## Statistical analysis of repeated outcomes of different types

Musoro, Z.J.

**Publication date**
2016
**Document Version**
Final published version

# Chapter 1

# General introduction

In health care, it is routine to monitor patients by collecting information regarding their health repeatedly over time. Commonly, both longitudinal response measurements and the time to an event of interest are collected for each patient. This has led to a growing need to adequately use such updated information (together with information collected only at baseline) to provide a better insight in patients' health status over time. This could help clinicians to make better decisions as regards to patient care. Patients could also benefit from being educated about possible health risks, which may motivate them to take more responsibility for their health. The motivating examples that will be revisited in more detail in this thesis include;

*I Post-kidney transplantation data* (Struijk *et al.*, 2010): In a kidney transplantation unit, immunological markers such as CD3+ T cells, B cells and natural killer cells were measured at several time points in order to monitor patients' immune status after transplant. Also, the times of repeated occurrences of opportunistic infections such as urinary tract, viral and upper respiratory infections were monitored. Follow-up could be terminated by death. Here, frailer patients exhibiting either deteriorating marker trajectories or experiencing more frequent recurrent events may benefit from a modification of their immunosuppressive therapy or be placed on preventive medication against future infections.

*II Intensive care (IC) data* (Toma *et al.*, 2007): In an intensive care unit (ICU), besides routinely collected baseline data, patients' Sequential Organ Failure Assessment (SOFA) scores were used for daily monitoring of organ dysfunction. Patients either died or were discharged from the hospital. The updated information in the daily SOFA scores may improve patients' prognosis at the end of hospital stay which is relevant, e.g., in assessing the quality of care in the ICU and for patient management.

*III Primary care data* (Siebeling *et al.*, 2011): In primary care, Chronic Obstructive Pulmonary Disease (COPD) patients were seen by their general practitioner (GP) at regular intervals and information about four quality of life (QoL) domains (dyspnoea, fatigue, emotional function, and mastery) and overall QoL were monitored. The time to death was also noted. In daily clinical practice, it is important to be able to show patients their expected course on (different domains of) their QoL. In case of potential decline in one or more domains, both the physician and the patient can discuss about prioritizing certain treatment decisions.

In such settings, when we desire to make inference on either a longitudinally measured variable (henceforth referred to as marker) or on the time to an event of interest (such as an infection or death), classical models such as a linear mixed effects model and a Cox proportional hazards model can be used respectively. But performing separate analyses for both the longitudinal and survival subprocesses may ignore possible underlying associations between the marker and time-to-event data. For instance, in modeling longitudinal data with missing marker values which are likely not missing at random, we may need to also model the times to dropout or death to avoid possible biases in the estimation of the longitudinal model. As another example, when fitting survival models which include a time-dependent covariate that may have been measured with error or contain missing values at event times, a longitudinal model for the time-dependent covariate is required to cater for the missing covariate values as well as for the measurement errors. In both examples, it is necessary to simultaneously evaluate both the longitudinal and the survival subprocesses using a joint model. In the first part of this thesis, we will consider the joint modelling framework (Tsiatis and Davidian, 2004) to study the association between longitudinal and survival data. We shall use joint models to perform dynamic predictions of survival probabilities for a new subject, using marker values that are accrued as time goes on. We shall also present an extension of the application of joint modelling to a setting with multiple markers and multi-type recurring events.

In the second part of this thesis, we shall look at landmarking (Van Houwelingen, 2007) as an alternative to joint modelling for performing dynamic predictions of survival probabilities. With the landmarking approach, a Cox model is fitted to a data set comprising a selection of subjects who are at risk at a landmark time point $t_s$ and treating their marker values at $t_s$ as a time fixed covariate. This approach circumvents possible computational complications of fitting time-dependent covariates, making it easier to compute survival probabilities compared to using joint models. Although a joint model (if correct) would yield a better predictive performance (Rizopoulos *et al.*,

2013), landmark models require fewer modelling assumptions, hence are more robust. Furthermore, landmark models can be easily extended to accommodate several markers. We shall present an extension of the application of landmark models to a setting with recurring events of the same type.

In the third part of this thesis, we shall look at construction and validation of prediction models in the presence of multiply imputed data. This was motivated by the fact that at certain moments in time, when we desired to make predictions on patients' health status, the number of predictors that were available per patient was larger than what clinicians would prefer to use in practice. Also, some of the predictors contained missing values. It was paramount to optimally select the "best" subset of predictors which were interpretable and practically useful. But the selection and validation procedure for this "best" subset of predictors was complicated by the presence of missing data. Here, we shall lay more emphasis on how to internally validate prediction models via bootstrap methods (Harrell *et al.*, 1996) after multiply imputing missing data. A pertinent question that we will answer is that of how resampling should be performed in the presence of multiply imputed data sets.

## Part I: Joint modelling

Joint models for longitudinal and survival data have gained a lot of attention in the literature (Tsiatis and Davidian, 2004; Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997; Xu and Zeger, 2001; Rizopoulos and Ghosh, 2011). In comparison to performing separate analysis for each outcome, joint modelling is commended to yield less biased and more efficient assessment of both the survival and longitudinal subprocess since it acknowledges that both subprocesses may be highly associated. A comprehensive overview of the subject is given by Tsiatis and Davidian (2004).

In recent years, the application of joint models have extended beyond the traditional one longitudinal response and a single event type setting to settings with; (*i*) multiple longitudinal outcomes and a single failure time (Rizopoulos and Ghosh, 2011; Sweeting and Thompson, 2011; Ibrahim *et al.*, 2004; Xu and Zeger, 2001), (*ii*) a single longitudinal outcome and multiple failure types (Gueorguieva *et al.*, 2012; Li *et al.*, 2010; Elashoff *et al.*, 2008), (*iii*) multiple longitudinal outcomes and multiple failure types (Chi and Ibrahim, 2006), and (*iv*) a single longitudinal outcome and recurrent events (Kim *et al.*, 2012; Liu and Huang, 2009; Liu *et al.*, 2008). In **chapter 2** of this thesis, we will further extend the joint modelling approach to include multiple longitudinal outcomes and multiple recurrent events. This was motivated by the post-kidney transplantation data which comprised patients who could experience up to ten

infection types, all at multiple times. Patients' immune status was monitored using five immunological markers. Our main goal was to evaluate the effect of the markers on the risk of each infection type, and to also measure the dependence *within* and *between* infection types. We proposed a joint model consisting of a multivariate mixed effects linear submodel and an infection-specific Cox submodel with a set of frailty terms to catch the *within* and *between* infection type dependence. Both submodels were linked by shared latent terms.

Maximization of the likelihood in a joint model requires integrating over the random effects distribution. This is usually computationally challenging, requiring for instance quadrature approximation techniques (Liu *et al.*, 2008) to cater for analytically intractable integrals or Bayesian approaches (Faucett and Thomas, 1996). However, in settings with multiple markers and multiple repeating infections as with the transplantation study, the dimensionality of the integral is high and these methods become too computationally expensive. In **chapter 3**, we propose to use a simulated maximum likelihood (SML) approach based on a quasi-Monte Carlo (QMC) integration technique to evaluate the likelihood of a joint model with high dimensions of random effects (Gouriéroux and Monfort, 1997; Lemieux, 2009). With the QMC technique, the integrals are evaluated using a deterministic point set. This is more computationally friendly, although the accuracy of the approximation depends on the size of the QMC point set. We illustrate the QMC approach using simulated data sets and the transplantation data.

The joint modeling framework can also be utilized to calculate and dynamically update individual survival probabilities based on information that is available at new time points. This presents physicians with an additional tool to optimize care at the patient level, which is in line with the growing trend of personalized medicine. For an overview and illustration of how joint models are used for dynamic prediction, as well as how to evaluate their predictive performance, see Rizopoulos (2011) and chapter 7 of Rizopoulos (2012). In **chapter 4**, motivated by the IC data of patients who were all observed to have died or discharged, we performed dynamic prediction of the risk of dying in the hospital. Despite the availability of daily measured SOFA scores, only data collected during the first 24 hours of admission to the ICU was often used to predict the survival status of patients at the end of hospital stay (Minne *et al.*, 2008). We applied the joint modelling framework which optimally uses all available information that is accrued over time. Our joint model comprised a linear mixed effects submodel for the development of longitudinal SOFA scores, and a proportional subdistribution hazards (Fine and Gray, 1999) submodel for death as end point, with

discharge as competing risk. The two parts were linked by shared latent terms. As opposed to the commonly used proportional cause-specific hazards formulation for competing risk data, the subdistribution hazard directly translates to event-specific survival probabilities. Also, since there was no censoring nor late entry that induced left truncation in our data, discharged patients could remain in the risk set with a weight of one. Hence, fitting our joint model was straightforward using free available software such as the JM package (Rizopoulos, 2010) within the R statistical software (R Core Team, 2013). However with respect to performing dynamic predictions within the JM package, it is only possible to calculate the risk of dying in the hospital for a patient given that he or she is still alive at given moments in time. This does not take into account that discharged patients are no longer at risk of dying in the hospital after their time of discharge. So additional routines were developed to perform dynamic predictions such that patients were no longer at risk after being discharged. We compared predictive values from our joint model with those obtained from an earlier modelling approach by Toma et al. (2007) which relied on patterns discovered in the SOFA scores over a given period of time.

## Part II: Landmarking

Though the joint modelling approach allows for flexibility in the description of the time updated marker subprocess, more modelling assumptions are often required and the modelling exercise could become computationally intensive. This is particularly the case when several longitudinal markers are used to make more accurate risk predictions. Alternatively, the landmarking paradigm (Van Houwelingen, 2007; Zheng and Heagerty, 2005) offers a more flexible and relatively simple way to characterize the association between a longitudinal marker and the time until an event. This does not require any assumption on the distribution of marker development. At landmark $s$, a survival model is fitted to data that comprise individuals who are at risk at landmark time point $t_s$. Only information that is available at $t_s$ is used and the value of the time varying marker at $t_s$ is handled as a time fixed variable. Conditional on updated information at $t_s$, it is easy to compute survival probabilities at a prediction horizon $t_{hor}$; $t_{hor} = t_s + w$, where $w$ is the width of the prediction window. The choice of $w$ depends on the length of follow-up and the overall prognosis.

An overview of existing methods for dynamic prediction based on landmarking has been provided by Van Houwelingen and Putter (2012). They restrict to single-type events (occurring once). Extensions to a competing risks scenario have also been addressed (Nicolaie *et al.*, 2013, 2012; Cortese and Andersen, 2010). In **chapter 5**, we generalize the landmarking approach to a setting with recurrent events of the same

type with application to the post-kidney transplantation data. At each landmark $s$, a Cox proportional hazards model with random effects for repeated infections was fitted since patients could experience more than one infection. The time-updated marker values at $t_s$ were handled as time-fixed covariates. Commonly, the last observed marker value prior to $t_s$ is carried forward to $t_s$. But for our case, since some patients had a large time difference between the last marker measurement and the landmark time point, it was more plausible to use fitted marker values at $t_s$ (according to a mixed effects model). This has the extra benefit of accounting for possible measurement error in the original marker measurements. For the sake of parsimony, we merged all landmark data sets to create a stacked set, and fitted supermodels that allow parameters to depend on the landmark in a smooth fashion. We described four ways to parameterize supermodels for recurrent event data, and settled for parameterizations that could be implemented within available software. Also, as opposed to settings with single-type events, where the stacked data set often comprise overlapping landmark periods (Van Houwelingen and Putter, 2012), it could be a problem to use overlapping landmark periods for recurrent events data since it will imply counting events multiple times over the overlapping landmark. So, we compared supermodels that were fitted on stacked data sets that comprise either overlapping or non-overlapping landmark periods using both our study data set and simulated data sets.

## Part III: Construction and validation of prediction models in the presence of multiply imputed data

In practice, besides being able to accurately predict patients' future health status, it is also important for prediction models to be as parsimonious as possible. Models with fewer predictors are often easier to implement and thus we are sometimes willing to sacrifice some predictive performance. For instance, most physicians will be unwilling to use large models that require collecting too much information. Furthermore, it is much easier to display a parsimonious model using visual decision tools such as nomograms which simplifies their potential application in everyday care. In the third part of this thesis, motivated by follow-up data of COPD patients in primary care, we aimed at constructing parsimonious models to predict four QoL domains, as well as overall QoL. The variable selection and validation procedures were complicated by the presence of missing data. Some strategies to go about this have been discussed in the literature (Vergouwe *et al.*, 2010; Vergouw *et al.*, 2010; Heymans *et al.*, 2007). For instance, Vergouw *et al.* (2010) and Heymans *et al.* (2007) proposed to combine multiple imputations (MI) with backward elimination (BE) and bootstrapping to obtain a parsimonious prediction model. Herein, we propose the least absolute shrinkage

and selection operator (lasso) technique (Tibshirani, 1996) for predictor selection and model fitting using data that has been multiply imputed, and bootstrap resampling for model validation. In **Chapter 6** we focus on the methodological aspects, especially on handling multiply imputed data sets when performing bootstrap resampling for model validation. A clinical application will be presented in **chapter 7**.

The lasso technique maximizes a function that combines the likelihood of the data with a penalty on the absolute value of the regression coefficients, such that parameter estimates are shrunk towards zero. This makes the lasso attractive for prognostic studies since it improves the quality of predictions compared to predictions based on a model fitted via unpenalized maximum likelihood. Parsimony is achieved as well since variables whose parameter estimates are exactly zero can be dropped. It is unclear whether the performance of a model fitted using the lasso still shows some optimism. Hence we investigated optimism of the lasso model via bootstrap resampling (Harrell, 2001). As opposed to traditional single-split data techniques where only portions of the data are used for model training and validation, bootstrap resampling makes full use of the data for model construction and gives nearly unbiased estimates of future model performance (Breiman, 1992; Harrell, 2001; Steyerberg *et al.*, 2001). This technique internally validates the original model fitting process by taking bootstrap samples from the original data set and reconstructing a new model using the same procedure as with the original data set. The performance of this new model is then checked on both the bootstrap sample and on the original data set, and any discrepancy is seen as evidence for optimism. The procedure is repeated at least 100 times to obtain a stable estimate of the expected value of optimism (Harrell, 2001). This procedure can be straightforwardly implemented in the absence of missing data. But when data are missing and MI is performed, it is unclear how resampling should be performed over the imputed data sets. We could either choose to: (*i*) perform resampling such that the same subjects are drawn in the bootstrap samples that are taken over the imputed data sets. Hence samples from the different imputed data sets differ only by imputed values, (*ii*) perform resampling such that the bootstrap samples taken over the imputed data sets differ by both the selected subjects and imputed values, (*iii*) perform the resampling process using only one of the imputed data sets, hence facilitating implementation, or (*iv*) resample the incomplete data set before performing MI. This issue had not been addressed in the literature.

In **chapter 8** we summarize our findings and conclusions and, give some directions for future research.