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

Bonneville, E.F.

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).