Maral Saadati
Dr. sc. hum.

Extending the Multistate Framework to Incorporate High-Dimensional Genomic Information

Fach/ Einrichtung: DKFZ
Doktormutter: Prof. Dr. Annette Kopp-Schneider

In clinical studies it is often necessary to analyze time-to-event data, where patients may encounter different event types. Consider, for example, patients newly diagnosed with acute myeloid leukemia, who enter a clinical trial and receive treatment. Some patients may reach remission, while others do not show sufficient response to treatment. Furthermore, as time progresses, patients who have achieved remission may die in remission, or relapse and die thereafter. Using statistical modeling and regression analysis it is possible to analyze such data appropriately. Due to the fact that different events can occur for an individual in the course of their disease, the data exhibit a complex structure with respect to the modeling endpoint. This process can be described by a multistate model that is interpreted as a series of competing risk models. Another complexity lies within the covariate structure: if molecular data are considered as covariates, the design matrix is high-dimensional (meaning the sample size is much smaller than the number of variables) and very noisy. Thus, detecting relevant associations between the time-to-event endpoint and molecular information can be challenging and particularly difficult to validate on independent data sets.

The overall goal of this research work was to incorporate high-dimensional covariate structures into the multistate modeling framework. Since competing risks are an integral part of multistate modeling, we first and foremost explored lasso-penalized competing risks models. Within the context of competing risks, the main focus was on the cause-specific hazard-based approach due to its strong conceptual advantages. Cause-specific hazards are the natural building blocks for competing risks modeling, and can be regarded as the quantities of choice when analyzing microarray data for understanding the underlying biological mechanisms. This is due to the fact that cause-specific hazards allow us to differentiate between covariate effects on different causes of failure. Finally, another major advantage is that penalized cause-specific hazards can offer extensions for multistate modeling in high dimensions.

In the context of penalized cause-specific hazard models, one needs to fit a model for each of the possible causes of failure. One of our main concerns was that fitting such high-dimensional models independently of each other might be prone to over-fitting or possibly be inefficient with respect to prediction. Thus we investigated the benefit of linking the independently penalized models with respect to minimal prediction error. The prediction error as defined by the Brier score is based on the cumulative incidence function of the event of interest. Thus, linking of the independently penalized models is achieved by choosing the penalization tuning parameters of the two models as the pair that minimizes the prediction error at a clinically relevant time.

We compared the performance of the independent and the linked cause-specific hazard model to the natural competitor, namely the lasso-penalized subdistribution hazard model. The subdistribution hazards approach requires fitting only one model and naturally links the cause-specific hazards by working on incidence level. Furthermore, subdistribution hazard approaches have proven useful for prediction purposes, but they suffer from interpretational challenges. The simulation results indicate that the linked cause-specific hazard approach can be useful in some situations. It can better detect variables that have moderate effect sizes of different signs on the two competing causes of failure and has higher true positive rates than the independently penalized model when the proportional hazards is violated. However, it also tends to choose somewhat larger models, leading to higher false discovery rates in many scenarios. Thus, linking the cause-specific hazards might be beneficial in some particular cases, but simple models using separately penalized cause-specific hazard are often justified.

It is in general not possible to fulfill proportionality assumptions for the subdistribution hazards model and both cause-specific hazards models simultaneously. Therefore, different scenarios were investigated with various settings. It was observed that variable selection performance and prediction accuracy are highly dependent on proportionality assumptions, and that model misspecification can have severe negative impact. In the interesting setting where proportionality is fulfilled for subdistribution hazards and cause-specific hazards of event 1 simultaneously, we observe very similar performance of the different methods with respect to variable selection and prediction accuracy. This was also the case in our application examples for the analysis of gene expression levels of patients with bladder carcinoma and patients with acute myeloid leukemia.

The work presented here contains some points for further improvement. For example, censoring could be incorporated into the simulation studies to investigate the impact of the censoring distribution on the results. Furthermore, penalty types other than the lasso may be more appropriate for some real-life data sets. Finally, one could consider standard competing risks approaches that are not hazard-based, such as pseudo value regression and direct binomial modeling.

All in all, our results indicate that penalized cause-specific hazard models are a viable solution for competing risks models in high dimensions, because they allow for interpretation of the transition hazards and have prediction accuracy similar to the subdistribution hazard method. This newly gained knowledge on competing risks processes was put to use for constructing Markovian multistate models in high dimensions. A penalized multistate model was fitted using a nested sequence of independently penalized competing risks models to analyze gene expression data of acute myeloid leukemia patients. The resulting high-dimensional multistate model revealed some genes that have already been reported in recent literature in connection with leukemia and some solid tumors. Furthermore, some genes were detected that have not been published yet and could be interesting candidates for further biomedical investigation.