the assumptions made by both models should not conflict with each other, e.g. the imputation model should at least feature the remaining substantive model covariates, as well as the outcome. We refer to an imputation model as being 'directly specified' when substantive model covariates and outcome variable(s), or any transformations thereof, are included explicitly as predictors in the imputation model. In settings with missingness spanning multiple covariates, the specification of a joint distribution via a set of directly specified imputation models is more commonly known as MICE (multivariate imputation by chained equations, van Buuren *et al.*, 1999).

In the context of cause-specific Cox proportional hazards models (Prentice *et al.*, 1978), it has been shown that the imputation model for a partially observed covariate should at least include as predictors the other covariates from the substantive model, together with the cause-specific cumulative hazard and event indicator for each competing risk (Bonneville *et al.*, 2022). Analogously to the standard single-event survival setting, this directly specified imputation model is generally only approximately compatible with the proportional hazards substantive model (White and Royston, 2009). Concretely, when the outcome model assumes proportional hazards, the conditional distribution of a partially observed covariate modelled using MICE is only an approximation of the 'true' (i.e. implied assuming the substantive model is correctly specified) conditional distribution of the partially observed covariate given the outcome and other substantive model covariates. If imputed values can instead be directly sampled from the latter distribution, it would ensure compatibility between analysis and imputation model. This alternative 'indirect' way of obtaining imputations is referred to as the substantive-model-compatible imputation (SMC-FCS, Bartlett *et al.*, 2015) approach, and it was extended by Bartlett and Taylor (2016) to accommodate cause-specific Cox substantive models. In terms of estimating cause-specific hazard ratios, simulation studies have shown that the SMC-FCS approach tends to outperform MICE in cases when the substantive model is correctly specified (Bartlett and Taylor, 2016; Bonneville *et al.*, 2022).

When a Fine–Gray subdistribution hazard model (Fine and Gray, 1999) is the substantive model of interest, there has to our knowledge been no research on how one should specify an imputation model for a missing covariate (Lau and Lesko, 2018). Nevertheless, MICE is still being used in the presence of missing covariates when the substantive model is a Fine–Gray model, particularly in the context of prediction models. While the structure of the imputation model is rarely reported, articles which do describe their imputation procedure appear to use different approaches. For example, in the prognostic Fine–Gray model presented by Archer *et al.* (2022) (where the primary outcome was time to serious fall resulting in hospital admission or death, with competing death due to other causes), the imputation model for a missing covariate contained the other substantive model covariates, and the cause-specific cumulative hazard and event indicator for each competing risk. In contrast, the MICE procedure reported as part of the prognostic models presented by Clift *et al.* (2020) used cumulative subdistribution hazards in the imputation model. Heuristically, it would seem the latter approach is more consistent with the substantive model, as the former imputes approximately compatibly with a cause-specific Cox model structure rather than the Fine–Gray model structure.

In this work, we extend the SMC-FCS approach for missing covariates to accommodate a Fine–Gray substantive model for one of the competing events. In the presence of independent and identically distributed censoring times that are stochastically independent of the competing risks process (i.e. the 'random censoring' assumption, Beyersmann *et al.*, 2012), the core idea is to multiply impute the potential censoring times for individuals failing from competing events in a first step (Ruan and Gray, 2008), and thereafter use existing SMC-FCS methodology (originally developed for the standard Cox model, Bartlett *et al.*, 2015) to impute the missing covariates in a second step.

The structure of the manuscript is as follows. We introduce competing risks notation in Section 5.2. In Section 5.3 we outline the proposed method, and thereafter assess its performance in a simulation study in Section 5.4. We also provide an illustrative analysis using a dataset from the field of allogeneic hematopoietic stem cell transplantation (alloHCT) in Section 5.5. Finally, findings are discussed in Section 5.6, together with recommendations on how to impute covariates in competing risks settings more generally.

5.2 Notation

We consider a setting in which individuals experience only one of K competing events. We denote the event time as $\tilde{T}$, and the competing event indicator as $\tilde{D} \in \{1, ..., K\}$. In practice, individuals are subject to some right-censoring time C, meaning we only observe realisations (t_i, d_i) of $T = \min(C, \tilde{T})$ and $D = I(\tilde{T} \leq C)\tilde{D}$, where $I(\cdot)$

is the indicator function and $D = 0$ indicates a right-censored observation. The cause-specific hazard for the k^{th} event is defined as

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq \tilde{T} < t + \Delta t, \tilde{D} = k \mid \tilde{T} \geq t)}{\Delta t}.$$

These hazards together make up the event-free survival function,

$$P(\tilde{T} > t) = \exp\left\{ -\sum_{k=1}^{K} \int_0^t h_k(u)\,\mathrm{d}u \right\} = \exp\left\{ -\sum_{k=1}^{K} H_k(t) \right\},$$

assuming the distribution of T is continuous, and $H_k(t)$ is the cause-specific cumulative hazard function for the k^{th} event. The cause-specific cumulative incidence function is then defined as

$$F_k(t) = P(\tilde{T} \leq t, \tilde{D} = k) = \int_0^t h_k(u) S(u-)\,\mathrm{d}u,$$

where $S(u-)$ is the event-free survival probability just prior to u.

The subdistribution hazard for the k^{th} event is defined as

$$\begin{aligned} \lambda_k(t) &= \frac{-\,\mathrm{d}\log\{1 - F_k(t)\}}{\mathrm{d}t}, \\ &= \frac{\mathrm{d}F_k(t)}{\mathrm{d}t} \times \{1 - F_k(t)\}^{-1}, \end{aligned}$$

which can be thought of as the hazard for the improper random variable $\tilde{V}_k = I(\tilde{D} = k) \times \tilde{T} + I(\tilde{D} \neq k) \times \infty$, for which we can write $F_k(t) = P(\tilde{V}_k \leq t)$ (Beyersmann *et al.*, 2012). The probability mass at infinity makes $\tilde{V}_k$ improper, i.e. that its density function does not integrate to one.

Suppose interest lies in modelling the cumulative incidence of one of the competing events, say $D = 1$, conditional on (time-fixed) covariates $\mathbf{Z}$. The Fine–Gray model for cause 1 can be written as

$$\lambda_1(t \mid \mathbf{Z}) = \lambda_{01}(t) \exp(\boldsymbol{\beta}^\top \mathbf{Z}),$$

with $\lambda_{01}(t)$ being the subdistribution baseline hazard function and $\boldsymbol{\beta}$ representing the effects of covariates $\mathbf{Z}$ on the subdistribution hazard. The cumulative incidence function for cause 1 can then be written as

$$F_1(t \mid \mathbf{Z}) = 1 - \exp\left\{ -\exp(\boldsymbol{\beta}^\top \mathbf{Z}) \int_0^t \lambda_{01}(u)\,\mathrm{d}u \right\},$$

where $\int_0^t \lambda_{01}(u)\,\mathrm{d}u = \Lambda_{01}(t)$ is the cumulative baseline subdistribution hazard. If we define a baseline cumulative incidence function $F_{01}(t) = 1 - \exp\{-\Lambda_{01}(t)\}$ (i.e. the cumulative incidence when $\mathbf{Z} = 0$), the model can also be written as

$$F_1(t \mid \mathbf{Z}) = 1 - \{1 - F_{01}(t)\}^{\exp(\boldsymbol{\beta}^\top \mathbf{Z})}. \tag{5.1}$$

In the presence of random right censoring, the Fine–Gray model is usually fitted by maximising a partial likelihood that uses time-dependent inverse probability of censoring weights (IPCW) (Fine and Gray, 1999).

5.3 MI approaches with a Fine–Gray substantive model

We consider a setting with p partially observed covariates $X = X_1, ..., X_p$, q fully observed covariates $Z = Z_1, ..., Z_q$, and $K = 2$ competing events. We assume that (possibly conditional on Z) censoring is independent of both X and the competing risks outcomes $\tilde{T}, \tilde{D}$. We furthermore let X^{obs} and X^{mis} respectively denote the observed and missing components of X for an individual, and let R be the vector of observation indicators (equal to 1 if the corresponding element of X is observed, or equal to 0 if it is missing).

The substantive model of interest is $\lambda_1(t \mid X, Z) = \lambda_{01}(t)\exp\{g(X, Z; \boldsymbol{\beta})\}$, which is a Fine–Gray model for cause 1, and where $g(X, Z; \boldsymbol{\beta})$ is a function of X and Z, parametrised by $\boldsymbol{\beta}$. In this section, we provide an overview of possible approaches for imputing each partially observed X_j. These imputation models can then be 'chained' together as described in Sections 4 and 5 of the work by Bartlett *et al.* (2015). In addition to an approach which imputes compatibly with the assumed substantive model, we also consider alternative methods which are either a) only approximately compatible with the substantive model; or b) impute assuming a different underlying competing risks structure (i.e. cause-specific proportional hazards). We require that the proposed approaches be valid under the missing-at-random (MAR) assumption, that is, $P(R \mid T, D, X, Z) = P(R \mid T, D, X^{\text{obs}}, Z)$.

5.3.1 MI based on cause-specific hazards models

5.3.1.1 CS-SMC

A first MI approach to consider is to impute compatibly with cause-specific Cox models, despite the substantive model of interest being a Fine–Gray model for cause 1. As

described by Bartlett and Taylor (2016), this method relies on the substantive-model-compatible imputation density for X_j, given by

$$f(X_j \mid T, D, X_{-j}, Z) \propto f(T, D \mid X, Z) f(X_j \mid X_{-j}, Z), \tag{5.2}$$

where X_{-j} refers to the components of X after removing X_j, and $f(\cdot)$ is a density function. For example, $f(T, D \mid X, Z)$ is used as shorthand notation for $f_{T,D \mid X,Z}(t, d \mid x, z)$, that is, the density function for the conditional distribution $T, D \mid X, Z$, evaluated at (t, d) for given values x and z.

In practice, the substantive model $f(T, D \mid X, Z; \psi)$ assumed for $f(T, D \mid X, Z)$ is a cause-specific Cox model (one for each competing risk). Therefore, ψ ($\psi \in \Psi$) contains the cumulative baseline hazards and log hazard ratios for each cause-specific hazard. A model $f(X_j \mid X_{-j}, Z; \phi)$ indexed by ϕ ($\phi \in \Phi$), is also assumed for $f(X_j \mid X_{-j}, Z)$. The idea is then to sample candidate imputed values for the missing X_j using $f(X_j \mid X_{-j}, Z; \phi)$, and accept these if they also represent draws from a density proportional to $f(T, D \mid X, Z; \psi) f(X_j \mid X_{-j}, Z; \phi)$. We refer to this method as the cause-specific SMC-FCS approach (CS-SMC).

5.3.1.2 CS-Approx

The approximately compatible analogue to the cause-specific SMC-FCS approach is described by Bonneville *et al.* (2022). As briefly described in the introduction, this approach involves directly specifying an imputation model $f(X_j \mid T, D, X_{-j}, Z; \alpha)$ for $f(X_j \mid T, D, X_{-j}, Z)$. In order to ensure approximate compatibility with assumed cause-specific Cox substantive models, the imputation model should include as predictors X_{-j}, Z, D (as a factor variable), and the (marginal, as obtained using the Nelson–Aalen estimator) cause-specific cumulative hazard for each cause $\hat{H}_k(T)$, evaluated at an individual's event or censoring time. We refer to this method as approximately compatible cause-specific MICE (CS-Approx).

5.3.1.3 MI based on the relation between the cause-specific and subdistribution hazards

The imputations generated by the CS-SMC and CS-Approx approaches will typically not be consistent with the assumption of proportional subdistribution hazards for cause 1 made by the substantive model of interest. This is because, for cause 1, proportionality on the cause-specific hazard scale will generally imply non-proportionality on the subdistribution hazard scale (Beyersmann *et al.*, 2012). One can derive the functional form of these time-varying covariate effects on the subdistribution hazard scale by using the relation between the subdistribution hazard and the cause-specific hazards (Putter *et al.*, 2020). The CS-SMC and CS-Approx approaches can therefore be thought

of as procedures to impute (approximately) compatibly with a Fine–Gray model with time-varying covariate effects, the functional form of which are determined by the assumptions made for the cause-specific Cox models of each competing event.

A relevant question at this point is whether the relation between cause-specific and subdistribution hazards can instead be used as part of a procedure to impute compatibly with proportional subdistribution hazards for cause 1. In order to motivate such a procedure, we first note that the conditional density of the observed outcome given covariates used in Equation 5.2 can be written both in terms of cause-specific hazards, and in terms of the cumulative incidence functions, as

$$\begin{aligned} f(T, D \mid X, Z) &= \{h_1(T \mid X, Z) S(T \mid X, Z)\}^{I(D=1)} \{h_2(T \mid X, Z) S(T \mid X, Z)\}^{I(D=2)} \\ &\quad \times S(T \mid X, Z)^{1-I(D=1)-I(D=2)}, \\ &= f_1(T \mid X, Z)^{I(D=1)} f_2(T \mid X, Z)^{I(D=2)} \\ &\quad \times \{1 - F_1(T \mid X, Z) - F_2(T \mid X, Z)\}^{1-I(D=1)-I(D=2)}, \end{aligned} \tag{5.3}$$

with $f_k(t \mid X, Z) = \mathrm{d}F_k(t \mid X, Z)/\,\mathrm{d}t$ known as the 'subdensity' for cause k (Gray, 1988). These subdensities, in turn, can be expressed in terms of the subdistribution hazard, as

$$\begin{aligned} f_k(t \mid X, Z) &= \lambda_k(t \mid X, Z)\{1 - F_k(t \mid X, Z)\}, \\ &= \lambda_k(t \mid X, Z) \exp\{-\Lambda_k(t \mid X, Z)\}, \end{aligned} \tag{5.4}$$

where $\Lambda_k(t \mid X, Z)$ is the cumulative subdistribution hazard for cause k conditional on X and Z. Specifying a Fine–Gray model for cause 1 is an assumption regarding only part of Equation 5.3, namely for any terms involving $f_1(T \mid X, Z)$. The practical implication of this is that Equation 5.2 cannot be used to impute the missing X_j without making assumptions about cause 2. One could for example assume (for imputation purposes) a cause-specific Cox model for cause 2, derive the implied $h_1(t \mid X, Z)$ using the relation between the subdistribution hazard and the cause-specific hazards, and then use both cause-specific hazards to evaluate $f(T, D \mid X, Z)$ in Equation 5.2.

Given that a Fine–Gray model is assumed for cause 1, some computational difficulties can be encountered while making assumptions for cause 2. For example, specifying a Fine–Gray model also for cause 2 in the imputation procedure could result in the total failure probability at an observed event time $F_1(T \mid X, Z) + F_2(T \mid X, Z)$ exceeding 1, meaning we would not be able to draw imputed values using Equation 5.2 for high-risk individuals (Austin *et al.*, 2021). An additional example concerns the approach described in the previous paragraph, where $h_1(t \mid X, Z)$ is derived based on $h_2(t \mid X, Z)$ and $\lambda_1(t \mid X, Z)$. The numerical integration step generally needed to compute $h_1(t \mid X, Z)$ could make the overall imputation procedure rather computationally inefficient. More details on potential issues when specifying a model for cause 2 when a Fine–Gray model is assumed for cause 1 can be found in Bonneville *et al.* (2024)

The above points mean that it is desirable to use an alternative approach which avoids having to specify a model for the cause-specific (or subdistribution) hazard of cause

2. In the next subsection, we propose a SMC-FCS approach assuming a Fine–Gray substantive model for cause 1, which avoids making explicit modelling assumptions concerning cause 2.

5.3.2 MI based on the Fine–Gray model

5.3.2.1 FG-SMC

Suppose for now that the potential censoring time C is known for all individuals. This is for example the case when there is a fixed end of study date (i.e. 'administrative' censoring), and no additional random right-censoring. Fine and Gray (1999) referred to these kind of data as 'censoring complete', since the subdistribution at-risk process is known. Equivalently, the 'observed' subdistribution random variable for cause 1 (henceforth referred to as 'subdistribution time'), $V = I(D = 1) \times T + I(D \neq 1) \times C$, is known for all individuals. In turn, this implies that (with complete covariate data), the Fine–Gray model can be estimated by fitting a standard Cox model with outcome V and event indicator $I(D = 1)$.

Consequently, an intuitive approach to imputing the missing X_j in our setting might therefore be to apply existing SMC-FCS methodology for standard Cox models (see section 6.3 of Bartlett *et al.*, 2015), but instead using V and $I(D = 1)$ as our outcome variables. We refer to this method as Fine–Gray SMC-FCS (FG-SMC). The substantive-model-compatible imputation density is now

$$f(X_j \mid V, D, Z) \propto f(V, D \mid X, Z) f(X_j \mid X_{-j}, Z), \tag{5.5}$$

where the conditional density of the observed outcome given the covariates can be written as

$$\begin{aligned} f(V, D \mid X, Z) &= f_1(V \mid X, Z)^{I(D=1)} \{1 - F_1(V \mid X, Z)\}^{I(D\neq 1)}, \\ &= [\lambda_1(V \mid X, Z) \exp\{-\Lambda_1(V \mid X, Z)\}]^{I(D=1)} \exp\{-\Lambda_1(V \mid X, Z)\}^{I(D=0)} \\ &\quad \exp\{-\Lambda_1(V \mid X, Z)\}^{I(D=2)}, \\ &= \lambda_1(V \mid X, Z)^{I(D=1)} \exp\{-\Lambda_1(V \mid X, Z)\}, \end{aligned} \tag{5.6}$$

using Equation 5.4 and the fact that $f_1(V \mid X, Z)^{I(D=1)} = f_1(T \mid X, Z)^{I(D=1)}$. Note that while Equation 5.5 and Equation 5.6 depend only on $I(D = 1)$, we still use D in the notation to make the contribution of those failing from cause 2 to the density explicit, which is relevant for the upcoming sections. Importantly, this procedure relies on a stronger MAR assumption (compared to the one introduced at the beginning of Section 5.3), namely $P\{R \mid V, I(D = 1), X, Z\} = P\{R \mid V, I(D = 1), X^{\text{obs}}, Z\}$. In essence, we ignore any terms involving $f_2(T \mid X, Z)$ in Equation 5.2 based on the assumption that missingness in X does not depend on either $I(D = 2)$ or the failure time for those failing from cause 2.

5.3.2.2 FG-Approx

The form of Equation 5.6 mirrors the likelihood in the standard Cox context, which can be obtained by replacing $\lambda_1(V \mid X, Z)$ with the hazard of a single event (in absence of competing risks). The practical implications of this for our MI context are that the findings of White and Royston (2009) in the single-event survival setting should in principle extend to the Fine–Gray context. Namely, that the (approximately compatible) directly specified imputation model $f(X_j \mid V, D, X_{-j}, Z; \alpha)$ for a partially observed X_j should contain as predictors at least X_{-j}, Z, the indicator for the competing event of interest $I(D = 1)$, and the cumulative subdistribution baseline hazard for the same event $\Lambda_{01}(V)$, evaluated at the individual subdistribution time V. Instead of the unknown true $\Lambda_{01}(V)$, one could use the estimated marginal cumulative subdistribution hazard $\hat{\Lambda}_1(V)$ instead, obtained using the Nelson–Aalen estimator using V and $I(D = 1)$ are outcome variables. We refer to this approximately compatible MICE approach as FG-Approx.

5.3.3 Accommodating random right-censoring

In addition to (deterministic) administrative censoring, random right-censoring may occur. In the presence of random right-censoring, the contribution of those failing from cause 2 to density Equation 5.6 is no longer evaluable, since we do not know their potential censoring time. Their subdistribution time has effectively been 'coarsened' by their cause 2 failure: we know only that the potential censoring time is later than the cause 2 failure time.

5.3.3.1 Via imputation of potential censoring times

One approach to estimate the parameters of a Fine–Gray model in the presence of random right censoring is to consider the potential censoring times for those failing from cause 2 as missing data, and multiply impute them. To this end, Ruan and Gray (2008) suggested the use of Kaplan–Meier (KM) imputation (Taylor *et al.*, 2002). Specifically, potential censoring times are randomly drawn from the conditional distribution with distribution function $1-P(C > t \mid C > T_i) = 1-\hat{G}(t-)/\hat{G}(T_i-)$, where $\hat{G}(t)$ is a KM estimate of the survival distribution of the censoring times $P(C > t)$ and T_i is the observed event time of an individual failing from a competing event. The imputation of these potential censoring times effectively produces multiple censoring complete datasets, in which a Fine–Gray model can be fit using standard software. Inference is then based on a pooled model, which combines the models fitted in each censoring complete dataset using Rubin's rules (Rubin, 1987).

We can make use of the above ideas in order to multiply impute covariates compatibly with a Fine–Gray model in the presence of random right random censoring. Specifically, we can apply the FG-SMC (or FG-Approx) method in each censoring complete dataset obtained after first imputing the potential censoring times for those failing from cause 2. In order to formalise this procedure, recall that $\boldsymbol{\beta}$ represents the parameters of the substantive model, and that $X = \{X^{\text{obs}}, X^{\text{mis}}\}$. We can similarly partition $V = \{V^{\text{obs}}, V^{\text{mis}}\}$, where V^{mis} is the vector of missing censoring times for those failing from cause 2.

From a Bayesian perspective, the goal is to estimate the conditional density of $\boldsymbol{\beta}$ given the observed data, namely

$$f(\boldsymbol{\beta} \mid X^{\text{obs}}, Z, V^{\text{obs}}, D) = \int_V \int_X f(\boldsymbol{\beta} \mid X^{\text{obs}}, X^{\text{mis}}, Z, V^{\text{obs}}, V^{\text{mis}}, D) \times f(X^{\text{mis}}, V^{\text{mis}} \mid X^{\text{obs}}, Z, V^{\text{obs}}, D) \, \mathrm{d}X^{\text{mis}} \, \mathrm{d}V^{\text{mis}}. \tag{5.7}$$

If we can sample imputed values M times from $f(X^{\text{mis}}, V^{\text{mis}} \mid X^{\text{obs}}, Z, V^{\text{obs}}, D)$, the integral above can be approximated by an average over $f(\boldsymbol{\beta} \mid X^{\text{obs}}, X^{\text{mis}}, Z, V^{\text{obs}}, V^{\text{mis}}, D)$ (the 'complete data' posterior density) evaluated at those M moments (Molenberghs *et al.*, 2014).

One option to sample from $f(X^{\text{mis}}, V^{\text{mis}} \mid X^{\text{obs}}, Z, V^{\text{obs}}, D)$, the joint posterior predictive density, is to use a sequential approach, where we factorise

$$\begin{aligned} f(X^{\text{mis}}, V^{\text{mis}} \mid X^{\text{obs}}, Z, V^{\text{obs}}, D) &= f(X^{\text{mis}} \mid X^{\text{obs}}, Z, V, D) f(V^{\text{mis}} \mid X^{\text{obs}}, Z, V^{\text{obs}}, D), \\ &= f(X^{\text{mis}} \mid X^{\text{obs}}, Z, V, D) f(V^{\text{mis}} \mid Z, V^{\text{obs}}, D). \end{aligned}$$

The above is valid as long as $C \perp X \mid Z$. Practically speaking, this involves imputing the potential censoring times (possibly in strata of Z) in a first step, and then imputing the missing X in a second step. This can be implemented easily using existing software packages in R: {kmi} for the imputation of censoring times (Allignol and Beyersmann, 2010), and {smcfcs} for the imputation of the missing covariates (Bartlett *et al.*, 2022)—see Figure 5.1 for an illustration of the workflow.

If the censoring process additionally depends on the partially observed X, we will need to iteratively sample from $f(V^{\text{mis}} \mid X, Z, V^{\text{obs}}, D)$ and $f(X^{\text{mis}} \mid X^{\text{obs}}, Z, V, D)$ with the following modifications:

1. A model for the censoring process is specified, which must condition X. When this model (which is used to impute the potential censoring times) is fitted during the imputation process, it must be fitted using the most recently imputed values of X. If X is continuous, this could be done using a Cox model for the censoring hazard.

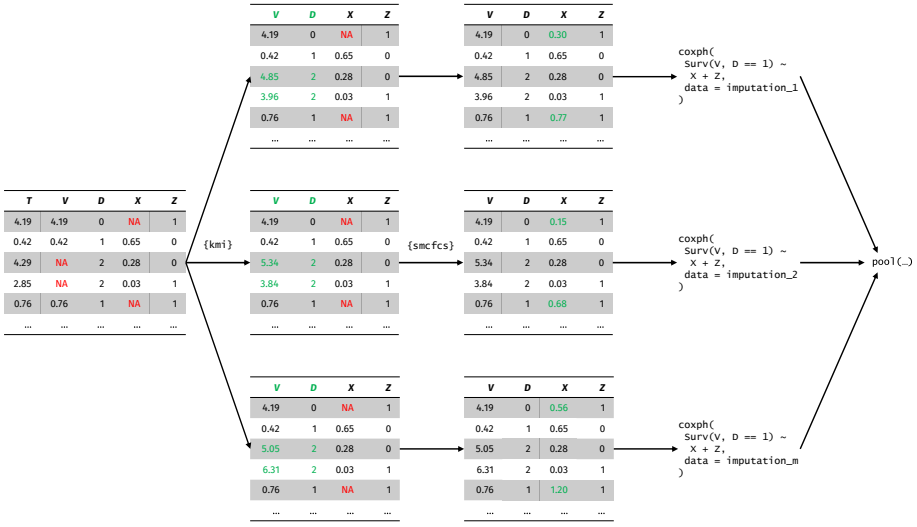

Figure 5.1: Sequential workflow for (compatible) covariate imputation and analysis for a Fine–Gray substantive model with two covariates X and Z, in the presence of random right-censoring. In the first step, the potential censoring times for those failing from cause 2 are multiple imputed using the {kmi} package. In the second step, the missing covariates are imputed using the {smcfcs} (or {mice}) package. This workflow is valid when the probability of being censored is independent of X and any Z related to the censoring process are modelled in {kmi}.

2. Since the censoring distribution partially depends on unobserved values of X, it cannot be ignored in the imputation model. That is, the probability density function of the censoring process no longer factors out of Equation 5.5 (as is the case under random censoring), and is hence non-ignorable or informative (Schluchter and Jackson, 1989). Therefore, the censoring process must be modelled as a cause-specific competing event when imputing the missing X (Borgan and Keogh, 2015). At each iteration, conditional on the most recently imputed potential censoring times, we impute X compatibly with two cause-specific Cox models using CS-SMC: one using V and $I(D = 1)$ as outcome variables (i.e. the Fine–Gray model), and the other using V and $I(D \neq 1)$ (i.e. the model for the censoring process, which must include X).

It is not yet possible to implement the above substantive-model-compatible procedure using existing software in a straightforward way. However, the extension to FG-Approx which accommodates censoring depending on X can be implemented in {mice} using custom imputation methods. The imputation model for X then includes X_{-j}, Z, $I(D = 1)$, $\hat{\Lambda}_1(V)$, and $\hat{H}_C(V)$. Here, $\hat{H}_C(V)$ is the marginal cumulative hazard for the censoring process, estimated based on the most recently imputed V and $I(D \neq 1)$. Based on simulations for other proportional hazards models, we expect that failing to account for non-ignorable censoring at the imputation phase in the present context would have a relatively mild effect on inferences, unless both the proportion of censored observations and the effect of X on the censoring process are large (Bartlett and Taylor, 2016; Borgan and Keogh, 2015; Qi *et al.*, 2010).

Note that the described KM-based procedure for imputing potential censoring times does not take into account any of the uncertainty in estimating $P(C > t)$. That is, the imputed potential censoring times do not involve an initial parameter draw, and are hence not proper Taylor *et al.* (2002). Ruan and Gray (2008) discussed using the non-parametric bootstrap to account for this uncertainty (i.e. each set of imputed censoring times is based on a censoring distribution estimated in a separate bootstrap sample) and improve estimation properties, and found similar results both with and without a bootstrap step. Alternatively, one could choose to specify a (flexible) parametric model for $P(C > t)$, with which we can easily draw from the posterior of all model parameters. In Appendix A, we visualise and give additional details concerning the imputation of potential censoring times.

5.3.3.2 Via censoring weights in the likelihood

Rather than multiply imputing the potential censoring times, an alternative approach is to incorporate inverse probability of censoring weights directly in Equation 5.6. If

we define time-dependent weights

$$w(t) = 1 \qquad \text{if } t \leq T_i,$$
$$w(t) = P(C > t \mid C > T_i) = \frac{G(t-)}{G(T_i-)} \qquad \text{if } t > T_i,$$

then the conditional density of the (subdistribution) outcome given the covariates can be written as

$$f(V, D \mid X, Z) = [\lambda_1(V \mid X, Z) \exp\{-\Lambda_1(V \mid X, Z)\}]^{I(D=1)} \exp\{-\Lambda_1(V \mid X, Z)\}^{I(D=0)} \times \exp\left\{ - \int_0^\infty w(u) \lambda_1(u \mid X, Z) \, \mathrm{d}u \right\}^{I(D=2)}, \tag{5.8}$$

where the term for those failing from the competing event involves integration in practice up to a maximum potential follow-up time t^*. As described by Lambert *et al.* (2017), this integral can be approximated by splitting time into intervals, in which the corresponding $w(t)$ is assumed to be constant.

The integration step needed for those failing from cause 2 in Equation 5.8 means that this approach cannot be implemented in a straightforward way with existing software, unlike the approach described in the previous subsection. The simulation study in this paper therefore focuses on the approach involving multiple imputation of potential censoring times.

5.3.4 Implementation of MI approaches

Methods CS-SMC, CS-Approx, FG-SMC, and FG-Approx can all be implemented using existing software packages in R. In this section, we summarise the steps needed to apply these methods in a given dataset in the presence of random right censoring (possibly in combination with administrative censoring). A minimal R code example can be found in supplementary material S1.

1. Add columns $\hat{H}_1(T)$ and $\hat{H}_2(T)$ to the original data, which are the marginal cause-specific cumulative hazards for each competing risk evaluated at an individual's event or censoring time (obtained using the Nelson–Aalen estimator).
2. Multiply impute the potential censoring for those failing from cause 2 using {kmi}, yielding m censoring complete datasets (i.e. with 'complete' V). The censoring distribution has support at both random and administrative censoring times. Any completely observed covariates that are known to affect the probability of being censored should be included as predictors in the model for the censoring process. {kmi} imputes based on stratified KM when Z are categorical, and based on a Cox model at least one of Z is continuous. If for example an individual's

time of entry into a study determines their maximum follow-up duration, this should be accounted for in the imputation procedure (e.g. by stratifying by year of entry).

3. In each censoring complete dataset, add an additional column $\hat{\Lambda}_1(V)$. This takes the value of the marginal cumulative subdistribution hazard for cause 1 at an individual's observed or imputed subdistribution time, obtained with the Nelson–Aalen estimator based on $I(D = 1)$ and imputed V.

4. In each censoring complete dataset (each with different V and $\hat{\Lambda}_1(V)$, but same $\hat{H}_1(T)$ and $\hat{H}_2(T)$), create a single imputed dataset using the desired covariate imputation method(s):

    - CS-SMC: use {smcfcs} to impute the missing covariate(s) compatibly with cause-specific Cox models. All covariates used in the Fine–Gray substantive model should feature in at least one of the specified cause-specific models.
    - CS-Approx: use {mice} to impute the missing covariate(s), where the imputation model contains as predictors the remaining substantive model covariates, D (as a factor variable), and both $\hat{H}_1(T)$ and $\hat{H}_2(T)$.
    - FG-SMC: use {smcfcs} to impute the missing covariate(s) compatibly with the Fine–Gray substantive model. This is done by using the imputation methods developed for the standard Cox model, but with as outcome variables $I(D = 1)$ and imputed V.
    - FG-Approx: use {mice} to impute the missing covariate(s), where the imputation model contains as predictors the remaining substantive model covariates, $I(D = 1)$, and $\hat{\Lambda}_1(V)$.

5. Fit the Fine–Gray substantive model in each imputed dataset (using standard Cox software with $I(D = 1)$ and imputed V as outcome variables), and pool the estimates using Rubin's rules.

5.4 Simulation study

We aim to evaluate the performance of different MI methods in the presence of missing covariate data when specifying a Fine–Gray model for the subdistribution hazard for one event of interest in the presence of one competing event. Specifically, we assume interest lies in the estimation (for cause 1) of both subdistribution hazard ratios, and the cumulative incidence for a particular individual at some future time horizon. We follow the ADEMP structure for the reporting of the simulation study (Morris *et al.*, 2019).

5.4.1 Data-generating mechanisms

We generate datasets of $n = 2000$ individuals, with two covariates X and Z. We assume $Z \sim \mathcal{N}(0,1)$ and $X \mid Z \sim \text{Bernoulli}\{(1+e^{-Z})^{-1}\}$.

We let $h_k(t \mid X, Z)$, $\lambda_k(t \mid X, Z)$ and $F_k(t \mid X, Z) = P(\tilde{T} \leq t, \tilde{D} = k \mid X, Z)$ respectively denote the cause-specific hazards, subdistribution hazards and cumulative incidence functions for cause k, conditional on X and Z. The competing event times will be generated following two mechanisms: one where the Fine–Gray model for cause 1 is correctly specified, and another where it is misspecified. These are detailed below, together with assumptions concerning both censoring and the missing data mechanisms.

5.4.1.1 Correctly specified Fine–Gray

For this mechanism, we simulate data using the 'indirect' method described in Beyersmann *et al.* (2012), and originally used in the simulations by Fine and Gray (1999). This approach involves first drawing the competing event indicator $\tilde{D}$, and then generating an event time for those with $\tilde{D} = 1$. The final step is to generate times of the competing event for the remaining individuals, who were assigned $\tilde{D} = 2$.

Here, we directly specify the cumulative incidence of cause 1 as

$$F_1(t \mid X, Z) = 1 - \left[1 - p\{1 - \exp(-b_1 t^{a_1})\}\right]^{\exp(\beta_1 X + \beta_2 Z)}.$$

The above expression corresponds to a Fine–Gray model, with as baseline cumulative incidence function a Weibull cumulative distribution function with shape a_1 and rate b_1 (parametrisation used in Klein and Moeschberger, 2003) multiplied by a probability p. Explicitly,

$$F_{01}(t) = p\{1 - \exp(-b_1 t^{a_1})\}.$$

With $\lim_{t\to\infty} F_{01}(t) = p$, we have that $P(\tilde{D} = 1 \mid X, Z) = 1 - (1-p)^{\exp(\beta_1 X + \beta_2 Z)}$, and $P(\tilde{D} = 2 \mid X, Z) = 1 - P(\tilde{D} = 1 \mid X, Z) = (1-p)^{\exp(\beta_1 X + \beta_2 Z)}$. These are the individual-specific cumulative incidences for each event at time infinity. Also note that the baseline subdistribution hazard for this mechanism can be obtained by $\{\mathrm{d}F_{01}(t)/\,\mathrm{d}t\} \times \{1 - F_{01}(t)\}^{-1}$.

The idea then is to generate the event times for cause 1 conditionally on the event indicator and covariates, using

$$\begin{aligned} P(\tilde{T} \leq t \mid \tilde{D} = 1, X, Z) &= \frac{P(\tilde{T} \leq t, \tilde{D} = 1 \mid X, Z)}{P(\tilde{D} = 1 \mid X, Z)} \\ &= \frac{1 - \left[1 - p\{1 - \exp(-b_1 t^{a_1})\}\right]^{\exp(\beta_1 X + \beta_2 Z)}}{1 - (1-p)^{\exp(\beta_1 X + \beta_2 Z)}}. \end{aligned} \tag{5.9}$$

To sample from the above, we first need to draw $\tilde{D} \sim \text{Bernoulli}\{(1-p)^{\exp(\beta_1 X+\beta_2 Z)}\}+1$. We can then use inverse transform sampling to draw failure times within the subset of individuals with $\tilde{D}=1$. Shortening $\exp(\beta_1 X+\beta_2 Z)=\exp(\eta)$, and with $u \sim \mathcal{U}(0,1)$, we can invert Equation 5.9 as

$$t=\left[-\frac{1}{b_1}\log\left[1-\frac{1-\left[1-u\{1-(1-p)^{\exp(\eta)}\}\right]^{1/\exp(\eta)}}{p}\right]\right]^{1/a_1}.$$

For the competing event, we can factorise the cumulative incidence function as

$$P(\tilde{T}\leq t, D=2\,|\,X,Z)=P(\tilde{T}\leq t\,|\,\tilde{D}=2,X,Z)P(\tilde{D}=2\,|\,X,Z).$$

A proportional hazards model can then be specified (for convenience) for

$$P(\tilde{T}\leq t\,|\,\tilde{D}=2,X,Z)=1-\exp\Big\{-H^*_{02}(t)\exp(\beta_1^* X+\beta_2^* Z)\Big\},$$

where $H^*_{02}(t)$ is the cumulative baseline hazard associated to the cumulative incidence function conditional on $\tilde{D}=2$. Since the event indicator is already drawn, the failure times can be drawn again using inverse transform sampling within the subset with $\tilde{D}=2$. Here, we specify a Weibull baseline hazard as $h^*_{02}(t)=a_2 b_2 t^{a_2-1}$.

We fix $\{\beta_1,\beta_2,\beta_1^*,\beta_2^*\}=\{0.75,0.5,0.75,0.5\}$, and the Weibull parameters used for both events as shape $\{a_1,a_2\}=0.75$ and rate $\{b_1,b_2\}=1$. We vary $p=\{0.15,0.65\}$, which is the expected proportion of event 1 failures for individuals with $X=0$ and $Z=0$.

5.4.1.2 Simulation based on cause-specific hazards (misspecified Fine–Gray)

In this data-generating mechanism (DGM), we assume proportionality on the cause-specific hazard scale, and simulate using latent failure times (Beyersmann *et al.*, 2009). We specify baseline Weibull hazards for both cause-specific hazards as

$$h_1(t\,|\,X,Z)=a_1 b_1 t^{a_1-1}\exp(\gamma_{11}X+\gamma_{12}Z),$$
$$h_2(t\,|\,X,Z)=a_2 b_2 t^{a_2-1}\exp(\gamma_{21}X+\gamma_{22}Z),$$

where $\{a_1,a_2\}$ and $\{b_1,b_2\}$ are respectively the shape and rate parameters. Under this DGM, a Fine–Gray model for cause 1 will be misspecified. Nevertheless, the coefficients resulting from the misspecified Fine–Gray model could still be interpreted as time-averaged effects on the (complementary log-log transformed) cumulative incidence function (Grambauer *et al.*, 2010).

We aim to have a scenario close to the one described in Section 5.4.1.1 (in terms of event proportions), where the main difference is that proportionality now holds on the cause-specific hazard scale. To fix the parameters in this DGM, we first simulate a large

dataset of one million individuals following the mechanism described in the previous subsection, where proportional subdistribution hazards hold. Parametric cause-specific proportional hazards models assuming baseline Weibull hazards are then fitted for each failure cause. The point estimates obtained from these models are used as the cause-specific data-generating parameters $\{a_1, b_1, \gamma_{11}, \gamma_{12}\}$ and $\{a_2, b_2, \gamma_{21}, \gamma_{22}\}$. These parameters will of course differ depending on $p = \{0.15, 0.65\}$, and also depending on the censoring distribution. While the cause-specific models fitted on this large dataset will be misspecified (cause-specific baseline hazards are not of Weibull shape, and covariates effects on the cause-specific hazards are non-proportional), the resulting 'least false' parameters are still useful.

Figure 5.2 summarises the DGMs, prior to the addition of any censoring. In the correctly specified Fine–Gray scenarios, the subdistribution log hazard ratio $\lambda_1(t \mid X = 1, Z)/\lambda_1(t \mid X = 0, Z)$ is time constant, while the cause-specific log hazard ratios are time-dependent. The reverse is true for the misspecified Fine–Gray scenarios. Overall, the correctly specified and misspecified Fine–Gray scenarios are very comparable in terms of (true) baseline hazards and cumulative incidences, for both values of p.

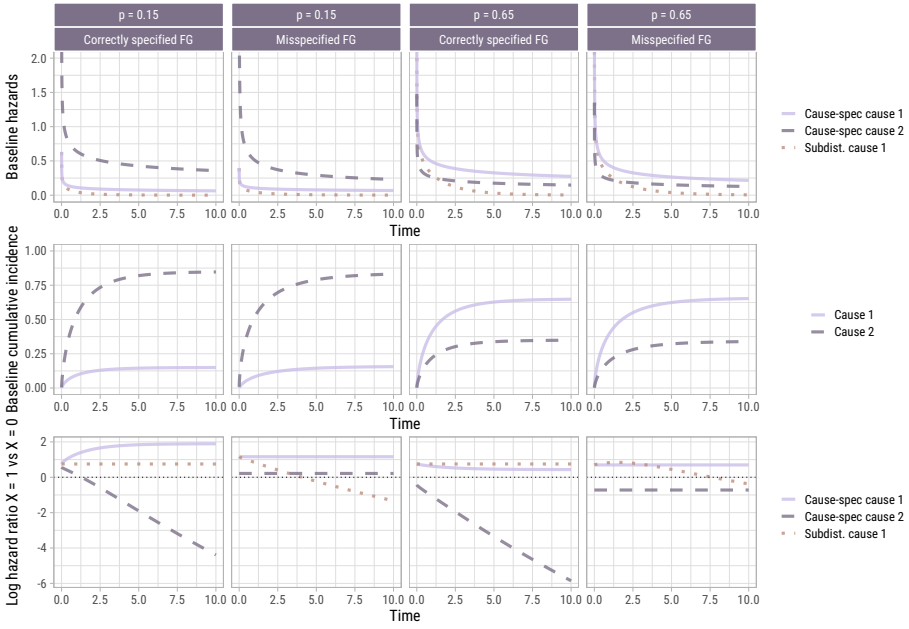

Figure 5.2: Summary of data-generating mechanisms prior to the addition of any censoring. For each value of p, both the correctly specified and misspecified Fine–Gray (FG) scenarios are very comparable in terms of (true) baseline hazards and cumulative incidences.

5.4.1.3 Censoring

The DGMs outlined above assume no loss to follow-up. As additional scenarios, we consider independent (i.e. not conditional on any covariates) right censoring where the censoring times are simulated from an exponential distribution with rate $\lambda_C = 0.49$, resulting in approximately 30% of censored observations. These censoring times will be considered as either: a) known (administrative censoring); or b) unknown (random censoring).

5.4.1.4 Covariate missingness

Missingness is induced in X, while Z remains fully observed. Let R_X be a binary variable indicating whether X is missing ($R_X = 0$) or observed ($R_X = 1$). We use a missing at random (MAR) mechanism conditional on Z, which was defined as $\text{logit}\, P(R_X = 0 \mid Z) = \eta_0 + \eta_1 Z$. We take $\eta_1 = 1.5$, a rather strong mechanism where higher values of Z are associated with more missingness in X. The value of η_0 is found via standard root solving, such that the average probability $P(R_X = 0) = \mathbb{E}\{P(R_X = 0 \mid Z)\}$ of being missing in a given dataset equals 0.4.

5.4.1.5 Summary

In summary, the simulation study varied

- Censoring type: no censoring, administrative and random censoring
- Relative occurrence of event 1, as low or high. This is done by varying the baseline cumulative incidence of event 1 (as $t \to \infty$) as $p = \{0.15, 0.65\}$.
- Failure time simulation methods, with a) directly specified cumulative incidence cause 1 (correctly specified Fine–Gray); b) cause-specific proportional hazards for both causes (misspecified Fine–Gray).

This adds up to 3 (censoring types) × 2 (relative occurence event 1) × 2 (failure time simulation methods) = 12 scenarios.

5.4.2 Estimands

The first estimands of interest are the subdistribution log hazard ratios β_1 and β_2 for X and Z, respectively. In the correctly specified Fine–Gray scenarios, these simply correspond to the data-generating parameters $\{\beta_1, \beta_2\} = \{0.75, 0.5\}$. In the misspecified Fine–Gray scenarios however, the target values (the 'least-false parameters'; time averaged subdistribution log hazard ratios $\{\tilde{\beta}_1, \tilde{\beta}_2\}$) are obtained by fitting a Fine–Gray model on a large simulated dataset of one million individuals, simulated as under the

second data-generating mechanism, after applying any censoring. For computational efficiency, the censoring times are assumed to be known when fitting the Fine–Gray model on this large dataset.

The second estimands of interest are the conditional cumulative incidence of event 1 at a grid of timepoints (between timepoints 0 and 5) for reference individuals $\{X, Z\} = \{0, 0\}$ (baseline) and $\{X, Z\} = \{1, 1\}$. In the correctly specified Fine–Gray scenarios, this corresponds to

$$F_1(t \mid X, Z) = 1 - \left[1 - p\{1 - \exp(-b_1 t^{a_1})\}\right]^{\exp(\beta_1 X + \beta_2 Z)},$$

while for the misspecified Fine–Gray scenarios, this corresponds to

$$\begin{aligned} F_1(t \mid X, Z) &= \int_0^t h_1(u \mid X, Z) \exp\Big\{ -H_1(u \mid X, Z) - H_2(u \mid X, Z) \Big\} \, \mathrm{d}u, \\ &= \int_0^t a_1 b_1 u^{a_1 - 1} \exp(\gamma_{11} X + \gamma_{12} Z) \\ &\qquad \times \exp\Big\{ -b_1 u^{a_1} \exp(\gamma_{11} X + \gamma_{12} Z) - b_2 u^{a_2} \exp(\gamma_{21} X + \gamma_{22} Z) \Big\} \, \mathrm{d}u, \end{aligned}$$

which is obtained via numerical integration.

5.4.3 Methods

The assessed methods are

- Full: analysis run on full data prior to missing values, as a benchmark for the best possible performance.
- CCA: complete-case analysis, as a 'lower' benchmark that the imputation methods need to outperform in order to be worthwhile.
- CS-SMC: MI, imputing compatibly with cause-specific Cox proportional hazards models. This method is described in Section 5.3.1.1. Both X and Z are used as predictors in each cause-specific model assumed by this procedure.
- CS-Approx: MI with both marginal cumulative cause-specific hazards (evaluated at the individual observed event or censoring time) and competing event indicator included as predictors in the imputation model, in addition to Z. This method is described in Section 5.3.1.2.
- FG-SMC: MI, imputing compatibly with a Fine–Gray model for cause 1 that has as covariates X and Z. This is the method described in Section 5.3.2.1.
- FG-Approx: MI with marginal cumulative subdistribution hazard (evaluated at the individual observed or imputed subdistribution time V) and indicator for event 1 included as predictors in the imputation model, in addition to Z. This method is described in Section 5.3.2.2.

The imputation methods are run with 30 imputed datasets. This was fixed following a pilot set of simulations with 50 imputed datasets, which showed that there was little reduction in empirical standard errors for the subdistribution log hazard ratios (and their Monte Carlo standard errors) beyond 30 imputed datasets. Approximately compatible MI methods CS-Approx and FG-Approx only require a single iteration because there is just one variable with missing values, while substantive-model-compatible (SMC) MI methods CS-SMC and FG-SMC are run with 20 iterations. The method used to model $f(X \mid V, D, Z; \alpha)$ for approximately compatible methods is logistic regression, while for SMC methods $f(X \mid Z; \psi)$ is specified as a logistic regression. We note that X was chosen to be binary as SMC methods do not require rejection sampling for variables with discrete sample space, thereby reducing simulation time. If X is chosen to be continuous, the performance of approximately compatible methods is expected to worsen, while no material impact is expected on performance of (correctly specified) SMC methods (Bonneville *et al.*, 2022).

For the scenarios with no or administrative censoring, the subdistribution time V is fully observed. While $V = T$ for those failing from cause 1, for those failing from cause 2, V is first set to either a) a large value greater than the largest observed event 1 time (in absence of censoring); or b) the known potential censoring time C (administrative censoring). The marginal cumulative subdistribution hazard used for the approximate subdistribution MI method is obtained using a marginal model with $I(D = 1)$ and the resulting V as outcome variables. The covariate MI methods are run once these V and $I(D = 1)$ variables have been created. In scenarios with random censoring, the potential censoring times for those failing from cause 2 are multiply imputed using the {kmi} R package with default settings: marginal non-parametric model for the censoring distribution, and no additional bootstrap layer. This yields 30 imputed datasets, each with a different V. In each of these datasets, the marginal cumulative subdistribution hazard is estimated in the same way as described above. Thereafter, the covariate MI methods are run in each of these datasets, yielding one imputed dataset for each imputed V (total of 30 imputed datasets), corresponding to the workflow in Figure 5.1.

For all methods, the Fine–Gray model for cause 1 is estimated using a Cox model with (known or imputed) V and $I(D = 1)$ as outcome variables. When the imputation methods are used (and for all methods when there is random right censoring), the estimated $\hat{\beta}_1$ and $\hat{\beta}_2$ are the results of coefficients pooled using Rubin's rules. Confidence intervals around these estimates are built as described in Section 2.4.2 in van Buuren (2018). For the cumulative incidences, the estimates for the two sets of reference values of X and Z are first made in *each* imputed dataset using Equation 5.1, and thereafter pooled using Rubin's rules after complementary log-log transformation—as described in Morisot *et al.* (2015) and recommended by Marshall *et al.* (2009). This predict-then-pool approach (rather than predicting using a pooled model) has been recommended by multiple authors (Mertens *et al.*, 2020; Wood *et al.*, 2015).

5.4.3.1 Performance measures

The primary measure of interest was bias in the estimated subdistribution log hazard ratios. In order to keep the Monte Carlo standard error (MCSE) of bias under a desired threshold of 0.01, we require $n_{\text{sim}} = 0.2^2/0.01^2 = 400$ replications per scenario, as we expect empirical standard errors to be under 0.2 for all scenarios (based on a pilot run). This number was rounded up to $n_{\text{sim}} = 500$. In addition to bias, we recorded empirical and estimated standard errors, and coverage probabilities. For the cumulative incidence estimates, we focused on both bias and root mean square error (RMSE).

5.4.3.2 Software

Analyses were performed using R version 4.3.1 (R Core Team, 2023). Core packages used were: {survival} version 3.5.7 (Therneau, 2023), {mice} version 3.16.0 (van Buuren and Groothuis-Oudshoorn, 2011), {smcfcs} version 1.7.1 (Bartlett *et al.*, 2022), {kmi} version 0.5.5 (Allignol and Beyersmann, 2010), and {rsimsum} version 0.11.3 (Gasparini, 2018).

5.4.4 Results

We summarise the main findings in this section, with full results available in a markdown file on the Github repository linked at the end of the present manuscript.

5.4.4.1 Subdistribution log hazard ratios

We focus on the results for β_1, together with its time-averaged analogue $\tilde{\beta}_1$ in the scenarios with time-dependent subdistribution hazard ratios. Results concerning bias are summarised in Figure 5.3 for all 12 scenarios, and presented on the relative scale (Monte Carlo standard errors were below the desired 0.01 for both bias and relative bias, across all methods and scenarios).

When the Fine–Gray model for cause 1 was correctly specified, the proposed FG-SMC approach was unbiased regardless of censoring type or (baseline) proportion of cause 1 failures. In contrast, imputing compatibly with the (incorrect) assumption of proportional cause-specific hazards showed strong biases, particularly when $p = 0.15$ in the absence of censoring (25% biased). In the presence of censoring however, this bias dropped to approximately 10%. The CS-Approx method showed consistent downward biases regardless of p and censoring type, while the FG-Approx method was only biased when $p = 0.65$. The latter finding is consistent with previous research in the simple survival setting; namely that the approximately compatible MI approach is expected to work well when cumulative incidence is low (White and Royston, 2009).

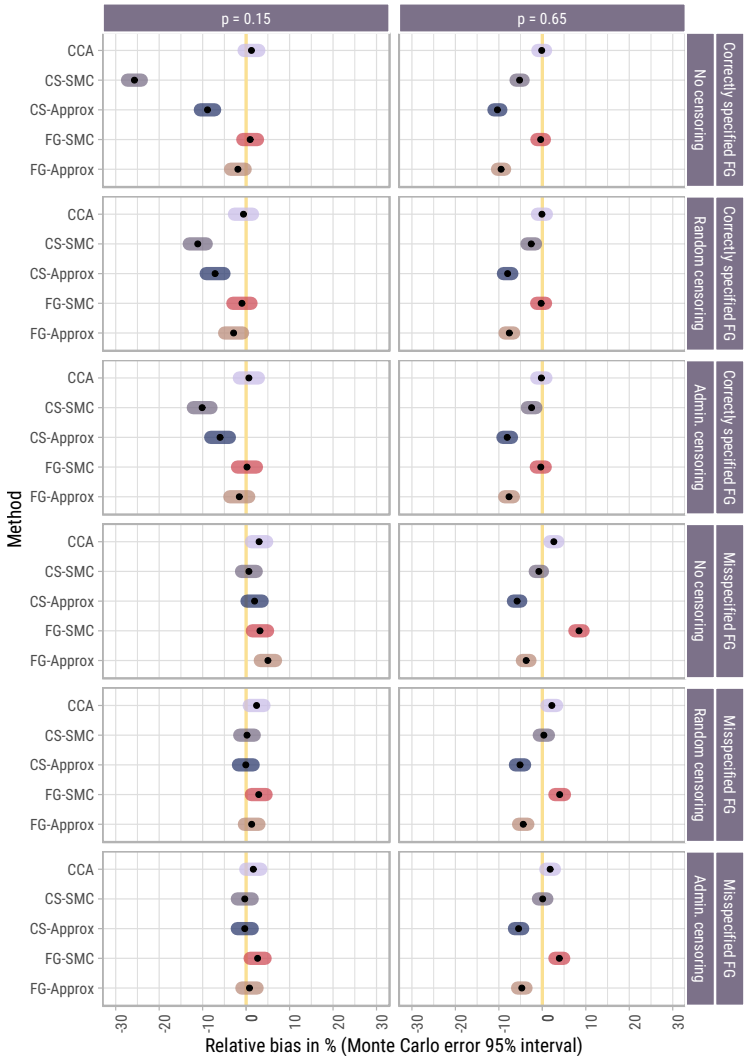

Figure 5.3: Relative bias (%) in estimating β_1, with corresponding 95% Monte Carlo confidence interval (constructed using the standard normal approximation). For each scenario and method, the distribution of $(\hat{\beta}_1 - \beta_1)/\beta_1$ across simulation replications was approximately normal. For the correctly specified Fine–Gray (FG) scenarios, $\beta_1 = 0.75$. In the misspecified FG scenarios, the value of the 'least-false' $\tilde{\beta}_1$ (time-averaged log subdistribution hazard ratio) depended on both p and the presence/absence of censoring. For $p = 0.15$, $\tilde{\beta}_1 \approx 0.76$ without censoring, and $\tilde{\beta}_1 \approx 0.93$ with censoring. For $p = 0.65$, $\tilde{\beta}_1 \approx 0.75$ both with and without censoring.

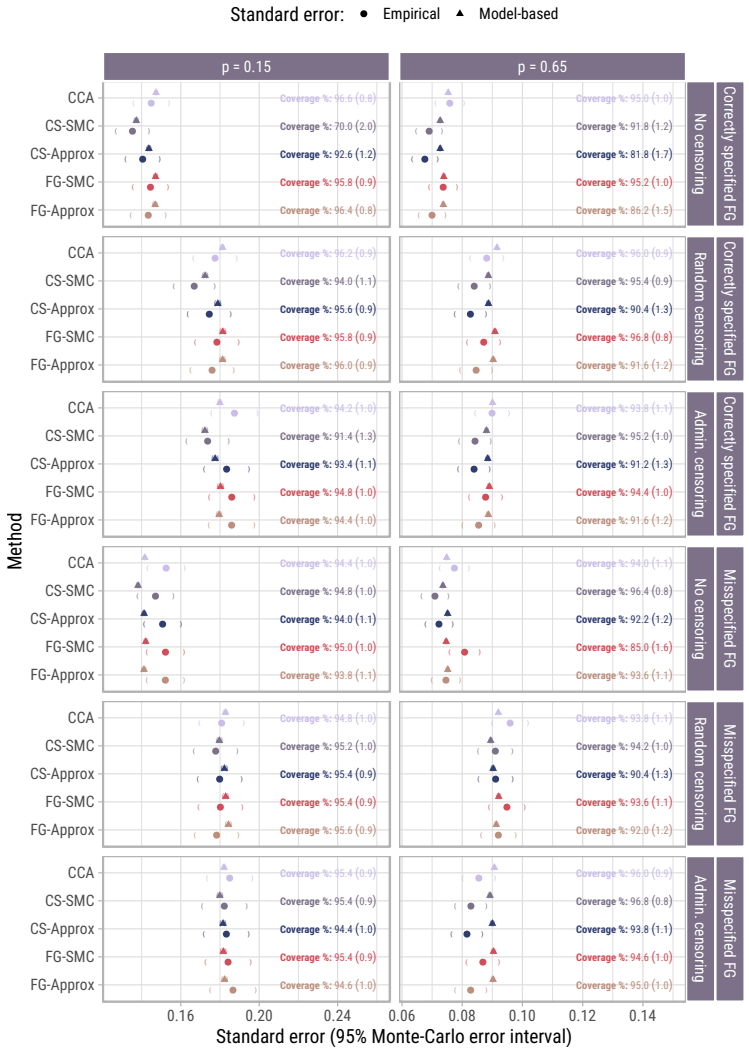

Figure 5.4: Summary of empirical and model-based standard errrors, together with coverage probabilities for β_1 (or $\tilde{\beta}_1$ in scenarios with non-proportional subdistribution hazards). Monte Carlo standard errors (MCSEs) are numerically given in brackets for the coverage probabilities, while the MCSEs for model-based/empirical standard errors are shown by 95% confidence intervals (using standard normal approximation) around a given point. The MCSEs for model-based standard errors are smaller than the graphical width of the point itself, and thus are only visible when zooming-in to the Figure.

When the DGM generated event times under proportional cause-specific hazards, the magnitude of any biases present were in general smaller (e.g. closer to the 5% mark for approximate MI approaches when $p = 0.65$). For the FG-SMC approach, bias was most noticeable when $p = 0.65$, and in the absence of censoring. CS-SMC was unbiased throughout these misspecified Fine–Gray scenarios.

Figure 5.4 summarises empirical and model-based standard errors, together with coverage probabilities for β_1 and $\tilde{\beta}_1$. The model-based standard errors were on average close to their empirical counterparts. CS-SMC appears to have a slight variance advantage over competing approaches, mainly when $p = 0.15$. Interestingly, there was no gain in efficiency when the censoring times were known compared to when they needed to be imputed. This is in line with simulation results in both Fine and Gray (1999) and Ruan and Gray (2008), that compared the censoring complete variance estimator (of subdistribution log hazard ratios) to estimators based on the weighted score function and KM imputation method, respectively. The FG-SMC approach showed good coverage (near the nominal 95% mark) when the Fine–Gray model was correctly specified, although there was slight over-coverage when imputation of censoring times was required. Using the non-parametric bootstrap when estimating $P(C > t)$, which was not investigated in the simulation study, is unlikely to correct for this over-coverage. Under-coverage showed by competing approaches was primarily due to biased estimates.

5.4.4.2 Individual-specific cumulative incidences

Figure 5.5 shows the true and average estimated baseline cumulative incidence function $F_{01}(t)$, the average difference between true and estimated $F_{01}(t)$, and the RMSE of the estimates. Figure 5.6 presents the same information instead for a patient with $\{X, Z\} = \{1, 1\}$. Scenarios where the censoring times are known are omitted from the Figure, as results were indistinguishable from scenarios where the censoring times needed to be imputed.

The cost of imputing compatibly with the wrong model (using CS-SMC when the Fine–Gray model was correctly specified, or FG-SMC when the DGM was based on cause-specific proportional hazards) when estimating $F_{01}(t)$ was only noticeable for the CS-SMC approach in the absence of censoring when $p = 0.15$, in terms of both absolute bias and RMSE. On the whole, the approximately compatible MI approaches performed comparably in terms of RMSE to the SMC approaches. In scenarios where the Fine–Gray model was misspecified, the effect of substantive model misspecification (post-imputation) was clear to see in terms of estimating $F_{01}(t)$ (over- and underestimation at different points in time). In these scenarios, even when CS-SMC is used (which results in the best possible imputations, since it is imputing compatibly with the true data-generating outcome model), assuming proportionality on the incorrect scale at the analysis phase results in biased estimates of the individual-specific cumulative

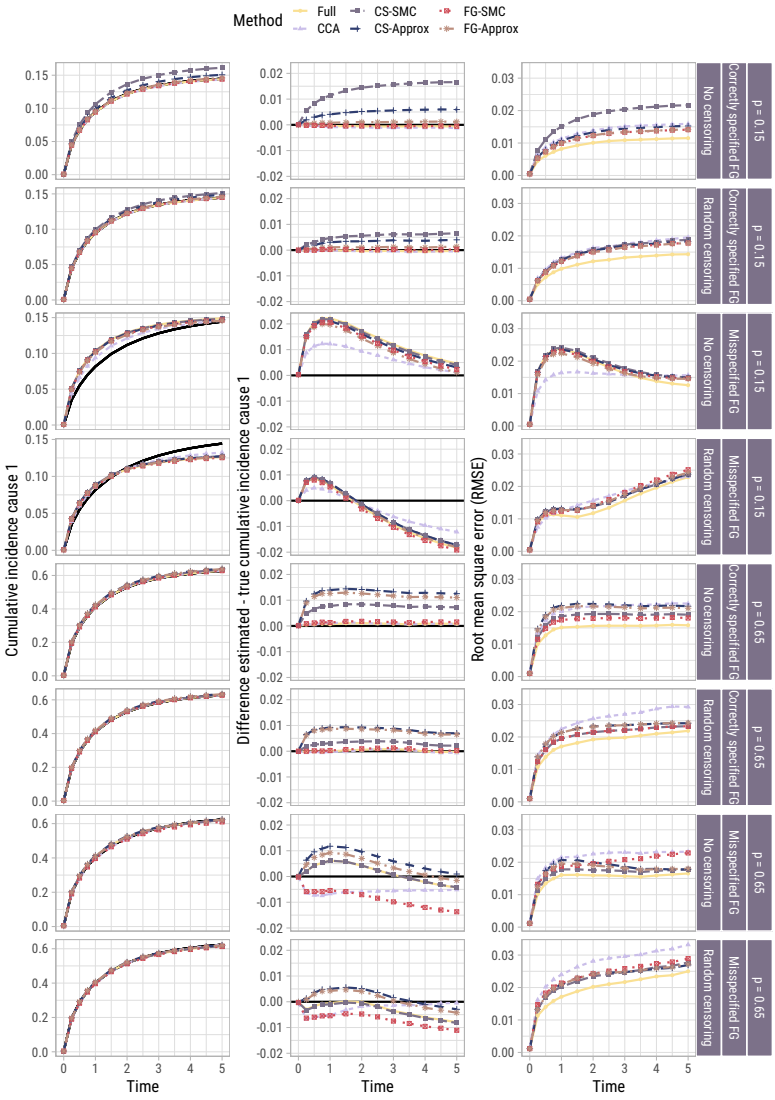

Figure 5.5: Per scenario (row) for a baseline individual $\{X, Z\} = \{0, 0\}$: true (black line) versus estimated cumulative incidence over time, averaged across the 500 replications per scenario (left column); difference between estimated and true (middle column); root mean square error (RMSE) of these estimates (right column). Results for scenarios with administrative censoring are omitted since they were indistinguishable from those with random censoring.

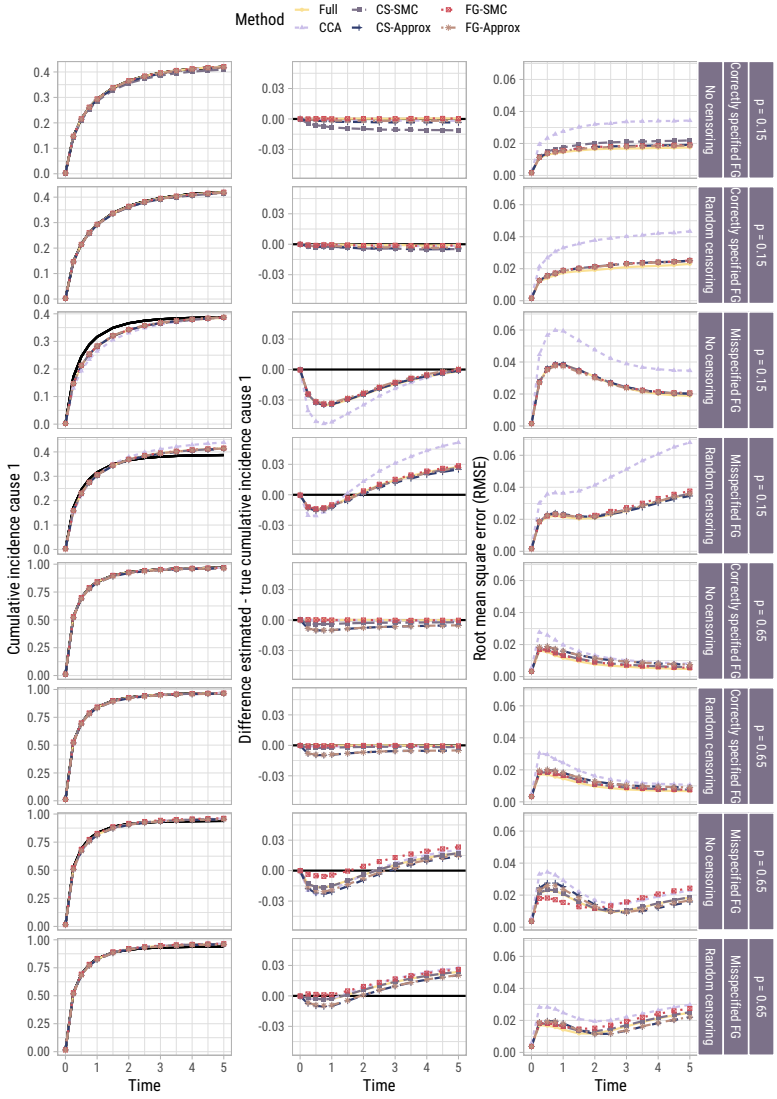

Figure 5.6: Per scenario (row) for individual $\{X, Z\} = \{1, 1\}$: estimated versus true (black line) cumulative incidence over time, averaged across the 500 replications per scenario (left column); difference between estimated and true (middle column); root mean square error (RMSE) of these estimates (right column). Results for scenarios with administrative censoring are omitted since they were indistinguishable from those with random censoring.

incidence function. When $\{X, Z\} = \{1, 1\}$, all imputation approaches outperformed CCA in terms of RMSE when estimating $F_1(t \mid X = 1, Z = 1)$, though to a lesser extent when $p = 0.65$. This can presumably be attributed to the efficiency gain in estimating β_2.

5.4.5 Additional simulations

Two additional sets of simulations were conducted, which build upon the correctly specified Fine–Gray data-generating mechanism with random censoring, with both $p = 0.15$ and $p = 0.65$. The objectives of these additional simulations were to assess the performance of the different imputation methods in settings where a) missingness depends on the observed competing risks outcomes; b) censoring depends on complete covariates, and the model used to impute the potential censoring times could potentially be misspecified.

Covariate imputation approaches were used as previously described in Section 5.4.3, and similarly these additional scenarios are each comprised of 500 simulation replications. The results of these simulations are presented in Appendix B, with a focus on the relative bias in estimating both β_1 and β_2, and further described below.

5.4.5.1 Outcome-dependent missingness

In the first set of simulations, the missingness in X was made to depend on the observed event time T as $\operatorname{logit} P(R_X = 0 \mid T) = \eta_0 + \eta_1 \log(T + 1)$, with $\eta_1 = -1.5$ and η_0 chosen such that 40% of observations in X are missing. This reflects a setting where baseline variables such as genetic information are retrospectively ascertained, and more likely to be available the longer an individual is in follow-up. Since this missingness mechanism depends partially on the failure times for those failing from cause 2, these simulations allow to assess the violation of the MAR assumption made by both FG-SMC and FG-Approx—see Section 5.3.2.1.

To briefly summarise, in Figure 5.9 (Appendix B), we see that the violation of the MAR assumption led to appreciable biases in the estimation of subdistribution log hazard ratios when the proportion of competing events was large (i.e. under $p = 0.15$). In this same scenario, CS-SMC outperformed other methods since it conditions also on the failure time from cause 2, but was still biased as it is imputing compatibly with cause-specific Cox models, which are the incorrect underlying outcome model. CCA was expectedly biased in these scenarios as missingness depended on the outcome.

5.4.5.2 Covariate-dependent censoring

In the second set of simulations, exponential censoring was made covariate-dependent with rate $\lambda_C = 0.49e^Z$, which yields an average censoring proportion which is comparable to the previously reported scenarios with random censoring (approximately 30% censored). In these scenarios, all covariate imputation approaches were applied after multiply imputing the potential censoring times using either a) a marginal KM estimate of the censoring distribution (misspecified censoring distribution); b) a Cox model for the censoring distribution, conditional on Z (correctly specified censoring distribution). The missingness in X also depended on Z, as outlined in Section 5.4.1.4.

In Figure 5.10 (Appendix B), we see that incorrectly specifying the model for the censoring distribution under covariate-dependent censoring led to large biases in the estimation of the subdistribution log hazard ratio for the variable related to the censoring mechanism (β_2 in these simulations). These biases were far less severe under $p = 0.65$, since there are fewer censoring times to impute. Interestingly, in these scenarios CS-SMC does not appear to pay a price for imputing compatibly with the incorrect underlying outcome model. FG-SMC was unbiased throughout when the model for the censoring was correctly specified.

5.5 Applied data example

We illustrate the methods assessed in the simulations study on a dataset of 3982 adult patients with primary and secondary myelofibrosis undergoing a hematopoietic stem cell transplantation (alloHCT) between 2009 and 2019, and registered with the European Society for Blood and Marrow Transplantation (EBMT) (Polverelli *et al.*, 2024). Myelofibrosis is a rare and chronic myeloproliferative neoplasm characterised by bone marrow fibrosis and extramedullary hematopoiesis, for which an alloHCT is the only treatment that can offer long term remission (Kröger *et al.*, 2024). In the original study, the primary objective was to evaluate the association between comorbidities at time of alloHCT and (cause-specific) death without prior relapse of the underlying disease, the so-called non-relapse mortality. In the present illustration, we instead assume that interest lies in developing a prognostic model for time to disease relapse in the first 60 months following an alloHCT. To this end, we developed a Fine–Gray model for relapse, with death prior to relapse as sole competing risk.

A set of 18 baseline predictors were chosen on the basis of substantive clinical knowledge, many of which had a considerable proportion of missing data (see Table 5.1 in Appendix C). These predictors included the 13 variables used in the multivariable models from the original study, and 5 additional variables that were either known to be predictive of disease relapse (use of T-cell depletion; presence of cytogenetic abnormalities), or provided relevant auxiliary information regarding the missing values

(year of transplantation; time between diagnosis and transplantation; and whether diagnosis was primary or secondary myelofibrosis). Note that since this is a model for (complementary log-log transformed) cumulative incidence of relapse, we want to make sure to include predictors known to be associated with the cause-specific hazards of *both* relapse and non-relapse mortality.

Since around 45% of patients were either event-free or censored within the first 60 months (see supplementary material S2.2, non-parametric curves), potential censoring times for those experiencing non-relapse mortality were first multiply imputed using the {kmi} package in strata defined by (completely observed) year of transplantation, yielding 100 datasets with 'complete' subdistribution time V but with partially observed covariate information. In each of these datasets, covariates were imputed once using each of the four imputation methods used in the simulation study, after 20 cycles across the covariates. The choice of 100 imputed datasets was motivated using von Hippel's quadratic rule (i.e. number of imputed datasets needed should increase approximately quadratically with increasing fraction of missing information), based on an initial set of 30 imputed datasets (von Hippel, 2020). Essentially, we sought to control the MCSEs of the standard errors of the estimated subdistribution log hazard ratios. Default imputation methods were used depending on the type of covariate: binary covariates using logistic regression, ordered categorical using proportional odds regression and nominal categorical using multinomial logistic regression. For continuous covariates, the default in {mice} is predictive mean matching, while linear regression is used for $f(X_j \mid X_{-j}, Z; \psi)$ in {smcfcs}. The imputation model for a given partially observed variable therefore contained as predictors all remaining fully and partially observed variables from the substantive model, together with the outcome. Each imputation approach differs mainly in how they incorporate the outcome in the imputation model: either by sampling directly from an assumed substantive model compatible distribution (FG-SMC and CS-SMC), or by including event indicator(s) and marginal cause-specific or subdistribution cumulative hazard(s) explicitly as additional predictors (FG-Approx and CS-Approx).

Figure 5.7 shows for all methods the estimated baseline cumulative incidence function, and the width of the corresponding confidence interval at each timepoint. As was the case in the simulation study, cumulative incidences are estimated in each imputed dataset, and pooled after complementary log-log transformation. The estimation procedure used for the standard errors of the cumulative incidences is described by Ozenne *et al.* (2017). The estimates using both FG-SMC and FG-Approx are virtually overlapping, which is consistent with the simulation study results when $p = 0.15$. Both CS-SMC and CS-Approx also yielded cumulative incidences that were close to those obtained by the subdistribution hazard based imputation approaches, which is in line with the results of the simulation study under random right censoring. The most stark differences were between CCA (which only uses 20% of patients) and the imputation approaches: the cumulative incidence of relapse at 60 months was almost 5% lower than the nearest MI-based curve, with confidence intervals that

were over twice as wide. For completeness, in supplementary material S2.3 we report the pooled subdistribution log hazard ratios, in addition to the pooled coefficients of cause-specific Cox models for relapse and non-relapse mortality (each containing the same predictors as the Fine–Gray model for relapse). The pooled coefficients of the Fine–Gray models were extremely similar between imputation approaches, and all differed considerably from the (much more variable) CCA. There were some noticeable differences between subdistribution hazard based and cause-specific hazard based imputation approaches when estimating the cause-specific Cox model for non-relapse mortality (see e.g. pooled coefficients for weight loss prior to transplantation, hemoglobin or high risk comorbidity score). Furthermore, the pooled subdistribution log hazard ratios were generally small in magnitude (none exceeding 0.5), a setting in which both SMC and approximately compatible approaches are expected to perform similarly.

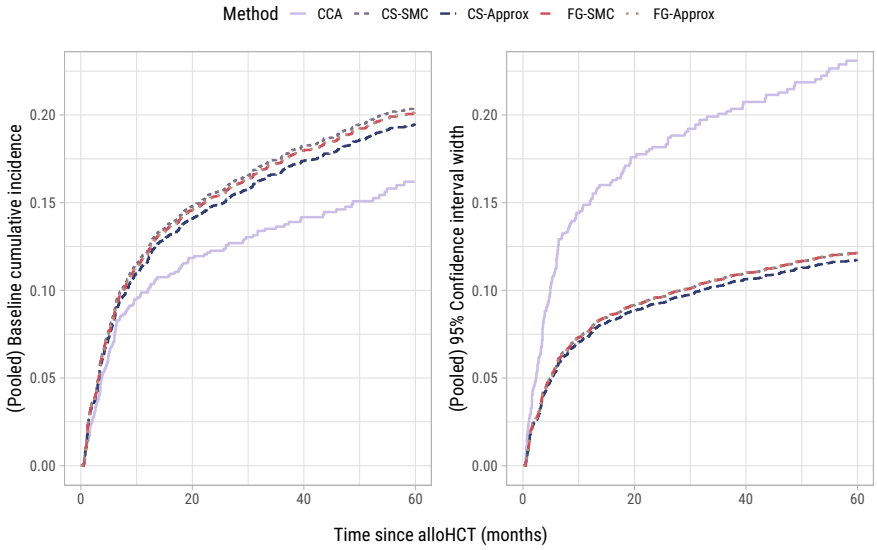

Figure 5.7: Pooled baseline cumulative incidence functions for relapse in the applied data example (left panel), and width of corresponding confidence intervals (right panel). These are the estimates for a patient aged 60, transplanted in 2019 immediately after diagnosis, with 10g/dL hemoglobin, 15 x10^9/L white blood cells, no peripheral blood blasts, and with reference levels for all categorical predictors (see Table 5.1).

The differences observed between point estimates obtained using the imputation based approaches and CCA are in large part explainable by the gulf in efficiency between the two approaches. Nevertheless, there are indications that the estimates obtained

using imputation methods could be less biased than their CCA counterparts in this example. An exploratory logistic model showed that the observed time to competing event and competing event indicator were both predictive of the probability of being a complete-case, after adjusting for other known important predictors of missingness such as year of transplantation (many variables recorded more often later on in time as their clinical relevance became clearer). Upon closer inspection, it appears that the probability of being a complete-case is significantly lower only for those censored earlier on in time. This seemingly unlikely association between future outcome and baseline complete-case indicator (outcome-dependent MAR, under which CCA is biased unless the missingness is related solely to the censoring process—see Rathouz, 2007) is likely confounded by transplant centre. That is, shorter follow-up times and missing values in covariates may both be symptomatic of a given centre's overall quality of data collection. Although ignored in the present analysis for simplicity, there is indeed heterogeneity in data completeness between EBMT affiliated transplant centres across and within different countries. The MI of potential censoring times would allow to model centre effects using standard software, for example by means of stratification or use of a frailty term.

5.6 Discussion

In this paper, we extended the SMC-FCS approach in order to impute missing covariates compatibly with a Fine–Gray substantive model. For a given competing event, the theory relies on using the subdistribution time V and the corresponding event-specific indicator as outcome variables. In the presence of random right-censoring, V is only partially observed, as the potential censoring times for those failing from competing events are unknown. These can be multiply imputed in a first step, after which covariates can be imputed by conditioning on the 'complete' outcome variables. The approach is straightforward to implement in R by making use of existing software packages {kmi} and {smcfcs}. While the imputation of potential censoring times appears underused in the subdistribution hazard modelling literature (relative to weighted approaches), it has inspired other methodological extensions e.g. enabling the use of deep learning in discrete time after single imputation of potential censoring times (Gorgi Zadeh *et al.*, 2022).

The simulation study compared the performance of the proposed method to competing MI approaches, including imputing compatibly with cause-specific proportional hazards models. The FG-SMC approach performed optimally (in terms of estimating both subdistribution log hazard ratios, and cumulative incidences) when the assumption of proportional subdistribution hazards held, and performed satisfactorily when this assumption did not hold. For cumulative incidence estimation, the choice of substantive model (i.e. cause-specific Cox vs. Fine–Gray) at the analysis phase appears

to be more important than the procedure used to impute the missing covariates. In terms of RMSE of these predictions, most imputation approaches outperform CCA. The applied data example also demonstrated the possible gain in efficiency when using MI instead of CCA.

One counterintuitive finding was that the presence of censoring seems to *improve* the performance of the misspecified SMC-FCS procedure (e.g. use of CS-SMC when underlying DGM assumes proportional subdistribution hazards). An explanation for this phenomenon is that the time-dependent factor relating the cause-specific and subdistribution hazards for cause 1 (the 'reduction factor,' Putter *et al.*, 2020) is closer to 1 earlier in time. Therefore (in the example with DGM assuming proportional subdistribution hazards), the violation of proportionality on the cause-specific hazard scale will appear to be less severe in earlier time-periods, thereby improving the performance of the misspecified SMC-FCS approach. This is also in line with earlier findings showing how similar the results of subdistribution and cause-specific hazards models can be in presence of heavy censoring (Grambauer *et al.*, 2010; van der Pas *et al.*, 2018). Notwithstanding, the additional simulations in Section 5.4.5 emphasise the importance of appropriately accounting for covariates related to the censoring process when modelling the subdistribution hazard (where in practice, completely random censoring is the default assumption—see Beyersmann *et al.*, 2012), as also discussed in previous work (Donoghoe and Gebski, 2017).

An advantage to the proposed approach is that it can be extended in various ways. For example, the approach can account for time-dependent effects, by making direct use of existing approaches developed in the context of standard Cox models (Keogh and Morris, 2018). Additionally, the proposed approach can be extended to accommodate interval censored outcomes, using the methodology described by Delord and Génin (2016), which relies on analogous principles: multiply impute interval censored V in order to work with simpler censoring complete data.

There are multiple limitations to the present work. The first is that the proposed SMC-FCS approach does not accommodate delayed entry (left truncation). Our current recommendation to impute approximately compatibly with a Fine–Gray model subject to delayed entry and right-censoring is to include $I(D = 1)$ and $\hat{\Lambda}_1(T)$ as predictors in the imputation model, in addition to other substantive model covariates. Here, $\hat{\Lambda}_1(t)$ is the estimated cumulative subdistribution hazard based on a marginal model that uses time-dependent weights in order to accommodate both left-truncation and right-censoring (Geskus, 2011). Note the proposed imputation model uses $\hat{\Lambda}_1(T)$ and not $\hat{\Lambda}_1(V)$, and therefore some downward bias is to be expected, as explained in Appendix A. Second, while FG-SMC does not require an explicit model for the competing risks, it does require the censoring distribution to be specified explicitly (e.g. non-parametrically using KM, or using a Cox model). Third, the proposed approach is geared towards imputing missing covariates when only one competing event is of interest. More generally, the strategy of estimating a Fine–Gray for each cause in turn

is not an approach the current authors endorse, based on both theoretical (Austin *et al.*, 2021; Beyersmann *et al.*, 2012) and simulation-based arguments (Bonneville *et al.*, 2024). When multiple competing events are of interest, we would instead recommend modelling the cause-specific hazards, or using the semiparametric approach suggested by Mao and Lin (2017) for joint inference on the cumulative incidence functions.

In conclusion, the proposed approach is most appropriate for imputing missing covariates in the context of prognostic modelling of only one event of interest. Based on the simulation study, imputing compatibly with cause-specific proportional hazards seems to be a good all-round strategy for a 'complete' competing risks analysis (investigating both the cause-specific hazards and cumulative incidence functions, see Latouche *et al.*, 2013), and can at the same time be used for prognostic modelling based on the cause-specific Cox models.

Supplementary materials

Supplementary materials for this work are available at https://arxiv.org/abs/2405.16602.

All R code (needed to reproduce simulation study, applied data example, and manuscript figures) is available at https://github.com/survival-lumc/FineGrayCovarMI. In addition to the minimal R code provided in the supplementary materials, a wrapper function for the proposed SMC-FCS Fine–Gray method is available inside the {smcfcs} R package.

Appendix A: Imputed censoring times, and resulting cumulative subdistribution hazards

As described in Section Section 5.3.3.1, the subdistribution time V is only partially observed in the presence of random right-censoring. Thus, the potential censoring times for those failing from cause 2 should first be multiply imputed, before imputing any missing covariates. This imputation of partially observed V is visualised more closely in Figure 5.8, using a simulated dataset of 2000 individuals following the parametrisation used in the simulation study scenario with correctly specified Fine–Gray, $p = 0.65$, and random exponential censoring. In this example, the potential censoring times for those failing from cause 2 were imputed $m = 10$ times.

The upper panel shows the imputed potential censoring times for a random selection of 20 individuals failing from cause 2, in addition to their cause 2 failure time and their true eventual censoring time. The lower panel shows the estimated marginal

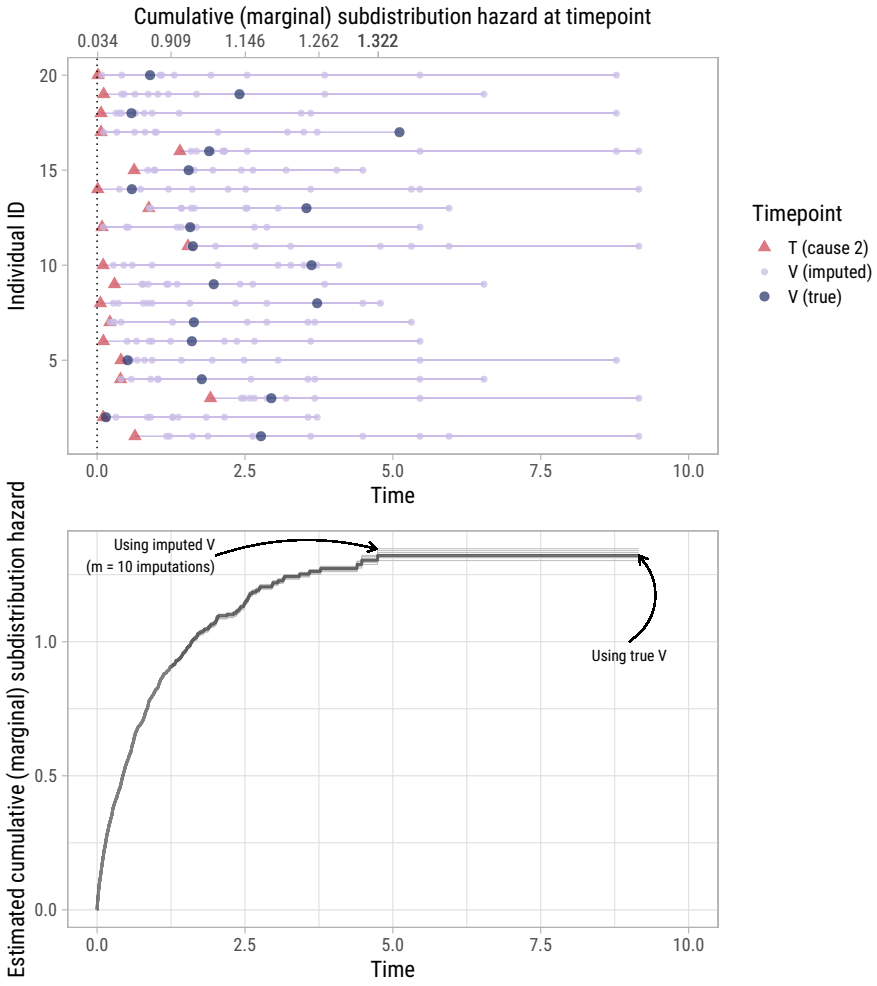

Figure 5.8: Based on a simulated dataset of $n = 2000$ (correctly specified Fine–Gray, $p = 0.65$, random censoring), we show the imputed ($m = 10$ imputations) potential censoring times for a random selection of 20 individuals failing from cause 2 (upper panel); and the estimated marginal cumulative subdistribution hazard function for cause 1 based on true V, and based on imputed V (lower panel).

cumulative subdistribution hazard function for $\hat{\Lambda}_1(t)$ resulting from using $I(D = 1)$ together with either the imputed or true V as outcomes in a marginal model. We used $\hat{\Lambda}_1(t)$ estimated using the true V to create the secondary x-axis in the upper panel, which shows the value of this function at a given timepoint. For example, the marginal cumulative subdistribution hazard was 1.146 at timepoint 2.5, and stayed constant at 1.322 after the last cause 1 event in this sample.

The upper panel in particular gives additional insights regarding the FG-Approx method, where $I(D = 1)$ and $\hat{\Lambda}_1(V)$ are included as predictors in the imputation model. Namely, the secondary x-axis shows the value of $\hat{\Lambda}_1(V)$ used in the imputation model for a missing X_j, for given imputed V. A first key point is that one should always use $\hat{\Lambda}_1(V)$ in the imputation model, and not $\hat{\Lambda}_1(T)$. Since the observed cause 2 failure time occurs before the eventual censoring time, $\hat{\Lambda}_1(T)$ will always be smaller than the marginal cumulative subdistribution hazard at the eventual censoring time. Since $\hat{\Lambda}_1(T)$ and $\hat{\Lambda}_1(V)$ are not proportional to each other, the imputation model will incur some bias. A second point is that in settings with fewer event 1 failures (e.g. $p = 0.15$ scenario in the simulation study), the corresponding secondary x-axis will have a smaller range, since the subdistribution hazard will be lower overall. Using $\hat{\Lambda}_1(T)$ instead of $\hat{\Lambda}_1(V)$ may therefore have a more limited impact. However, as evidenced by the simulations in Section 5.4.5, misspecification the censoring distribution impacts inferences more when $p = 0.15$, since there are more censoring times to impute.

The lower panel shows that the estimated $\hat{\Lambda}_1(t)$ varies very little between imputed datasets, with differences only being noticeable later on in follow-up as risk sets become smaller and associated cumulative hazard jumps more pronounced. Note also that while $\hat{\Lambda}_1(t)$ based on the true V appears in this dataset to be a kind of 'average' of the functions based on imputed V, this will not be the case in general, especially with smaller sample sizes. The $\hat{\Lambda}_1(t)$ based on the weighted estimator (Geskus, 2011) will however coincide with the 'average' of the functions based on imputed V, as will using the negative log of one minus the Aalen–Johansen estimate of the marginal cumulative incidence function.

Appendix B: Additional simulations

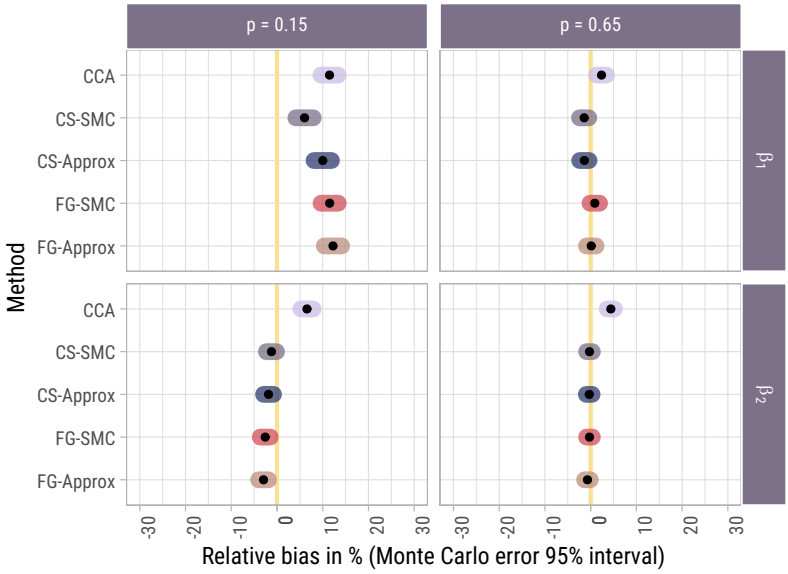

Figure 5.9: Relative bias (%) in estimating $\{\beta_1, \beta_2\} = \{0.75, 0.5\}$, with corresponding 95% Monte Carlo confidence interval (constructed using the standard normal approximation). These are additional simulations under the correctly specified Fine–Gray data-generating mechanism with random censoring, with both $p = 0.15$ and $p = 0.65$. The missingness in X was made to depend on the observed event time T as $\text{logit}\, P(R_X = 0 \mid T) = \eta_0 + \eta_1 \log(T + 1)$, with $\eta_1 = -1.5$ and η_0 chosen such that 40% of observations in X are missing.

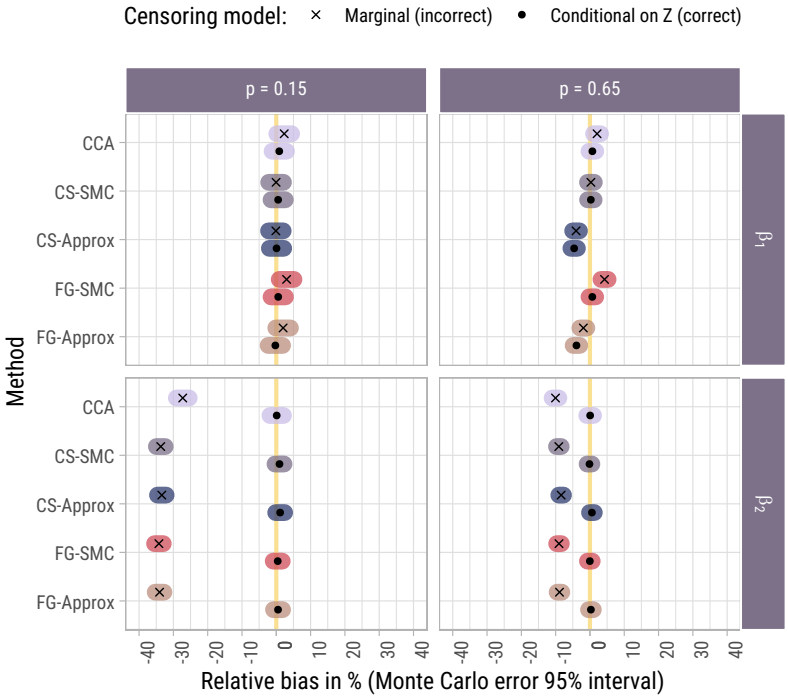

Figure 5.10: Relative bias (%) in estimating $\{\beta_1, \beta_2\} = \{0.75, 0.5\}$, with corresponding 95% Monte Carlo confidence interval (constructed using the standard normal approximation). These are additional simulations under the correctly specified Fine–Gray data-generating mechanism with random censoring, with both $p = 0.15$ and $p = 0.65$. The censoring was made covariate-dependent with rate $\lambda_C = 0.49e^Z$, and all covariate imputation approaches were applied after multiply imputing the potential censoring times using either a) a marginal (incorrect) Kaplan–Meier estimate of the censoring distribution; b) a Cox model for the censoring distribution, conditional on Z (correct). The missingness in X here also depended on Z.

Appendix C: Data dictionary

Table 5.1: Data dictionary. The median and interquartilerange are reported for continuous variables, and for categorical variables the counts and proportion per level are reported. The difference between this cohort (n = 3982) and the one reported in Chapter 4 (n = 4086), is that the present cohort is selected on available outcome data (i.e. no missing relapse times). CMV: cytomegalovirus; HLA: human leukocyte antigen; HCT-CI: Hematopoietic stem cell transplantation-comorbidity index; MF: myelofibrosis.

Variable or level	N = 3982
Patient age (years)	58 (52, 64)
Patient/donor CMV match	
Patient negative/Donor negative	1,142 (30%)
Other	2,715 (70%)
(Missing)	125
Donor type	
HLA identical sibling	1,183 (30%)
Other	2,795 (70%)
(Missing)	4
Hemoglobin (g/dL)	9.10 (8.10, 10.40)
(Missing)	1,873
HCT-CI risk category	
Low risk (0)	1,674 (54%)
Intermediate risk (1 – 2)	743 (24%)
High risk (≥ 3)	674 (22%)
(Missing)	891
Interval diagnosis-transplantation (years)	3 (1, 9)
Karnosfky performance score	
≥ 90	2,475 (66%)
80	986 (26%)
≤ 70	267 (7.2%)
(Missing)	254
Patient sex	
Female	1,484 (37%)
Male	2,498 (63%)
Peripheral blood (PB) blasts (%)	1.0 (0.0, 3.0)
(Missing)	2,323
Conditioning	
Standard	1,373 (35%)
Reduced	2,553 (65%)
(Missing)	56
Ruxolitinib given	
No	1,832 (66%)
Yes	931 (34%)
(Missing)	1,219
Disease subclassification	
Primary MF	2,912 (73%)
Secondary MF	1,070 (27%)
Night sweats	
No	1,256 (70%)
Yes	529 (30%)

(Missing)	2,197
T-cell depletion (in- or ex-vivo)	
No	1,012 (26%)
Yes	2,905 (74%)
(Missing)	65
Cytogenetics	
Normal	1,318 (59%)
Abnormal	910 (41%)
(Missing)	1,754
White blood cell count (WBC, $x10^9$/L)	7 (4, 14)
(Missing)	1,884
>10% Weight loss prior to transplantation	
No	1,329 (73%)
Yes	492 (27%)
(Missing)	2,161
Year of transplantation	2,015.0 (2,012.0, 2,018.0)

Chapter 6

Why you should avoid using multiple Fine–Gray models: insights from (attempts at) simulating proportional subdistribution hazards data

Chapter based on: **Bonneville, E. F.**, de Wreede, L. C. and Putter, H. (2024) Why you should avoid using multiple Fine–Gray models: insights from (attempts at) simulating proportional subdistribution hazards data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae056. DOI: 10.1093/jrsssa/qnae056.

Abstract

Studies considering competing risks will often aim to estimate the cumulative incidence functions conditional on an individual's baseline characteristics. While the Fine–Gray subdistribution hazard model is tailor-made for analysing only one of the competing events, it may still be used in settings where multiple competing events are of scientific interest, where it is specified for each cause in turn. In this work, we provide an overview of data-generating mechanisms where proportional subdistribution hazards hold for at least one cause. We use these to motivate why the use of multiple Fine–Gray models should be avoided in favour of better alternatives such as cause-specific hazard models.

6.1 Introduction

Competing risks are ubiquitous across medical studies, where patients can experience one of several distinct events, such as death due to different causes. One of the typical aims for a study considering competing risks is to estimate the model-based cumulative probabilities of experiencing one or more of the competing events by a certain time for specific patients (the predicted cumulative incidence functions). Presently, given a set of baseline covariates and outcome data, cumulative incidence functions are generally estimated using either the Fine–Gray subdistribution hazard model (Fine and Gray, 1999), or by fitting and thereafter combining cause-specific Cox proportional hazards models for each event (Putter *et al.*, 2007). The former is often the tool of choice when developing prognostic models for a single event of interest, as it does not require explicitly specifying models for the competing events. It is also arguably simpler to externally validate, requiring only cumulative subdistribution baseline hazard estimates at relevant timepoints together with the estimated regression coefficients (instead of the full cause-specific cumulative hazards for all events).

Even when only one of the endpoints is of primary interest, multiple authors have argued in favour of a more holistic approach to competing risks analyses, suggesting that *all* events should be studied together (Andersen *et al.*, 2012; Gerds *et al.*, 2012; Latouche *et al.*, 2013). For example, as emphasised by Latouche *et al.* (2007), the effect of a given covariate on the cumulative incidence of one event should not be considered in isolation from its effect on the cumulative incidence of competing events. Consider patients with a malignant haematological disease, that are at risk of (competing) disease relapse and non-relapse mortality after undergoing an allogeneic stem cell transplantation (alloSCT). A less intensive pre-transplantation conditioning regimen may not be able to sufficiently control the disease compared with a more intensive regimen (i.e. increased risk of relapse), but it will be less toxic for the patient (i.e. reduced risk of non-relapse mortality)—see Shimoni *et al.* (2016) for an example of opposing effects of conditioning regimen on the cumulative incidence functions of relapse and non-relapse mortality.

In a context where more than one of the competing events are of (possibly equal) interest, one may opt to fit a Fine–Gray model for *each* competing event in turn. Since these models are fitted independently, it is possible that the sum of the estimated cumulative incidence functions (the total failure probability, TFP) for given covariate values at certain timepoints exceeds 1. This was recently illustrated by Austin *et al.* (2021) with data considering cardiovascular and non-cardiovascular death, where the TFP exceeded 1 for 5% of patients at 5 years after hospital admission. This known issue of the TFP exceeding 1 occurs partly as a result of at least one of the specified Fine–Gray models being incorrect (Beyersmann *et al.*, 2012). That is, the assumption of proportional subdistribution hazards fails to hold for at least one of the models.

A useful starting point for further understanding these issues when using multiple Fine–Gray models is to consider a simplified context with two competing risks, and suppose that the Fine–Gray model has been correctly specified for one event (cause 1). The objective of this article is to outline the implications (i.e. the implied assumptions) of specifying a Fine–Gray model for cause 1 on the cumulative incidence function for cause 2. To do so, we provide an overview of data-generating mechanisms (DGMs) where the Fine–Gray model is correctly specified for at least cause 1. For these DGMs, we (a) give example specifications (e.g. choice of distributions), (b) refer to instances where they have been used across the methodological competing risks literature (e.g. as part of simulation studies), if at all, and (c) touch upon potential difficulties from a simulator's perspective. While these DGMs are particularly relevant for methodological researchers aiming to simulate competing risks data under different assumptions, they provide additional insights for applied researchers seeking to motivate their choice of analysis method(s) in a more principled way. In the discussion, we therefore reflect on what the characteristics of the outlined DGMs imply for the use of multiple Fine–Gray models in practice, and argue in favour of cause-specific hazard models for cumulative incidence prediction.

6.2 Competing risks and the Fine–Gray model

In a competing risks setting, we assume that individuals can experience only one of K distinct events or, phrased differently, that only the first event of interest is observed. We denote the failure time as T, and the competing event indicator as $D \in \{1, ..., K\}$. In practice, individuals are subject to a right-censoring time C (generally assumed independent of T and D), and we thus only observe realisations of $\tilde{T} = \min(T, D)$ and $\tilde{D} = I(T \leq C)D$. The cause-specific hazard for the k^{th} event is defined as

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t}.$$

These hazards fully define the event-free survival function,

$$P(T > t) = \exp\left(- \sum_{k=1}^{K} \int_0^t h_k(u)\, \mathrm{d}u \right) = \exp\left(- \sum_{k=1}^{K} H_k(t) \right),$$

assuming the distribution of T is continuous, and $H_k(t)$ is the cause-specific cumulative hazard function for the k^{th} event. The cause-specific cumulative incidence function is then defined as

$$F_k(t) = P(T \leq t, D = k) = \int_0^t h_k(u) S(u-)\, \mathrm{d}u,$$

where $S(u-)$ is the event-free survival probability just prior to u.

A relevant question is whether we can model $F_k(t)$ directly, without needing to model all cause-specific hazards. In order to do so, the idea is to specify a hazard for the k^{th} event, $\lambda_k(t)$, that satisfies

$$P(T \leq t, D = k) = 1 - \exp\left(-\int_0^t \lambda_k(u)\,\mathrm{d}u\right),$$

analogously to the standard single-event survival setting. Rearranging the above yields

$$\begin{aligned}\lambda_k(t) &= \frac{-\,\mathrm{d}\log\{1 - F_k(t)\}}{\mathrm{d}t}, \\ &= \frac{\mathrm{d}F_k(t)}{\mathrm{d}t} \times \{1 - F_k(t)\}^{-1},\end{aligned}$$

which is the commonly known expression for the subdistribution hazard. It can also be written as $\lambda_k(t) = f_k(t)/\{1 - F_k(t)\}$, where $f_k(t) = \mathrm{d}F_k(t)/\,\mathrm{d}t$ is referred to as the subdensity function (Gray, 1988). $F_k(t)$ is not a true distribution function since $F_k(\infty) = P(D = k) < 1$, and is hence known as a 'subdistribution' function. The cause-specific hazard can also be written in terms of the subdensity function, as $h_k(t) = f_k(t)/S(t)$. Thus, the cause-specific hazard conditions on being event-free by t, while the subdistribution hazard conditions on not having failed by event k by t.

The Fine–Gray model is a semiparametric model that assumes proportionality on the subdistribution hazard scale. Using covariate vector $\mathbf{Z}$, the Fine–Gray model for cause k can be written as

$$\lambda_k(t \,|\, \mathbf{Z}) = \lambda_{k0}(t) \exp(\boldsymbol{\beta}_k^\top \mathbf{Z}),$$

with $\lambda_{k0}(t)$ being the subdistribution baseline hazard function and $\boldsymbol{\beta}_k$ representing the effects of covariates $\mathbf{Z}$ on the subdistribution hazard. The cumulative incidence function for the k^{th} event can then be written as

$$F_k(t \,|\, \mathbf{Z}) = 1 - \exp\left\{-\exp(\boldsymbol{\beta}_k^\top \mathbf{Z}) \int_0^t \lambda_{k0}(u)\,\mathrm{d}u\right\},$$

which corresponds to modelling the cumulative incidence function with a complementary log-log transformation. Furthermore, let $F_{k0}(t) = 1 - \exp(-\int_0^t \lambda_{k0}(u)\,\mathrm{d}u)$ be the baseline cumulative incidence function, i.e. the cumulative incidence when $\mathbf{Z} = 0$. We can then also write the Fine–Gray model as

$$1 - F_k(t \,|\, \mathbf{Z}) = \{1 - F_{k0}(t)\}^{\exp(\boldsymbol{\beta}_k^\top \mathbf{Z})},$$

which is a similar relation to that of the survival functions in a Cox model.

6.3 Data-generating mechanisms

For the sake of simplicity, we restrict ourselves to $K = 2$ competing events, a single time-constant covariate X, and assume that cause 1 is of primary interest. Results can be generalised to more complex settings. We let $h_k(t \mid X)$, $\lambda_k(t \mid X)$ and $F_k(t \mid X) = P(T \leq t, D = k \mid X)$, respectively, denote the cause-specific hazard, subdistribution hazard and cumulative incidence function for cause k, conditional on X. Furthermore, we let β_k and γ_k represent the effect of X on the subdistribution hazard and cause-specific hazard of cause k, respectively. In what follows, we present DGMs for which a Fine–Gray model correctly holds for cause 1. When illustrating the different DGMs, X is assumed to be binary.

In essence, the task is to specify a joint density $f(T, D \mid X)$ where the Fine–Gray model is correctly specified for cause 1. With $S(T \mid X) = \exp\{-H_1(T \mid X) - H_2(T \mid X)\}$, we have

$$\begin{aligned} f(T, D \mid X) &= \{h_1(T \mid X)S(T \mid X)\}^{I(D=1)}\{h_2(T \mid X)S(T \mid X)\}^{I(D=2)}S(T \mid X)^{1-I(D=1)-I(D=2)}, \\ &= f_1(T \mid X)^{I(D=1)} f_2(T \mid X)^{I(D=2)}\{1 - F_1(T \mid X) - F_2(T \mid X)\}^{1-I(D=1)-I(D=2)}, \end{aligned}$$

as written by Andersen and Ravn (2023). Since the Fine–Gray model provides an expression for $F_1(t \mid X)$, we need only think about what assumptions to make regarding cause 2. Additionally, the hazard functions comprising the above density should fulfil various restrictions, which are outlined in the work of Haller and Ulm (2014). Namely,

1. All hazard functions must be non-negative for all time points $t > 0$.
2. The cause-specific and subdistribution hazards (for the event of interest) should be identical before the occurrence of the first competing event. Therefore, $h_1(t \mid X) = \lambda_1(t \mid X)$ at $t = 0$.
3. $F_1(t \mid X)$ must converge to $P(D = 1 \mid X)$ as $t \to \infty$. If $P(D = 1 \mid X) < 1$, this in turn implies that $\lim_{t\to\infty} \lambda_1(t \mid X) = 0$ and even that the cumulative subdistribution hazard $\Lambda_1(t \mid X)$ should not go to infinity for $t \to \infty$.

6.3.1 Using the reduction factor

A first approach to specifying $f(T, D \mid X)$ is to make assumptions about all cause-specific hazard functions. That is, we would like to select a set of cause-specific hazard functions for which proportionality holds on the subdistribution hazard of event 1. To do so, we can make use of the link between $h_1(t \mid X)$ and $\lambda_1(t \mid X)$, which is given by

$$\lambda_1(t \mid X) = h_1(t \mid X)\frac{S(t \mid X)}{1 - F_1(t \mid X)}, \tag{6.1}$$

with the latter expression referred to as the *reduction factor* by Putter *et al.* (2020). In the book by Beyersmann *et al.* (2012) (Equation 5.3.9), this has also been written as

$$h_1(t \mid X) = \lambda_1(t \mid X) \left\{ 1 + \frac{F_2(t \mid X)}{S(t \mid X)} \right\},$$

which holds since $S(t \mid X) = 1 - \sum_{k=1}^{2} F_k(t \mid X)$. The above expressions allow to simulate data by specifying $\lambda_1(t \mid X)$ and one of $h_1(t \mid X)$, $h_2(t \mid X)$, or $h_1(t \mid X) + h_2(t \mid X)$, and thereafter deriving the implied cause-specific hazard(s). With both cause-specific hazards being defined, one should be able simulate using standard methods, i.e. with latent times or using the all-cause hazard function (Beyersmann *et al.*, 2009).

6.3.1.1 Specifying $\lambda_1(t \mid X)$ and $h_2(t \mid X)$

Since we assume that the Fine–Gray model is correctly specified for $\lambda_1(t \mid X)$, we can express our assumptions regarding cause 2 for instance by specifying a model for $h_2(t \mid X)$, which can be any hazard-based regression model (e.g. cause-specific Cox, additive hazards, or other), and derive the implied $h_1(t \mid X)$. By re-arranging Equation 6.1 and thereafter integrating with respect to t, we can write

$$\overbrace{h_1(t \mid X) \exp\{-H_1(t \mid X) - H_2(T \mid X)\}}^{f_1(t \mid X)} = \lambda_1(t \mid X) \exp\{-\Lambda_1(t \mid X)\},$$

$$\exp\{-H_1(t \mid X)\} = 1 - \int_0^t \lambda_1(u \mid X) \exp\{-\Lambda_1(u \mid X) + H_2(u \mid X)\} \, \mathrm{d}u.$$

It then follows that, given choices of $\lambda_1(t \mid X)$ and $h_2(t \mid X)$, the implied cause-specific hazard for event 1 is given by

$$h_1(t \mid X) = \frac{\lambda_1(t \mid X) \exp\{-\Lambda_1(t \mid X) \ + H_2(t \mid X)\}}{1 - \int_0^t \lambda_1(u \mid X) \exp\{-\Lambda_1(u \mid X) \ + H_2(u \mid X)\} \, \mathrm{d}u}. \tag{6.2}$$

The above expression has the advantage of naturally ensuring that $\lambda_1(t \mid X) = h_1(t \mid X)$ at $t = 0$. However, depending on the choices of $\lambda_1(t \mid X)$ and $h_2(t \mid X)$, the implied $h_1(t \mid X)$ may become negative at certain timepoints. Specifically, this occurs when the integral in the denominator (corresponding to $1 - \exp\{-H_1(t \mid X)\}$, the 'net risk' for cause 1) in Equation 6.2 exceeds 1. Another way of looking at these potentially negative hazard values is to express the event-free survival $S(t \mid X)$ in terms of $\lambda_1(t \mid X)$ and $h_2(t \mid X)$, as

$$S(t \mid X) = \left[1 - \int_0^t \lambda_1(u \mid X) \exp\{-\Lambda_1(u \mid X) \ + H_2(u \mid X)\} \, \mathrm{d}u \right] \times \exp\{-H_2(t \mid X)\}. \tag{6.3}$$

Therefore, when the implied net risk of cause 1 exceeds one (generally as a result of excessively large cause-specific hazard for cause 2), the event-free survival itself become negative, meaning that $1 - S(t \mid X)$ (the TFP) exceeds 1.

To illustrate this DGM, we specify

$$\lambda_1(t \mid X) = \nu_1 e^{\kappa_1 t} \exp(\beta_1 X),$$

where κ_1 and ν_1 respectively are the shape and rate parameters of a Gompertz baseline hazard. Fixing $\beta_1 = 0.5$, and choosing a negative shape $\kappa_1 = -2$, and rate $\nu_1 = 0.5$, we have that $P(D = 1 \mid X) = 1 - [1 - \{1 - \exp(\nu_1/\kappa_1)\}]^{\exp(\beta_1 X)}$. This asymptote of the Gompertz cumulative distribution function, which is less than 1 when the shape parameter is negative, has made it an attractive choice of distribution in work investigating direct parametric modelling of cumulative incidence functions (Jeong and Fine, 2006). For cause 2, we assume cause-specific proportional hazards as

$$h_2(t \mid X) = a_2 b_2 t^{a_2 - 1} \exp(\gamma_2 X),$$

where a_2 and b_2 are respectively the shape and rate parameters of a Weibull baseline hazard. We fix $\{a_2, b_2, \gamma_2\} = \{0.5, 1.25, 0.25\}$, and derive $h_1(t \mid X)$ using Equation 6.2. Figure 6.1 depicts the true (obtained with numerical integration) stacked cumulative incidence functions and cause-specific hazards (conditional on $X = 1$) for both causes, as well as the implied subdistribution hazard ratios. We see that this mechanism (when $X = 1$) is only properly defined prior to $t \approx 3.20$, after which $h_1(t \mid X = 1)$ is negative. The corresponding cumulative incidence functions demonstrate that beyond this timepoint, there is no probability space left to fill, as the TFP has already reached one. Additionally, panel C from Figure 6.1 emphasises that proportional subdistribution hazards do not hold for cause 2.

Making appropriate use of this approach for simulation purposes means paying attention to the fact that the choice $h_2(t \mid X)$ will have to respect the remaining probability space left over by the proportional subdistribution hazards structure assumed by cause 1. Practically speaking, this means specifying a $h_2(t \mid X)$ such that the implied event-free survival in Equation 6.3 does not become negative. One may also choose to set a maximum follow-up time, before which all hazards behave appropriately for all X and the TFP is smaller than 1. In Figure 6.1, this could be achieved by setting a maximum follow-up time smaller or equal to 3.20 (or adjusting the parameters of the hazards function in order to allow a larger maximum follow-up time). An example use of this DGM is found in the work of Lambert *et al.* (2017), as part of a simulation study assessing the performance of a proposed flexible parametric approach for modelling the subdistribution hazard of one event. Both $\lambda_1(t \mid X)$ and $h_2(t \mid X)$ assumed proportional hazards with mixture Weibull baseline hazards, and the maximum (simulated) follow-up time was set to 5 years.

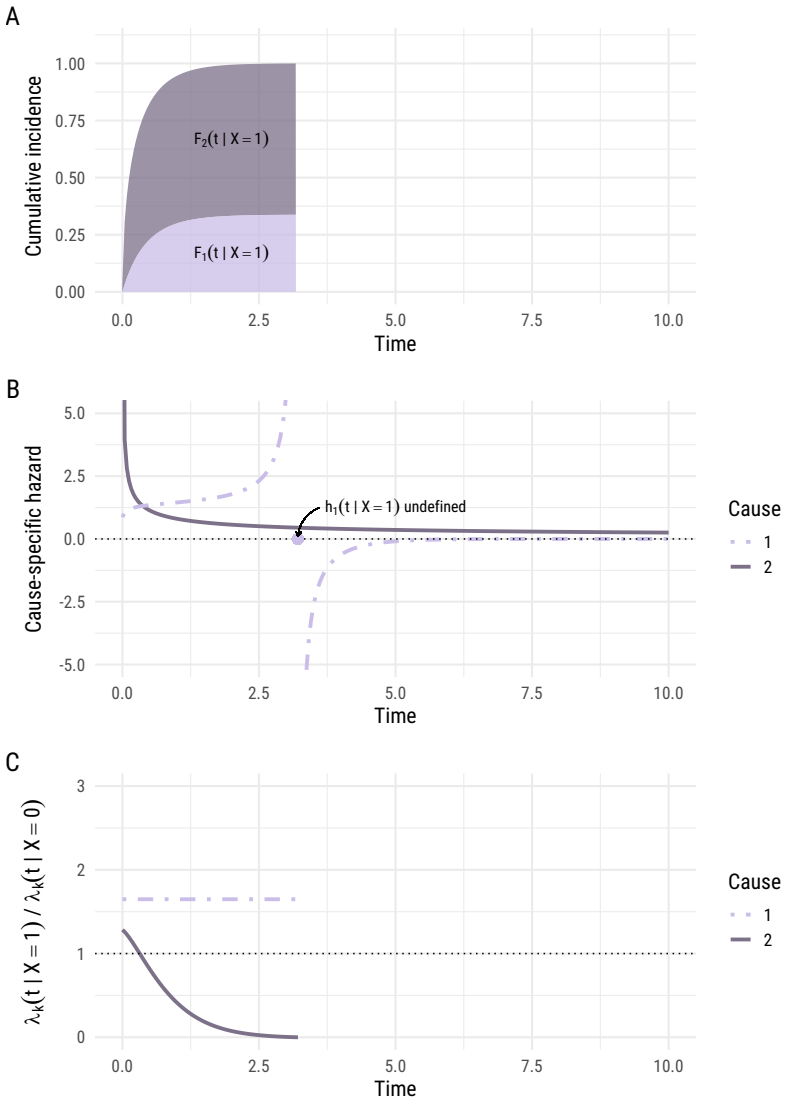

Figure 6.1: True stacked cumulative incidence functions (panel A) and cause-specific hazards (panel B) conditional on $X = 1$, and the subdistribution hazard ratios (panel C) for both causes, under DGM described in Section 6.3.1.1. This DGM assumes a Fine–Gray model for cause 1 with Gompertz baseline subdistribution hazard, and cause-specific Cox model for cause 2 with a Weibull baseline hazard.

6.3.1.2 Specifying $\lambda_1(t \mid X)$ and $h_1(t \mid X)$

If we instead choose to specify both the subdistribution and cause-specific hazards for event 1, the cause-specific hazard for cause 2 can be derived by re-arranging Equation 6.1 as

$$h_2(t \mid X) = \lambda_1(t \mid X) - h_1(t \mid X) - \frac{\mathrm{d}}{\mathrm{d}t} \log\left\{ \frac{\lambda_1(t \mid X)}{h_1(t \mid X)} \right\}. \tag{6.4}$$

While a Gompertz baseline hazard could again be specified for $\lambda_1(t \mid X)$, it is important to specify $h_1(t \mid X)$ such that $h_1(t \mid X) = \lambda_1(t \mid X)$ at $t = 0$. From a simulator's point of view, it means paying attention to the fact that due to the form of the reduction factor (a time-dependent weight), proportionality generally cannot hold for both $\lambda_1(t \mid X)$ and $h_1(t \mid X)$ simultaneously. If the Fine–Gray model holds for $\lambda_1(t \mid X)$, the cause-specific hazard ratio $h_1(t \mid X = 1)/h_1(t \mid X = 0)$ should usually be time-dependent.

An exception to the above is found in the work of Saadati *et al.* (2018). There, a DGM is presented where $h_2(t \mid X)$ is chosen such that proportionality holds on both the subdistribution and cause-specific hazard scale for cause 1. The chosen $h_2(t \mid X)$ is based on setting the reduction factor to 1 for all covariate patterns and timepoints, which also implies that the covariate effects on the subdistribution and cause-specific hazard for event 1 need to be equal to each other.

Similarly to Equation 6.2, the implied $h_2(t \mid X)$ when specifying $\lambda_1(t \mid X)$ and $h_1(t \mid X)$ may also become negative. To understand an instance of where this can occur, note that we can write the all-cause cumulative hazard as a function of $\lambda_1(t \mid X)$ and $h_1(t \mid X)$ using Equation 6.1, as

$$-\log\{S(t \mid X)\} = \frac{\lambda_1(t \mid X) \exp\{-\Lambda_1(t \mid X)\}}{h_1(t \mid X)}.$$

As an example, we can use the same Gompertz parametrisation for $\lambda_1(t \mid X)$ as in the previous sub-section, namely $\{\kappa_1, \nu_1, \beta_1\} = \{-2, 0.5, 0.5\}$. Suppose now we also decide to use a Gompertz baseline hazard for $h_1(t \mid X)$, using the same parametrisation (i.e. base rate also equal to ν_1 and $\beta_1 = \gamma_1$, ensuring $\lambda_1(t \mid X) = h_1(t \mid X)$ at $t = 0$), but instead setting the shape parameter to 2 (exponentially increasing). By solving $-\log\{S(t \mid X)\} - H_1(t \mid X) = 0$, one can find the timepoint at which $H_1(t \mid X)$ starts to exceed the all-cause cumulative hazard. Prior to this timepoint, $H_2(t \mid X)$ will be forced to decrease in order to maintain $-\log\{S(t \mid X)\} = H_1(t \mid X) + H_2(t \mid X)$, implying negative $h_2(t \mid X)$. Note also that this is not the fault of the Gompertz distribution: it is perfectly possible to simulate competing risks data with baseline cause-specific Gompertz hazards.

From a simulator's point of view, a DGM based on directly specified $\lambda_1(t \mid X)$ and $h_1(t \mid X)$ is rather tedious to implement given (a) the restriction that $\lambda_1(t \mid X) = h_1(t \mid X)$

at $t = 0$ for all X, (b) the (generally) time-dependent nature of the cause-specific hazard ratio for event 1. Even when $\beta_1 = \gamma_1$ (same covariate effects on cause-specific and subdistribution hazard of cause 1), which is unlikely to be the case in practice unless there are relatively few cause 2 failures, specifying an adequate $h_1(t \mid X)$ is not very flexible. As part of work on simulating proportional subdistribution hazard data with time-varying effects, Haller and Ulm (2014) provide an example of simulating from a DGM based on specifying both $h_1(t \mid X)$ and $\lambda_1(t \mid X)$. There, $h_1(t \mid X)$ is chosen to be time constant, with rate equal to $\lambda_1(t \mid X)$ at $t = 0$.

6.3.1.3 Specifying $\lambda_1(t \mid X)$ and a model for the all-cause hazard

The reduction factor could also form the basis for a DGM if a model is specified for the all-cause hazard $\sum_{k=1}^{K} H_k(t \mid X) = -\log\{S(t \mid X)\}$, together with the Fine–Gray model for $\lambda_1(t \mid X)$. One can derive the implied $h_1(t \mid X)$ from Equation 6.1, and subtract it from the all-cause hazard to obtain $h_2(t \mid X)$. Proportional hazards on the all-cause scale however typically implies that the cause-specific hazards will be non-proportional. Note that this DGM requires that $-\log\{S(t \mid X)\} > \Lambda_1(t \mid X)$ at all timepoints and for all X. That is, that the all-cause cumulative hazard is always greater than the cumulative subdistribution hazard of event 1, otherwise the implied cause-specific hazard for cause 2 is forced to be negative. When simulating from this DGM, this could for example occur if the specified covariate effects differ substantially between the subdistribution hazard model for cause 1 and the all-cause model (e.g. pushing $\Lambda_1(t \mid X)$ above $-\log\{S(t \mid X)\}$ for $X = 1$). Nevertheless, as long as precautions are taken when specifying $-\log\{S(t \mid X)\}$, this DGM again represents a valid way to specify $f(T, D \mid X)$ such that proportional subdistribution hazards hold for cause 1. To the best of our knowledge, this mechanism has not been used in articles simulating proportional subdistribution hazards data.

6.3.2 Squeezing

Instead of specifying the various hazard functions, we can also work with the cumulative incidence functions directly. Recall that the Fine–Gray model for cause 1 can be expressed as

$$1 - F_1(t \mid X) = \{1 - F_{10}(t)\}^{\exp(\beta_1 X)}.$$

The idea is now to specify $F_{10}(t)$ directly. Since $F_k(\infty) = P(D = k) < 1$, we have to first pick some proper cumulative distribution $\tilde{F}_{10}(t)$ (e.g. exponential or Weibull cumulative distribution function, CDF) with $\lim_{t\to\infty} \tilde{F}_{10}(t) = 1$ and scale it down by a factor $0 < p < 1$ (since $p = 0$ or $p = 1$ would imply no competing risks). This leaves

$$1 - F_1(t \mid X) = \left[1 - p\{\tilde{F}_{10}(t)\}\right]^{\exp(\beta_1 X)}.$$

Note that $\lim_{t\to\infty} F_{10}(t) = p$, and $p_1(x) = P(D = 1 \mid X) = 1 - (1-p)^{\exp(\beta_1 X)}$. The probability of experiencing cause 2 therefore needs to be 'squeezed' into the remaining probability space $p_2(x) = P(D = 2 \mid X) = 1 - P(D = 1 \mid X) = (1-p)^{\exp(\beta_1 X)}$. Since $p_2(x)$ is determined by $p_1(x)$, it is guaranteed that $p_1(x) + p_2(x) = 1$. The second cumulative incidence function takes the form

$$P(T \leq t, D = 2 \mid X) = P(T \leq t \mid D = 2, X)P(D = 2 \mid X),$$

where $P(T \leq t \mid D = 2, X)$ can be chosen to be any standard CDF, which is then scaled down by $P(D = 2 \mid X)$. When simulating using this DGM, it is convenient to first generate the competing event indicator, and thereafter draw event times conditional on this indicator e.g. for event 1, drawing from $P(T \leq t \mid D = 1, X)$. For more details, see section 5.3.6 of Beyersmann *et al.* (2012). This DGM is arguably the most commonly used approach to simulate proportional subdistribution hazard data, as it ensures the TFP remains below or equal to 1. Multiple simulation studies, along with the original article proposing the Fine–Gray model, have simulated data in this way (Austin *et al.*, 2021; Bellach *et al.*, 2019; Fine and Gray, 1999; Saadati *et al.*, 2018).

To illustrate this mechanism, we use Weibull-type distribution functions and set

$$\tilde{F}_{10}(t) = 1 - \exp(-b_1 t^{a_1}),$$
$$P(T \leq t \mid D = 2, X) = 1 - \exp\{-b_2 t^{a_2} \exp(\beta_2^* X)\},$$

with $\{a_1, b_1, \beta_1, p\} = \{1.25, 1, 0.5, 0.2\}$ and $\{a_2, b_2, \beta_2^*\} = \{1.5, 1, 0.5\}$. β_2^* is denoted as such since it is not a subdistribution log hazard ratio, but instead denotes the effect of X on $P(T \leq t \mid D = 2, X)$. Figure 6.2 shows the baseline hazards and hazard ratios ($X = 1$ relative to $X = 0$) over time for the cause-specific and subdistribution hazards of both events. For this DGM, these functions are arguably more interesting to show, since they are only implicitly specified e.g. $h_1(t \mid X)$ is obtained by dividing $\mathrm{d}F_1(t \mid X)/\,\mathrm{d}t$ by $1 - F_1(t \mid X) - F_2(t \mid X)$. Note that in panels B and D (and also in panel C in Figure 6.1), no logarithmic transformation was applied to the y-axis (as would normally be the case for hazard ratio plots) due to the cause-specific and subdistribution hazard ratios for cause 2 going to zero as $t \to \infty$.

This DGM provides a clear picture on why one may choose to avoid Fine–Gray models for more than one cause. As other authors have similarly noted (Beyersmann *et al.*, 2012), a Fine–Gray model being specified for cause 1 effectively constrains the remaining probability space available to cause 2 to a maximum of $1 - P(D = 1 \mid X)$. Equivalently, this is a constraint on $\Lambda_2(t \mid X) = -\log\{1 - F_2(t \mid X)\}$, the cumulative subdistribution hazard for cause 2 conditional on X. In the case of a binary X, this translates to the coefficient of a Fine–Gray model for the competing cause needing to be *determined* by the Fine–Gray model for cause 1. Explictly, we can express the

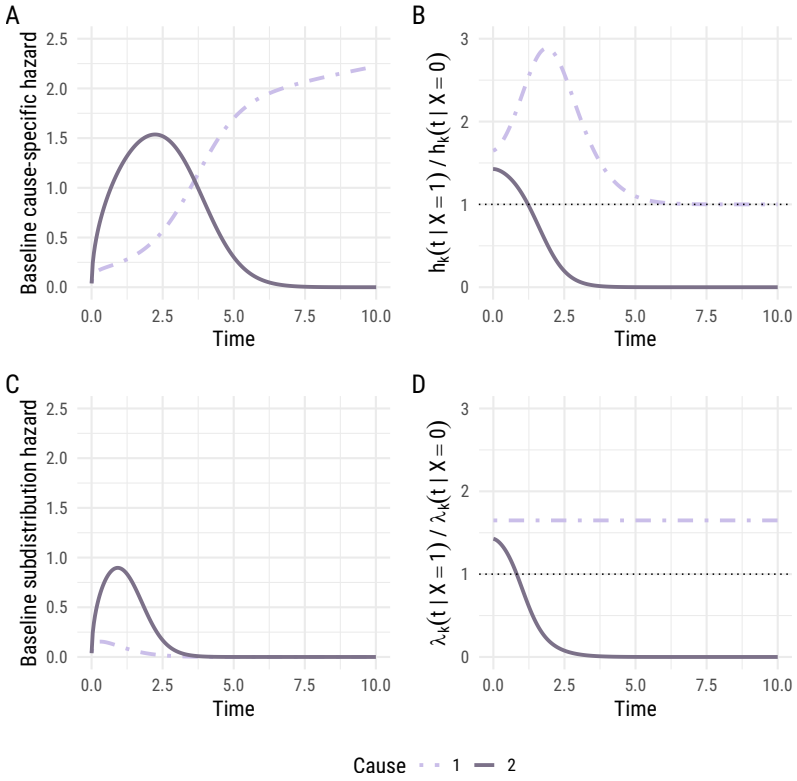

Figure 6.2: Baseline hazards (panels A and C) and hazard ratios $X = 1$ relative to $X = 0$ (panels B and D) over time for the cause-specific and subdistribution hazards of both events, based on the 'squeezing' DGM.

cumulative subdistribution hazard ratio for cause 2 as $t \to \infty$ for this DGM as

$$\begin{aligned}\lim_{t\to\infty} \frac{\Lambda_2(t \mid X=1)}{\Lambda_2(t \mid X=0)} &= \frac{-\log\{1-P(D=2 \mid X=1)\}}{-\log\{1-P(D=2 \mid X=0)\}}, \\ &= \frac{\log\{P(D=1 \mid X=1)\}}{\log\{P(D=1 \mid X=0)\}}, \\ &= \frac{\log\{1-(1-p)^{\exp(\beta_1)}\}}{\log(p)},\end{aligned}$$

by using $P(D=2 \mid X) = 1-\exp\{-\lim_{t\to\infty}\Lambda_2(t \mid X)\}$ and $P(D=2 \mid X) = 1-P(D=1 \mid X)$. Therefore, the cumulative subdistribution hazard ratio for cause 2 as $t \to \infty$ should be completely determined by β_1 and p. Since running a Fine–Gray model for the second cause does not account for this restriction, it is typically misspecified, leading to issues such as the TFP exceeding 1. As shown in Figure 6.2, the extent of this misspecification can be alarming: the true subdistribution hazard ratio for the competing cause is severely time-dependent, for which the time-averaged subdistribution hazard ratio (see Grambauer *et al.*, 2010) obtained from a Fine–Gray model is perhaps a suboptimal summary.

6.3.3 Two Fine–Gray models

The previous sub-section may suggest that it is impossible for proportional subdistribution hazards to hold for more than one competing event. In fact, if we choose to also directly specify the cumulative incidence for cause 2 in the same style as in the 'squeezing' mechanism (instead of having it determined by cause 1), we can achieve proportional subdistribution hazards for both events. Suppose we have

$$\begin{aligned}F_1(t \mid X) &= 1 - \left[1 - p_{10}\{\tilde{F}_{10}(t)\}\right]^{\exp(\beta_1 X)}, \\ F_2(t \mid X) &= 1 - \left[1 - p_{20}\{\tilde{F}_{20}(t)\}\right]^{\exp(\beta_2 X)}.\end{aligned}$$

If we let $P(D=k \mid X) = 1-(1-p_{k0})^{\exp(\beta_k X)} = p_k(x)$, then we can write that as $t \to \infty$, $\text{TFP}(x) = p_1(x) + p_2(x)$. To determine the event indicator when simulating, we would draw $u \sim \mathcal{U}(0,1)$, and set

$$D = \begin{cases} 1, & \text{if } u \le p_1(x), \\ 2, & \text{if } p_1(x) < u \le p_2(x). \end{cases}$$

The event times can then be drawn as in the preceding subsection, by inverting $P(T \le t \mid D=k, X)$. Those with $u > p_1(x) + p_2(x)$ are technically considered 'cured', that is, never at risk of any of the competing events. To illustrate this DGM, we set

$$\begin{aligned}\tilde{F}_{10}(t) &= 1 - \exp(-b_1 t^{a_1}), \\ \tilde{F}_{20}(t) &= 1 - \exp(-b_2 t^{a_2}),\end{aligned}$$

with $\{a_1, b_1, \beta_1, p_{10}\} = \{0.75, 1, 0.5, 0.25\}$ and $\{a_2, b_2, \beta_2, p_{20}\} = \{0.75, 1, 0.5, 0.5\}$. Figure 6.3 shows the baseline cumulative incidence functions, and those conditional on $X = 1$. The chosen baseline cumulative incidence functions and subdistribution hazard ratios result in the TFP exceeding 1 from $t \approx 3$ when $X = 1$. While this mechanism can be flexible when baseline hazard rates are small and covariate effects are modest, it by design cannot guarantee a TFP ≤ 1 for any timepoint. Indeed, if X is not binary, and instead continuous and unbounded, the TFP is guaranteed to exceed 1 (though for small t, X will need to be very extreme for this to occur) (Austin *et al.*, 2021).

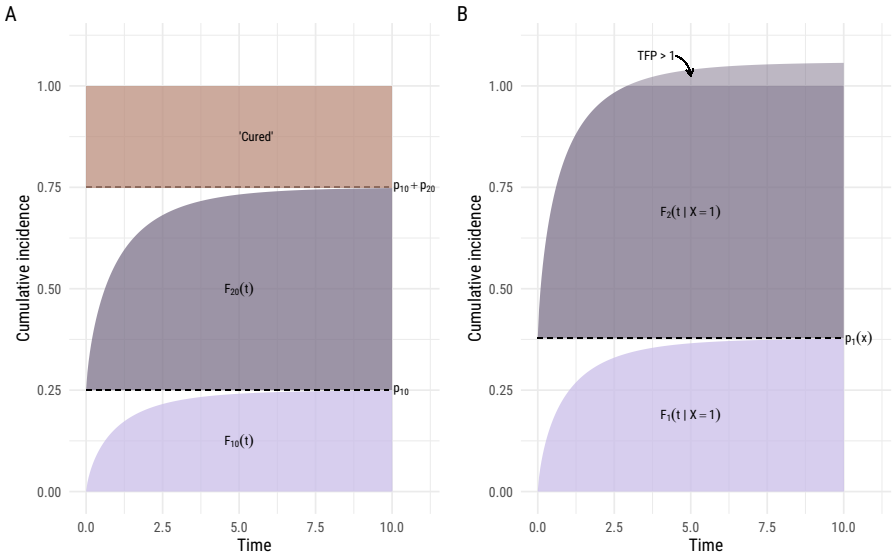

Figure 6.3: Stacked cumulative incidence functions conditional on $X = 0$ (baseline, panel A) and those conditional on $X = 1$ (panel B), for a DGM in which proportional subdistribution hazards can hold for both causes.

Two other useful implementations of (variants of) this DGM are found in the simulations of Mao and Lin (2017), and Mozumder *et al.* (2018)—both investigating the performance of different approaches (semiparametric and flexible parametric, respectively) for direct modelling of the cumulative incidence functions. Mao and Lin directly specify Gompertz cumulative subdistribution hazards $\Lambda_k(t \mid X)$ for both events, and thereafter invert $P(T \leq t \mid D = k, X)$ analogously using $P(T \leq t, D = k \mid X) = 1 - \exp\{-\Lambda_k(t \mid X)\}$ and its limit as $t \to \infty$. Mozumbder et al. specify mixture Weibull baseline subdistribution hazards for both events, and then derive the implied cause-specific hazards using the reduction factor, and use these for simulating.

An important point is that both of the above approaches specify a maximum follow-up

time in their simulations. Indeed, as pointed out by Latouche *et al.* (2013), 'it is possible that such models may hold over restricted time ranges, which has practical implications for studies with limited longitudinal follow-up', with 'such models' referring to multiple proportional subdistribution hazard models. Note also that specifying a maximum follow-up time τ means that the $1 - F_1(\tau \mid X) - F_2(\tau \mid X)$ proportion of individuals that did not experience the event by τ are simply considered as censored at τ.

6.4 Discussion

In this work, we have outlined various ways of specifying a joint density $f(T, D \mid X)$ in which a Fine–Gray model for cause 1 is correctly specified. The goal was to outline the possible assumptions that can be made regarding the (cumulative incidence of) cause 2, given that a Fine–Gray model is correctly specified for cause 1. From a simulator's perspective, all DGMs are fundamentally aiming to do the same thing, which is to fill up the probability space leftover from the assumed Fine–Gray model for cause 1 i.e. $1 - F_1(t \mid X)$. The 'squeezing' DGM does this in the most explicit way, by making sure cause 2 fills up all of the remaining space, in turn ensuring that the TFP remains below or equal to 1 for all X and at all timepoints. This makes it the approach with the fewest restrictions when simulating data for which proportional subdistribution hazards holds for one cause only. Using the reduction factor in contrast can be quite inflexible, possibly producing negative cause-specific hazard values without great care in choices of parameter values and distributions. Note also that the described DGMs can be readily adapted in order to simulate under link functions other than the complementary log-log, discussed for example by Gerds *et al.* (2012). More generally, there is no DGM for which proportional subdistribution hazards holds simultaneously for both causes, unless one assumes finite follow-up and a bounded covariate space.

Suppose that we have a dataset where proportional subdistribution hazards perfectly hold for both causes up to some maximum follow-up time, e.g. a simulated dataset, for which we know the DGM. Here, fitting a Fine–Gray model for each cause in turn (i.e. using the exact models the data were generated from) can still lead to the TFP exceeding 1. This can occur more generally in finite samples, since the Fine–Gray models for each cause are estimated separately. As a result, alternative approaches have been developed (based on the complete data likelihood) to facilitate simultaneous modelling of all cumulative incidence functions, while incorporating this TFP constraint (Mao and Lin, 2017; Shi *et al.*, 2013) The parametric approach suggested by Shi *et al.* (2013) actually does so by explicitly incorporating the 'squeezing' of the second cause (i.e. that covariates effects on cause 2 depend on the asymptote of the cumulative incidence function for cause 1) into the likelihood. Note also that modelling the cause-specific hazards and combining these to obtain predicted cumulative incidence functions ensures that the TFP remains below 1 for any timepoint and covariate

combination (Austin *et al.*, 2021). The cause-specific approach also extends naturally (i.e. without TFP issues or tedious algebra) to settings with more than two competing events.

In applied settings, proportionality assumptions for either event on any of cause-specific and subdistribution hazard scales, such as those made in the outlined DGMs, will never hold exactly. Using alloSCT data, Gerds *et al.* (2012) compared the performance of cause-specific hazard and Fine–Gray models (as well as other transformation models) for predicting competing events relapse and non-relapse mortality. Both predictive accuracy (based on cross-validated Brier score) and individual predictions were similar for both approaches. Wolbers *et al.* (2009) reported that the cause-specific and Fine–Gray approaches showed comparable calibration when predicting coronary heart disease, though they did not consider calibration of the competing event. Kantidakis *et al.* (2023) also reported similar predictive performance of the cause-specific and Fine–Gray approaches when applied on a dataset of patients with extremity soft-tissue sarcoma. This was as part of a broader comparison with machine learning techniques, with the goal of predicting competing events disease progression and death.

One explanation for this comparable performance is that both models make use of their non-parametric baseline hazard to compensate, to some extent, for misspecified (i.e. non-proportional) covariate effects—allowing them to still predict the cumulative incidence functions fairly accurately (noted in Shi *et al.*, 2013 for Fine–Gray models). Differences in performance may also be modest in settings with a shorter follow-up period or when there is heavy censoring: the time-dependent weight relating the cause-specific and subdistribution hazards (the reduction factor) is less influential earlier in time. Nevertheless, there are situations in which one would expect (multiple) Fine–Gray models to underperform with respect to cause-specific hazard approaches. First, when the estimated TFP exceeds 1 in a non-negligible proportion of patients, as in the example by Austin *et al.* (2021), the predictions for one or more of the competing events must by definition be partly miscalibrated (due to risk overestimation). Second, when the true cumulative incidence curves for an event (e.g. relapse probabilities for two conditioning regimens) cross each other, the predicted curves by the Fine–Gray model will not be allowed to cross. Poythress *et al.* (2020) therefore suggest to compare predicted curves to their non-parametric counterparts as a possible diagnostic, and note that cause-specific hazards models are much more suited to capturing more complex cumulative incidence function shapes. For example, the cause-specific hazards could be modelled using flexible parametric approaches, which naturally accommodate time-varying effects (Hinchliffe and Lambert, 2013; Kipourou *et al.*, 2019).

In conclusion, the described DGMs outline the variety of ways in which proportional subdistribution hazards could hold for at least one of two event types. In terms of cumulative incidence prediction for both causes, we argue that cause-specific hazard models should be preferred over multiple Fine–Gray models, as they (a) by design ensure that the TFP does not exceed 1, (b) are able to capture complex shapes for the

cumulative incidence functions (although in the Fine–Gray context, one could technically include time by covariate interactions), and (c) additionally provide inference on the cause-specific hazards, which are the 'natural building blocks for competing risks modeling' (Saadati *et al.*, 2018). The {riskRegression} R package in particular provides useful functions for developing and validating prediction models based on cause-specific hazards (Ozenne *et al.*, 2017). While using a Fine–Gray model for one cause only may still be defendable (e.g. for prediction purposes, when other causes are truly a nuisance), it does go against the holistic approach to competing risks analyses described in the introduction, where all causes should ideally be studied together. Cause-specific hazard models, which are often misunderstood to be less suitable for prediction compared to Fine–Gray models (see e.g. D'Amico *et al.*, 2018), should perhaps also be the preferred approach also in settings where predicting a single cause is of interest. When the main goal is simultaneous inference on the cumulative incidence functions, the proposed semiparametric approach by Mao and Lin (2017) is a promising alternative to multiple Fine–Gray models, as it (a) provides more efficient inference, (b) allows the use of different link functions (e.g. accommodates non-proportional hazards, and allows odds ratio interpretation of parameters) for different events, (c) does not need to model the censoring distribution. For inference at specific timepoints, one may also consider to specify models using pseudovalues as the outcome variable (Klein and Andersen, 2005).

Supplementary materials

The R code to reproduce the figures for the described DGMs is available at https://github.com/survival-lumc/FineGrayDGM.

Chapter 7

Joint models quantify associations between immune cell kinetics and allo-immunological events after allogeneic stem cell transplantation and subsequent donor lymphocyte infusion

Chapter based on: Koster, E. A. S., **Bonneville, E. F.**, Borne, P. A. von dem, et al. (2023) Joint models quantify associations between immune cell kinetics and allo-immunological events after allogeneic stem cell transplantation and subsequent donor lymphocyte infusion. *Frontiers in Immunology*, 14. DOI: 10.3389/fimmu.2023.1208814.

Abstract

Alloreactive donor-derived T-cells play a pivotal role in alloimmune responses after allogeneic hematopoietic stem cell transplantation (alloSCT); both in the relapse-preventing Graft-versus-Leukemia (GvL) effect and the potentially lethal complication Graft-versus-Host-Disease (GvHD). The balance between GvL and GvHD can be shifted by removing T-cells via T-cell depletion (TCD) to reduce the risk of GvHD, and by introducing additional donor T-cells (donor lymphocyte infusions [DLI]) to boost the GvL effect. However, the association between T-cell kinetics and the occurrence of allo-immunological events has not been clearly demonstrated yet. Therefore, we investigated the complex associations between the T-cell kinetics and alloimmune responses in a cohort of 166 acute leukemia patients receiving alemtuzumab-based TCD alloSCT. Of these patients, 62 with an anticipated high risk of relapse were scheduled to receive a prophylactic DLI at 3 months after transplant. In this setting, we applied joint modelling which allowed us to better capture the complex interplay between DLI, T-cell kinetics, GvHD and relapse than traditional statistical methods. We demonstrate that DLI can induce detectable T-cell expansion, leading to an increase in total, CD4+ and CD8+ T-cell counts starting at 3 months after alloSCT. CD4+ T-cells showed the strongest association with the development of alloimmune responses: higher CD4 counts increased the risk of GvHD (hazard ratio 2.44, 95% confidence interval 1.45–4.12) and decreased the risk of relapse (hazard ratio 0.65, 95% confidence interval 0.45–0.92). Similar models showed that natural killer cells recovered rapidly after alloSCT and were associated with a lower risk of relapse (HR 0.62, 95%-CI 0.41–0.93). The results of this study advocate the use of joint models to further study immune cell kinetics in different settings.

7.1 Introduction

The curative potential of allogeneic stem cell transplantation (alloSCT) in the treatment of hematological malignancies depends on the introduction of donor-derived alloreactive T-cells (Horowitz *et al.*, 1990). These T-cells recognise non-self antigens on patient-derived cells and can, once activated, expand and eliminate those cells. Targeting antigens on lymphohematopoietic cells including the malignant cells leads to the desired Graft-versus-Leukemia (GvL) effect and prevents relapse. However, when other tissues of the patient are targeted, Graft-versus-Host-Disease (GvHD) may develop (Falkenburg and Jedema, 2017). Natural killer (NK) cells may discriminate between healthy and non-healthy (e.g., virus-infected or malignant) cells by acting on signals from inhibitory and activating receptors that bind to the target cell. In the setting of alloSCT, early NK cell recovery can protect against relapse and viral infections (Dunbar *et al.*, 2008; Minculescu *et al.*, 2016). However, NK cells do not appear to be important effector cells in GvHD (Simonetta *et al.*, 2017).

To reduce the risk of severe GvHD, donor T-cell depletion (TCD) can be applied, although this will decrease the GvL effect (Busca and Aversa, 2017). In order to restore the GvL effect to prevent relapse, TCD alloSCT can be combined with the administration of donor lymphocyte infusions (DLIs) after transplant (Eefting *et al.*, 2014; Eefting *et al.*, 2016; Falkenburg and Jedema, 2017). DLI as part of a pre-emptive strategy is administered to patients with detectable minimal residual disease (MRD) or with residual patient hematopoiesis: mixed chimerism (MC). DLI as part of a prophylactic strategy is given to all patients in whom no GvHD has developed as sign of alloreactivity. The alloreactive potential of DLI decreases over time after alloSCT: both the efficacy (GvL effect) and toxicity (GvHD) are highest early after alloSCT (Krishnamurthy *et al.*, 2013; Yun and Waller, 2013). Therefore, administration preferably starts a few months after alloSCT to allow for sufficient GvL without severe GvHD (Falkenburg *et al.*, 2019).

Since T-cells are pivotal for alloimmune responses, several groups have investigated T-cell kinetics after alloSCT and their impact on the development of GvHD or relapse. However, as shown in the recent review by Yanir *et al.* (2022), the reported results are inconsistent, and their interpretation is complicated by several factors. First, T-cells can be patient- or donor-derived, while only donor-derived T-cells are responsible for GvHD and GvL. Second, the T-cell changes following alloSCT are the combined result of de novo T-cell generation from infused hematopoietic stem cells starting at least 6 months after alloSCT, homeostatic proliferation of T-cells present in the patient or graft, T-cell expansion during infections and expansion of alloreactive T-cells responsible for GvL and GvHD. Especially cytomegalovirus (CMV) reactivations are common during the first 3 months after alloSCT and strongly affect the kinetics of both T-cells and NK cells after alloSCT (Bosch *et al.*, 2012; Hassan *et al.*, 2022; Stern *et al.*, 2022). This may distort the association between the kinetics of the main T-cell

subsets and specific alloimmune responses, i.e., the presence of GvHD and the absence of relapse as a result of the GvL effect. Third, factors that could influence both the T-cell kinetics and the risks of GvHD and relapse, such as the conditioning regimen, donor type and the use and method of TCD, should be properly accounted for. Finally, ignoring clinical events or interventions during follow-up can also be problematic: over time, the patients that have not yet experienced an event like relapse, death or the development of GvHD, become less representative of the population at the beginning of follow-up. As death by definition prevents further T-cell measurements and the possibility of experiencing subsequent GvHD and relapse, bias is created by considering the patients who died as having non-informatively dropped out (i.e. that their measurements could have been measured if kept under follow-up). Likewise, DLI and the use of posttransplant prophylactic immunosuppression are known to affect the risks of relapse and GvHD, but may also affect the T-cell kinetics (Bellucci *et al.*, 2002; Gooptu *et al.*, 2019; Guillaume *et al.*, 2012; Lewalle *et al.*, 2003; Nikiforow *et al.*, 2016; Schmaelter *et al.*, 2021; Schultze-Florey *et al.*, 2021; Toor *et al.*, 2015). To fully understand the complex interplay between all these factors, sophisticated statistical methods are required that properly model the T-cell kinetics themselves, along with their association with GvHD or relapse. Joint modelling captures the T-cell trajectories and the clinical events simultaneously, accounting for informative dropout, as well as the measurement error and heterogeneity in individual trajectories (Rizopoulos, 2012).

In this study, we performed joint modelling to investigate the complex associations between the immune cell kinetics and alloreactivity in a cohort of 166 patients receiving an alloSCT for acute leukemia or myelodysplastic syndrome (MDS). All patients received an alemtuzumab-based TCD alloSCT after nonmyeloablative conditioning without any posttransplant prophylactic immunosuppression. Patients with an anticipated high risk of relapse were scheduled to receive an early low-dose DLI prophylactically at 3 months after alloSCT, while prophylactic DLI administration for the other patients started at 6 months. In this unique setting we investigated the impact of the early low-dose DLI on the T-cell and NK cell kinetics during the first 6 months after transplant and the association between these kinetics and the development of clinical events.

7.2 Methods

7.2.1 Study population

This retrospective study included all adult patients with acute myeloid leukemia, acute lymphoblastic leukemia or MDS in complete morphologic remission after intensive induction therapy who received their first alloSCT from a 9 or 10 out of 10 HLA

matched donor using nonmyeloablative conditioning and alemtuzumab-based TCD (von dem Borne *et al.*, 2009) between March 2008 and December 2019 at Leiden University Medical Center (LUMC, Leiden, The Netherlands). Two patients who were transplanted while receiving systemic immunosuppression for a non-transplant indication (polymyalgia rheumatica and cryptogenic organising pneumonia) were excluded because of the potential impact of the ongoing systemic immunosuppression on the immune cell recovery after alloSCT. All patients signed informed consent for data collection and analysis. Data were analysed as of July 2021.

7.2.2 Transplantation and DLI strategy

As conditioning regimen patients received either fludarabine (6 days 50 mg/m^2 orally or 30 mg/m^2 intravenously) and busulfan (2 days 4x0.8 mg/kg intravenously), or the FLAMSA regimen: fludarabine (5 days 30 mg/m^2 intravenously), cytarabine (4 days 2000 mg/m^2 intravenously), amsacrine (4 days 100 mg/m^2 intravenously) and busulfan (4 days 4x0.8 mg/kg intravenously). In both regimens, TCD was performed by adding 20 mg alemtuzumab (Sanofi Genzyme, Naarden, The Netherlands) to the graft before infusion and by administering 15 mg alemtuzumab intravenously on days -4 and -3. Patients with an unrelated donor (UD) received rabbit-derived anti-thymocyte globulin (ATG; Sanofi Genzyme) additionally on day -2 (until April 2010 2mg/kg and thereafter 1mg/kg). None of the patients received posttransplant GvHD prophylaxis.

The dose of unmodified pre-emptive and prophylactic DLIs was based on donor type and timing after alloSCT. Standard DLIs given at 6 months after alloSCT contained 3x10^6 or 1.5x10^6 T-cells/kg for patients with a related donor (RD) or an UD, respectively. Early low-dose DLIs given at 3 months after alloSCT contained 0.3x10^6 or 0.15x10^6 T-cells/kg for patients with a RD or an UD, respectively. Since May 2010, all patients without any relapse and without GvHD requiring systemic immunosuppressive treatment at 6 months after alloSCT prophylactically (i.e., irrespective of chimerism or posttransplant MRD status) were planned to receive the standard DLI. Patients who were considered to have a high risk of relapse based on the disease characteristics or MRD status at time of alloSCT or who received the FLAMSA regimen were also scheduled to receive the early low-dose DLI prophylactically at 3 months after alloSCT. All patients, including those transplanted before May 2010, could receive pre-emptive DLIs in case of MC or MRD positivity, starting from 3 months after alloSCT. Additionally, as part of several clinical trials, patients could receive modified T-cell products prophylactically or virus-specific T-cell infusions to treat severe viral infections.

7.2.3 Monitoring of CMV and absolute numbers of circulating immune cells

CMV serostatus was assessed in all patients and donors before alloSCT. After transplant CMV was monitored routinely by PCR on peripheral blood samples in all patients. Absolute numbers of circulating total (CD3+), CD4+CD8- and CD4-CD8+ T-cells, B cells and NK cells were measured routinely at predefined timepoints on anticoagulated fresh venous blood by flow cytometry with bead calibration (Trucount tubes, BD Biosciences). Samples were measured either on a FACSCalibur using anti-CD3-APC, anti-CD4-FITC, anti-CD8-PE, and anti-CD45-PerCP or with anti-CD3-FITC, anti-CD16-PE, anti-CD19-APC, anti-CD45-PerCP, and anti-CD56-PE, or on a FACSCanto using anti-CD3-APC, anti-CD4-PB, anti-CD8-FITC, anti-CD16-PE, anti-CD19-PE Cy7, anti-CD45-PerCP, and anti-CD56-PE (all from BD). The lower detection limit was $0.5x10^6$ cells/L.

7.2.4 Definitions of events

Relapse was defined as the recurrence of at least 5% blasts on cytomorphologic bone marrow examination or at least 1% blasts in peripheral blood (if possible, confirmed by BM biopsy). We defined clinically significant GvHD as the start of therapeutic systemic immunosuppression for GvHD (Koster *et al.*, 2023). We defined 'other failure' as the occurrence of an adverse event with a potential impact on the immune cell kinetics: death, graft failure, start of systemic immunosuppression for a non-GvHD indication, and virus-specific T-cell infusion for a severe viral infection (whichever occurred first). Graft failure was defined as the occurrence of >95% patient BM chimerism in all lineages tested or refractory granulopenia (granulocyte count $<0.5x10^9$/l) in the absence of relapse or ongoing myelotoxic medication.

For this study we analysed the T-cell and NK cell kinetics and events during the first 6 months after alloSCT, during which the early immunological recovery and most CMV reactivations take place. Furthermore, during this period the impact of the early low-dose DLI can be assessed, as the standard DLI is given to all eligible patients around 6 months after alloSCT. As part of the analyses assessing the net impact of the early low-dose DLI on the T-cell and NK cell kinetics and clinical events, patients receiving a standard DLI or modified T-cell product as part of a clinical trial were censored at 7 days after this infusion. We considered this to be non-informative censoring, since these interventions were prophylactic and not driven by the clinical course of the patient. For the T-cell kinetics we considered the circulating cell counts of the total (CD3+) T-cell population and the two major T-cell subpopulations: the CD4+CD8- and the CD4-CD8+ T-cells.

7.2.5 Statistical analyses

Probabilities of overall survival (OS) and relapse-free survival (RFS) after alloSCT with associated 95% confidence intervals (95%-CI) were calculated by the Kaplan-Meier method. The cumulative incidences of clinically significant GvHD and relapse from time of alloSCT were estimated by means of the Aalen–Johansen method, treating other failure (as described in the previous section) as a third competing risk.

To study the complex interplay between the immune cell kinetics, DLI and clinically relevant endpoints (GvHD and relapse), two joint models were developed; model I starting at time of alloSCT and model II at time of the early low-dose DLI.

Shared-parameter joint models consist of two components: a longitudinal submodel, and a time-to-event submodel (Rizopoulos, 2012). The former often takes the form of a mixed-effects regression model, and the latter is generally assumed to follow a proportional hazards structure, similar to a Cox model (for one or possibly multiple endpoints such as GvHD or relapse). The mixed-effects model allows to model cell count trajectories over time, while appropriately accounting for both the heterogeneity in subject-specific trajectories (using random effects) and measurement error. These two submodels are linked together via an association structure. Practically speaking, this allows the hazard of a particular event to depend on characteristics of an individual's specific trajectory, such as the 'true' underlying (i.e. in absence of measurement error) value over time. In turn, this enables the estimation of an association between a longitudinal marker (e.g. CD3 counts) and the risk of a clinical event (e.g. GvHD). In the presence of an association, the estimated trajectories themselves will be corrected for bias related to the measurements being terminated by the occurrence of endpoints (generally known as 'informative dropout').

Below follows a concise description of the joint models developed for the present application. Detailed explanation of the statistical models and the underlying rationale can be found in the Statistical Supplement (Appendix A). For all models, absolute cell counts were analysed on the log scale after setting measurements under the detection limit to 0.5. This only occurred at earliest timepoints where because of the lymphodepletion by the conditioning regimen and TCD, the counts are expected to be around zero.

7.2.5.1 Model I (starting from alloSCT)

To investigate the effect of early low-dose DLI on the kinetics of the T-cell and NK cell counts after TCD alloSCT, we performed an intention-to-treat (ITT) analysis with a baseline group distinguishing between those scheduled for early low-dose DLI because of a high anticipated risk of relapse (henceforth 'high risk' group) and those who were not ('non-high risk' group). We chose this approach instead of a per-protocol analysis

since we could not properly define a control group of patients who did not receive early DLI but could have been candidates as we did not know for each patient who was not scheduled for early DLI whether he/she would have been able to receive it.

Figure 7.1 A shows a schematic overview of joint model I. The model was run separately for each T-cell subset, respectively using CD3, CD4 or CD8 counts, and the total NK counts. All patients started at time of alloSCT and were followed-up until 6 months after alloSCT or until the occurrence of an earlier endpoint (GvHD, relapse or other failure), whichever occurred first. The longitudinal submodel was a linear mixed-effects model, which used restricted cubic splines to flexibly model the log counts over time. The baseline covariates included in this submodel were disease risk (non-high risk or high risk), donor type (RD or UD with ATG-containing conditioning regimen) and patient/donor CMV status (both seronegative [CMV -/-] or not). The patient/donor CMV status was included as simple fixed effect, and both disease risk and donor type were included as part of a three-way interaction with time. This was in order to both properly accommodate the expected slower lymphocyte recovery in patients treated with ATG, and to evaluate a difference in trajectories between the disease risk groups. The time-to-event submodel comprised three cause-specific proportional hazards models, with GvHD, relapse and other failure as competing events. As predictors, they each contained the time-dependent current value (i.e. the underlying 'true' value at a given timepoint, as estimated by the longitudinal submodel) of the log immune cell count, as well as the baseline factors donor type and disease risk. The latter was omitted as a covariate from the model for 'other failure' due to the limited number of events.

To investigate whether the current slope (i.e. rate of increase or decrease of counts at a given moment) of the T-cell counts was associated with the development of GvHD, we also extended the models by adding the current slope of the log counts in addition to the current value to the time-to-event submodel (so-called 'time-dependent slopes' parametrisation).

7.2.5.2 Model II (starting from early low-dose DLI)

To further investigate the T-cell kinetics after the early low-dose DLI, we constructed a joint model including only the patients who actually received the early low-dose DLI without any prior event of interest (Figure 7.1 B). Since NK cells recover rapidly after alloSCT (Elfeky *et al.*, 2019) (expected before the administration of early low-dose DLI in this study), they were not considered for model II. The time-scale was taken from DLI instead of from alloSCT, and follow-up was restricted to 3 months after this DLI, until administration of a second DLI, or until the occurrence of a terminating event, whichever occurred first. The disease risk factor was omitted since all included patients belonged to the high risk group. Since only 7 patients had a non-GvHD event within 3 months after the early low-dose DLI (Supplementary Figure 1), relapse and

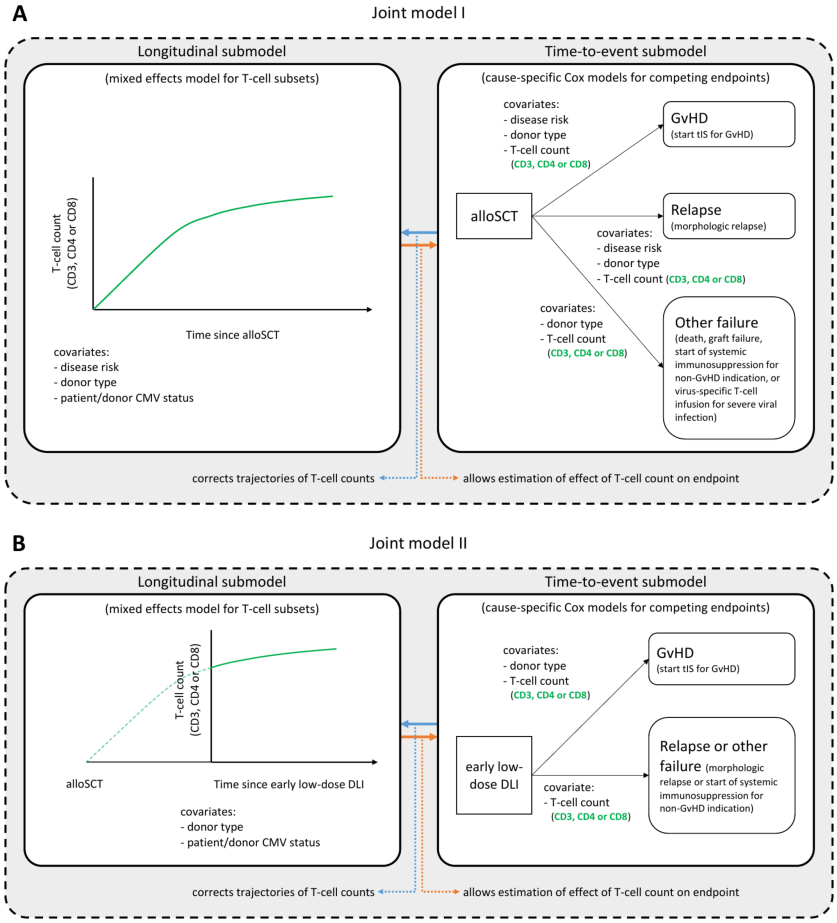

Figure 7.1: *Structure of the joint models.* Graphical description of the two joint models. Joint model I (A) starts at time of alloSCT, joint model II (B) at time of the early low-dose DLI. Each model consists of a longitudinal and a time-to-event submodel and was run in turn for each T-cell subset, considering either the CD3+, CD4+ or CD8+ T-cell counts, and the NK cell counts. These are the outcome of the longitudinal submodel and a time-dependent covariate in the time-to-event submodel. All other variables in each submodel are baseline covariates. Per endpoint of the time-to-event submodels, the clinical events that occurred during the relevant time period (first 6 months after alloSCT or first 3 months after the early low-dose DLI) are described. The NK cells were only analysed in model I. See the Statistical Supplement for a detailed description of the model structures.

other failure were combined into one composite endpoint to compete with GvHD and the donor type factor was omitted for this composite endpoint.

7.2.6 Software

All analyses were performed in R version 4.2.1 using the packages JM (version 1.5-2, Rizopoulos, 2010), survival (version 3.4.0, Therneau, 2023) and nlme (3.1-157, Pinheiro *et al.*, 2023). Full code needed to reproduce the results of the present work is available at https://github.com/survival-lumc/ImmuneReconstJM, and structured using the targets (0.14.0, Landau, 2021) package.

7.3 Results

7.3.1 Population

166 patients were included in this study. Baseline characteristics are presented in Table 7.1. All surviving patients had at least 12 months follow-up since alloSCT. OS and RFS at 6 months after alloSCT were 77% (95%-CI 71–83) and 70% (95%-CI 64–77), respectively. A total of 62 patients were considered to have a high risk of relapse and were scheduled for an early low-dose DLI, of whom 42 actually received it after a median interval of 3.1 months (range: 2.7–4.4) without any prior event of interest (Supplementary Figure 1). Twenty patients did not receive an early low-dose DLI: 10 because of early relapse, 9 because of early other failures (death [n=1], graft failure [n=2], start of systemic immunosuppression for a non-GvHD indication [n=4], or administration of a virus-specific T-cell infusion [n=2]), and 1 patient did not receive the early low-dose DLI because of mild skin GvHD requiring topical treatment. All 19 events occurred within 4 months after alloSCT. The patient with mild skin GvHD remained event-free for at least 51 months after alloSCT. None of the 104 non-high risk patients received an early low-dose DLI. At 6 months after alloSCT, the cumulative incidence of clinically significant GvHD was 26% (95%-CI 15–37) and 5% (95%-CI 0–9) for the high risk patients scheduled for early low-dose DLI and the non-high risk patients, respectively (Supplementary Figure 2). All clinically significant GvHD in the high risk patients occurred after administration of the early low-dose DLI (but before standard DLI) of which 88% occurred in patients receiving DLI from an UD after an ATG-containing conditioning regimen.

Table 7.1: *Baseline characteristics.* Intention for early low-dose DLI is based on the anticipated high risk of relapse after alloSCT. DLI, donor lymphocyte infusion; alloSCT, allogeneic stem cell transplantation; AML, acute myeloid leukemia; ALL, acute lymphoblastic leukemia; MDS, myelodysplastic syndrome; Flu, fludarabine; Bu, busulfan; Ara-C, cytarabine; Amsa, amsacrine; RD, related donor; UD, unrelated donor; GCSF, granulocyte-colony stimulation factor; PBSC, peripheral blood stem cells; BM, bone marrow. *One patient had not received a second consolidation course before transplant and received 2 days cyclophosphamide 750 mg/m^2 intravenously additionally to the conditioning regimen.

	Total cohort (n = 166)	Intention for early low-dose DLI (n = 62)	No intention for early low-dose DLI (n = 104)
Age at alloSCT (years)			
median (range)	63 (28-78)	64 (31-78)	63 (28-73)
Disease			
AML	133 (80%)	46 (74%)	87 (84%)
ALL	17 (10%)	10 (16%)	7 (7%)
MDS	16 (10%)	6 (10%)	10 (10%)
Nonmyeloablative conditioning			
Flu/Bu	150 (90%)	46 (74%)	104 (100%)*
Flu/Bu/Ara-C/Amsa (FLAMSA)	16 (10%)	16 (26%)	0
Donor			
RD, 10/10 HLA matched	57 (34%)	20 (32%)	37 (36%)
UD, 10/10 HLA matched	101 (61%)	39 (63%)	62 (60%)
UD, 9/10 HLA matched	8 (5%)	3 (5%)	5 (5%)
Graft source			
G-CSF mobilized PBSC	165 (99%)	62 (100%)	103 (99%)
BM	1 (1%)	0	1 (1%)
CMV serostatus patient/donor			
Patient +/Donor +	79 (48%)	32 (52%)	47 (45%)
Patient +/Donor -	25 (15%)	8 (13%)	17 (16%)
Patient -/Donor +	11 (7%)	4 (6%)	7 (7%)
Patient -/Donor -	51 (31%)	18 (29%)	33 (32%)
Main reason for intention for early low-dose DLI			
FLAMSA regimen	-	16 (26%)	-
MRD+ at time of alloSCT	-	14 (23%)	-
AML/MDS: EVI1 overexpression	-	9 (15%)	-
AML: monosomal karyotype	-	8 (13%)	-
AML: ASXL mutation, only one remission induction course, or persisting underlying disease	-	4 (6%)	-
ALL: t(9;22)	-	4 (6%)	-
ALL: hypodiploidy, no CR1, or t(4;11)	-	4 (6%)	-
Therapy-related AML	-	2 (3%)	-
AML: progression before alloSCT	-	1 (2%)	-

7.3.2 T-cell trajectories after alloSCT and DLI

7.3.2.1 DLI-related increase of T-cell counts after 3 months after alloSCT observed in patients with an unrelated donor

To investigate whether administration of the early low-dose DLI increased the numbers of circulating T-cells during the first 6 months after alloSCT, we performed an ITT analysis using model I (see Methods) to compare the 62 high risk patients who were scheduled for early low-dose DLI with the 104 non-high risk patients who were not. All patients had at least 2 T-cell measurements with a median of 6 measurements per patient (interquartile range: 5–8). Although patients showed very different T-cell kinetics over time (Supplementary Figure 3), the model was flexible enough to capture the different shapes of patient-specific trajectories (Figure 7.2). Patients who were CMV seropositive or who had a CMV seropositive donor had significantly higher CD3 and CD8 counts during the first 6 months after TCD alloSCT compared to CMV seronegative patients with a CMV seronegative donor, corresponding to a significant increase on the log scale of 0.49 (95%-CI 0.31–0.67) and 0.45 (95%-CI 0.08–0.80) for CD3+ and CD8+ T-cells, respectively. For instance, the model-based CD3 count at 6 months for a non-high risk patient with a RD was $425x10^6$/l if CMV -/- compared to $694x10^6$/l for any other CMV serostatus combination. The model-based CD8 count at this time was $222x10^6$/l compared to $347x10^6$/l, respectively, suggesting expansion of CMV-specific T-cells. A same trend was observed for the CD4 counts (increase of 0.11 on the log scale, 95%-CI 0–0.23). As shown in Figure 7.3, patients with an UD had lower T-cell counts during the first 3 months after TCD alloSCT than patients with a RD, illustrating the enduring effect of the additional ATG that was given to all patients with an UD. We observed no significant difference in the cell count trajectories between the disease risk groups for patients with a RD. In contrast, in patients with an UD the CD4 trajectories started to diverge at 3 months after alloSCT, resulting in higher cell counts in the high risk patients intended to receive an early low-dose DLI at 3 months. The CD3 and CD8 counts showed similar trends. Taken together, these data show that a strategy of early low-dose DLI can lead to T-cell expansion.

7.3.2.2 CD3, CD4 and CD8 counts increase after early low-dose DLI

To investigate whether the T-cell counts increased after the early low-dose DLI as the ITT-analysis suggested, we used model II including only the 42 patients who actually received this DLI without any prior event and modelled the kinetics during the first 3 months after DLI. One of the 42 patients did not have any T-cell measurement during this period and was excluded. Baseline characteristics of the 41 included patients are described in Supplementary Table 1. These patients had at least one T-cell measurement during the 3-month period after early low-dose DLI with a median of 4 measurements (interquartile range: 2–5). Again, a flexible model was constructed to capture the

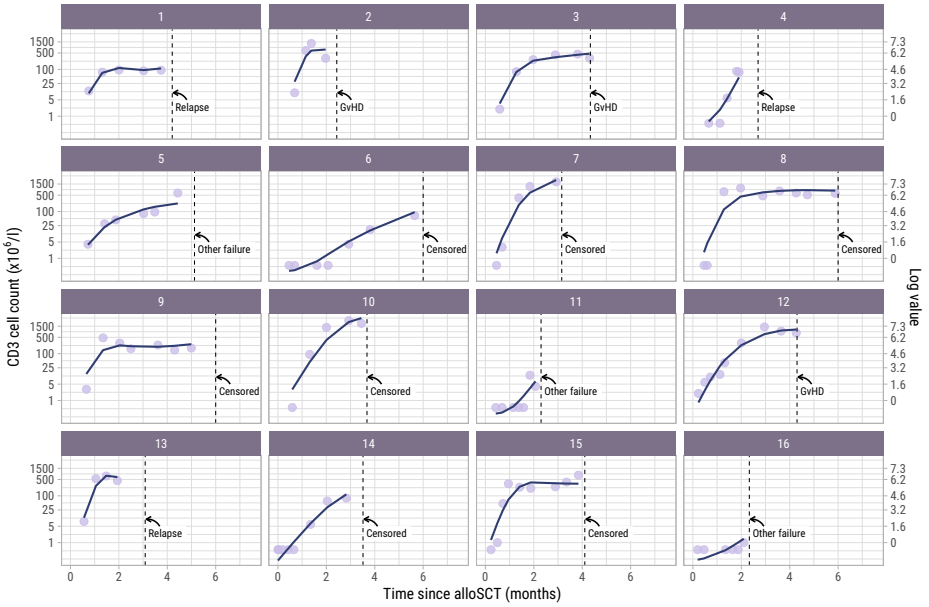

Figure 7.2: *Observed versus estimated CD3 counts from alloSCT.* Observed (dots) and estimated subject-specific trajectories (solid line) of a random subset of 16 patients in the dataset. The estimated trajectories are based on the longitudinal submodel of model I. Dotted lines show the time of terminating event or administrative censoring because of administration of a modified T-cell product or standard DLI. The secondary axis shows the cell counts on the log scale, which is the scale used for modelling. For example, a cell count of 1 on the primary axis corresponds to log(1) = 0 on the secondary axis.

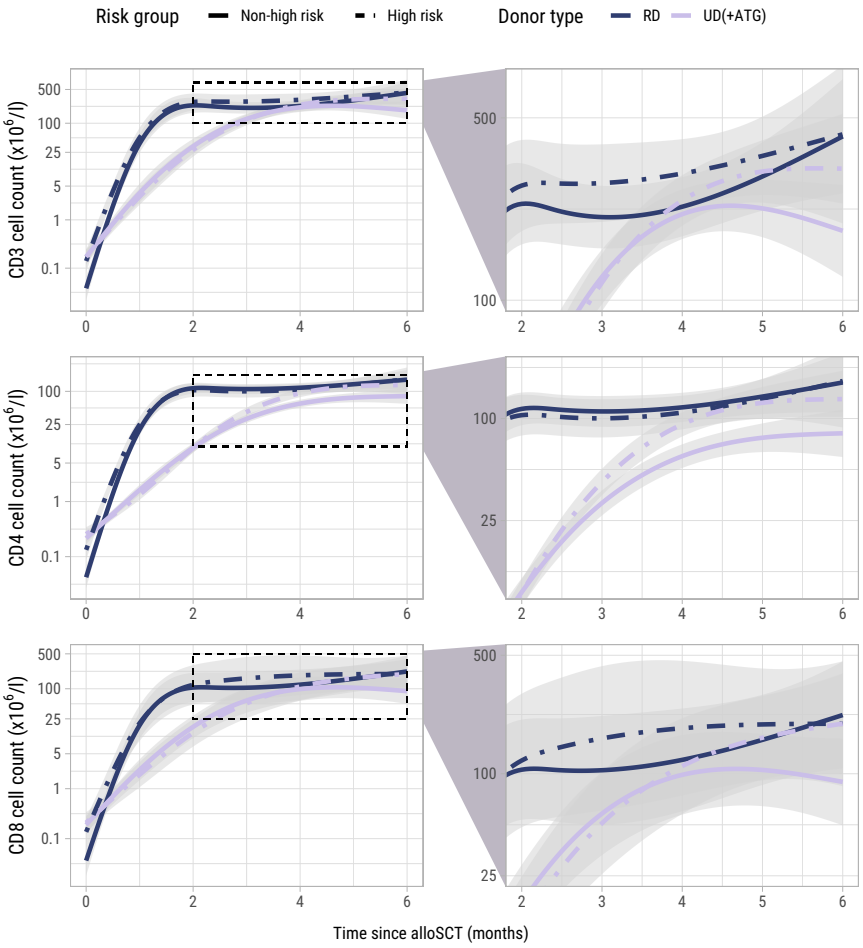

Figure 7.3: *Model-based T-cell count trajectories after alloSCT*. Predicted average trajectories of the total, CD4+ and CD8+ T-cell counts during the first 6 months after alloSCT, based on the longitudinal submodel of model I. For all predicted trajectories, the patient/donor CMV status was set to -/-. 95% confidence intervals are shown in grey. The right column zooms in on a specific part of the total trajectory.

different shapes of the T-cell kinetics of the included patients (Supplementary Figure 4 and Supplementary Figure 5). The model-based trajectories of the total, CD4+ and CD8+ T-cell counts (Figure 7.4) showed increasing T-cell counts after DLI, with similar effects of the patient/donor CMV serostatus and donor type on the T-cell counts as in the earlier models.

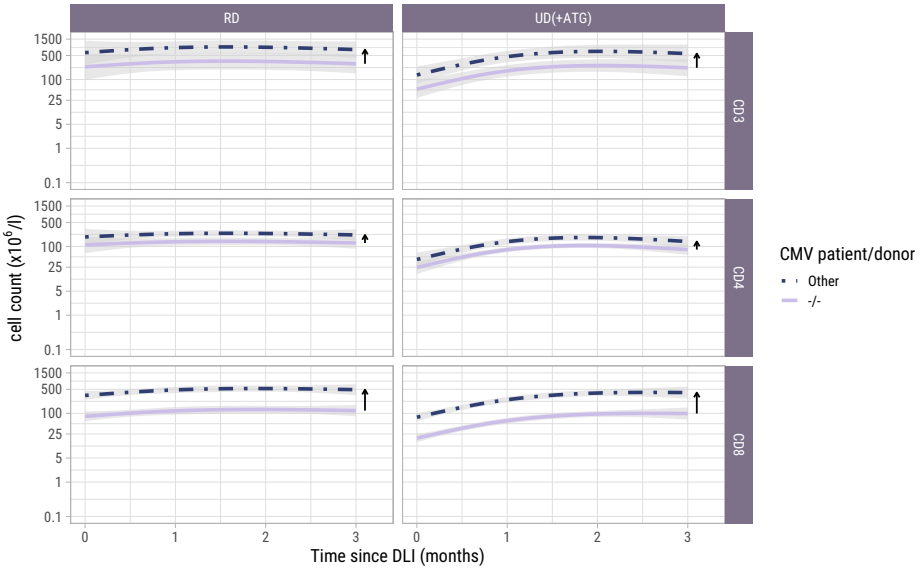

Figure 7.4: Model-based T-cell count trajectories after early low-dose DLI. Predicted average trajectories of the total, CD4+ and CD8+ T-cell counts during the first 3 months after early low-dose DLI. These are based on the longitudinal submodel of model II. 95% confidence intervals are shown in grey. The distance between the two lines in each panel (and further visualised by the adjacent arrows) corresponds to the CMV patient/donor effect on the trajectories. Namely, higher cell counts are predicted for patient/donor pairs where at least one is CMV seropositive, relative to a pair where both are CMV seronegative.

7.3.3 Associations between T-cell kinetics and alloimmune responses after alloSCT and DLI

7.3.3.1 Higher CD3 and CD4 counts are associated with a higher risk of GvHD

To study the association between the T-cell kinetics and the development of GvHD or relapse after TCD alloSCT and DLI, we added disease risk and donor type as time-fixed

covariates alongside the time-dependent T-cell counts in the cause-specific submodels (with GvHD, relapse and other failure as competing events) of model I. As shown in Figure 7.5, donor type showed no significant association with the risk of GvHD, although in the CD4 model a trend for higher risk in patients with an UD despite the ATG in the conditioning regimen was observed (hazard ratio [HR] 2.7, 95%-CI 1.0–7.4). High risk patients, who were scheduled for early low-dose DLI, had a considerably higher risk of GvHD compared to non-high risk patients with HRs ranging between 6.3 (CD8 model, 95%-CI 2.1–18.8) and 7.3 (CD4 model, 95%-CI 2.4–22.2), indicating an alloimmune effect of the early low-dose DLI in this setting. The current values of the log CD4 and CD3 counts significantly increased the risk of GvHD (HR 2.4, 95%-CI 1.4–4.1) and HR 1.5 (95%-CI 1.0–2.3) for CD4+ T-cells and CD3+ T-cells, respectively), while CD8+ T-cells showed a similar trend (HR 1.3, 95%-CI 0.9–1.8). These HRs represent the relative increase in GvHD risk for an increase of one in the log counts, assuming same disease risk and donor type. These results indicate that the absolute total numbers of circulating CD4+ and CD3+ T-cells after alloSCT and DLI are informative for the development of GvHD.

We hypothesised that not only the current value but also the slope of the T-cell counts would be associated with the development of an alloimmune response. To investigate this, we extended the time-to-event submodel of model I by additionally including the current slope of the T-cell counts as a covariate for all endpoints. However, we observed no association between the slope of any of the T-cell subsets and the development of GvHD (*p*-values 0.59–0.87). We therefore retained the simpler version of model I with only the current value.

7.3.3.2 Protective effect of CD4+ T-cells against relapse and other failure

To investigate whether higher T-cell counts were associated with a lower risk of relapse, we examined the risk factors for relapse in the time-to-event submodel of model I. Despite the ATG, patients with an UD had a significantly lower risk of relapse than patients with a RD (HRs ranging between 0.2 (95%-CI 0.1–0.5) and 0.3 (95%-CI 0.1–0.8), Figure 7.5). A trend was observed for higher relapse risk in the high risk patients (HR 2.1 in all models, 95%-CI for CD4+ T-cells: 0.9–5.0, respectively), suggesting that the addition of early low-dose DLI to the strategy did not completely compensate for the higher relapse risk. While CD3+ and CD8+ T-cells showed no significant association with relapse, higher CD4 counts decreased the risk of relapse significantly (HR 0.6, 95%-CI 0.5–0.9).

Of the 36 patients who experienced other failures, 6 died, 8 developed graft failure, 18 required systemic immunosuppression for a non-GvHD indication (of whom 9 received rituximab for EBV) and 4 received a virus-specific T-cell infusion for a severe viral infection. Only in the CD8 model a trend was observed for a higher risk of other failure in patients with an UD receiving an ATG-containing conditioning regimen (HR 2.6,

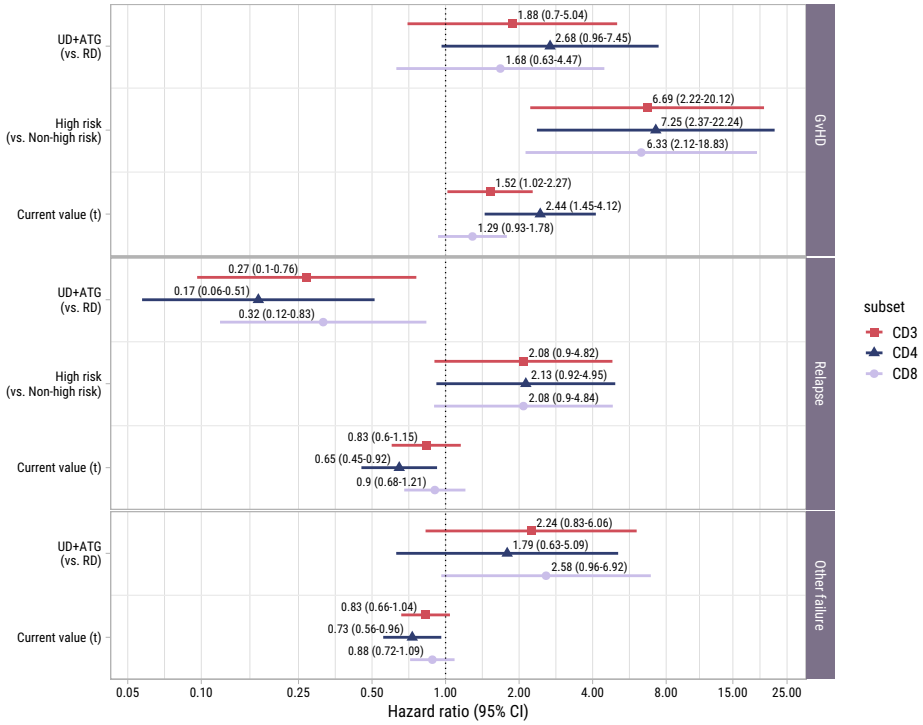

Figure 7.5: *Forest plot for ITT analysis.* Hazard ratios with associated 95% confidence intervals for donor type, disease risk and current value of the log of total, CD4+ or CD8+ T-cell counts on the events of interest. These are based on the time-to-event submodel of model I (see Figure 7.1 A).

95%-CI 1.0–6.9). Higher CD4+ T-cell counts significantly lowered the hazard of the composite endpoint other failure (HR 0.7, 95%-CI 0.6–1.0).

7.3.3.3 T-cell counts after early low-dose DLI retain their association with the development of GvHD

To investigate whether the T-cell kinetics were also associated with the development of alloimmune responses in the postDLI setting, we used the time-to-event submodel of model II starting from early low-dose DLI with GvHD and non-GvHD events as competing events. We observed no significant association between the current values and the very heterogenous composite endpoint of relapse and other failure (Figure 7.6). However, patients with an UD had a considerably higher risk of GvHD with HRs ranging between 7.0 (CD8+ T-cells, 95%-CI 1.5–32.1) and 22.5 (CD4+ T-cells, 95%-CI 3.7–138.9) compared to patients with a RD. For all T-cell subsets, higher current values increased the risk of GvHD with HRs ranging between 1.6 (CD8+ T-cells, 95%-CI 1.0–2.6) and 6.7 (CD4+ T-cells, 95%-CI 2.1–21.5). These data show that in the subset of patients receiving early low-dose DLI, total CD3+, CD4+ and CD8+ T-cell counts after DLI are associated with the development of GvHD.

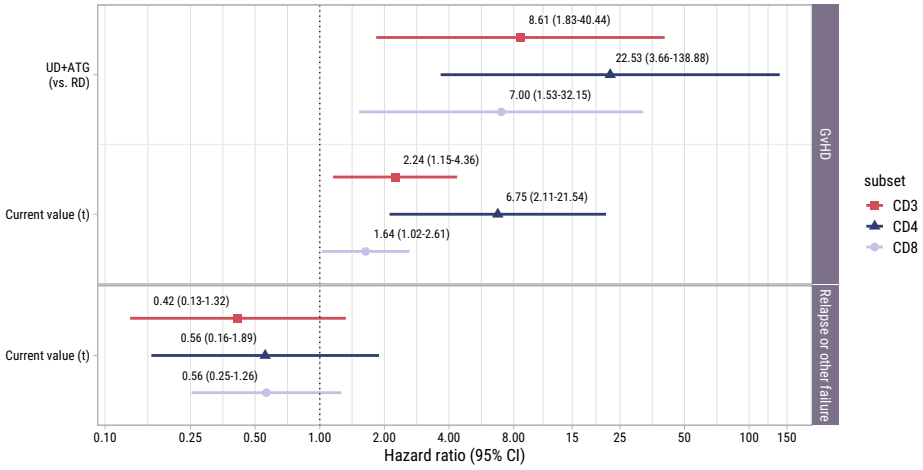

Figure 7.6: *Forest plot for postDLI models.* Hazard ratios with associated 95% confidence intervals for donor type and current value of the log of total, CD4+ or CD8+ T-cell counts on the events of interest. These are based on the time-to-event submodel of model II (see Figure 7.1 B).

7.3.4 NK cell kinetics and associations with alloimmune responses after alloSCT

To investigate the NK cell kinetics and their association with GvHD and relapse, we returned to model I starting at alloSCT. As shown in Supplemental Figure 6, the NK cell counts recovered rapidly, reaching the normal levels of 40–390x10^6 NK cells/l for almost all patients within 2 months, before the time of administration of the early low-dose DLI. As shown in Figure 7.7, CMV seropositive patients or patients with a CMV seropositive donor had significantly higher NK counts than CMV -/- patients, as was seen for the T-cell subsets. In contrast to T-cell kinetics, patients with an UD and ATG did not have a slower recovery of NK counts compared to patients with a RD and no ATG. Furthermore, there was no association between the risk group and NK counts, indicating that there was no impact of DLI on the NK cell kinetics. Higher current NK counts were associated with a higher risk of GvHD (HR 1.95 per unit log count increase, 95%-CI 1.10–3.47) and a lower risk of relapse (HR 0.62, 95%-CI 0.41–0.93) but had no significant association with the risk of other failure. We hypothesised that the observed association between the NK count and GvHD may not be due to a direct effect of the NK cells, but instead reflected the high correlation between the NK and CD4 count trajectories, the latter being expected to be the main driver of GvHD. We therefore ran a cause-specific Cox model for GvHD, which included disease risk and donor type as time-fixed covariates, and both CD4 and NK counts as time-dependent covariates. In this model, CD4 counts were significantly associated with the development of GvHD (HR 2.08, 95%-CI 1.16–3.74) while the HR for the NK cell counts was 1.07 (p-value 0.83), supporting that the CD4+ T-cells were the important drivers for the development of GvHD.

7.4 Discussion

In this study we investigated the interplay between immune cell kinetics and alloimmune responses after both TCD alloSCT and subsequent DLI using joint modelling. In the ITT analysis we observed significantly more GvHD in the high risk patients intended to receive an early low-dose DLI and an increase in T-cell counts starting at 3 months after alloSCT in high risk patients with an UD receiving an ATG-containing conditioning regimen. The ITT allocation was solely based on the disease characteristics of the patients. Since all patients were in complete remission at time of alloSCT, the TCD strategy was similar between the disease risk groups, and all GvHD in the high risk group only occurred after DLI, the only plausible explanation for both the higher risk of GvHD and the associated T-cell expansion is the administration of the early low-dose DLI. We also observed significant associations between the CD4 counts and alloimmune responses after TCD alloSCT and DLI: an increase in CD4+ T-cells was associated with a higher risk of GvHD and at the same time a lower risk of relapse

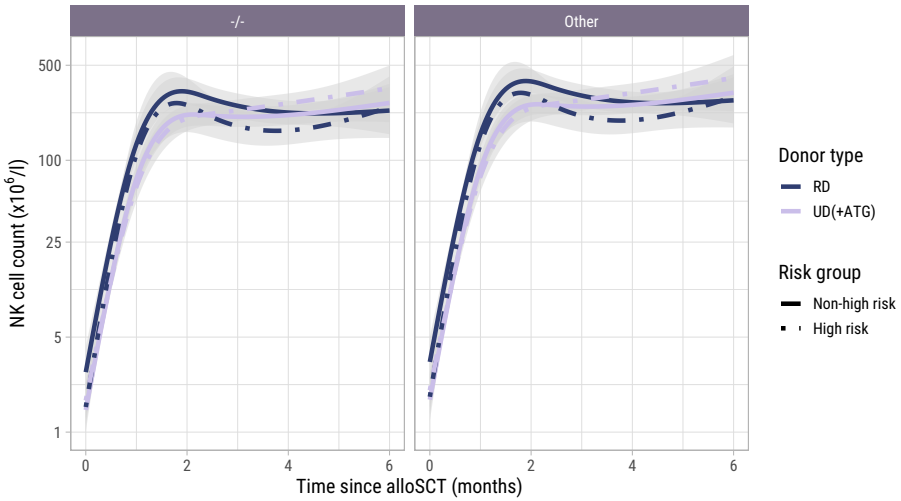

Figure 7.7: Model-based NK cell count trajectories after alloSCT. Predicted average trajectories of the NK cell counts during the first 6 months after alloSCT, based on the longitudinal submodel of model I. The left panel shows the predicted trajectories for CMV seronegative patients with a CMV seronegative donor, the right panel the predicted trajectories for patients with any other patient/donor CMV serostatus combination. 95% confidence intervals are shown in grey.

suggesting establishment of a GvL effect. Interestingly, we only observed DLI-induced T-cell expansion in patients transplanted using an UD. This likely reflects an alloimmune response as GvHD was mainly seen in patients with an UD after receiving a DLI, and the T-cell counts after DLI were associated with the development of GvHD. The alloreactive T-cell expansion may have been more easily detectable in patients with an UD compared to RD because of the deeper lymphopenia at time of DLI due to the long-lasting immunosuppressive effect of ATG that patients with an UD received (Bosch *et al.*, 2012). In addition, the high prevalence of HLA-DP mismatches, targeted by CD4+ T-cells, in patients with an UD (Fleischhauer *et al.*, 2012; Mariano *et al.*, 2019; Shaw *et al.*, 2007) could contribute to the strong association between CD4+ T-cells and the development of GvHD. In contrast to T-cells, NK cells recovered early after alloSCT and were not significantly influenced by donor type and TCD, consistent with previous studies (Bosch *et al.*, 2012; Ito *et al.*, 2020; Penack *et al.*, 2008), nor by DLI. As previously reported (Dunbar *et al.*, 2008; McCurdy *et al.*, 2022), higher NK counts were associated with a lower risk of relapse. The joint model also suggested that higher NK counts were associated with a higher risk of GvHD. However, in an exploratory cause-specific Cox model, this association between NK cells and GvHD disappeared after adjusting for the CD4 counts, indicating that the CD4+ T-cells were the important drivers for GvHD.

Our results suggest a DLI-induced T-cell expansion measurable in total numbers of the major T-cell subsets where others did not observe a significant effect of DLI on the T-cell kinetics (Bellucci *et al.*, 2002; Guillaume *et al.*, 2012; Nikiforow *et al.*, 2016; Schultze-Florey *et al.*, 2021). This may be due to several factors. Our comparatively larger cohort size (other studies usually included less than 25 patients) allowed for detection of more subtle differences. Furthermore, the strategy of administering early prophylactic DLI to a subset of patients based on their relapse risk provided an intervention and control group who were treated according to the same transplantation strategy. Lastly, conclusions drawn can be influenced by the choice of the statistical method. For example, matched pair analysis as used by Guillaume *et al.* (2012) and Schultze-Florey *et al.* (2021) only allowed them to compare the cells counts between two timepoints. The repeated measures analysis used by Nikiforow *et al.* (2016) and the mixed model used by Bellucci *et al.* (2002) allowed to compare the trajectories over time but could not account for informative dropout. Because we used joint modelling, we could flexibly model the T-cell trajectories over a longer period of time and properly account for informative dropout and random variation. To our knowledge, thus far only a single study used joint modelling to study T-cell kinetics after alloSCT (Salzmann-Manrique *et al.*, 2018). We now have used this technique to investigate the immunological effects of DLI.

There are several limitations to our study. The total CD3, CD4 and CD8 counts are crude measures for potentially alloreactive T-cells, as only donor-derived T-cells can induce GvHD and GvL and the counts are not informative about the subpopulations, activation status or kinetics of specific T-cell clones. Thus, if we had measured the

chimerism status and clonality, we might have expected to find stronger associations between the T-cell kinetics and the clinical events. Moreover, our ITT approach attenuated the observed effects of DLI on the T-cell kinetics and clinical endpoints as not all high risk patients received the early low-dose DLI and most patients who did receive this DLI did not receive it at exactly the same time after transplant. Therefore, we constructed model II starting from early low-dose DLI to see whether similar associations were observed. Joint modelling requires substantial numbers of both clinical events and longitudinal measurements to estimate associations with sufficient accuracy. Despite our comparatively larger sample size, the modest numbers of clinical events limited both the accurate estimation of association parameters (between T-cell counts and the endpoints), as well as the inclusion of additional risk factors for each endpoint. This was especially noticeable in our models focusing on the subset of the patients actually receiving an early low dose DLI. Due to the limited number of events, we used suboptimal composite endpoints such as 'other failure' and 'relapse and other failure', which hampered estimation of the association between the T-cell kinetics and these endpoints.

Further studies are necessary to assess the clinical implications of the findings from the present work. Aside from validation of our findings, larger studies must be performed to investigate the predictive utility of the T-cell and NK cell counts. While these counts are crude measures, they are often measured standardly and therefore attractive biomarkers for predicting alloimmune responses in patients receiving alloSCT and/or DLI. Further investigation of the immune cell kinetics in other alloSCT settings is needed to see whether similar associations between the T-cell and NK cell kinetics and alloimmune responses can be observed when using joint modelling. For instance, the recent machine learning analysis by McCurdy et al also suggested important roles of CD4+ T-cells in the development of acute GvHD and of NK cells in the development of relapse after alloSCT with posttransplant cyclophosphamide (McCurdy *et al.*, 2022). For DLI, we would suggest to perform a prospective study where the T-cell counts are measured at time of DLI and every week after DLI during the first 6 weeks. Most GvHD develops within this period and by measuring more often, dynamic prediction tools (i.e. updated personalised probabilities of GvHD given measurement history) could be developed (Andrinopoulou *et al.*, 2015). In order to develop such tools however, one would ideally need to model the T-cell subsets and NK cells jointly as part of a multivariate joint model, which will account for the correlation between each subset, but may be complicated to fit and will require larger sample sizes. In our study, we were not able to present such a multivariate joint model because of both sample size and software limitations. Nevertheless, results from the exploratory time-dependent cause-specific Cox model for GvHD with both the CD4 and NK counts hints at the importance of modelling immune subsets jointly.

Generally speaking, further characterisation of the circulating T-cell subsets, differentiation and metabolic fitness could provide valuable additional insight in future studies on T-cell kinetics (Dekker *et al.*, 2020; Uhl *et al.*, 2020).

In summary, joint modelling allowed us to capture the associations between DLI, T-cell and NK cell counts, GvHD and relapse in a very complex clinical setting, even with modest numbers of patients and events. NK cells recover early after alloSCT and may have a protective effect against relapse. We demonstrate that DLI can induce detectable T-cell expansion and observe that the CD4+ T-cells show the strongest association with the development of alloimmune responses. Higher CD4 counts increase the risk of GvHD and decrease the risk of relapse.

Supplementary materials

Supplementary tables and figures are available online at https://doi.org/10.3389/fimmu.2023.1208814.

Appendix A: Statistical supplement

Joint models only consider measurements taken prior to the occurrence of the clinical events of interest. Occasionally, the measurement time and event time coincide: for example, T-cell counts may be recorded on the same day as the start of therapeutic systemic immunosuppression for Graft-versus-Host-Disease (GvHD). In order to retain the information of the measurements taken at event times, we set the time of these measurements to one day earlier, which assumes that the measurement at the event time was representative of the T-cell counts the day before the event. However, we excluded measurements at time of relapse, since the presence of blasts in the peripheral blood could lead to incorrect counts of the normal T-cells. We also excluded measurements at time of autologous recovery, as donor-derived T-cells were no longer present, and therefore also no potentially alloreactive T-cells capable of inducing GvHD or Graft-versus-leukemia (GvL) effect.

Joint model I

Model formulation

The longitudinal submodel assumes that the true underlying (log) immune cell counts (either CD3, CD4, CD8, or NK) for the i^{th} patient are given by

$$\begin{aligned} m_i(t) = \beta_0 + \sum_{q=1}^{3}(\beta_q + b_{iq})B_q(t) + \sum_{q=1}^{3}\beta_{q+3}\{B_q(t) \times \text{Risk}_i\} + \sum_{q=1}^{3}\beta_{q+6}\{B_q(t) \times \text{Donor}_i\} \\ + \sum_{q=1}^{3}\beta_{q+9}\{B_q(t) \times \text{Risk}_i \times \text{Donor}_i\} + \beta_{13}\text{CMV}_i + \beta_{14}\text{Risk}_i + \beta_{15}\text{Donor}_i \\ + \beta_{16}\{\text{Risk}_i \times \text{Donor}_i\}, \end{aligned}$$

with random effects vector $b_i \sim \mathcal{N}(0, D)$. The observations for the i^{th} patient at timepoints t_{ij} $(j = 1, \ldots, n_i)$ are given by

$$y_{ij} = m_i(t_{ij}) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are independent random error terms.

Risk_i, Donor_i and CMV_i respectively represent the dummy variables for baseline disease risk (the intention-to-treat variable, high-risk compared to non-high risk), donor type (unrelated compared to related donor) and patient/donor Cytomegalovirus (CMV) serostatus at baseline (any one of patient or donor positive, compared to patient and donor both negative).

Time since allogeneic stem cell transplantation (alloSCT) was modelled flexibly assuming restricted (natural) cubic splines with two internal knots placed at the 33.3% and 66.7% percentiles of the measurement times. This is represented above by $B_q(t)$, corresponding to the q^{th} basis function of the spline. The fixed effects part of the model posits a three-way interaction between time, donor type and baseline disease risk, as well as a main effect of patient/donor CMV status. The three-way interaction was constructed to a) capture the slower expected average trajectory of patients with an unrelated donor, due to the use of anti-thymocyte globulin (ATG) in this group; and b) to test for a difference in average trajectories between baseline disease risk groups.

In terms of random effects, this models assumes random slopes b_{iq} (one for each basis function), and a fixed intercept. This fixed intercept was justified given that this cohort underwent T-cell depleted (TCD) alloSCT, and all patients were therefore expected to start follow-up with immune cell counts close to zero. The random slopes were assumed to be normally distributed with mean zero, with unstructured covariance matrix D.

The time-to-event submodel was composed of multiple cause-specific proportional hazards models as

$$h_{1i}(t) = h_{10}(t)\exp\left\{\gamma_{11}\text{Donor}_i + \gamma_{12}\text{Risk}_i + \alpha_1 m_i(t)\right\},$$
$$h_{2i}(t) = h_{20}(t)\exp\left\{\gamma_{21}\text{Donor}_i + \gamma_{22}\text{Risk}_i + \alpha_2 m_i(t)\right\},$$
$$h_{3i}(t) = h_{30}(t)\exp\left\{\gamma_{31}\text{Donor}_i + \alpha_3 m_i(t)\right\},$$

where the $h_{ki}(t)$ for $k \in \{1, 2, 3\}$ respectively represent the cause-specific hazards of GvHD, relapse, and other failures. The cause-specific baseline hazards $h_{k0}(t)$ were approximated on the log scale using cubic B-splines with three internal knots. The above corresponds to the 'current value' parametrisation of the joint model, where the $\exp(\alpha_k)$ would represent the hazard ratio (for cause k) when comparing two patients (with same covariates) whose 'true' (model-based) underlying log immune cell values at a particular timepoint $m_i(t)$ differ by one. The γ_{kp} coefficients are interpreted analogously to main effects in standard cause-specific Cox proportional hazards models.

In addition to the current value parametrisation, we also ran the models assuming a time-dependent slopes association structure as $\alpha_{k1} m_i(t) + \alpha_{k2}\{\mathrm{d}m_i(t)/\,\mathrm{d}t\}$.

Goodness of fit

Figure 7.8 presents standardised residuals plots for Joint Model I, which summarise how well the model fits the data overall (i.e. across all observations)—both for the average and subject-specific trajectories. The fitted (i.e. log immune cell counts predicted by the model) values are plotted against the standardised distance between the observed measurement and the predicted value. The blue line is a smoothed average of the standardised residuals as a function of the fitted values, and should ideally be horizontal at 0.

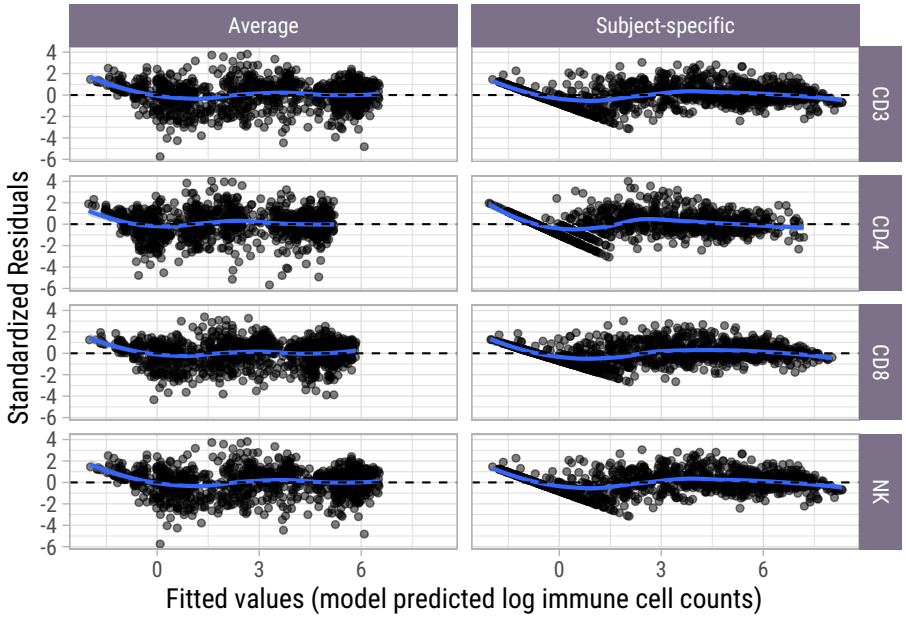

Figure 7.8: Standardised residuals plots for Joint Model I (separate joint models for each cell type).

Joint model II

Model formulation

For model II, the time scale was no longer from alloSCT, but instead from time of early low-dose donor lymphocyte infusion (DLI). Therefore, this model was only run among the subset that *did* in fact receive an early low-dose DLI before the occurrence of other competing events. Furthermore, some patients did not have a T-cell measurement on the day of DLI but only a few days prior. For these patients, we used the measurement closest to DLI taken within the last week before DLI as the measurement at time of DLI (time 0).

The longitudinal submodel was again a linear mixed-effects model, where the true underlying log T-cell counts are given by

$$m_i(t) = (\beta_0 + b_{i0}) + \sum_{q=1}^{2}(\beta_q + b_{iq})B_q(t) + \sum_{q=1}^{2}\beta_{q+2}\{B_q(t) \times \text{Donor}_i\} + \beta_5 \text{CMV}_i,$$

with random effects vector $b_i \sim \mathcal{N}(0, D)$. Observations for i^{th} patient are again given by

$$y_{ij} = m_i(t_{ij}) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are independent random error terms.

Time was again modelled with restricted cubic splines, but in contrast to model I, we used a single internal knot. The focus on a shorter timespan resulted in a reduced sample size, and fewer measurements per person. For consistency with model I, this average trajectory was allowed to differ across donor types (two-way interaction). In this model, disease risk at baseline was redundant as we ran the model among those having actually received an early low-dose DLI. A fixed effect for patient/donor CMV serostatus was also added to the model. This model comprised both random intercepts b_{i0} and random slopes b_{iq}, assumed to follow normal distributions with mean zero and unstructured covariance matrix.

Due to a limited number of events, relapse and other failures were merged into a composite endpoint. The time-to-event submodel was therefore specified as

$$h_{1i}(t) = h_{10}(t) \exp\left\{\gamma_{11}\text{Donor}_i + \alpha_1 m_i(t)\right\},$$
$$h_{2i}(t) = h_{20}(t) \exp\left\{\alpha_2 m_i(t)\right\},$$

where the $h_{ki}(t)$ for $k \in \{1, 2\}$ respectively represent the cause-specific hazards of GvHD and the composite of relapse and other failures for subject i. The cause-specific baseline hazards $h_{k0}(t)$ were approximated on the log scale using cubic B-splines with two internal knots. In this joint model, only the current value parametrisation was explored.

Goodness of fit

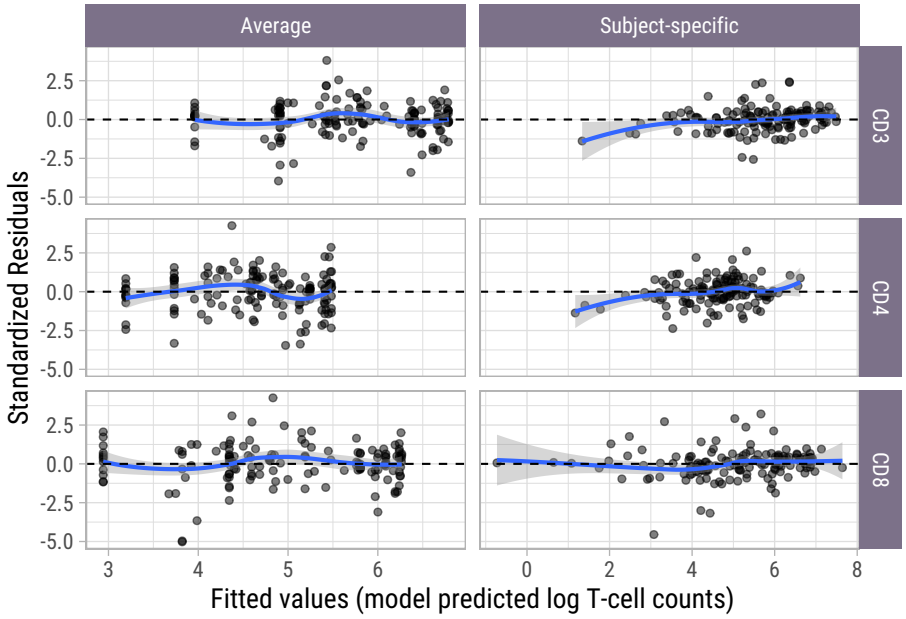

Figure 7.9: Standardised residuals plots for Joint Model II (separate joint models for each cell type).

Chapter 8

Conclusions

In this doctoral dissertation, we developed and assessed statistical methodology for handling missing data in competing risks settings. This was mainly in the context of observational data, where missing values are often both more prevalent and non-random compared to the clinical trials setting. The primary focus was on the use of multiple imputation (MI) for dealing with partially observed baseline covariates for both cause-specific Cox models, and Fine–Gray subdistribution hazard models. Shared-parameter competing risks joint models were also applied, in part to deal with missing values in longitudinal covariates. Three separate datasets of patients undergoing an allogeneic stem cell transplantation (alloSCT) were used to illustrate and test the various methodologies. In what follows, we summarise the key takeaways from the different chapters and potential avenues for further research.

Assessing the impact of missing data in a particular study, as a reader, means being at the mercy of how well these missing values, their possible causes, and the methods used to handle them have been *reported* in the first place. For example, reporting the proportion of missing data on a per-variable basis in a descriptives table is not sufficient to determine the number of individuals involved in a regression model restricted to complete-cases when combinations of variables can be missing per individual. Additionally, the rather ominous absence of any missing data in a descriptives table often suggests these were excluded at a pre-processing stage, meaning that the resulting sample may no longer be representative of the original target population.

Chapter 2 represents a first systematic assessment, to our knowledge, of how missing covariate data are reported and handled in clinical studies in haematology. In this systematic review, almost two-thirds (195 of 299) included articles published in 2021 across major haematological journals reported missing values in one of more covariates

from one or more of the multivariable (proportional hazards, including competing risks outcomes) models presented. While this number suggests missing covariate data are prevalent across clinical haematological studies, it is not straightforward to interpret. For example, researchers may choose to completely exclude a particular variable from a multivariable model because it is deemed to have too many missing values. Assuming the remaining covariates are complete, the resulting model (from a reviewer's perspective) is technically free of missing data concerns. Furthermore, the absence of missing data on one or more variables can be used as an inclusion criterion for a study, which may (in 80 of 299 articles) or may not be explicitly reported. Ideally, missing data should never be used as an exclusion criterion: it usually results in some selection bias unless there is high confidence the missing data used to exclude individuals are MCAR. This controversial practice has been reported in similar reviews (Baker *et al.*, 2024; Carroll *et al.*, 2020; Mainzer *et al.*, 2024), and literature has also investigated the approach of first multiply imputing missing values prior to applying any inclusion criteria in each imputed dataset (Austin *et al.*, 2023).

The systematic review also showed that a minority of articles explicitly reported the missing data handling method, and among these only 6 used MI, with complete-case analysis and the missing indicator method (MIM) being the most popular approaches. This is at odds with previous publications reporting that MI is being increasingly used (Hayati Rezvan *et al.*, 2015), although it should be noted that the present review only included publications from 2021, making it ill-suited for evaluating recent trends in the use of missing data methods across haematological journals. Furthermore, future systematic reviews should aim to extract both the number of individuals included in a given analysis, if reported at all, and the study design (e.g. clinical trial, or retrospective study). This would allow to give more context to the limited use of MI in clinical haematology: a) there is little added benefit to using MI instead of complete-case analysis when there are few missing values; b) other methods such as the MIM are often preferred over MI for non-observational designs such as clinical trials (Sullivan *et al.*, 2018; White and Thompson, 2005).

If MI is to be used, the method(s) used to impute should be reported together with various details such as the number of imputations and the variables included in the imputation model(s) (Sterne *et al.*, 2009). The latter is particularly important, since the imputation model should ideally be specified such that it is consistent, or *compatible*, with the assumptions made by the analysis model. This means the outcome should be included in the imputation model, but also that any non-linearities in the analysis model structure (e.g. interaction terms) should also be accounted for.

An imputation model (in the context of MI using chained equations, MICE) is motivated, for a given variable with missing data, by deriving the conditional distribution of that variable given the outcome and the covariates of the analysis model, and eventually also auxiliary variables. The main goal of Chapter 3 was to mathematically motivate imputation models for partially observed covariates when the analysis model of interest

is one or more cause-specific Cox proportional hazards models. In this context, we showed that a directly specified imputation for a partially observed covariate should at least include the remaining covariates from the analysis model, the competing event indicator as a factor variable, and the marginal cumulative cause-specific hazards for each competing event obtained using the Nelson–Aalen estimator (and evaluated at an individual's event or censoring time).

The aforementioned model is only *approximately* compatible with the analysis model, as it represents a linear approximation of a conditional distribution which is in fact non-linear, as analogously reported for the standard single-event Cox model (White and Royston, 2009). Additionally including the interactions of all cumulative hazards with the remaining analysis model covariates in the imputation model improves the accuracy of the approximation, but comes at the potential cost of an over-parametrised imputation model. Instead of directly specified, approximately compatible imputation models (referred to as MICE in this discussion), one could opt to use the substantive-model-compatible fully conditional specification (SMC-FCS) approach introduced by Bartlett *et al.* (2015), and adapted for cause-specific Cox outcome models by Bartlett and Taylor (2016). This 'indirect' approach samples imputed values using the 'true' (i.e. implied assuming the analysis model is correctly specified) conditional distribution of a partially observed covariate given the outcome and remaining analysis model covariates, without needing to rely on any approximations.

The large-scale simulation study in Chapter 3 showed that SMC-FCS outperformed MICE across a range of scenarios, particularly in terms of bias when estimating cause-specific hazard ratios. Increased bias by using MICE approaches (with and without inclusion of cumulative cause-specific hazard interactions with covariates) was caused by larger covariate effects, higher proportion of missing values, weaker missingness mechanisms, 'different' baseline hazard shapes, and the partially observed covariate being continuous. Note that ignoring competing risks at the imputation stage (i.e. by omitting the indicator and cause-specific cumulative hazard of the competing event) was the approach that performed the poorest. Furthermore, the impact of different missing data handling approaches on the estimation of individual-specific cumulative incidence functions was also a novel contribution of this work. Here, all imputation-based approaches performed more comparably.

In Chapter 4, we applied the approximately compatible MICE approach from Chapter 3 to a multi-centre cohort of 4086 patients with myelofibrosis undergoing an alloSCT, where the research question primarily concerned the impact of partially observed comorbidities (as summarised by the hematopoietic cell transplantation comorbidity index, HCT-CI, developed in Sorror *et al.*, 2005) and body mass index (BMI) on the cause-specific hazard of non-relapse mortality (NRM). We primarily chose to use the MICE approach over SMC-FCS in order to impute skewed continuous covariates, such as peripheral blood blasts, using predictive mean matching (Kleinke, 2017; Lee and Carlin, 2017). However, there is also a lack of research regarding the use of SMC-FCS

when multiple substantive models are of interest, and when auxiliary covariates are used—both of which played a role in this study.

Furthermore, both BMI and HCT-CI are so-called *derived* variables: BMI is calculated based on an individual's weight and height, while the HCT-CI is a weighted summary of the presence or absence of multiple comorbidities. This was a non-standard situation where two derived variables had to be imputed, which: a) are of different types, since BMI is a ratio while HCT-CI is additive; b) partly overlap, since BMI over 35 is a constituent of HCT-CI; c) have different missing data patterns for their constituents, as those with missing BMI generally missed both height and weight, while the patterns for the constituents of HCT-CI was more varied. Based on the work in Clements (2022) (see Figure 6.1), we chose to impute BMI directly as a continuous covariate, and impute the individual constituents of the HCT-CI.

Chapter 5 in turn focused on the development of MI methodology when the analysis model is a Fine–Gray subdistribution hazard model. The proposed SMC-FCS and MICE approaches are tailored for a single competing event of interest, and rely on the parallels between the Fine–Gray model and the standard Cox model. Specifically, the potential censoring times for those failing from competing events are multiply imputed in a first step (Ruan and Gray, 2008), and then covariates are imputed using the resulting 'censoring-complete' datasets. We showed that approximately compatible imputation models should include as predictors the remaining covariates of the outcome model, the indicator for the competing event of interest, and the marginal cumulative subdistribution hazard for the event of interest (evaluated at an individual's actual or imputed *subdistribution time*). The proposed SMC-FCS approach was integrated into the open source R package {smcfcs} (Bartlett *et al.*, 2022).

The simulation study assessed the performance of the proposed MI approaches, additionally comparing them to imputing (approximately) compatibly with cause-specific Cox models, in scenarios where proportional hazards hold on either of the cause-specific hazard or subdistribution hazard scales. The SMC-FCS approach which imputes compatibly with the correct underlying outcome model was always unbiased. The bias of competitor MI approaches depended on both the (baseline) proportion of failures from the cause of interest, and the presence/absence of any censoring. Interestingly, the presence of censoring *improved* the performance of the misspecified SMC-FCS approach, as it appears to 'soften' the violation of the proportionality assumption at the imputation stage. Furthermore, the simulation study corroborated the results of Chapter 3 in terms of individual-specific cumulative incidence estimation: the differences between imputation approaches were much less pronounced.

Chapter 3 and Chapter 5, in conjunction with previous simulation studies (Bartlett and Taylor, 2016), suggest that SMC-FCS should be the go-to for imputing with missing (at-random) covariate data for major competing risks regression models. SMC-FCS can reflect the proportional hazards structure of both cause-specific Cox and Fine–Gray

models, and any assumed interactions or other non-linear effects are automatically accounted for. Nevertheless, there are several caveats to consider.

First, the aforementioned simulation studies predominantly considered settings where a) the analysis model was well specified; b) the covariate space was low-dimensional; c) missingness was mainly univariate and MAR; d) covariate effects were rather extreme (e.g. log-hazard ratio of 1 for continuous covariates). Points a) and d) in particular put SMC-FCS at an advantage with respect to MICE, particularly when method 'norm' (standard linear regression) is used to impute using MICE. The strong non-linear relationship between covariates and outcome in simulated proportional hazards data with large covariate effects implies that MICE using predictive mean matching, classification and regression trees (CART), or random forests may be more appropriate comparator methods. Regarding point c), subsequent simulations by Austin (2024) considered more realistic (i.e. based on real data) multivariate missing data patterns and covariates effects, however the cause-specific Cox analysis models were still correctly specified. In the aforementioned simulations, no imputation-based approach uniformly outperformed other approaches. There is a clear need for simulation studies which assess the performance of these imputation approaches in more complex settings where analysis models will not be correctly specified. One could for example use non-parametric data synthesising approaches to generate simulated datasets (Nowok *et al.*, 2016), or use resampling-based methods (e.g. as done in Shah *et al.*, 2014).

Second, Chapter 3 and Chapter 5 both applied the various MI approaches to different alloSCT datasets. In both cases, both MICE and SMC-FCS methods performed extremely comparably. Similar results were found in the case studies in Bartlett and Taylor (2016) and Austin (2024). It would seem that in real datasets, where covariate effects are usually modest, the specified analysis model (usually with linear effects) is not the 'true' data-generating model, and missingness will never completely be at-random, that MICE and SMC-FCS are expected to perform similarly. When interactions or spline terms are included in the analysis model, and their effects are non-negligible, it is arguably preferable to use SMC-FCS. Note that for both approaches, the gain in efficiency relative to a complete-case analysis can be substantial, even without additional auxiliary variables.

Third, SMC-FCS and MICE approaches are both subject to possible biases (when missingness is multivariate) due to *mutual incompatibility* of imputation models, which can occur when covariates are non-linearly related to each other. Joint modelling MI approaches, such as using a latent multivariate normal model (e.g. Quartagno and Carpenter, 2019) or a fully Bayesian approach using sequential factorisation (e.g. Erler *et al.*, 2016), circumvent this issue and can also impute compatibly with a specified analysis model. Neither approach has yet been extended to accommodate competing risks outcomes. In sum, based especially on Chapter 3 and the simulations in Bartlett and Taylor (2016), a more concrete advice is to make sure not to ignore competing events when imputing missing covariates (whether that be using SMC-FCS or MICE), even if

only scientifically interested in one event. Note that only including the subdistribution hazard and event indicator for only one event is not equivalent to ignoring competing risks, since the subdistribution hazard (specifically, the subdensity in the numerator) does depend on the cause-specific cumulative hazard of the competing event.

Chapter 3 and Chapter 5 both describe MI approaches for imputing missing covariates consistently with popular regression models for competing risks outcomes. As described in Lee *et al.* (2021) with focus on observational studies, the analysis model of interest should usually first be specified without consideration of any missing data. Chapter 6 examines this question of analysis model choice for competing risks from a data-generating perspective. Specifically, we provided an overview of data-generating mechanisms in which the Fine–Gray model is correctly specified for at least one cause. A core conclusion was that specifying a Fine–Gray model for each competing event should be avoided, since there is no data-generating mechanism for which the assumption of proportional subdistribution hazards holds simultaneously for all causes, unless finite follow-up and a bounded covariate space are assumed. When interested in more than one competing event, cause-specific hazard models usually represent a more flexible alternative.

The SMC-FCS approach presented in Chapter 5 is tailored for one competing event of interest, and as such avoids having to specify a model for other competing events. Nevertheless, an alternative SMC-FCS approach could be developed which *does* specify a model for the competing events. In this regard, Chapter 6 provides an overview of possible assumptions that can be made regarding the competing event (at the imputation stage) when one assumes proportional subdistribution hazards for an event of interest. A potential advantage of a SMC-FCS approach based for example on the 'squeezing' data-generating mechanism, is that it would rely on a broader MAR assumption than the assumption made in Chapter 5 (namely, the same one as in Chapter 3, where missingness is allowed to depend on the observed outcomes for individuals experiencing competing events).

Instead of baseline covariates, the focus of Chapter 7 is on *longitudinal* covariates. Specifically, we modelled the trajectories of immune cell counts for a single-center cohort of 166 acute leukaemia patients in the first 6 months following a T-cell depleted alloSCT using joint modelling. Furthermore, we quantified the associations of various immune cell counts with the cause-specific hazards of competing events graft-versus-host disease (GvHD), disease relapse, and other failures (e.g. death). A previously published joint model also assessing immune cell kinetics after alloSCT did not consider post-baseline interventions such as donor lymphocyte infusions (DLI) or competing events in the time-to-event submodel (Salzmann-Manrique *et al.*, 2018), while other work in the domain of alloSCT considered different longitudinal measurements such as minimal residual disease (Huang *et al.*, 2021) or donor chimerism (Tang *et al.*, 2014). We were able to capture the shape of highly non-linear immune cell count trajectories with flexible modelling of time using splines, and the impact of a scheduled early low-dose

DLI was evaluated using a three-way interaction in the longitudinal submodel. The estimated opposing effects of (current value) CD4+ cell counts on GvHD and disease relapse highlighted the importance of accounting for competing risks in this setting.

Extensions to the work in Chapter 7 will require larger sample sizes, more frequent longitudinal measurements, and advances in statistical software. For example, one could consider a multivariate longitudinal submodel, to account for the correlations between CD4+, CD8+ and NK cells, rather than modelling them separately. Additionally, one could specify a multi-state structure for the time-to-event submodel, in order to also account for deaths after relapse or GvHD. While both of these extensions can be implemented using the {JMbayes2} package (Rizopoulos *et al.*, 2023), the limited number of patients and events in the cohort from Chapter 7 implies that fitting such a model would require any one or a combination of a) greatly simplifying the covariate structure of the longitudinal submodels; b) reducing flexibility in the transition-specific baseline hazards; c) applying regularisation. It would additionally be useful for the current application to compare in detail the joint model estimates to those obtained with mixed models and time-varying Cox models. This comparison would also allow to assess how sensitive the estimated immune cell count trajectories are to the different missing data assumptions made (see Rouanet *et al.*, 2019 for more general discussion), and to quantify the added benefit of using joint models over time-varying Cox models when estimating association parameters.

In conclusion, this dissertation developed and assessed statistical methodology for handling missing data in competing risks settings. Various applications using datasets of patients undergoing an alloSCT were presented, and underline the importance of appropriately accounting for competing risks at both the analysis phase (e.g. existence of opposing effects on cause-specific hazards) and when dealing with missing data. Broader application of the methodologies for imputing missing covariate data will first and foremost rely on collaboration between clinical investigators, data managers, and statisticians, in order to establish plausible reasons why data are missing, or ideally ensure more robust data collection for future studies. In addition to more principled application and thorough reporting of MI in clinical studies, its use should (where appropriate) be supplemented with sensitivity analyses that assess violations of the MAR assumption (e.g. using delta-adjustment as in Tompsett *et al.*, 2018).

Furthermore, throughout this thesis, we placed a particular emphasis on Open Science and its principles (Vicente-Saez and Martinez-Fuentes, 2018). Analyses for the different chapters are implemented using the free software R (R Core Team, 2024), and documented code is openly available at https://github.com/survival-lumc. All of the introduced statistical methodology in this thesis is accompanied by minimal code which seeks to leverage existing, regularly maintained, open source R packages. In particular, we hope that the sharing of simulation study code helps stimulate both future a) 'neutral' method comparison studies (Boulesteix *et al.*, 2013); b) replication

studies, which are notoriously rare in methodological research (Boulesteix *et al.*, 2020; Lohmann *et al.*, 2022).

Bibliography

Aalen, O., Borgan, O. and Gjessing, H. (2008) *Survival and Event History Analysis: A Process Point of View*. DOI: 10.1007/978-0-387-68560-1.

Aalen, O. O. and Johansen, S. (1978) An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, **5**, 141–150. Available at: https://www.jstor.org/stable/4615704.

Acosta-Medina, A. A., Baranwal, A., Johnson, I. M., et al. (2023) Comparison of Pretransplantation Prediction Models for Nonrelapse Mortality in Patients with Myelofibrosis Undergoing Allogeneic Stem Cell Transplantation. *Transplantation and Cellular Therapy*, **29**, 360.e1–360.e8. DOI: 10.1016/j.jtct.2023.02.002.

Adès, L., Itzykson, R. and Fenaux, P. (2014) Myelodysplastic syndromes. *The Lancet*, **383**, 2239–2252. DOI: 10.1016/S0140-6736(13)61901-7.

Allignol, A. and Beyersmann, J. (2010) Software for fitting nonstandard proportional subdistribution hazards models. *Biostatistics*, **11**, 674–675. DOI: 10.1093/biostatistics/kxq018.

Andersen, P. K. and Keiding, N. (2012) Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, **31**, 1074–1088. DOI: 10.1002/sim.4385.

Andersen, P. K. and Ravn, H. (2023) *Models for Multi-State Survival Data: Rates, Risks, and Pseudo-Values*.

Andersen, P. K., Abildstrom, S. Z. and Rosthøj, S. (2002) Competing risks as a multi-state model. *Statistical Methods in Medical Research*, **11**, 203–215. DOI: 10.1191/0962280202sm281ra.

Andersen, P. K., Geskus, R. B., de Witte, T., et al. (2012) Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology*, **41**, 861–870.

DOI: 10.1093/ije/dyr213.

Andrinopoulou, E.-R., Rizopoulos, D., Geleijnse, M. L., et al. (2015) Dynamic prediction of outcome for patients with severe aortic stenosis: Application of joint models for longitudinal and time-to-event data. *BMC Cardiovascular Disorders*, **15**, 28. DOI: 10.1186/s12872-015-0035-z.

Antunes, L., Mendonça, D., Bento, M. J., et al. (2021) Dealing with missing information on covariates for excess mortality hazard regression models – Making the imputation model compatible with the substantive model. *Statistical Methods in Medical Research*, **30**, 2256–2268. DOI: 10.1177/09622802211031615.

Archer, L., Koshiaris, C., Lay-Flurrie, S., et al. (2022) Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment: Retrospective cohort study. *BMJ*, **379**, e070918. DOI: 10.1136/bmj-2022-070918.

Armand, P., Gibson, C. J., Cutler, C., et al. (2012) A disease risk index for patients undergoing allogeneic stem cell transplantation. *Blood*, **120**, 905–913. DOI: 10.1182/blood-2012-03-418202.

Armand, P., Kim, H. T., Logan, B. R., et al. (2014) Validation and refinement of the Disease Risk Index for allogeneic stem cell transplantation. *Blood*, **123**, 3664–3671. DOI: 10.1182/blood-2014-01-552984.

Austin, P. C. (2024) Multiple imputation with competing risk outcomes. *Computational Statistics*. DOI: 10.1007/s00180-024-01518-w.

Austin, P. C., Steyerberg, E. W. and Putter, H. (2021) Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Statistics in Medicine*, **40**, 4200–4212. DOI: 10.1002/sim.9023.

Austin, P. C., Giardiello, D. and van Buuren, S. (2023) Impute-then-exclude versus exclude-then-impute: Lessons when imputing a variable used both in cohort creation and as an independent variable in the analysis model. *Statistics in Medicine*, **42**, 1525–1541. DOI: 10.1002/sim.9685.

Baker, W. L., Moore, T., Baron, E., et al. (2024) A Systematic Review of Reporting and Handling of Missing Data in Observational Studies Using the UNOS Database. *The Journal of Heart and Lung Transplantation*. DOI: 10.1016/j.healun.2024.10.023.

Bakoyannis, G., Siannis, F. and Touloumi, G. (2010) Modelling competing risks

data with missing cause of failure. *Statistics in Medicine*, **29**, 3172–3185. DOI: 10.1002/sim.4133.

Bartlett, J. W. and Taylor, J. M. G. (2016) Missing covariates in competing risks analysis. *Biostatistics*, **17**, 751–763. DOI: 10.1093/biostatistics/kxw019.

Bartlett, J. W., Seaman, S. R., White, I. R., et al. (2015) Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, **24**, 462–487. DOI: 10.1177/0962280214521348.

Bartlett, J. W., Keogh, R. H. and Bonneville, E. F. (2022) *Smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification*. Manual.

Beesley, L. J., Bartlett, J. W., Wolf, G. T., et al. (2016) Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Statistics in Medicine*, **35**, 4701–4717. DOI: 10.1002/sim.7048.

Bell, M. L., Fiero, M., Horton, N. J., et al. (2014) Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*, **14**, 118. DOI: 10.1186/1471-2288-14-118.

Bellach, A., Kosorok, M. R., Rüschendorf, L., et al. (2019) Weighted NPMLE for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, **114**, 259–270. DOI: 10.1080/01621459.2017.1401540.

Bellucci, R., Alyea, E. P., Weller, E., et al. (2002) Immunologic effects of prophylactic donor lymphocyte infusion after allogeneic marrow transplantation for multiple myeloma. *Blood*, **99**, 4610–4617. DOI: 10.1182/blood.V99.12.4610.

Beyersmann, J., Latouche, A., Buchholz, A., et al. (2009) Simulating competing risks data in survival analysis. *Statistics in Medicine*, **28**, 956–971. DOI: 10.1002/sim.3516.

Beyersmann, J., Allignol, A. and Schumacher, M. (2012) *Competing Risks and Multistate Models with R*. Use R! DOI: 10.1007/978-1-4614-2035-4.

Blake, H. A., Leyrat, C., Mansfield, K. E., et al. (2020) Estimating treatment effects with partially observed covariates using outcome regression with missing indicators. *Biometrical Journal*, **62**, 428–443. DOI: 10.1002/bimj.201900041.

Bonneville, E. F., Resche-Rigon, M., Schetelig, J., et al. (2022) Multiple imputation for cause-specific Cox models: Assessing methods for estimation and prediction. *Statistical Methods in Medical Research*, **31**, 1860–1880. DOI: 10.1177/09622802221102623.

Bonneville, E. F., Schetelig, J., Putter, H., et al. (2023) Handling missing covariate data in clinical studies in haematology. *Best Practice & Research Clinical Haematology*, **36**, 101477. DOI: 10.1016/j.beha.2023.101477.

Bonneville, E. F., de Wreede, L. C. and Putter, H. (2024) Why you should avoid using multiple Fine–Gray models: Insights from (attempts at) simulating proportional subdistribution hazards data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae056. DOI: 10.1093/jrsssa/qnae056.

Borgan, Ø. and Keogh, R. (2015) Nested case–control studies: Should one break the matching? *Lifetime Data Analysis*, **21**, 517–541. DOI: 10.1007/s10985-015-9319-y.

Bosch, M., Dhadda, M., Hoegh-Petersen, M., et al. (2012) Immune reconstitution after anti-thymocyte globulin-conditioned hematopoietic cell transplantation. *Cytotherapy*, **14**, 1258–1275. DOI: 10.3109/14653249.2012.715243.

Boulesteix, A.-L., Lauer, S. and Eugster, M. J. A. (2013) A Plea for Neutral Comparison Studies in Computational Sciences. *PLOS ONE*, **8**, e61562. DOI: 10.1371/journal.pone.0061562.

Boulesteix, A.-L., Hoffmann, S., Charlton, A., et al. (2020) A Replication Crisis in Methodological Research? *Significance*, **17**, 18–21. DOI: 10.1111/1740-9713.01444.

Breslow, N. E. (1972) Contribution to discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B*, **34**, 216–217.

Burton, A. and Altman, D. G. (2004) Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer*, **91**, 4–8. DOI: 10.1038/sj.bjc.6601907.

Busca, A. and Aversa, F. (2017) In-vivo or ex-vivo T cell depletion or both to prevent graft-versus-host disease after hematopoietic stem cell transplantation. *Expert Opinion on Biological Therapy*, **17**, 1401–1415. DOI: 10.1080/14712598.2017.1369949.

Carpenter, J. R. and Smuk, M. (2021) Missing data: A statistical framework for practice. *Biometrical Journal*, **63**, 915–947. DOI: 10.1002/bimj.202000196.

Carpenter, J. R., Bartlett, J. W., Morris, T. P., et al. (2023) *Multiple Imputation and Its Application.* 2nd ed. Statistics in practice series.

Carroll, O. U., Morris, T. P. and Keogh, R. H. (2020) How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. *BMC Medical Research Methodology*, **20**, 134. DOI: 10.1186/s12874-020-01018-7.

Chiang, C. L. (1961) A Stochastic Study of the Life Table and Its Applications. III. The Follow-Up Study with the Consideration of Competing Risks. *Biometrics*, **17**, 57–78. DOI: 10.2307/2527496.

Clements, L. Z. (2022) *Multiple imputation of a derived variable in a survival analysis context*. PhD thesis.

Clift, A. K., Coupland, C. A. C., Keogh, R. H., et al. (2020) Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: National derivation and validation cohort study. *BMJ*, **371**, m3731. DOI: 10.1136/bmj.m3731.

Cornfield, J. (1957) Estimation of the Probability of Developing a Disease in the Presence of Competing Risks. *American Journal of Public Health and the Nations Health*, **47**, 601–607. DOI: 10.2105/ajph.47.5.601.

Cox, D. R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276. DOI: 10.1093/biomet/62.2.269.

D'Amico, G., Morabito, A., D'Amico, M., et al. (2018) Clinical states of cirrhosis and competing risks. *Journal of Hepatology*, **68**, 563–576. DOI: 10.1016/j.jhep.2017.10.020.

de Wreede, L. C., Fiocco, M. and Putter, H. (2010) The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, **99**, 261–274. DOI: 10.1016/j.cmpb.2010.01.001.

de Wreede, L. C., Fiocco, M. and Putter, H. (2011) Mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, **38**, 1–30. DOI: 10.18637/jss.v038.i07.

Dekker, L., de Koning, C., Lindemans, C., et al. (2020) Reconstitution of T Cell Subsets Following Allogeneic Hematopoietic Cell Transplantation. *Cancers*, **12**, 1974. DOI: 10.3390/cancers12071974.

Delgado, J., Pereira, A., Villamor, N., et al. (2014) Survival analysis in hematologic malignancies: Recommendations for clinicians. *Haematologica*, **99**, 1410–1420. DOI: 10.3324/haematol.2013.100784.

Delord, M. and Génin, E. (2016) Multiple imputation for competing risks regression

with interval-censored data. *Journal of Statistical Computation and Simulation*, **86**, 2217–2228. DOI: 10.1080/00949655.2015.1106543.

Demirtas, H. and Hedeker, D. (2016) Computing the Point-biserial Correlation under Any Underlying Continuous Distribution. *Communications in Statistics - Simulation and Computation*, **45**, 2744–2751. DOI: 10.1080/03610918.2014.920883.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., et al. (2006) Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, **59**, 1087–1091. DOI: 10.1016/j.jclinepi.2006.01.014.

Donoghoe, M. W. and Gebski, V. (2017) The importance of censoring in competing risks analysis of the subdistribution hazard. *BMC Medical Research Methodology*, **17**, 52. DOI: 10.1186/s12874-017-0327-3.

Du, H., Alacam, E., Mena, S., et al. (2022) Compatibility in imputation specification. *Behavior Research Methods*. DOI: 10.3758/s13428-021-01749-5.

Dunbar, E. M., Buzzeo, M. P., Levine, J. B., et al. (2008) The relationship between circulating natural killer cells after reduced intensity conditioning hematopoietic stem cell transplantation and relapse-free survival and graft-versus-host disease. *Haematologica*, **93**, 1852–1858. DOI: 10.3324/haematol.13033.

Eefting, M., Halkes, C. J. M., de Wreede, L. C., et al. (2014) Myeloablative T cell-depleted alloSCT with early sequential prophylactic donor lymphocyte infusion is an efficient and safe post-remission treatment for adult ALL. *Bone Marrow Transplantation*, **49**, 287–291. DOI: 10.1038/bmt.2013.111.

Eefting, M., de Wreede, L. C., Halkes, C. J. M., et al. (2016) Multi-state analysis illustrates treatment success after stem cell transplantation for acute myeloid leukemia followed by donor lymphocyte infusion. *Haematologica*, **101**, 506–514. DOI: 10.3324/haematol.2015.136846.

Elfeky, R., Lazareva, A., Qasim, W., et al. (2019) Immune reconstitution following hematopoietic stem cell transplantation using different stem cell sources. *Expert Review of Clinical Immunology*, **15**, 735–751. DOI: 10.1080/1744666X.2019.1612746.

Enders, C. K. (2022) *Applied Missing Data Analysis, 2nd Ed.* Applied missing data analysis, 2nd ed.

Erler, N. S., Rizopoulos, D., Rosmalen, J. van, et al. (2016) Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, **35**, 2955–2974. DOI: 10.1002/sim.6944.

Erler, N. S., Rizopoulos, D. and Lesaffre, E. M. E. H. (2021) JointAI: Joint Analysis and Imputation of Incomplete Data in R. *Journal of Statistical Software*, **100**, 1–56. DOI: 10.18637/jss.v100.i20.

Falcaro, M., Nur, U., Rachet, B., et al. (2015) Estimating Excess Hazard Ratios and Net Survival When Covariate Data Are Missing: Strategies for Multiple Imputation. *Epidemiology*, **26**, 421–428. DOI: 10.1097/EDE.0000000000000283.

Falkenburg, J. H. F. and Jedema, I. (2017) Graft versus tumor effects and why people relapse. *Hematology*, **2017**, 693–698. DOI: 10.1182/asheducation-2017.1.693.

Falkenburg, J. H. F., Schmid, C., Kolb, H. J., et al. (2019) Delayed transfer of immune cells or the art of donor lymphocyte infusion. *The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies*, 443–448. DOI: 10.1007/978-3-030-02278-5_59.

Fine, J. P. and Gray, R. J. (1999) A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, **94**, 496–509. DOI: 10.1080/01621459.1999.10474144.

Fleischhauer, K., Shaw, B. E., Gooley, T., et al. (2012) Effect of T-cell-epitope matching at HLA-DPB1 in recipients of unrelated-donor haemopoietic-cell transplantation: A retrospective study. *The Lancet Oncology*, **13**, 366–374. DOI: 10.1016/S1470-2045(12)70004-9.

Gasparini, A. (2018) Rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, **3**, 739. DOI: 10.21105/joss.00739.

Gerds, T. A., Scheike, T. H. and Andersen, P. K. (2012) Absolute risk regression for competing risks: Interpretation, link functions, and prediction. *Statistics in Medicine*, **31**, 3921–3930. DOI: 10.1002/sim.5459.

Geskus, R. B. (2011) Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring. *Biometrics*, **67**, 39–49. DOI: 10.1111/j.1541-0420.2010.01420.x.

Geskus, R. B. (2024) Competing Risks: Concepts, Methods, and Software. *Annual Review of Statistics and Its Application*, **11**, 227–254. DOI: 10.1146/annurev-statistics-040522-094556.

Gooptu, M., Kim, H. T., Howard, A., et al. (2019) Effect of Sirolimus on Immune Reconstitution Following Myeloablative Allogeneic Stem Cell Transplantation: An Ancillary Analysis of a Randomized Controlled Trial Comparing Tacrolimus/Sirolimus

and Tacrolimus/Methotrexate (Blood and Marrow Transplant Clinical Trials Network/BMT CTN 0402). *Biology of Blood and Marrow Transplantation*, **25**, 2143–2151. DOI: 10.1016/j.bbmt.2019.06.029.

Gorgi Zadeh, S., Behning, C. and Schmid, M. (2022) An imputation approach using subdistribution weights for deep survival analysis with competing events. *Scientific Reports*, **12**, 3815. DOI: 10.1038/s41598-022-07828-7.

Grambauer, N., Schumacher, M. and Beyersmann, J. (2010) Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine*, **29**, 875–884. DOI: 10.1002/sim.3786.

Gray, R. J. (1988) A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of Statistics*, **16**, 1141–1154. DOI: 10.1214/aos/1176350951.

Groenwold, R. H. H., White, I. R., Donders, A. R. T., et al. (2012) Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *CMAJ*, **184**, 1265–1269. DOI: 10.1503/cmaj.110977.

Guillaume, T., Gaugler, B., Chevallier, P., et al. (2012) Escalated lymphodepletion followed by donor lymphocyte infusion can induce a graft-versus-host response without overwhelming toxicity. *Bone Marrow Transplantation*, **47**, 1112–1117. DOI: 10.1038/bmt.2011.231.

Haensch, A.-C., Bartlett, J. and Weiß, B. (2022) Multiple imputation of partially observed covariates in discrete-time survival analysis. *Sociological Methods & Research*, 00491241221140147. DOI: 10.1177/00491241221140147.

Haller, B. and Ulm, K. (2014) Flexible simulation of competing risks data following prespecified subdistribution hazards. *Journal of Statistical Computation and Simulation*, **84**, 2557–2576. DOI: 10.1080/00949655.2013.793345.

Han, S. Y., Mrozek K., Voutsinas J., et al. (2021) Secondary cytogenetic abnormalities in core-binding factor AML harboring inv(16) vs t(8;21). *Blood Advances*, **5**, 2481–2489. DOI: 10.1182/BLOODADVANCES.2020003605.

Hansen, D. K., Kim J., Thompson Z., et al. (2021) ELN 2017 Genetic Risk Stratification Predicts Survival of Acute Myeloid Leukemia Patients Receiving Allogeneic Hematopoietic Stem Cell Transplantation. *Transplantation and Cellular Therapy*, **27**, 256.e1–256.e7. DOI: 10.1016/j.jtct.2020.12.021.

Hassan, N., Eldershaw, S., Stephens, C., et al. (2022) CMV reactivation initiates

long-term expansion and differentiation of the NK cell repertoire. *Frontiers in Immunology*, **13**. DOI: 10.3389/fimmu.2022.935949.

Hayati Rezvan, P., Lee, K. J. and Simpson, J. A. (2015) The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, **15**, 30. DOI: 10.1186/s12874-015-0022-1.

Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and Coarse Data. *The Annals of Statistics*, **19**, 2244–2253. DOI: 10.1214/aos/1176348396.

Hickey, G. L., Philipson, P., Jorgensen, A., et al. (2016) Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology*, **16**, 117. DOI: 10.1186/s12874-016-0212-5.

Hickey, G. L., Philipson, P., Jorgensen, A., et al. (2018) A Comparison of Joint Models for Longitudinal and Competing Risks Data, with Application to an Epilepsy Drug Randomized Controlled Trial. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **181**, 1105–1123. DOI: 10.1111/rssa.12348.

Hinchliffe, S. R. and Lambert, P. C. (2013) Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Medical Research Methodology*, **13**, 13. DOI: 10.1186/1471-2288-13-13.

Horowitz, M., Gale, R., Sondel, P., et al. (1990) Graft-versus-leukemia reactions after bone marrow transplantation. *Blood*, **75**, 555–562. DOI: 10.1182/blood.V75.3.555.555.

Horowitz, M. M. (2008) The role of registries in facilitating clinical research in BMT: Examples from the Center for International Blood and Marrow Transplant Research. *Bone Marrow Transplantation*, **42**, S1–S2. DOI: 10.1038/bmt.2008.101.

Huang, A., Chen, Q., Fei, Y., et al. (2021) Dynamic prediction of relapse in patients with acute leukemias after allogeneic transplantation: Joint model for minimal residual disease. *International Journal of Laboratory Hematology*, **43**, 84–92. DOI: 10.1111/ijlh.13328.

Hughes, R. A., Heron, J., Sterne, J. A. C., et al. (2019) Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, **48**, 1294–1304. DOI: 10.1093/ije/dyz032.

Inoue, Y., Nakano N., Fuji S., et al. (2021) Impact of conditioning intensity and regimen on transplant outcomes in patients with adult T-cell leukemia-lymphoma. *Bone Marrow Transplantation*, **56**, 2964–2974. DOI: 10.1038/s41409-021-01445-0.

Ito, A., Kitano, S., Tajima, K., et al. (2020) Impact of low-dose anti-thymocyte globulin on immune reconstitution after allogeneic hematopoietic cell transplantation. *International Journal of Hematology*, **111**, 120–130. DOI: 10.1007/s12185-019-02756-1.

Jeong, J.-H. and Fine, J. (2006) Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**, 187–200. DOI: 10.1111/j.1467-9876.2006.00532.x.

Kalbfleisch, J. D. and Prentice, R. L. (2011) *The Statistical Analysis of Failure Time Data*. Wiley series in probability and statistics.

Kantidakis, G., Putter, H., Litière, S., et al. (2023) Statistical models versus machine learning for competing risks: Development and validation of prognostic models. *BMC Medical Research Methodology*, **23**, 51. DOI: 10.1186/s12874-023-01866-z.

Kaplan, E. L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457–481. DOI: 10.2307/2281868.

Keogh, R. H. and Morris, T. P. (2018) Multiple imputation in Cox regression when there are time-varying effects of covariates. *Statistics in Medicine*, **37**, 3661–3678. DOI: 10.1002/sim.7842.

Kipourou, D.-K., Charvat, H., Rachet, B., et al. (2019) Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Statistics in Medicine*, **38**, 3896–3910. DOI: 10.1002/sim.8209.

Klein, J. P. and Andersen, P. K. (2005) Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function. *Biometrics*, **61**, 223–229. DOI: 10.1111/j.0006-341X.2005.031209.x.

Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data (2nd Edition)*.

Kleinke, K. (2017) Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. *Journal of Educational and Behavioral Statistics*, **42**, 371–404. DOI: 10.3102/1076998616687084.

Koster, E. A. S., von dem Borne, P. A., van Balen, P., et al. (2023) Competitive Repopulation and Allo-Immunologic Pressure Determine Chimerism Kinetics after T Cell-Depleted Allogeneic Stem Cell Transplantation and Donor Lympho-

cyte Infusion. *Transplantation and Cellular Therapy*, **29**, 268.e1–268.e10. DOI: 10.1016/j.jtct.2022.12.022.

Krishnamurthy, P., Potter, V. T., Barber, L. D., et al. (2013) Outcome of Donor Lymphocyte Infusion after T Cell–depleted Allogeneic Hematopoietic Stem Cell Transplantation for Acute Myelogenous Leukemia and Myelodysplastic Syndromes. *Biology of Blood and Marrow Transplantation*, **19**, 562–568. DOI: 10.1016/j.bbmt.2012.12.013.

Kröger, N., Bacigalupo, A., Barbui, T., et al. (2024) Indication and management of allogeneic haematopoietic stem-cell transplantation in myelofibrosis: Updated recommendations by the EBMT/ELN International Working Group. *The Lancet Haematology*, **11**, e62–e74. DOI: 10.1016/S2352-3026(23)00305-8.

Lambert, P. C., Wilkes, S. R. and Crowther, M. J. (2017) Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in Medicine*, **36**, 1429–1446. DOI: 10.1002/sim.7208.

Landau, W. M. (2021) The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, **6**, 2959. DOI: 10.21105/joss.02959.

Latouche, A., Boisson, V., Chevret, S., et al. (2007) Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine*, **26**, 965–974. DOI: 10.1002/sim.2600.

Latouche, A., Allignol, A., Beyersmann, J., et al. (2013) A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, **66**, 648–653. DOI: 10.1016/j.jclinepi.2012.09.017.

Lau, B. and Lesko, C. (2018) Missingness in the Setting of Competing Risks: From Missing Values to Missing Potential Outcomes. *Current Epidemiology Reports*, **5**, 153–159. DOI: 10.1007/s40471-018-0142-3.

Lee, K. J. and Carlin, J. B. (2017) Multiple imputation in the presence of non-normal data. *Statistics in Medicine*, **36**, 606–617. DOI: 10.1002/sim.7173.

Lee, K. J., Tilling, K. M., Cornish, R. P., et al. (2021) Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology*, **134**, 79–88. DOI: 10.1016/j.jclinepi.2021.01.008.

Lee, K. J., Carlin, J. B., Simpson, J. A., et al. (2023) Assumptions and analysis planning in studies with missing data in multiple variables: Moving beyond the

MCAR/MAR/MNAR classification. *International Journal of Epidemiology*, **52**, 1268–1275. DOI: 10.1093/ije/dyad008.

Lennon, H., Sperrin, M., Badrick, E., et al. (2016) The Obesity Paradox in Cancer: A Review. *Current Oncology Reports*, **18**, 56. DOI: 10.1007/s11912-016-0539-4.

Lewalle, P., Triffet, A., Delforge, A., et al. (2003) Donor lymphocyte infusions in adult haploidentical transplant: A dose finding study. *Bone Marrow Transplantation*, **31**, 39–44. DOI: 10.1038/sj.bmt.1703779.

Little, R. J. (2021) Missing Data Assumptions. *Annual Review of Statistics and Its Application*, **8**, 89–107. DOI: 10.1146/annurev-statistics-040720-031104.

Little, R. J. (2024) Missing Data Analysis. *Annual Review of Clinical Psychology*, **20**, 149–173. DOI: 10.1146/annurev-clinpsy-080822-051727.

Little, R. J. and Rubin, D. B. (2019) *Statistical Analysis with Missing Data*.

Little, R. J. A. (1995) Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, **90**, 1112–1121. DOI: 10.1080/01621459.1995.10476615.

Little, R. J. A. and Rubin, D. B. (1987) Statistical analysis with missing data. *New York: Wiley*.

Liu, J., Gelman, A., Hill, J., et al. (2014) On the stationary distribution of iterative imputations. *Biometrika*, **101**, 155–173. DOI: 10.1093/biomet/ast044.

Lohmann, A., Astivia, O. L. O., Morris, T. P., et al. (2022) It's time! Ten reasons to start replicating simulation studies. *Frontiers in Epidemiology*, **2**. DOI: 10.3389/fepid.2022.973470.

Madley-Dowd, P., Hughes, R., Tilling, K., et al. (2019) The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, **110**, 63–73. DOI: 10.1016/j.jclinepi.2019.02.016.

Mainzer, R. M., Moreno-Betancur, M., Nguyen, C. D., et al. (2024) Gaps in the usage and reporting of multiple imputation for incomplete data: Findings from a scoping review of observational studies addressing causal questions. *BMC Medical Research Methodology*, **24**, 193. DOI: 10.1186/s12874-024-02302-6.

Mao, L. and Lin, D. Y. (2017) Efficient Estimation of Semiparametric Transformation Models for the Cumulative Incidence of Competing Risks. *Journal of the*

Royal Statistical Society Series B: Statistical Methodology, **79**, 573–587. DOI: 10.1111/rssb.12177.

Mariano, L., Zhang, B. M., Osoegawa, K., et al. (2019) Assessment by Extended-Coverage Next-Generation Sequencing Typing of DPA1 and DPB1 Mismatches in Siblings Matching at HLA-A, -B, -C, -DRB1, and -DQ Loci. *Biology of Blood and Marrow Transplantation*, **25**, 2507–2509. DOI: 10.1016/j.bbmt.2019.07.033.

Marini, M. M., Olsen, A. R. and Rubin, D. B. (1980) Maximum-Likelihood Estimation in Panel Studies with Missing Data. *Sociological Methodology*, **11**, 314–357. DOI: 10.2307/270868.

Marshall, A., Altman, D. G., Holder, R. L., et al. (2009) Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Medical Research Methodology*, **9**, 57. DOI: 10.1186/1471-2288-9-57.

Marshall, A., Altman, D. G. and Holder, R. L. (2010) Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study. *BMC Medical Research Methodology*, **10**, 112. DOI: 10.1186/1471-2288-10-112.

Martinussen, T. and Stensrud, M. J. (2023) Estimation of separable direct and indirect effects in continuous time. *Biometrics*, **79**, 127–139. DOI: 10.1111/biom.13559.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC monographs on statistics and applied probability series.

McCurdy, S. R., Radojcic, V., Tsai, H.-L., et al. (2022) Signatures of GVHD and relapse after posttransplant cyclophosphamide revealed by immune profiling and machine learning. *Blood*, **139**, 608–623. DOI: 10.1182/blood.2021013054.

Meng, X.-L. (1994) Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9**, 538–558. DOI: 10.1214/ss/1177010269.

Mertens, B. J. A., Banzato, E. and de Wreede, L. C. (2020) Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation. *Biometrical Journal*, **62**, 724–741. DOI: 10.1002/bimj.201800289.

Minculescu, L., Marquart, H. V., Friis, L. S., et al. (2016) Early Natural Killer Cell Reconstitution Predicts Overall Survival in T Cell–Replete Allogeneic Hematopoietic Stem Cell Transplantation. *Biology of Blood and Marrow Transplantation*, **22**, 2187–2193. DOI: 10.1016/j.bbmt.2016.09.006.

Molenberghs, G. and Fitzmaurice, G. (2008) Incomplete data: Introduction and overview. In *Longitudinal Data Analysis*.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., et al. (2014) *Handbook of Missing Data Methodology*.

Monterrubio-Gómez, K., Constantine-Cooke, N. and Vallejos, C. A. (2024) A review on statistical and machine learning competing risks methods. *Biometrical Journal*, **66**, 2300060. DOI: 10.1002/bimj.202300060.

Morisot, A., Bessaoud, F., Landais, P., et al. (2015) Prostate cancer: Net survival and cause-specific survival rates after multiple imputation. *BMC Medical Research Methodology*, **15**, 54. DOI: 10.1186/s12874-015-0048-4.

Morris, T. P., White, I. R. and Crowther, M. J. (2019) Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**, 2074–2102. DOI: 10.1002/sim.8086.

Mozumder, S. I., Rutherford, M. and Lambert, P. (2018) Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach. *Statistics in Medicine*, **37**, 82–97. DOI: 10.1002/sim.7498.

Murray, J. S. (2018) Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, **33**, 142–159. DOI: 10.1214/18-STS644.

Nikiforow, S., Kim, H. T., Daley, H., et al. (2016) A phase I study of CD25/regulatory T-cell-depleted donor lymphocyte infusion for relapse after allogeneic stem cell transplantation. *Haematologica*, **101**, 1251–1259. DOI: 10.3324/haematol.2015.141176.

Nowok, B., Raab, G. M. and Dibben, C. (2016) Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, **74**, 1–26. DOI: 10.18637/jss.v074.i11.

Ozenne, B., Sørensen, A. L., Scheike, T., et al. (2017) riskRegression: Predicting the risk of an event using cox regression models. *The R Journal*, **9**, 440–460. DOI: 10.32614/RJ-2017-062.

Papageorgiou, G. and Rizopoulos, D. (2021) An alternative characterization of MAR in shared parameter models for incomplete longitudinal data and its utilization for sensitivity analysis. *Statistical Modelling*, **21**, 95–114. DOI: 10.1177/1471082X20927114.

Papageorgiou, G., Mauff, K., Tomer, A., et al. (2019) An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes. *Annual Review of Statistics and Its Application*, **6**, 223–240. DOI: 10.1146/annurev-statistics-030718-105048.

Passamonti, F., Cervantes, F., Vannucchi, A. M., et al. (2010) A dynamic prognostic model to predict survival in primary myelofibrosis: A study by the IWG-MRT (International Working Group for Myeloproliferative Neoplasms Research and Treatment). *Blood*, **115**, 1703–1708. DOI: 10.1182/blood-2009-09-245837.

Paz, D. L., Riou J., Verger E., et al. (2021) Genomic analysis of primary and secondary myelofibrosis redefines the prognostic impact of ASXL1 mutations: A FIM study. *Blood Advances*, **5**, 1442–1451. DOI: 10.1182/bloodadvances.2020003444.

Penack, O., Fischer, L., Stroux, A., et al. (2008) Serotherapy with thymoglobulin and alemtuzumab differentially influences frequency and function of natural killer cells after allogeneic stem cell transplantation. *Bone Marrow Transplantation*, **41**, 377–383. DOI: 10.1038/sj.bmt.1705911.

Pinheiro, J., Bates, D. and R Core Team (2023) *Nlme: Linear and Nonlinear Mixed Effects Models*. Manual.

Polverelli, N., Tura, P., Battipaglia, G., et al. (2020) Multidimensional geriatric assessment for elderly hematological patients (≥60 years) submitted to allogeneic stem cell transplantation. A French–Italian 10-year experience on 228 patients. *Bone Marrow Transplantation*, **55**, 2224–2233. DOI: 10.1038/s41409-020-0934-1.

Polverelli, N., Mauff, K., Kröger, N., et al. (2021) Impact of spleen size and splenectomy on outcomes of allogeneic hematopoietic cell transplantation for myelofibrosis: A retrospective analysis by the chronic malignancies working party on behalf of European society for blood and marrow transplantation (EBMT). *American Journal of Hematology*, **96**, 69–79. DOI: 10.1002/ajh.26020.

Polverelli, N., Hernández-Boluda, J. C., Czerw, T., et al. (2023) Splenomegaly in patients with primary or secondary myelofibrosis who are candidates for allogeneic hematopoietic cell transplantation: A Position Paper on behalf of the Chronic Malignancies Working Party of the EBMT. *The Lancet Haematology*, **10**, e59–e70. DOI: 10.1016/S2352-3026(22)00330-1.

Polverelli, N., Bonneville, E. F., de Wreede, L. C., et al. (2024) Impact of comorbidities and body mass index on the outcomes of allogeneic hematopoietic cell transplantation in myelofibrosis: A study on behalf of the Chronic Malignancies Working Party of EBMT. *American journal of hematology*, **99**, 993—996. DOI: 10.1002/ajh.27262.

Poythress, J. c., Lee, M. Y. and Young, J. (2020) Planning and analyzing clinical trials with competing risks: Recommendations for choosing appropriate statistical methodology. *Pharmaceutical Statistics*, **19**, 4–21. DOI: 10.1002/pst.1966.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., et al. (1978) The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics*, **34**, 541–554. DOI: 10.2307/2530374.

Putter, H., Fiocco, M. and Geskus, R. B. (2007) Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389–2430. DOI: 10.1002/sim.2712.

Putter, H., Schumacher, M. and van Houwelingen, H. C. (2020) On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: The Fine–Gray model revisited. *Biometrical Journal*, **62**, 790–807. DOI: 10.1002/bimj.201800274.

Qi, L., Wang, Y.-F. and He, Y. (2010) A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine*, **29**, 2592–2604. DOI: 10.1002/sim.4016.

Quartagno, M. and Carpenter, J. R. (2019) Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal*, **61**, 1003–1019. DOI: 10.1002/bimj.201800222.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. Manual.

R Core Team (2023) *R: A Language and Environment for Statistical Computing*. Manual.

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. Manual.

Rathouz, P. J. (2007) Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics*, **8**, 345–356. DOI: 10.1093/biostatistics/kxl014.

Resche-Rigon, M., White, I. and Chevret, S. (2012) Imputing missing covariate values in presence of competing risk. In: *International Society for Clinical Biostatistics Conference*, Bergen, Norway, 19-23 August 2012, P22.10, August 2012.

Rizopoulos, D. (2010) JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**, 1–33. DOI: 10.18637/jss.v035.i09.

Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-Event Data*. 1st edition.

Rizopoulos, D., Papageorgiou, G. and Miranda Afonso, P. (2023) *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data*. Manual.

Rouanet, A., Helmer, C., Dartigues, J.-F., et al. (2019) Interpretation of mixed models

and marginal models with cohort attrition due to death and drop-out. *Statistical Methods in Medical Research*, **28**, 343–356. DOI: 10.1177/0962280217723675.

Ruan, P. K. and Gray, R. J. (2008) Analyses of cumulative incidence functions via non-parametric multiple imputation. *Statistics in Medicine*, **27**, 5709–5724. DOI: 10.1002/sim.3402.

Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys*. DOI: 10.1002/9780470316696.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592. DOI: 10.1093/biomet/63.3.581.

Rücker, G. and Schwarzer, G. (2014) Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, **14**, 129. DOI: 10.1186/1471-2288-14-129.

Saadati, M., Beyersmann, J., Kopp-Schneider, A., et al. (2018) Prediction accuracy and variable selection for penalized cause-specific hazards models. *Biometrical Journal*, **60**, 288–306. DOI: 10.1002/bimj.201600242.

Saccardi, R., Putter, H., Eikema, D.-J., et al. (2023) Benchmarking of survival outcomes following Haematopoietic Stem Cell Transplantation (HSCT): An update of the ongoing project of the European Society for Blood and Marrow Transplantation (EBMT) and Joint Accreditation Committee of ISCT and EBMT (JACIE). *Bone Marrow Transplantation*, **58**, 659–666. DOI: 10.1038/s41409-023-01924-6.

Salzmann-Manrique, E., Bremm, M., Huenecke, S., et al. (2018) Joint Modeling of Immune Reconstitution Post Haploidentical Stem Cell Transplantation in Pediatric Patients With Acute Leukemia Comparing CD34+-Selected to CD3/CD19-Depleted Grafts in a Retrospective Multicenter Study. *Frontiers in Immunology*, **9**. DOI: 10.3389/fimmu.2018.01841.

Schetelig, J., Wreede, L. C. de, Gelder, M. van, et al. (2019) Late treatment-related mortality versus competing causes of death after allogeneic transplantation for myelodysplastic syndromes and secondary acute myeloid leukemia. *Leukemia*, **33**, 686–695. DOI: 10.1038/s41375-018-0302-y.

Schluchter, M. D. and Jackson, K. L. (1989) Log-Linear Analysis of Censored Survival Data with Partially Observed Covariates. *Journal of the American Statistical Association*, **84**, 42–52. DOI: 10.1080/01621459.1989.10478737.

Schmaelter, A.-K., Waidhauser, J., Kaiser, D., et al. (2021) Alterations of Peripheral Blood T Cell Subsets following Donor Lymphocyte Infusion in Patients after Allogeneic

Stem Cell Transplantation. *Hemato*, **2**, 692–702. DOI: 10.3390/hemato2040046.

Schultze-Florey, C. R., Kuhlmann, L., Raha, S., et al. (2021) Clonal expansion of CD8+ T cells reflects graft-versus-leukemia activity and precedes durable remission following DLI. *Blood Advances*, **5**, 4485–4499. DOI: 10.1182/bloodadvances.2020004073.

Seaman, S., Galati, J., Jackson, D., et al. (2013) What Is Meant by "Missing at Random"? *Statistical Science*, **28**, 257–268. DOI: 10.1214/13-STS415.

Shah, A. D., Bartlett, J. W., Carpenter, J., et al. (2014) Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, **179**, 764–774. DOI: 10.1093/aje/kwt312.

Sharma, A., Huang S., Li Y., et al. (2021) Outcomes of pediatric patients with therapy-related myeloid neoplasms. *Bone Marrow Transplantation*, **56**, 2997–3007. DOI: 10.1038/s41409-021-01448-x.

Shaw, B. E., Gooley, T. A., Malkki, M., et al. (2007) The importance of HLA-DPB1 in unrelated donor hematopoietic cell transplantation. *Blood*, **110**, 4560–4566. DOI: 10.1182/blood-2007-06-095265.

Shi, H., Cheng, Y. and Jeong, J.-H. (2013) Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal*, **55**, 82–96. DOI: 10.1002/bimj.201200011.

Shimoni, A., Labopin, M., Savani, B., et al. (2016) Long-term survival and late events after allogeneic stem cell transplantation from HLA-matched siblings for acute myeloid leukemia with myeloablative compared to reduced-intensity conditioning: A report on behalf of the acute leukemia working party of European group for blood and marrow transplantation. *Journal of Hematology & Oncology*, **9**, 118. DOI: 10.1186/s13045-016-0347-1.

Simonetta, F., Alvarez, M. and Negrin, R. S. (2017) Natural Killer Cells in Graft-versus-Host-Disease after Allogeneic Hematopoietic Cell Transplantation. *Frontiers in Immunology*, **8**. DOI: 10.3389/fimmu.2017.00465.

Snowden, J. A., Saccardi, R., Orchard, K., et al. (2020) Benchmarking of survival outcomes following haematopoietic stem cell transplantation: A review of existing processes and the introduction of an international system from the European Society for Blood and Marrow Transplantation (EBMT) and the Joint Accreditation Committee of ISCT and EBMT (JACIE). *Bone Marrow Transplantation*, **55**, 681–694. DOI: 10.1038/s41409-019-0718-7.

Sorror, M. L., Maris, M. B., Storb, R., et al. (2005) Hematopoietic cell transplantation (HCT)-specific comorbidity index: A new tool for risk assessment before allogeneic HCT. *Blood*, **106**, 2912–2919. DOI: 10.1182/blood-2005-05-2004.

Stern, L., McGuire, H. M., Avdic, S., et al. (2022) Immunoprofiling reveals cell subsets associated with the trajectory of cytomegalovirus reactivation post stem cell transplantation. *Nature Communications*, **13**, 2603. DOI: 10.1038/s41467-022-29943-9.

Sterne, J. A., Hernán, M. A., Reeves, B. C., et al. (2016) ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, **355**, i4919. DOI: 10.1136/bmj.i4919.

Sterne, J. A. C., White, I. R., Carlin, J. B., et al. (2009) Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, **338**, b2393. DOI: 10.1136/bmj.b2393.

Sullivan, T. R., Yelland, L. N., Lee, K. J., et al. (2017) Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. *Clinical Trials*, **14**, 387–395. DOI: 10.1177/1740774517703319.

Sullivan, T. R., White, I. R., Salter, A. B., et al. (2018) Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical Methods in Medical Research*, **27**, 2610–2626. DOI: 10.1177/0962280216683570.

Sweeting, M. J. and Thompson, S. G. (2011) Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, **53**, 750–763. DOI: 10.1002/bimj.201100052.

Tang, X., Alatrash, G., Ning, J., et al. (2014) Increasing Chimerism after Allogeneic Stem Cell Transplantation Is Associated with Longer Survival Time. *Biology of Blood and Marrow Transplantation*, **20**, 1139–1144. DOI: 10.1016/j.bbmt.2014.04.003.

Taylor, J. M. G., Murray, S. and Hsu, C.-H. (2002) Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, **58**, 221–232. DOI: 10.1016/S0167-7152(02)00030-5.

Therneau, T. M. (2023) *A Package for Survival Analysis in R*. Manual.

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., et al. (2018) On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statistics in Medicine*, **37**, 2338–2353. DOI: 10.1002/sim.7643.

Toor, A. A., Sabo, R. T., Roberts, C. H., et al. (2015) Dynamical System Modeling of

Immune Reconstitution after Allogeneic Stem Cell Transplantation Identifies Patients at Risk for Adverse Outcomes. *Biology of Blood and Marrow Transplantation*, **21**, 1237–1245. DOI: 10.1016/j.bbmt.2015.03.011.

Uhl, F. M., Chen, S., O'Sullivan, D., et al. (2020) Metabolic reprogramming of donor T cells enhances graft-versus-leukemia effects in mice and humans. *Science Translational Medicine*, **12**, eabb8969. DOI: 10.1126/scitranslmed.abb8969.

van Buuren, S. (2018) *Flexible Imputation of Missing Data, Second Edition*.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1–67. DOI: 10.18637/jss.v045.i03.

van Buuren, S., Oudshoorn, K. and TNO Preventie en Gezondheid (1999) *Flexible Multivariate Imputation by MICE*. January.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., et al. (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049–1064. DOI: 10.1080/10629360600810434.

van der Pas, S., Nelissen, R. and Fiocco, M. (2018) Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events. *Acta Orthopaedica*, **89**, 145–151. DOI: 10.1080/17453674.2018.1427314.

van Houwelingen, H. and Putter, H. (2012) *Dynamic Prediction in Clinical Survival Analysis*. DOI: 10.1201/b11311.

Vandenbroucke, J. P., Elm, E. von, Altman, D. G., et al. (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLOS Medicine*, **4**, e297. DOI: 10.1371/journal.pmed.0040297.

Vicente-Saez, R. and Martinez-Fuentes, C. (2018) Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, **88**, 428–436. DOI: 10.1016/j.jbusres.2017.12.043.

von dem Borne, P. A., Starrenburg, C. W. J. I., Halkes, S. J. M., et al. (2009) Reduced-intensity conditioning allogeneic stem cell transplantation with donor T-cell depletion using alemtuzumab added to the graft ('Campath in the bag'). *Current Opinion in Oncology*, **21 Suppl 1**, S27–29. DOI: 10.1097/01.cco.0000357472.76337.0e.

von Hippel, P. T. (2020) How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociological Methods & Research*, **49**, 699–718. DOI:

10.1177/0049124117747303.

White, I. R. and Carlin, J. B. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **29**, 2920–2931. DOI: 10.1002/sim.3944.

White, I. R. and Royston, P. (2009) Imputing missing covariate values for the Cox model. *Statistics in Medicine*, **28**, 1982–1998. DOI: 10.1002/sim.3618.

White, I. R. and Thompson, S. G. (2005) Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*, **24**, 993–1007. DOI: 10.1002/sim.1981.

White, I. R., Pandis, N. and Pham, T. M. (2022) Missing data, part 7. Pitfalls in doing multiple imputation. *American Journal of Orthodontics and Dentofacial Orthopedics*, **162**, 975–977. DOI: 10.1016/j.ajodo.2022.08.013.

Wolbers, M., Koller, M. T., Witteman, J. C. M., et al. (2009) Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology*, **20**, 555–561. DOI: 10.1097/EDE.0b013e3181a39056.

Wood, A. M., Royston, P. and White, I. R. (2015) The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, **57**, 614–632. DOI: 10.1002/bimj.201400004.

Yanir, A., Schulz, A., Lawitschka, A., et al. (2022) Immune Reconstitution After Allogeneic Haematopoietic Cell Transplantation: From Observational Studies to Targeted Interventions. *Frontiers in Pediatrics*, **9**. DOI: 10.3389/fped.2021.786017.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., et al. (2020) A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, **39**, 1199–1236. DOI: 10.1002/sim.8471.

Yun, H. D. and Waller, E. K. (2013) Finding the Sweet Spot for Donor Lymphocyte Infusions. *Biology of Blood and Marrow Transplantation*, **19**, 507–508. DOI: 10.1016/j.bbmt.2013.02.005.

Zhu, J. and Raghunathan, T. E. (2015) Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. *Journal of the American Statistical Association*, **110**, 1112–1124. DOI: 10.1080/01621459.2014.948117.

Nederlandse samenvatting

Vroeg of laat zullen ontbrekende data een uitdaging voor elke observationele medische studie vormen. Dit is vooral het geval bij het bestuderen van de tijd tot een bepaalde gebeurtenis, vanaf een klinisch relevant beginpunt. We kunnen bijvoorbeeld geïnteresseerd zijn in de tijd tot een infectie na een allogene hematopoietische stamceltransplantatie (alloSCT), een behandeling die voornamelijk wordt gegeven aan mensen met bloedkanker. Een studie moet uiteindelijk beëindigd worden, en op dat moment zullen sommige patiënten waarschijnlijk nog geen infectie hebben gehad. Omdat we niet weten of en wanneer ze geïnfecteerd zouden zijn als de onderzoeksperiode langer was geweest, is hun tijd tot infectie onbekend en worden ze aan het einde van de studie aangeduid als *rechtgecensureerd*. De tijd tot infectie zou ook onbekend zijn als een patiënt tijdens het onderzoek overlijdt voordat hij wordt geïnfecteerd. Overlijden wordt hier beschouwd als een *concurrerend risico*, omdat het een toekomstige infectie uitsluit.

Bovendien kan het interessant zijn om de relatie tussen patiëntspecifieke kenmerken en het optreden van een of meer concurrerende gebeurtenissen te bestuderen met behulp van statistische modellen. We kunnen bijvoorbeeld onderzoeken of de cumulatieve kans op infectie een jaar na een alloSCT kan worden voorspeld door de leeftijd van een patiënt bij alloSCT (een *baseline* covariaat), of door het aantal immuuncellen dat in de loop van de tijd in hun bloed circuleert (een *longitudinale* covariaat). Ontbrekende data in een of meer (soorten van) covariaten vormen een bedreiging voor de geldigheid van de modellen die worden gebruikt om de bovengenoemde vragen te beantwoorden. Naïeve uitsluiting van individuen met ontbrekende data kan de statistisch onderscheidingsvermogen drastisch verminderen en kan vertekende schattingen van doelgrootheden opleveren als de mogelijke oorzaken van de ontbrekende gegevens niet grondig worden overwogen. Adequate behandeling van ontbrekende waarden met behulp van principiële methoden zoals *multipele imputatie* (MI), in de context van regressiemodellen die rekening houden met concurrerende risico's, heeft weinig aandacht gekregen in de statistische literatuur. Dit proefschrift probeert deze onderzoekskloof te adresseren door het ontwikkelen, beoordelen en toepassen van statistische methodologie voor het omgaan met ontbrekende data in de context van alloSCT-studies met concurrerende risico's.

Hoofdstuk 2 geeft een hedendaags overzicht van hoe er momenteel wordt omgegaan met ontbrekende covariaatdata in de belangrijkste hematologische tijdschriften. Deze systematische review toonde aan dat ontbrekende covariaatdata veel voorkomen in klinische studies in de hematologie, maar dat bestaande richtlijnen met betrekking tot de behandeling en rapportage van ontbrekende data over het algemeen niet worden gevolgd. Bovendien werd MI, in tegenstelling tot eenvoudiger methoden zoals complete-case analyse en de missing indicator methode, zelden gebruikt.

In **Hoofdstuk 3** werden covariaat imputatiemodellen afgeleid die ongeveer 'compatibel' zijn met een of meer oorzaak-specifieke Cox proportional hazards analysemodellen. Een simulatiestudie toonde aan dat deze imputatiemodellen wat betreft het schatten van oorzaak-specifieke hazard ratio's over het algemeen slechter presteerden dan een eerder voorgestelde zogeheten substantive-model-compatible MI methode (SMC-FCS). Wat betreft het schatten van individu-specifieke cumulatieve incidentiefuncties presteerden beide methoden vergelijkbaar. De bovengenoemde ongeveer-compatibele MI methode werd toegepast in **Hoofdstuk 4** op een dataset van patiënten met myelofibrose die een alloSCT hebben ondergaan, waarbij het doel was om de invloed van gedeeltelijk ontbrekend body mass index en comorbiditeiten op het oorzaak-specifieke risico van overlijden voor ziekteterugval te beoordelen. Dit hoofdstuk vormt ook een casestudy met betrekking tot de imputatie van zogenaamde 'afgeleide' covariaten (d.w.z. de combinatie van twee of meer direct gemeten variabelen).

In **Hoofdstuk 5** werd nieuwe SMC-FCS methodologie ontwikkeld die het mogelijk maakt om ontbrekende covariaten compatibel te imputeren met een Fine–Gray analysemodel, zonder een model te hoeven specificeren voor de concurrerende gebeurtenis(sen). Deze methode werd beoordeeld in een simulatiestudie, waarin er scenario's waren waarbij de imputatieprocedure compatibel was met het onjuiste data-genererende mechanisme (d.w.z. waar proportionaliteit gold op de oorzaak-specifieke hazard schaal in plaats van op de subdistributie hazard schaal). Hier bleek censurering een belangrijke rol te spelen, door de impact van analysemodel-mispecificatie in de imputatiefase te verzachten.

Hoofdstuk 6 geeft een beknopt overzicht van data-genererende mechanismen waarbij een Fine–Gray model correct is voor ten minste één concurrerende gebeurtenis. De belangrijkste conclusie van dit hoofdstuk was dat men de voorkeur moet geven aan oorzaak-specifieke hazard modellen boven meerdere Fine–Gray modellen wanneer meer dan één concurrerende gebeurtenis van belang is, omdat er geen data-genererend mechanisme is waarvoor de aanname van proportionele subdistributie hazards tegelijkertijd geldt voor alle gebeurtenissen (tenzij aanvullende aannames worden gemaakt). Dit hoofdstuk biedt ook inzichten met betrekking tot SMC-FCS met een Fine–Gray analysemodel, waarbij een model voor de concurrerende gebeurtenis alleen voor imputatiedoeleinden zou kunnen worden gespecificeerd.

Ten slotte toont **Hoofdstuk 7** het gebruik van zogenoemde *joint models* voor het analyseren van de trajecten van het aantal immuuncellen voor een cohort acute

leukemiepatiënten in de eerste 6 maanden na een alloSCT met T-celdepletie. De resultaten onderstreepten het belang van het in aanmerking nemen van concurrerende risico's in het tijd-tot-gebeurtenis submodel, waarbij met name de 'huidige waarde' van het aantal CD4+ cellen tegengestelde effecten heeft op de oorzaak-specifieke risico's van graft-versus-hostziekte (GvHD) en ziekte terugval.

Summary

Sooner or later, incomplete data will pose a challenge to any observational study in medicine. This rings especially true when studying the time from a clinically relevant starting point, to a particular event. For example, we may be interested in the time to an infection after an allogeneic haematopoietic stem cell transplantation (alloSCT), a treatment primarily given to individuals with blood cancer. A study has to end eventually, at which point some patients will likely not yet have experienced an infection. Since we do not know whether or when they would have been infected had the study period been longer, their time to infection is unknown, and they are said to be *right-censored* at the end of the study. The time to infection would also be unknown if a patient dies during the study before becoming infected. Death here is considered to be a *competing risk*, since it precludes any future infection.

Moreover, it may be of interest to study the relation between patient-specific characteristics and the occurrence of one or more competing events, using statistical models. For instance, we may investigate whether the cumulative probability of infection one year after an alloSCT can be predicted by the age of a patient at alloSCT (a *baseline* covariate), or by the number of immune cells circulating in their blood over time (a *longitudinal* covariate). Missingness in one of more (types of) covariates will pose a threat to the validity of the models used to answer the aforementioned questions. Naive exclusion of individuals with missing data can dramatically reduce statistical power, and can yield biased estimates of targeted quantities when the potential causes underlying the missing data are not thoroughly considered. Appropriate handling of missing values using principled methods such as *multiple imputation* (MI), in the context of regression models that account for competing risks, has received little attention in the statistical literature. The present dissertation seeks to fill this research gap, by developing, assessing, and applying statistical methodology for dealing with incomplete data in the context of alloSCT studies with competing risks.

Chapter 2 provides a contemporary overview of how missing covariate data are currently being handled across major haematological journals. This systematic review showed that missing covariate data are prevalent in clinical studies in haematology, but that existing guidelines regarding the handling and reporting of missing data are

generally not being followed. Additionally, in contrast to simpler approaches such as complete-case analysis and the missing indicator method, MI was rarely used.

In **Chapter 3**, approximately compatible covariate imputation models were derived for a setting where one or more cause-specific Cox proportional hazards models were the substantive model of interest. A simulation study showed that, in terms of estimating cause-specific hazard ratios, these imputation models generally underperformed relative to a previously proposed substantive-model-compatible MI approach (SMC-FCS). In terms of estimating individual-specific cumulative incidence functions, both methods performed comparably. The aforementioned approximately compatible MI approach was applied in **Chapter 4** on a dataset of patients with myelofibrosis who have undergone an alloSCT, where the aim was to assess the impact of partially observed body mass index and comorbidities on the cause-specific hazard of non-relapse mortality. This chapter also represents a case study regarding the imputation of so-called 'derived' covariates (i.e. the combination of two or more directly measured variables).

In **Chapter 5**, novel SMC-FCS methodology was developed allowing to impute missing covariates compatibly with a Fine–Gray substantive model, without needing to specify a model for the competing event(s). Its performance was assessed in a simulation study, which included settings which involved imputing compatibly with the incorrect data-generating mechanism (i.e. where proportionality held on the cause-specific rather than on the subdistribution hazard scale). Here, censoring was found to play an important role, softening the impact of substantive model misspecification at the imputation stage.

Chapter 6 provides a concise overview of data-generating mechanisms where a Fine–Gray model correctly holds for at least one competing event. The core conclusion of this chapter was that one should favour cause-specific hazard models over multiple Fine–Gray models when more than one competing event is of interest, since there is no data-generating mechanism for which the assumption of proportional subdistribution hazards simultaneously holds for all events (unless additional assumptions are made). This chapter also provides insights regarding SMC-FCS with a Fine–Gray substantive model, where a model for the competing event could be specified solely for imputation purposes.

Finally, **Chapter 7** showcases the use of *joint modelling* for analysing the trajectories of immune cell counts for a cohort of acute leukaemia patients in the first 6 months following a T-cell depleted alloSCT. The results underlined the importance of accounting for competing risks in the time-to-event submodel, with in particular the 'current value' of CD4+ cell counts having opposing effects on the cause-specific hazards of graft-versus-host disease (GvHD) and disease relapse.

List of publications

- **Bonneville, E. F.**, Resche-Rigon, M., Schetelig, J., Putter, H., de Wreede, L. C. (2022) Multiple imputation for cause-specific Cox models: Assessing methods for estimation and prediction. *Statistical Methods in Medical Research*, 31, 1860–1880. DOI: 10.1177/09622802221102623.
- **Bonneville, E. F.**, Schetelig, J., Putter, H., de Wreede, L. C. (2023) Handling missing covariate data in clinical studies in haematology. *Best Practice & Research Clinical Haematology*, 36, 101477. DOI: 10.1016/j.beha.2023.101477.
- Koster, E. A. S.[†], **Bonneville, E. F.**[†], von dem Borne, P. A., van Balen, P., Marijt, E. W. A., Tjon, J. M. L., Snijders, T. J. F., van Lammeren, D., Veelken, H., Putter, H., Falkenburg, J. H. F., Halkes, C. J. M., de Wreede, L. C. (2023) Joint models quantify associations between immune cell kinetics and allo-immunological events after allogeneic stem cell transplantation and subsequent donor lymphocyte infusion. *Frontiers in Immunology*, 14. DOI: 10.3389/fimmu.2023.1208814.
- Polverelli, N.[†], **Bonneville, E. F.**[†], de Wreede, L. C., Koster, L., Kröger, N. M., Schroeder, T., Peffault de Latour, R., Passweg, J., Sockel, K., Broers, A. E. C., Clark, A., Dreger, P., Blaise, D., Yakoub-Agha, I., Petersen, S.L., Finke, J., Chevallier, P., Helbig, G., Rabitsch, W., Sammassimo, S., Arcaini, L., Russo, D., Drozd-Sokolowska, J., Raj, K., Robin, M., Battipaglia, G., Czerw, T., Hernández-Boluda, J. C., McLornan, D. P. (2024) Impact of comorbidities and body mass index on the outcomes of allogeneic hematopoietic cell transplantation in myelofibrosis: A study on behalf of the Chronic Malignancies Working Party of EBMT. *American Journal of Hematology*, 99, 993—996. DOI: https://doi.org/10.1002/ajh.27262
- **Bonneville, E. F.**, de Wreede, L. C. and Putter, H. (2024) Why you should avoid using multiple Fine–Gray models: Insights from (attempts at) simulating proportional subdistribution hazards data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae056. DOI: 10.1093/jrsssa/qnae056

[†]These authors contributed equally to this work and share first authorship.

- **Bonneville, E. F.**, Beyersmann, J., Keogh, R. H., Bartlett, J. W., Morris, T. P., Polverelli, N., de Wreede, L. C., Putter, H. (2024) Multiple imputation of missing covariates when using the Fine–Gray model. *arXiv preprint* arXiv:2405.16602. DOI: 10.48550/arXiv.2405.16602.

Acknowledgements

As part of the journey that was writing this thesis, I had the privilege of collaborating with and being surrounded by a number of fantastic people, to whom I would like to express my gratitude.

First and foremost, I would like to thank my supervisors Liesbeth de Wreede and Hein Putter, for their unwavering support and mentorship during the past years. Liesbeth, the consistently high scientific standards by which you work have been an absolute inspiration. Hein, I will always admire how you are able to communicate difficult statistical concepts with ease and enthusiasm. I have learned a lot from the both of you, and will always keep a fond memory of our weekly meetings, which rarely felt complete without culturo-linguistic insights and coffee. Hartelijk dank.

This thesis would likely not have been about missing data at all had it not been for Matthieu Resche-Rigon. Thank you for motivating me to build on your earlier ideas on multiple imputation in competing risks settings. Relatedly, I would like to thank Jacco Wallinga and Marta Fiocco, for encouraging me during my master's thesis to consider a PhD position in the first place.

Furthermore, I had the pleasure of working with several statistical and clinical collaborators during the PhD trajectory. Thank you Ruth Keogh and Jan Beyersmann for welcoming me in London and Ulm respectively, and for the enjoyable collaboration that ensued also together with Tim Morris and Jonathan Bartlett. On the clinical side, I would like to thank Johannes Schetelig and Nicola Polverelli for their contributions to several articles in this thesis, as well as Constantijn Halkes and Frederik Falkenburg for the stimulating collaboration throughout the joint models project. Special thanks to Eva Koster, for being a joy to work with, and for your patience in repeatedly explaining the basics of allogeneic stem cell transplantation (alloSCT) to me.

The present thesis relies on various datasets of patients who have undergone an alloSCT, registered with the European Society for Bone and Marrow Transplantation (EBMT). I thank both the EBMT for permitting the use of these data, and EBMT data managers Linda Koster and Laurien Baaij for their help in preparing the datasets. I would also like to express my gratitude to all patients and centres involved in the original studies. Moreover, I would like to acknowledge all the EBMT statisticians

that I have had the opportunity of learning from during the past years: Dirk-Jan, Luuk, Jarl, Paddy, Katerina, Simona, Gloria, Giulia, Katya, Richard, Jacques-Emmanuel, Christophe, Ariane, Maud, and Myriam.

Next, I would like to show my appreciation to my colleagues from Medical Statistics (MSTAT) at the Leiden University Medical Center (LUMC): Ningning, Mitra, Irene, Susanne, Alexandros, Xu, Mirko, Junfeng, Jelle, Saskia, Mar, Erik, Bart, Roula, Stefan, Szymon, Ramin, Ghislaine, Lu, Leonie, Lies, Ewout, and Nan. Ewout, thank you for getting me interested in prediction models, and for the many spontaneous conversations in your office. Nan, it was a pleasure to work together on the validation of competing risks prediction models project.

Needless to say, completing this thesis would have been a mere fraction of the fun without being surrounded by such a brilliant group of MSTAT PhD students: Chiara, Yongxi, Anna Kaal; Riccardo, Angela, Anna Vesely, Kaya, Andrea, Mari, Damjan (les visiteurs); Toby, Jasper, Georgy, Frank (the honourable gentlemen from office 42); Marije, Doranne, Lars, and Ilaria (the OGs). Thank you all for being absolute legends.

To my paranymphs Lars and Ila. Lars, beyond the many times you have made me laugh, I thank you in particular for inspiring me to continue to pursue interests and hobbies outside of work. Ila, aside from the consistent provision of Italian snacks, I thank you for being such a dependable friend.

Outside of MSTAT, several groups of PhD students positively impacted my time at the LUMC. Thank you to the 'Hacky Hour' folks for sparking my interest in Open Science early on in my PhD: Kim, Linda, Anna Lohmann, Daniela, Tariq, and Xante. Thank you also to the lovely people of the LUMC Association for PhD Candidates (LAP) for the welcome distraction that was organising events together: Daniele, Merian, Naomi, Alice, and Nienke.

To my parents and brother. Papá, Mamá, les agradezco de todo corazón su apoyo y todos los sacrificios que han hecho por mí. Je vous aime fort. Charlie, eres un crack—gracias por siempre hacerme reír, y por ser una parte tan positiva de mi vida.

Finally, my deepest thanks go to my other half Sarah (with the long name) von Grebmer zu Wolfsthurn. Not least for actually showing me that there was an end in sight to the PhD after you completed yours, but for your boundless patience, love, and support. Thank you being the best part of my life, and for always pushing me to become a better version of myself.

Curriculum vitae

Edouard Francis Bonneville was born on July 20th 1995 in Madrid (Spain). After completing his secondary education in Colomiers (France) in 2013, he pursued a BSc in Psychology at the University of Bristol (United Kingdom). In 2016, he moved to Leiden (the Netherlands) for the MSc Statistical Science for the Life and Behavioural Sciences, which is now known under the name 'Statistics and Data Science'. As part of his master thesis, he spent three months at the RIVM Dutch National Institute for Public Health and the Environment working on a Bayesian approach for forecasting infection disease epidemics.

In 2019, he started his PhD in Biostatistics at the Department of Biomedical Data Sciences at Leiden University Medical Center (LUMC, the Netherlands) under the supervision of Dr. Liesbeth de Wreede and Prof. Dr. Hein Putter. The results of his research, which focused on statistical methodology at the intersection of competing risks and missing data, are outlined in this thesis. During his PhD programme, he visited both the Department of Medical Statistics at the London School of Hygiene & Tropical Medicine (United Kingdom), and the Institute of Statistics at Ulm University (Germany). Alongside his PhD, he was part of the LUMC Association for PhD Candidates (LAP) board for two years, during which he was treasurer and co-organised several events.

Between 2022 and 2024, he also worked part-time as a study statistician for the European Society for Bone and Marrow Transplantation (EBMT). Finally, he has served as a reviewer for the journals *Statistics in Medicine*, *Statistical Methods in Medical Research*, *BMC Medical Research Methodology*, *Biometrical Journal* and *Journal of Computational and Graphical Statistics*.