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

Bonneville, E.F.

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.