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

Bonneville, E.F.

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,

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

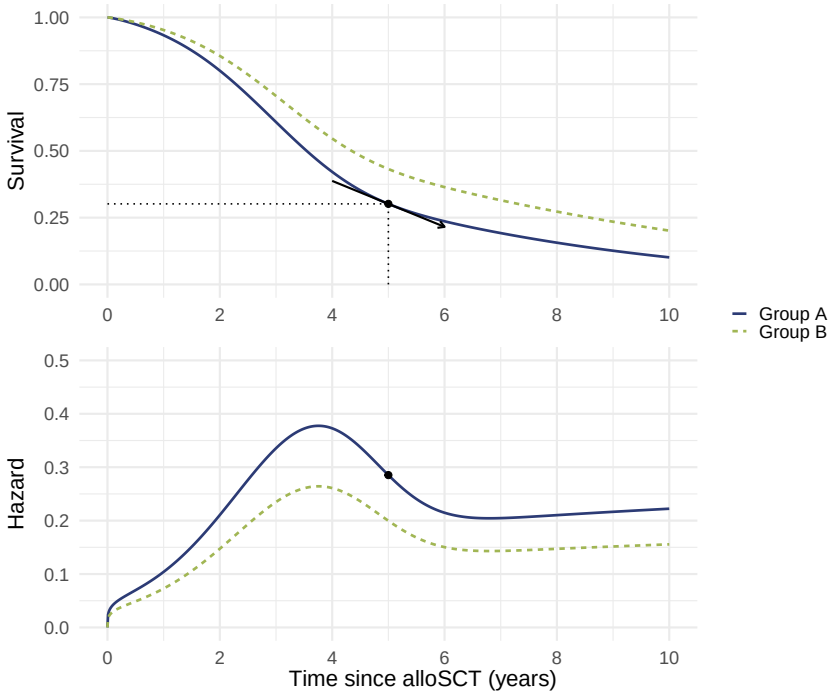

Figure 1.1: Example (all-cause) hazard and survival functions for two subgroups from a given population. The hazard is based on two aspects of the survival function at a particular timepoint: its value (i.e. the proportion of individuals still alive), and its slope (i.e. how rapidly the curve is decreasing, 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.

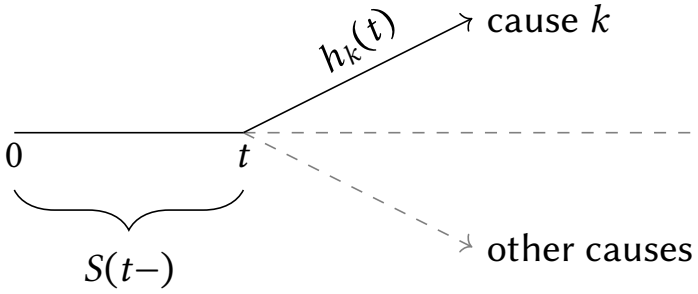

Figure 1.2: Summary of the competing risks process, adapted from Figure 1 in Geskus (2024).

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.