KU LEUVEN

Literature Review

Thesis title:

A Study on the Effect of Misspecifying the Association Structure in Dynamic Predictions Obtained from Joint Models for Longitudinal and Survival Data

Adam Mohammad
adam.mohammad@student.kuleuven.be
Student Number: r0478542

Supervisors: Geert Verbeke (KU Leuven),
Dimitris Rizopoulos (Erasmus MC, Rotterdam)

Programme: Master of Statistics

Academic year: 2016/2017

Introduction

In biomedical sciences follow-up studies are frequently conducted. In such studies, the sample units, typically patients, are followed over time and measured on different types of outcomes. These outcomes most often include longitudinal responses and time-to-event outcomes which tend to be related to each other.

Depending on the research question, the two outcomes can be analysed either separately or jointly. The choice between joint and separate analysis depends on the assumption about missing longitudinal observations in the study. Little and Rubin (2002) describe three missingness mechanisms: First, missing completely at random (MCAR) assumes that reason for missingness does not depend on any longitudinal observations. Second, missing at random (MAR) assumes that missingness can depend on collected, or observed, longitudinal observations, but not on missing, or unobserved, longitudinal observations. Third, missing not at random (MNAR), assumes that missingness depends on observed and missing longitudinal observations. When MCAR or MAR are assumed, separate analysis can be conducted.

Separate analysis is also performed when the outcomes are not related to each other and when the time-varying covariates are exogenous, or external (Exogenous covariates come from predictable processes. An example of such a covariate is time of the day, season of the year, etc. On the other hand, endogenous time-varying covariates are not predictable and therefore cannot be analysed separately). When separate analysis is conducted, continuous longitudinal outcomes are usually analysed using linear mixed models (Verbeke and Molenberghs, 2000), while time-to-event outcomes can be analysed by a variety of methods, such as a proportional hazard model, an accelerated failure time model, or an additive hazards model (Kleinbaum and Klein, 2012). In the class of proportional hazards models, the most well known model is the Cox model, which is a semiparametric method for modelling survival data. The extended version of the Cox model (Kleinbaum and Klein, 2012) allows to include exogenous time-varying covariates.

Joint models for longitudinal and time-to-event outcomes were developed for situations when the research question requires joint analysis of both outcomes. Diggle et al. (2009) describe three situations that require joint analysis: First, when the missingness mechanism is missing not at random (MNAR). Second, when the survival outcome is analysed and an endogenous time-varying covariate, such as a continuous biomarker, needs to be included into the model as a predictor. Third, when the association between the two outcomes is of interest and predictions need to be obtained.

The focus of this thesis is on joint models and more specifically, on the association structure between the two outcomes and obtaining predictions. In today's world of personalised medicine, it is of high interest to tailor medical decision making to the characteristics of patients, and patients themselves are invited to make decisions about their treatment. For that reason, both patients and physicians are interested in knowing patient-specific predictions of biomarker levels and the probability of experiencing an event of interest in a pre-specified time-frame. Hence, both parties require that the predictions obtained from joint models are as precise as possible. The association between the two outcomes can take different functional forms and each of these forms can yield different values of predictions. The purpose of this thesis is to study how the misspecification of the association structure in joint models for longitudinal and time-to-event outcomes affects dynamic predictions.

The Aortic Valve dataset (Bekkers et al., 2011) will be used in this thesis. The dataset contains information about 286 patients with aortic valve or aortic root disease who received a human tissue

valve at the Department of Cardio-Thoracic Surgery of the Erasmus Medical Center in Rotterdam between 1987 and 2008. Sub-coronary implantation was provided to 77 patients and the remaining 209 patients received a root replacement. Patients were then followed prospectively via telephone interviews and echocardiographic assessment. At each echocardiographic assessment, measurements of aortic gradient were taken. Due to the fact that human tissue valves are susceptible to degenerations, re-operations are often required to prevent death. Hence, the purpose is to analyse the association between aortic gradient and risk of death or re-operation. In order to delay death and plan re-operations in advance, cardiologists are also interested in accurately predicting patient's risk of death or re-operation.

Joint Modelling of Longitudinal and Time-to-Event Data

Estimation Techniques

Joint models, also called shared parameter models, consist of longitudinal and survival submodels, and use random effects to link the longitudinal and survival outcomes. More specifically, a mixed effects model is used to model the underlying longitudinal profile of each subject and this longitudinal profile is then included in the survival submodel as a time-varying covariate.

Initial approaches to estimate joint models were based on two-stage procedures, in which random effects were estimated in the first step and then included into the Cox model for the survival outcome in the second stage. More specifically, random effects for each event time were estimated by empirical Bayes methodology and then included into partial likelihood for the Cox model. Sweeting and Thompson (2011) have shown that this approach does not perform well under model misspecification, nor is it able to estimate well joint model parameters. Therefore, maximum likelihood and Markov chain Monte Carlo (MCMC) approaches turned preferable.

Later, joint likelihood for joint models was developed together with an Expectation-Maximisation (EM) algorithm that is used to obtain all parameters of interest. The basic assumption of this approach is that given random effects, the joint likelihood can be factorised into the likelihood corresponding to a longitudinal submodel and the likelihood corresponding to a survival submodel. The basic framework of joint modelling that uses maximisation of the joint likelihood of both outcomes was laid out by Wulfsohn and Tsiatis (1997). Their model was semiparametric as they did not have any distributional assumptions on the baseline hazard in the survival submodel. Thereafter, fully parametric joint likelihood was introduced, where the baseline hazard distribution is specified. The baseline hazard in the survival submodel is usually modelled by piecewise constant model or by using B-spline basis functions (Rizopoulos, 2012b).

Due to fact that the joint likelihood contains random effects, the integrals in the likelihood and score vectors do not have analytical solutions and numerical integration techniques have to be applied. Adaptive Gauss-Hermite quadrature is the standard choice to approximate such integrals. In order to decrease the computational burden, Rizopoulos (2012a) proposed a pseudo-adaptive Gauss-Hermite quadrature rule that uses the position of quadrature points from separately fitted linear mixed model that do not get updated when fitting the joint model. However, if the random effects structure is high-dimensional, i.e., if more than 5 random effects are included, the extension of the method to this setting is not straightforward. A possible alternative is to use Laplace approximation (Rizopoulos

et al., 2009).

The fully Bayesian approach has also been used to estimate joint model parameters and has become more popular in recent years, as it requires less assumptions and tends to fit more complex models faster. A Bayesian approach towards estimation of parameters in joint models uses MCMC techniques and was considered, for example, by Wang and Taylor (2001), Xu and Zeger (2001), Brown and Ibrahim (2003), Rizopoulos and Ghosh (2011).

Several extensive overviews of joint modelling of longitudinal and time-to-event data methodology can be found in the literature, e.g., by Tsiatis and Davidian (2004), Diggle et al. (2009), Rizopoulos (2012b), Ye and Yu (2013), and Rizopoulos (2013). Concerning software, these models can be fitted by using packages *JM* (Rizopoulos, 2010) or *JMbayes* (Rizopoulos, 2016) in `R` (R Core Team, 2016). The former package uses maximum likelihood estimation while the latter uses MCMC techniques.

Modelling Nonlinear Longitudinal Profiles

The longitudinal submodel often contains trajectories that are highly nonlinear. One approach to deal with such nonlinearities is to use higher-order polynomials, but splines tend to have better numerical properties (Ruppert et al., 2003). Brown et al. (2005) used B-splines for the longitudinal submodel with multidimensional random effects and Rizopoulos and Ghosh (2011) used natural cubic splines. A different perspective on modelling a nonlinear longitudinal outcome uses an additional stochastic term in the mixed model specification of a longitudinal submodel. In this way, the remaining serial correlation is modelled. For example, Henderson et al. (2000) included latent stationary Gaussian process and Wang and Taylor (2001) considered integrated Ornstein-Uhlenbeck (IOU) process. The choice between splines and stochastic process approaches is rather a philosophical issue as the choice depends on the belief about the true underlying biological process of the longitudinal outcome (Tsiatis and Davidian, 2004). However, including an additional stochastic process is more difficult to implement in software.

Joint Model Extensions

Several extensions of joint models to multiple longitudinal and multiple time-to-event outcomes have been proposed in the statistical literature. Models with multiple longitudinal outcomes were considered by Brown et al. (2005) and Rizopoulos and Ghosh (2011). Multiple time-to-event outcomes within joint modelling framework were considered, for example, by Elashoff et al. (2007). More specifically, they used competing risk model for multiple time-to-event outcomes. Other extensions of joint models, such as incorporating accelerated failure time model for time-to-event outcome, recurrent events, or handling categorical longitudinal outcomes, are discussed in Rizopoulos (2012b) and Ye and Yu (2013).

Modelling Association Structure

As noted above, the association between the two outcomes is one of the research questions for which joint models are used. Rizopoulos (2012b), Rizopoulos et al. (2014), and Rizopoulos (2016) present different possibilities of modelling the association structure between longitudinal and time-to-event outcomes. In its original form, the predicted current value of a longitudinal outcome is used as a predictor in a survival submodel. However, this is not always meaningful, as the hazard of experiencing an event may depend on previous values of the longitudinal outcome as well. Some other possible associations could be: including just random effects for a particular subject, lagged effects, value of the slope, cumulative effects, etc.

Dynamic Predictions

The dynamic prediction of the survival probability for a future time point can be described as the predicted probability that a subject will not experience the event of interest until that future time point, provided that he/she has not experienced the event until the current time point at which the prediction is made. Similarly, a future value of the longitudinal outcome can be predicted. The fact that new observations for both outcomes are recorded with every new visit, allows the predictions to be updated with every visit. Hence, these predictions are called dynamic. Dynamic predictions and prospective accuracy are discussed, for example, by Rizopoulos (2012b), Rizopoulos (2011) or Ye and Yu (2013).

The first approach to obtain dynamic predictions was the Landmarking approach. This approach uses the extended version of the Cox model with the longitudinal outcome as time-varying covariate. The survival probabilities for a new subject are estimated using the Breslow estimator. The major problem of this approach is that it uses the "last observation carried forward" approach for the longitudinal covariate, which is not meaningful for endogenous time-varying covariates. The landmarking approach also assumes that the missing longitudinal observations are MCAR, that the visiting process is non-informative, and that censoring is non-informative. More details about this approach can be found in van Houwelingen and Putter (2012) and Putter (2013).

On the other hand, joint modelling approach towards dynamic predictions allows the censoring and the visiting processes to be informative, i.e., to depend on the longitudinal history, and is also suitable for handling endogenous time-varying covariates. Rizopoulos (2011) introduced the Monte Carlo algorithm to estimate predictions for both outcomes.

As different values of dynamic predictions can be obtained from joint models with different association structures, it might be difficult to know which prediction is the best unless the true underlying association between the two outcomes is known in advance. Rizopoulos et al. (2014) considered combining, or averaging, these predictions from models with different association structure by using Bayesian Model Averaging.

Purpose of the Thesis

The purpose of this thesis is to study how misspecification of the association structure between longitudinal and survival outcomes in joint models affects dynamic predictions.

As written above, there are different functional forms of the association structure between the two outcomes in joint models. Several studies have been conducted on which features of longitudinal profiles are most strongly associated with the risk of an event. However, not much research has been conducted on how the functional form affects predictions and what is the effect of misspecification of the functional form.

The Aortic Valve dataset (Bekkers et al., 2011) will be used in the thesis to study the effect of misspecification. The dataset will be randomly divided into 2 parts in order to perform validation, i.e., one part of the dataset will be used to estimate the joint model parameters with different association functional forms and the other part will be used to obtain dynamic predictions. This will be done in an iterative way. A simulation study will also be conducted in a similar manner.

The estimated predictions will then be compared with the true values by using discrimination and

calibration which are discussed in more detail by Rizopoulos (2016). The discrimination approach measures how well the biomarker can discriminate between subjects at high risk of event and low risk of event. The approach is based on sensitivity (the probability of correctly estimating an event in pre-specified time frame) and specificity (the probability of correctly estimating that the subject will not experience an event). The discriminative ability of the model is then measured by the area under the receiver operating curve (ROC) and the dynamic concordance index. The calibration approach measures how accurately the biomarker is able to predict future events. This is done either at one future timepoint by the prediction error (PE), or in future time interval by an integrated prediction error (IPE).

References

Bekkers, J. A., Klieverik, L. M., Raap, G. B., Takkenberg, J. J. and Bogers, A. J. (2011). "Re-operations for aortic allograft root failure: experience from a 21-year single-center prospective follow-up study". *European Journal of Cardio-Thoracic Surgery* **40**(1), pp. 35–42. DOI: 10.1016/j.ejcts.2010.11.025.

Brown, E. R. and Ibrahim, J. G. (2003). "A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data". *Biometrics* **59**(2), pp. 221–228. DOI: 10.1111/1541-0420.00028.

Brown, E. R., Ibrahim, J. G. and DeGruttola, V. (2005). "A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival". *Biometrics* **61**(1), pp. 64–73. DOI: 10.1111/j.0006-341x.2005.030929.x.

Diggle, P., Henderson, R. and Philipson, P. (2009). "Random-effects models for joint analysis of repeated-measurements and time-to-event outcomes". In: *Handbook of Longitudinal Data Analysis*. Ed. by Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. Chapman & Hall/CRC, pp. 349–366.

Elashoff, R. M., Li, G. and Li, N. (2007). "A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types". *Biometrics* **64**(3), pp. 762–771. DOI: 10.1111/j.1541-0420.2007.00952.x.

Henderson, R., Diggle, P. and Dobson, A. (2000). "Joint modelling of longitudinal measurements and event time data". *Biostatistics* **1**(4), pp. 465–480. DOI: 10.1093/biostatistics/1.4.465.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis*. New York: Springer. DOI: 10.1007/978-1-4419-6646-9.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd. Hoboken, NJ: John Wiley & Sons. DOI: 10.1002/9781119013563.

Putter, H. (2013). "Landmarking". In: *Handbook of Survival Analysis*. Ed. by Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Roton, FL: Chapman and Hall/CRC, pp. 441–456.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rizopoulos, D. (2010). "JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data". *Journal of Statistical Software* **35**(1), pp. 1–33. DOI: 10.18637/jss.v035.i09.

Rizopoulos, D. (2011). "Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data". *Biometrics* **67**(3), pp. 819–829. DOI: 10.1111/j.1541-0420.2010.01546.x.

Rizopoulos, D. (2012a). "Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule". *Computational Statistics & Data Analysis* **56**(3), pp. 491–501. DOI: 10.1016/j.csda.2011.09.007.

Rizopoulos, D. (2012b). *Joint models for longitudinal and time-to-event data: With applications in R*. Boca Roton, FL: Chapman & Hall/CRC. DOI: 10.1201/b12208.

Rizopoulos, D. (2013). "Joint Modeling of Longitudinal and Time-to-Event Data: Challenges and Future Directions". In: *Advances in Theoretical and Applied Statistics*. Ed. by Torelli, N., Pesarin, F. and Bar-Hen, A. Springer, pp. 199–209. DOI: 10.1007/978-3-642-35588-2_19.

Rizopoulos, D. (2016). "The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC". *Journal of Statistical Software* **72**(1), pp. 1–46. DOI: 10.18637/jss.v072.i07.

Rizopoulos, D. and Ghosh, P. (2011). "A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event". *Statistics in Medicine* **30**(12), pp. 1366–1380. DOI: 10.1002/sim.4205.

Rizopoulos, D., Hatfield, L. A., Carlin, B. P. and Takkenberg, J. J. M. (2014). "Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging". *Journal of the American Statistical Association* **109**(508), pp. 1385–1397. DOI: 10.1080/01621459.2014.931236.

Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009). "Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(3), pp. 637–654. DOI: 10.1111/j.1467-9868.2008.00704.x.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press. 404 pp. DOI: 10.1017/cbo9780511755453.

Sweeting, M. J. and Thompson, S. G. (2011). "Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture". *Biometrical Journal* **53**(5), pp. 750–763. DOI: 10.1002/bimj.201100052.

Tsiatis, A. A. and Davidian, M. (2004). "Joint Modeling of Longitudinal and Time-to-Event Data: An Overview". *Statistica Sinica* **14**(3), pp. 809–834.

van Houwelingen, H. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer. DOI: 10.1007/978-1-4419-0300-6.

Wang, Y. and Taylor, J. M. G. (2001). "Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome". *Journal of the American Statistical Association* **96**(455), pp. 895–905. DOI: 10.1198/016214501753208591.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). "A Joint Model for Survival and Longitudinal Data Measured with Error". *Biometrics* **53**(1), pp. 330–339. DOI: 10.2307/2533118.

Xu, J. and Zeger, S. L. (2001). "Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **50**(3), pp. 375–387. DOI: 10.1111/1467-9876.00241.

Ye, W. and Yu, M. (2013). "Joint Models of Longitudinal and Survival Data". In: *Handbook of Survival Analysis*. Ed. by Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Roton, FL: Chapman and Hall/CRC, pp. 523–547.