Joint longitudinal and time-to-event models: development, implementation and applications in health research

---

Samuel L. Brilleman

Bachelor of Science (Hons.), Master of Science (Medical Statistics)

A thesis submitted for the degree of Doctor of Philosophy
at Monash University in 2018

Department of Epidemiology and Preventive Medicine
School of Public Health and Preventive Medicine
Faculty of Medicine, Nursing and Health Sciences

Table of Contents

List of tables

(This page has been intentionally left blank)

List of figures

(This page has been intentionally left blank)

List of abbreviations

AFT	Accelerated failure time
ANZDATA	Australia and New Zealand Dialysis and Transplant Registry
AUC	Area under the (receiver operating characteristic) curve
BMI	Body mass index
CDF	Cumulative distribution function
CI	Confidence interval
CRAN	Comprehensive R Archive Network
CrI	(Bayesian) credible interval
ctDNA	Circulating tumor DNA
EM	Expectation-maximisation
ESKD	End-stage kidney disease
FEMA	Federal Emergency Management Agency
HIV	Human immunodeficiency virus
HMC	Hamiltonian Monte Carlo
HR	Hazard ratio
HRS	Health and Retirement Study
IPD	Individual patient data (meta-analysis)
MICE	Multiple imputation using chained equations
MCMC	Markov chain Monte Carlo
NHMRC	National Health and Medical Research Council
NSCLC	Non-small cell lung cancer
PSA	Prostate-specific antigen
RRT	Renal replacement therapy

(This page has been intentionally left blank)

Abstract

Joint modelling of longitudinal and time-to-event data is a fertile area of research. The last two decades has seen significant attention paid to the development of the methodology. More recently, there has been increased uptake of the methods in applied research. Nonetheless, there is still scope for the wider use of joint modelling approaches. This thesis contributes to the uptake of joint modelling in applied health research through three key areas: methodological development, software implementation, and application. Chapter 2 of the thesis provides a review of the methodological framework for the joint modelling of longitudinal and time-to-event data. Chapters 3 and 4 of the thesis present research projects demonstrating the application of the two most common joint modelling approaches. First, in Chapter 3, a shared parameter joint model with competing risks is applied in an epidemiological and public health project related to disasters, death and disability. Second, in Chapter 4, a latent class joint model is applied in a clinical research project related to end-stage kidney disease. Chapter 5 of the thesis describes the development of new software for estimating joint longitudinal and time-to-event models under a Bayesian framework. The software package provides a user-friendly interface for fitting a variety of shared parameter joint models, including univariate (one longitudinal outcome) or multivariate (more than one longitudinal outcome) joint models, different longitudinal data types (continuous, binary, counts), multilevel data structures, and a range of association structures. In Chapter 6, the development of novel methodology for fitting joint models in the presence of multilevel data structures is described in detail. Lastly, in Chapter 7, the overall findings from the thesis and potential avenues for future work are discussed.

(This page has been intentionally left blank)

Acknowledgements

I would first like to thank my amazing supervisors Prof Rory Wolfe, Dr Margarita Moreno-Betancur, and Dr Michael Crowther. They have each been extremely generous with their time over the last few years and I have learnt much from them.

I would also like to acknowledge Jacki Buros Novik and Eric Novik, who have assisted me greatly during the PhD and with whom I have enjoyed collaborating. In addition, I would like to acknowledge a number of other key collaborators on these projects; Kevan Polkinghorne, Sergine Lo, Nidal Al-Huniti, Jim Dunyak, Rob Fox, Daniel Lee, Jonah Gabry and Ben Goodrich, all of whom have assisted me in some way, shape, or form during the last few years.

I would also like to acknowledge the biostatistics group at the School of Public Health and Preventive Medicine and the Victorian Centre for Biostatistics (ViCBiostat). Both have been supportive and enjoyable environments in which to work and be involved. I am also thankful to the various colleagues who have supported me on my research journey prior to Monash University, including those from my time in Bristol and Leicester. I would also like to acknowledge the PhD coordinators at the School of Public Health and Preventive Medicine for their support over the course of my PhD; Kathryn Daly, Liz Douglas, Renee Conroy, and Natalie Pekin.

Of course, I would like to thank my family, Deb, Ed, Rebecca, and Dion, and my friends both inside and outside work, for being supportive throughout this PhD journey and, at times, seeming even more interested in my PhD than I was.

Lastly, I would like to spend one of my biggest thanks on my partner, Anita, for her patience, support and all-round awesomeness. Although the PhD has been a relatively smooth journey, she has nonetheless been the one sitting in the carriage next to me on the few dips in the rollercoaster. I hereby promise I will start cooking again and stop working late.

(This page has been intentionally left blank)

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes one original paper published in a peer-reviewed journal, two submitted publications, and one peer-reviewed paper published as part of conference proceedings. The core theme of the thesis is the joint modelling of longitudinal and time-to-event data and its application in clinical, public health, or epidemiological studies. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Public Health and Preventive Medicine under the supervision of Professor Rory Wolfe (Monash University), Dr Margarita Moreno-Betancur (Murdoch Childrens Research Institute, University of Melbourne), and Dr Michael Crowther (University of Leicester).

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. In the case of Chapters 3, 4, 5, and 6, my contribution to the work is summarised in Table 1. None of the co-authors included in any of the works were Monash University students. I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: [redacted] **Date:** 5th March 2018

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature: [redacted] **Date:** 5th March 2018

Table 1. Author contributions for publications included in the thesis.

Thesis chapter	Publication title	Publication status[1]	Nature and % of student contribution	Co-author name(s) and nature of their contribution
3	Associations between community-level disaster exposure and individual-level changes in disability and risk of death for older Americans	Published	75%. Project planning, data management, statistical analysis, first draft of paper, contribution to final manuscript.	TJI: Chief investigator, project conception, obtained data and funding. RW, MMB, TJI: Project planning, advice on analysis. AES, KML, YL, ELDB, LR: Assisted with project conception and grant application. RW, MMB, AES, KML, YL, ELDB, LR, TJI: Comments on written drafts, approval of final manuscript.
4	Longitudinal changes in body mass index and the competing outcomes of death and transplant in patients undergoing hemodialysis	Submitted	75%. Data management, statistical analysis, first draft of paper, contribution to final manuscript.	JT: Preliminary analyses. MMB, KRP, MJC, RW: Advice on analysis. MMB, KRP, MJC, SPM, JT, RW: Comments on written drafts, approval of final manuscript.

5	Joint longitudinal and time-to-event models via Stan	Published (conference paper)	85%. Writing code, statistical analysis for application, preparation of manuscript.	MJC, MMB, JBN, RW: Advice on analysis, comments on written drafts, approval of final manuscript.
6	Joint longitudinal and time-to-event models for multilevel hierarchical data	Submitted	75%. Project planning, data management, statistical analysis, first draft of paper, contribution to final manuscript.	JBN: Data cleaning. MJC, MMB, JBN, NAH, RF, JD, JH, RW: Project planning, advice on analysis, comments on written drafts, approval of final manuscript.

[1] Available categories: published, in press, accepted or returned for revision.

(This page has been intentionally left blank)

Research outcomes during enrolment

Peer-reviewed journal articles

1. Moreno-Betancur M, Carlin J, **Brilleman SL**, Tanamas S, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics.* 2017 (Epub ahead of print)
2. Karim Md N, Reid CM, Huq M, **Brilleman SL**, Cochrane A, Tran L, Billah B. Predicting long-term survival after coronary artery bypass graft surgery. *Interactive Cardiovascular and Thoracic Surgery*. 2017 (Epub ahead of print)
3. **Brilleman SL**, Howe LD, Wolfe R, Tilling K. Bayesian piecewise linear mixed models with a random change point: an application to BMI rebound in childhood. *Epidemiology*. 2017;28(6):827-833.
4. Chimeddamba O, Gearon E, **Brilleman SL**, Tumenjargal E, Peeters A. Increases in waist circumference independent of weight in Mongolia over the last decade: the Mongolian STEPS surveys. *BMC Obesity*. 2017;4:19.
5. **Brilleman SL**, Wolfe R, Moreno-Betancur M, Sales AE, Langa KM, Li Y, Daugherty Biddison EL, Rubinson L, Iwashyna TJ. Associations between community-level disaster exposure and individual-level changes in disability and risk of death for older Americans. *Social Science & Medicine*. 2017;173:118-125.
6. **Brilleman SL**, Crowther MJ, May M, Gompels M, Abrams K. Joint longitudinal hurdle and time-to-event models: an application related to viral load and duration of the first treatment regimen in HIV patients initiating therapy. *Statistics in Medicine*. 2016;35(20):3583-3594.
7. **Brilleman SL**, Metcalfe C, Peters TJ, Hollingworth W. The reporting of treatment non-adherence and its associated impact on economic evaluations conducted alongside randomised trials: a systematic review. *Value in Health*. 2016;19(1):99-108.
8. McClean S, **Brilleman S**, Wye L. What is the perceived impact of Alexander technique lessons on health status, costs and pain management in the real life setting of an English

hospital? The results of a mixed methods evaluation of an Alexander technique service for those with chronic back pain. *BMC Health Services Research*. 2015;15:293.

Submitted journal articles

1. **Brilleman SL**, Crowther MJ, Moreno-Betancur M, Buros Novik J, Dunyak J, Al-Huniti N, Fox R, Hammerbacher J, Wolfe R. Joint longitudinal and time-to-event models for multilevel hierarchical data. *Submitted for publication*.
2. **Brilleman SL**, Moreno-Betancur M, Polkinghorne KR, McDonald SP, Crowther MJ, Thomson J, Wolfe R. Longitudinal changes in body mass index and the competing outcomes of death and transplant in patients undergoing hemodialysis: a joint latent class mixed model approach. *Submitted for publication.*
3. Flehr A, Coles J, Dixon J, Gibson S, **Brilleman SL**, Harris M, Loxton, D. Epidemiology of trauma history and pain outcomes: a retrospective study of community based Australian women. *Submitted for publication*.

Software

1. **Brilleman SL** (2017) simsurv: Simulate survival data. R package version 0.2.0. https://CRAN.R-project.org/package=simsurv
2. Stan Development Team (2018) rstanarm: Bayesian applied regression modeling via Stan. R package version 2.17.0. http://mc-stan.org/
3. **Brilleman SL** (2018) simjm: Simulate joint longitudinal and survival data. R package version 0.1.0. https://github.com/sambrilleman/simjm

Peer-reviewed conference papers

1. Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J, Wolfe R. Joint longitudinal and time-to-event models via Stan. In: *Proceedings of StanCon 2018*. 10-12 Jan 2018. Pacific Grove, California, USA. https://github.com/stan-dev/stancon_talks

Contributed conference talks

1. **Brilleman SL,** Crowther MJ, Moreno-Betancur M, Buros Novik J, Wolfe R. Joint longitudinal and time-to-event models via Stan. *StanCon 2018.* 10-12 Jan 2018. Pacific Grove, California, USA.

2. **Brilleman SL**, Crowther MJ, Moreno-Betancur M, Lo S & Wolfe R. Bayesian joint models for multiple longitudinal biomarkers and a time-to-event outcome: software development and a melanoma case study. *38th Annual Conference of the International Society for Clinical Biostatistics.* 9-13 July 2017. Vigo, Spain.
3. **Brilleman SL**, Crowther MJ, Moreno-Betancur M & Wolfe R. Bayesian joint models for multiple longitudinal biomarkers and time-to-event data: methods and software development. *Australian Statistical Conference 2016*. 5-9 Dec 2017. Canberra, Australia.
4. **Brilleman SL**, Iwashyna TJ, Moreno-Betancur M & Wolfe R. Joint longitudinal and survival models for investigating the association between natural disasters and disability whilst accounting for non-random dropout due to death. *Conference of the International Biometric Society: Australasian Region.* 29 Nov – 3 Dec 2015. Hobart, Australia.

(This page has been intentionally left blank)

Chapter 1: Overview of the thesis

1.1 Introduction and motivations

Joint, or simultaneous, modelling of longitudinal and time-to-event data generating processes is the focus of this thesis. This refers to methodology for modelling the evolution of a repeatedly measured clinical biomarker marker alongside an event process. Commonly, this is achieved by specifying a joint likelihood formulation for longitudinal and time-to-event outcomes.

There are a number of potential motivations for using a joint modelling approach. First, one may be interested in the effect of the longitudinal biomarker on the event but wishes to allow for the discrete-time, imperfect measurement of the biomarker. Second, one may be interested in the longitudinal process, but wishes to account for informative dropout due to occurrence of the event. Third, one may be interested in exploring the structure of the association between the longitudinal and event processes. Lastly, one may wish to develop a prognostic model for the risk of the event whilst utilising the time-varying information related to the biomarker. Each of these potential motivations will be revisited and described in greater detail in Chapter 2.

Joint modelling is an extremely fertile area of research. The earliest research papers describing joint modelling predominantly focussed on a single continuous longitudinal outcome and a single time-to-event outcome (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000). However, since those early papers, there have been a vast array of methodological developments. Joint modelling approaches for longitudinal and time-to-event data have been extended to handle a wide variety of complex data structures and research questions. This includes, for example, competing events (Li et al., 2009; Williamson et al., 2008; Andrinopoulou et al., 2014; Lu, 2017), recurrent events (Król et al., 2016), non-normally distributed longitudinal biomarkers (Faucett et al., 1998;

Brilleman et al., 2016), accelerated failure time models (Tseng et al., 2005), multi-state models (Dantan et al., 2011; Ferrer et al., 2016), latent class joint models (Garre et al., 2007; Lin et al., 2002; Proust-Lima et al., 2014), multiple longitudinal biomarkers (Proust-Lima et al., 2009; Rizopoulos and Ghosh, 2011), a variety of association structures (Crowther, Lambert, et al., 2013; Mauff et al., 2017), and more.

Although the majority of joint modelling research has been methodological in nature, there has also been a growing uptake of the methodology in applied research. Joint modelling approaches are now being used in applied studies published in both clinical and epidemiological journals. Nonetheless, there is still much scope for the wider use of joint modelling approaches in applied health research. In a relatively recent review of joint modelling methods and software Lawrence Gould et al. (2015) described several clinical applications of joint models. However, they also highlighted the important issue that most published articles relating to, or using, joint models have so far appeared in the statistical literature, thereby limiting their exposure to applied researchers.

Accordingly, this thesis will aim to contribute to the wider uptake of joint modelling methodology in applied health research. This aim will be achieved through three main objectives, however, before defining those objectives, the motivations underlying them will be described.

The first objective of this thesis is motivated by the need to contribute to a wider awareness of joint modelling methods amongst applied researchers. Many applied researchers may not yet be aware of joint modelling methodology since the majority of publications have appeared in the statistical literature. As readers of epidemiological and clinical medicine journals become more familiar with the methodology, and aware of the benefits that joint modelling approaches can provide, they are more likely to consider using the methods in their own studies. By publishing applied projects in clinical, epidemiological or public health journals this thesis will help demonstrate how joint modelling methods can be used to answer research questions of importance to health.

The second objective of this thesis is motivated by the need for further methodological developments in the area of joint modelling. Joint modelling methodology has been developing at a rapid pace. Nonetheless, there are still a number of scenarios encountered in practice for which existing methods may not be appropriate. One such scenario is the presence of multilevel data structures. For example, in an oncology trial, repeated

measurements of a clinical biomarker might be observed for multiple tumour lesions clustered within each patient. Alternatively, in the analysis of a registry-based dataset, longitudinal and event-time data might be measured on patients clustered within clinics or hospitals. In both of these settings, there are additional clustering factors beyond just that of the patient. Such multilevel data structures are commonly encountered in health research but have not yet been considered in the joint modelling context. Accordingly, there is scope for the development of new methodology for the joint modelling of longitudinal and time-to-event data in the presence of complex data structures commonly encountered in health research.

The third objective of this thesis is motivated by the need to provide a broader range of user-friendly software implementations for joint models, particularly for those methods that extend beyond the standard joint model formulation. Many of the proposed methodological extensions to standard joint models have not been implemented in user-friendly software. Some examples of methodological developments for which there are currently only limited, or recently available, user-friendly software implementations are: non-normally distributed biomarkers, multivariate joint models, multilevel data structures, and complex association structures (e.g. cumulative effects, lagged effects, or interactions between biomarkers). If novel methods are not implemented in user-friendly software, then there is a much smaller likelihood that they will be adopted by applied researchers.

1.2 Aims

The broad aim of this thesis is to help facilitate, as well as actively contribute to, the wider uptake of joint modelling methodology in applied health research.

1.3 Objectives

The aim of the thesis will be achieved through three main objectives:

- Promote the use of joint modelling by undertaking applied research projects that use joint modelling methodology to answer important health research questions, and seek to publish those analyses in clinical, epidemiological or public health journals.
- Contribute to the development of novel joint modelling methodology by extending the methods to data structures commonly encountered in applied health research but that are not yet accommodated within the published joint modelling literature.

- Facilitate the ease of access to joint modelling methods by developing user-friendly, open-source, software to fit joint models under a Bayesian framework.

1.4 Structure of the thesis

This thesis is organised as follows. The following chapter (Chapter 2) introduces the methodological framework for the joint modelling of longitudinal and time-to-event data. It highlights the key methodological developments and gaps in the literature, focussing on those that are most relevant to the content of later chapters of this thesis.

Chapters 3 and 4 each present an applied research project that used joint modelling methodology. The main content of each chapter is an applied research paper describing the project. The first applied project, in Chapter 3, is based on a shared parameter joint modelling approach and explored the associations between disaster exposure, disability, and risk of death in a cohort of older Americans. The project had both an epidemiological and public health focus and was published in 2016 in *Social Science and Medicine*. The second applied project, in Chapter 4, is based on a latent class joint modelling approach and explored associations between longitudinal changes in body mass index (BMI) and the risk of death or transplant in Australian and New Zealand end-stage kidney disease patients undergoing haemodialysis. The paper for this project was recently submitted for publication.

Chapter 5 describes the development of new functionality within the **rstanarm** (Brilleman et al., 2018; Stan Development Team, 2017a) R package for Bayesian estimation of joint longitudinal and time-to-event models. These developments were motivated by a lack of user-friendly software for fitting multivariate joint models (i.e. more than one longitudinal biomarker), non-normally distributed longitudinal biomarkers (e.g. binary or counts), or data with a multilevel structure (i.e. clustering factors beyond just that of the individual). The primary content of the chapter is a published conference paper that describes the main features of the software package. The chapter also includes a qualitative comparison of the software with the features included in other packages currently available for fitting multivariate joint models. Lastly, the chapter includes a simulation study to assess the performance of the new joint modelling software. The simulation study required the development of two additional R packages; one for simulating complex time-to-event (i.e. survival) data based on the method of Crowther and Lambert (2013), and another for

simulating joint longitudinal and time-to-event data. Therefore, the development of these two new R packages is also described.

Chapter 6 describes a methodological project. The project was motivated by a clinical study in which the aim was to explore the association between tumour burden and progression-free survival in East-Asian non-small cell lung cancer (NSCLC) patients initiating treatment. In this study, the tumour burden for a patient was defined as a summary of the sizes of the target tumour lesions. Since a patient could have more than one tumour lesion, and each lesion was tracked over time, the data for the study had a multilevel hierarchical structure; observation times (level 1) were clustered within lesions (level 2) which were clustered within patients (level 3). The study therefore required the development of novel methodology (and software) for the joint modelling of longitudinal and time-to-event data when there are clustering factors beyond just that of the patient. The main content of the chapter is a paper that has been recently submitted and is currently under review.

Chapter 7, the final chapter of the thesis, contains a discussion of the overall findings and outlines possible avenues for future work.

Chapter 2: Background to the joint modelling of longitudinal and time-to-event data

2.1 Chapter introduction

This chapter provides an introduction to the methodological framework for the joint modelling of longitudinal and time-to-event data. The intention of this chapter is not to provide an exhaustive review of the joint modelling literature, since a number of reviews have been published elsewhere by previous authors (Tsiatis and Davidian, 2004; Rizopoulos, 2012b; Lawrence Gould et al., 2015; Proust-Lima et al., 2014; Hickey et al., 2016). Rather, the intention of this chapter is to introduce the methods underpinning the estimation of joint models and to highlight important developments and gaps in the literature, focussing on those *that are of most relevance to the content of this thesis*.

Section 2.2 begins by describing the context for joint modelling; that is, the situations in which we might observe both longitudinal and time-to-event data and the types of research questions that might benefit from modelling both outcome types simultaneously. Section 2.3 describes the formulation of a shared parameter joint model, starting with the simplest specification and then followed by a number of possible extensions that have been discussed in the literature. Section 2.4 introduces latent class joint models, and explains their differences to the shared parameter joint modelling approach. Section 2.5 summarises the various estimation approaches that have been used for joint models including classical, Bayesian, and novel two-stage estimation approaches.

2.2 Context

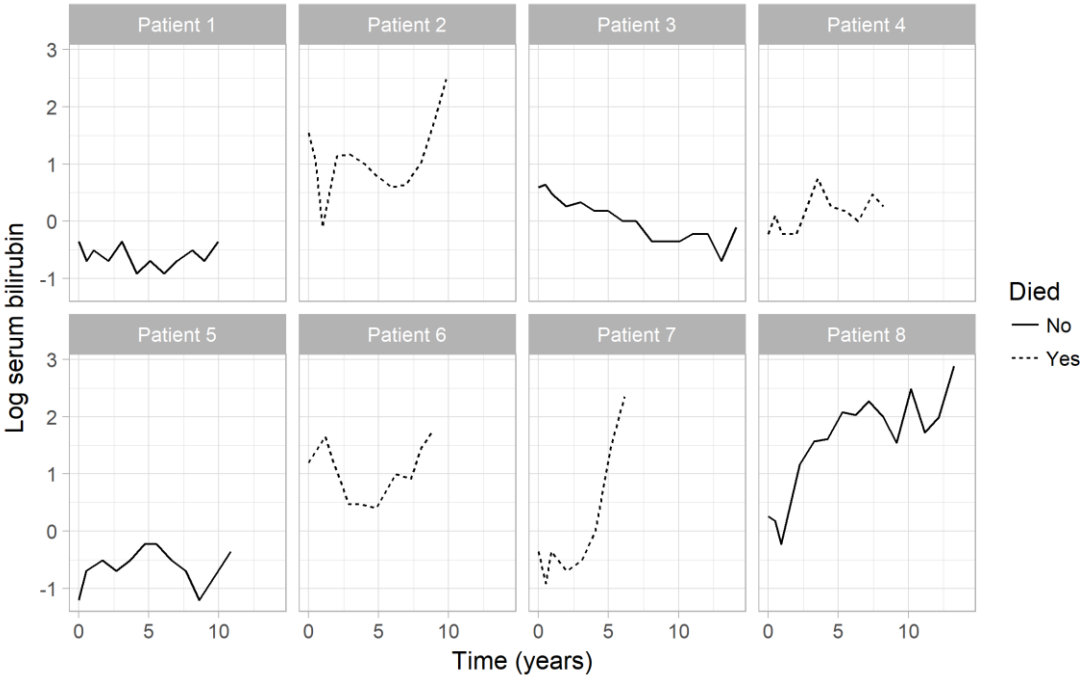
Let us assume we observe repeated measurements of a clinical biomarker, over time, for a sample of patients drawn from some target population of interest. Figure 1, for example, shows observed longitudinal measurements (i.e. observed "trajectories") of log serum

bilirubin for a sample of eight patients with primary biliary cirrhosis (the data are taken from the PBC dataset that will be introduced in Chapter 5). Along with observing the longitudinal biomarker measurements for these patients, we may also observe the patient-specific time from a defined origin (e.g. diagnosis of the disease) until a terminating clinical event such as death. Accordingly, solid lines in Figure 1 are used for representing those patients who were still alive at the end of the follow up period, while dashed lines are used for those patients who died.

An important characteristic of Figure 1 is that we observe between-patient variation in the longitudinal trajectories for log serum bilirubin, with some patients showing an increase in the biomarker over time, others decreasing, and some remaining stable. Moreover, there is variation between patients in terms of the frequency and timing of the longitudinal measurements.

From the perspective of clinical risk prediction, we may be interested in asking whether between-patient variation in the log serum bilirubin trajectories provides meaningful

Figure 1. Observed longitudinal trajectories of log serum bilirubin for eight primary biliary cirrhosis patients.



Notes. Solid lines are used for those patients who were still alive at the end of follow up, while dashed lines are used for those patients who died.

prognostic information that can help us differentiate patients with regard to some clinical event of interest, such as death. Alternatively, from an epidemiological perspective we may wish to explore the potential for etiological associations between changes in log serum bilirubin and mortality. Joint modelling approaches provide us with a framework under which we can begin to answer these types of clinical and epidemiological questions.

More formally, the motivations for undertaking a joint modelling analysis of longitudinal and time-to-event data might include one or more of the following:

- One may be interested in how *underlying changes in the biomarker influence the occurrence of the event*. However, including the observed biomarker measurements directly in a time-to-event model as time-varying covariates poses several problems. For example, if the widely used Cox proportional hazards model is assumed for the time-to-event model then biomarker measurements need to be available for all patients at all failure times, which is unlikely to be the case (Tsiatis and Davidian, 2004). If simple methods of imputation are used to deal with this missing data challenge, such as the "last observation carried forward" method, then these are likely to induce bias (Liu and Gould, 2002). Furthermore, the observed biomarker measurements may be subject to measurement error and therefore their inclusion as time-varying covariates may result in biased and inefficient estimates. In most cases, the measurement error will result in parameter estimates which are shrunk towards the null (Prentice, 1982). On the other hand, joint modelling approaches allow us to estimate the association between the biomarker (or some function of the biomarker trajectory, such as rate of change in the biomarker) and the risk of the event, whilst allowing for both the discrete time and measurement-error aspects of the observed biomarker.
- One may be interested primarily in the evolution of the clinical biomarker but *may wish to account for what is known as informative dropout*. If the values of future (unobserved) biomarker measurements are associated with the occurrence of the terminating event, then those unobserved biomarker measurements will be "missing not at random" (Baraldi and Enders, 2010; Philipson et al., 2008). In other words, biomarker measurements for patients who have an event will differ from those who do not have an event. Under these circumstances, inference based solely on observed measurements of the biomarker, under a “missing at random” (ignorability) assumption will be subject to bias. A joint modelling approach can help to adjust for informative dropout and has been shown to reduce bias in estimates of parameters quantifying

longitudinal changes in the biomarker (Henderson et al., 2000; Pantazis et al., 2005; Philipson et al., 2008). Nonetheless, it is worth highlighting that in practice this benefit of joint modelling is likely to depend strongly on untestable assumptions, such as a correctly specified model for the non-ignorable missing data mechanism.

- Joint models are naturally suited to the task of *dynamic risk prediction*. For example, joint modelling approaches have been used to develop prognostic models where predictions of event risk can be updated as new longitudinal biomarker measurements become available. Taylor et al. (2013) jointly modelled longitudinal measurements of prostate-specific antigen (PSA) and time to clinical recurrence of prostate cancer. The joint model was then used to develop a web-based calculator which could provide real-time predictions of the probability of recurrence based on a patient's up to date PSA measurements.

2.3 Shared parameter joint models

2.3.1 Specification

The most common joint modelling approach appearing in the literature to date is the so-called *shared parameter joint model*. A standard shared parameter joint model consists of two related submodels, one for each of the longitudinal and time-to-event outcomes. These are therefore commonly referred to as the *longitudinal submodel* and the *event submodel*. The submodels are linked through shared individual-specific parameters, modelled as individual-level random effects, which can be incorporated into the model in a number of ways.

The most common formulation of a shared parameter joint model is one in which the log hazard (i.e. instantaneous rate) of the time-to-event outcome at time t is assumed to be linearly associated with the expected value of the longitudinal biomarker at time t. The two submodels under this formulation are described next.

Longitudinal submodel

Assume $y_{ij} = y_i(t_{ij})$ corresponds to the longitudinal values of a biomarker observed at the j^{th} $(j = 1, \dots, n_i)$ time point, t_{ij}, for the i^{th} $(i = 1, \dots, N)$ individual in a random sample of N individuals. Although the longitudinal biomarker is observed at discrete time points, these are assumed to be error-prone measurements of an underlying continuously evolving

process. Therefore, the biomarker is modelled using a linear mixed model that can be specified as

$$\begin{aligned} y_i(t) &= \mu_i(t) + \varepsilon_i(t) \\ \mu_i(t) &= \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i \end{aligned} \tag{1}$$

where $\mu_i(t)$ is the expected value of the biomarker for individual i at time t, $\varepsilon_i(t)$ is a random (measurement) error term assumed to be drawn from a normal distribution with mean zero and variance σ_{y}^2, and $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ are vectors of covariates, possibly time-dependent, with associated vectors of population-level (i.e. fixed effects) parameters $\boldsymbol{\beta}$ and individual-specific (i.e. random effects) parameters $\boldsymbol{b}_i$. The individual-specific parameters are assumed to be normally distributed, with $\boldsymbol{b}_i \sim N(0, \boldsymbol{\Sigma}_{\mathrm{b}})$ for some unstructured variance-covariance matrix $\boldsymbol{\Sigma}_b$.

Event submodel

Let T_i^* denote the event time for individual i, which may or may not be observed due to right-censoring. Therefore, in practice, we observe $T_i = \min(T_i^*, C_i)$ for individual i, where C_i is the right-censoring time. An event indicator can be defined as $d_i = I(T_i^* \leq C_i)$, where $I(.)$ is the indicator function taking a value of 1 if $T_i^* \leq C_i$ or 0 otherwise.

The hazard of the event at time t is the instantaneous rate of occurrence for the event at time t. Mathematically, it is defined as

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_i^* < t + \Delta t \mid T_i^* > t)}{\Delta t} \tag{2}$$

where Δt is the width of some small time interval. The numerator in equation (2) is the conditional probability of the individual experiencing the event during the time interval $[t, t + \Delta t)$, given that they were still at risk of the event at time t. The denominator in the equation converts the conditional probability to a rate per unit of time. As Δt approaches the limit, the width of the interval approaches zero and the instantaneous event rate is obtained (Rodríguez, 2007).

The most common approach in the joint modelling literature has been to model the hazard of the event using a proportional hazards model of the form

$$h_i(t) = h_0(t)\exp\{\boldsymbol{w}_i'(t)\boldsymbol{\gamma} + \alpha\mu_i(t)\} \quad (3)$$

where $h_i(t)$ is the hazard of the event for individual i at time t, $h_0(t)$ is the baseline hazard at time t (i.e. the hazard for individuals with a value of zero for all covariates in the model), and $\boldsymbol{w}_i(t)$ is a vector of covariates, possibly time-dependent, with associated vector of population-level (i.e. fixed effects) parameters $\boldsymbol{\gamma}$. The parameter α is also a population-level (i.e. fixed effect) parameter termed the *association parameter* in joint modelling because it quantifies the magnitude of association between the longitudinal and event outcomes.

Association structure

The term $\alpha\mu_i(t)$ in equation (3) is commonly referred to as the *association structure*, since it induces an association between the longitudinal and event processes. Specifically, the association structure in equation (3) posits an association between the (log) hazard of the event at time t and the current expected value of the biomarker, also evaluated at time t; hence it is often referred to as a *current value association structure*.

It is noteworthy that the formulation in equation (3) assumes an association between the log hazard of the event and the *expected value* of the biomarker, rather than the *observed value* of the biomarker. This is important for two reasons. First, although the biomarker is observed intermittently at discrete observation times, the joint model formulation assumes it evolves in continuous time and therefore the *expected value* of the biomarker is defined at all possible values of t. Second, using the *expected value* ensures that the measurement error component of the observed biomarker measurements is excluded from the biomarker-event association. This is important since it has been shown that measurement error in an observed covariate in a proportional hazards model leads to bias towards the null for the estimated log hazard ratio (Prentice, 1982). Therefore, if the longitudinal submodel is appropriately specified, the joint model postulates a meaningful association between the hazard of the event and the underlying "true" error-free individual-specific value of the biomarker.

In contrast, if we were to include the biomarker as a time-varying covariate in a time-dependent Cox model then we must obtain an *observed value* of the biomarker for every individual at each unique event time. In practice, however, such observed values of the biomarker are generally not available and therefore the so-called "missing" biomarker

measurements must be imputed. Often the method of imputation used in the time-dependent Cox model, be it implicitly or explicitly, is the last observation carried forward; a method which has been shown to lead to biased inferences (Andersen and Liestøl, 2003).

Lastly, although the current value association structure has been the most common type of association structure appearing in the joint modelling literature, a variety of other (sometimes more complex) association structures are also possible and these will be discussed in greater detail in Section 2.3.4.

Likelihood

Under a set of assumptions that will be described in greater detail in Section 2.3.5, the likelihood function for a single individual i under the shared parameter joint model can be written as

$$L_i = \int \left(\prod_{j=1}^{n_i} p\left(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \right) p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i \qquad (4)$$

where the contribution to the likelihood from the event submodel is

$$p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = h_i(T_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})^{d_i} \exp\left(- \int_0^{T_i} h_i(s \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, ds \right) \qquad (5)$$

and where $p(.)$ denotes probability density functions, and $\boldsymbol{\theta}$ denotes the combined vector of all remaining unknown parameters in the model. Estimation of the model based on this likelihood will be discussed in Section 2.5.

2.3.2 Extensions to the longitudinal submodel

2.3.2.1 Multivariate generalised linear mixed models

With regard to the longitudinal submodel, two main extensions are of relevance to this thesis. The first extension is a generalisation of the longitudinal submodel so that it can accommodate different types of outcome data, for example, repeatedly measured binary outcomes or counts (Faucett et al., 1998; Li et al., 2009). The second extension is to allow for more than one longitudinal outcome (i.e. multiple biomarkers), commonly known as a multivariate joint model.

Both of these aforementioned extensions can be accommodated within a single joint modelling framework. This can be achieved by specifying the longitudinal submodel as a multivariate generalised linear mixed model, which includes the multivariate linear mixed model as a special case. Dependence between the multiple longitudinal outcomes can be captured by allowing for correlations between individual-specific parameters across the outcome models.

Assume $y_{ijm} = y_{im}(t_{ijm})$ corresponds to the value of the m^{th} longitudinal biomarker observed at the j^{th} $(j = 1, \dots, n_{ijm})$ time point, t_{ijm}, for the i^{th} $(i = 1, \dots, N)$ individual. A multivariate generalised linear mixed model assumes that each longitudinal biomarker can be modelled in continuous time where $Y_{im}(t)$ follows a distribution in the exponential family with mean $\mu_{im}(t)$ and linear predictor

$$\eta_{im}(t) = g_m\big(\mu_{im}(t)\big) = \boldsymbol{x}'_{im}(t)\boldsymbol{\beta}_m + \boldsymbol{z}'_{im}(t)\boldsymbol{b}_{im} \tag{6}$$

where $\boldsymbol{x}_{im}(t)$ and $\boldsymbol{z}_{im}(t)$ are vectors of covariates, possibly time-dependent, with associated vectors of population-level (i.e. fixed effects) parameters $\boldsymbol{\beta}_m$ and individual-specific (i.e. random effects) parameters $\boldsymbol{b}_{im}$, and $g_m(.)$ is some known link function.

The distributional family and link function are allowed to differ over the M longitudinal submodels. Specific choices of link function and distributional family lead to, for example, a logistic (logit link, Bernoulli family) or Poisson (log link, Poisson family) regression submodel. As previously mentioned, dependence between the different longitudinal biomarkers is captured via correlated individual-specific parameters. Specifically, a multivariate normal distribution for the individual-specific parameters can be assumed; that is

$$\begin{pmatrix} \boldsymbol{b}_{i1} \\ \vdots \\ \boldsymbol{b}_{iM} \end{pmatrix} = \boldsymbol{b}_i \sim N(0, \boldsymbol{\Sigma}_b) \tag{7}$$

for some unstructured variance-covariance matrix $\boldsymbol{\Sigma}_b$.

Assuming that the M different longitudinal outcomes are independent, conditional on the individual-specific parameters $\boldsymbol{b}_i$ (i.e. conditional independence), the joint distribution of the biomarkers at time t can be written as

$$p(y_{i1}(t), \dots, y_{iM}(t) \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \prod_{m=1}^{M} p(y_{im}(t) \mid \boldsymbol{b}_i, \boldsymbol{\theta}) \tag{8}$$

and accordingly, the likelihood function for the multivariate shared parameter joint model as

$$L_i = \int \left(\prod_{m=1}^{M} \prod_{j=1}^{n_{im}} p\left(y_{ijm} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \right) p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i \tag{9}$$

where the contribution to the likelihood from the event submodel remains the same as was specified in equation (5).

Due to its flexibility and relative computational ease, this specification has been the most common approach for handling multiple longitudinal biomarkers in the joint modelling literature to date (Hickey et al., 2016). However, assuming a correlated random effects structure is not the only possible approach to allow for dependence between multiple, correlated, longitudinal outcomes (Verbeke et al., 2014). An example of a different approach is a conditional specification of outcomes, as in the "product-normal" model (Spiegelhalter, 2002; Cooper et al., 2007). Another example is the use of copulas (Lambert and Vandenhende, 2002). However, these approaches will not be discussed further as they are less common and are not of direct relevance to this thesis.

The use of a multivariate longitudinal submodel also increases the potential complexity for the specification of the joint model association structure. Following on from Section 2.3.1, a natural association structure for the multivariate joint model would be to assume that the log hazard of the event at time t is linearly related to the current value of the linear predictor for each biomarker. That is, to assume an event submodel of the form

$$h_i(t) = h_0(t) \exp\left(w_i'(t)\gamma + \sum_{m=1}^{M} \alpha_m \eta_{im}(t) \right) \tag{10}$$

However, more complicated association structures are also possible and these will be discussed in Section 2.3.4.

In terms of extensions to the longitudinal submodel, this thesis will primarily concentrate on the multivariate generalised linear mixed model. However, a variety of other extensions

to the longitudinal submodel have been considered in the literature and these will be summarised briefly in the following sections.

2.3.2.2 Two-part models

A small number of studies have used joint modelling approaches with two-part models for the longitudinal outcome. Brilleman et al. (2016) (Brilleman et al., 2016) used a two-part "hurdle" model for the longitudinal response, consisting of both logistic and Gamma mixed effect regression model components. Their approach was motivated by longitudinal biomarker data (HIV RNA viral load) that was subject to a lower detection limit. Hatfield et al. (2012) (Hatfield et al., 2012) used a two-part "hurdle" model with logistic and Beta mixed effects regression model components to model repeatedly measured patient-reported outcomes. Their patient-reported outcomes were on a bounded scale between 0 and 100 and were therefore rescaled to follow the support of the Beta distribution. Their two-part model allowed for excess zeroes in the observed outcome measurements. Other authors have also considered two-part modelling approaches for the longitudinal submodel (Liu, 2009; Rizopoulos et al., 2008).

2.3.2.3 Mechanistic models

Another approach used for the longitudinal submodel has been to specify a non-linear mixed effects model or a set of non-linear ordinary differential equations. This approach has primarily been used in pharmacometric-based applications, where researchers can specify mechanistic models which they believe accurately mimic the underlying biological processes. Desmee et al. (2017b) used a set of non-linear ordinary differential equations to model changes in PSA during chemotherapy for metastatic prostate cancer and considered a number of different association structures. In further work (Desmée et al., 2017a) they discuss methods for generating dynamic predictions from these types of mechanistic joint models. Mbogning et al. (2015) proposed a joint model with a non-linear mixed effect longitudinal submodel and either a terminating, or recurrent, event submodel. Moreover, they described how researchers can estimate their models using the Monolix software (Lixoft SAS, 2018). The main disadvantage of these types of joint models is that in general they are computationally intensive. However, it has been suggested that the use of a stochastic approximation expectation–maximization (SAEM) algorithm may alleviate some of the computational burden (Mbogning et al., 2015).

2.3.3 Extensions to the event submodel

As previously discussed, the most common formulation of a shared parameter joint model has included a proportional hazards regression submodel for modelling the hazard of the event outcome. However, this standard formulation has been extended in a number of ways, including methods to handle cause-specific events (i.e. competing risks) (Li et al., 2009; Williamson et al., 2008), recurring events (Król et al., 2016), or non-proportional hazards. Similarly, some authors have proposed alternatives that do not require an assumption of proportional hazards at all; for example, through the use of additive hazards or an accelerate failure time (AFT) specification. These extensions are each discussed in the following subsections.

2.3.3.1 Competing risks

A competing risks event submodel with a current value association structure can be specified as

$$h_{ik}(t) = h_{0k}(t) \exp\left(\boldsymbol{w}'_{ik}(t)\boldsymbol{\gamma}_k + \sum_{m=1}^{M} \alpha_{mk}\eta_{im}(t) \right) \tag{11}$$

where $h_{ik}(t)$ is the cause-specific hazard for individual i at time t for event type k ($k = 1, \dots, K$), $h_{0k}(t)$ is the cause-specific baseline hazard at time t for event type k, $\boldsymbol{w}_{ik}(t)$ is a vector of covariates, possibly time-dependent, for individual i and assumed to be related to event type k through the vector of population-level parameters $\boldsymbol{\gamma}_k$, and $\eta_{im}(t)$ is the linear predictor for the m^{th} longitudinal outcome evaluated for individual i at time t.

The model in equation (11) posits a cause-specific association between the linear predictor from the longitudinal submodel and the log hazard of each event type. The estimated cause-specific association parameter, α_{mk}, is interpreted in the same way as in the model without competing risks. That is, given that individual i is still alive at time t, then α_{mk} quantifies the association between a one unit increase in the linear predictor for the m^{th} biomarker and the log cause-specific hazard of event type k (assuming all other covariates in the model are held constant).

To accommodate the cause-specific hazard functions in the estimation of the model the likelihood for the event submodel is written as

$$p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \prod_{k=1}^{K} h_{ik}(T_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})^{d_{ik}} \exp\left(-\int_0^{T_i} h_{ik}(s \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, ds\right) \quad (12)$$

where d_{ik} takes the value 1 if individual i experienced event type k, and 0 otherwise.

However, under a competing risks model the one-to-one correspondence between the cause-specific hazard rate for an event type and the cause-specific risk of the event no longer holds (Andersen et al., 2012). That is, in the absence competing risks, the risk of the event occurring for individual i up to time t is given by

$$F_i(t) = 1 - S_i(t) = 1 - P(T_i^* \geq t) = 1 - \exp(-H_i(t)) \quad (13)$$

where $S_i(t)$ is the probability of individual i being event free at time t (i.e. the survival probability), and $H_i(t) = \int_{s=0}^{t} h_i(s)ds$ is the cumulative hazard for individual i at time t. In the presence of the competing events, the risk of individual i experiencing any event prior to time t depends on the cause-specific hazard rates for all $k = 1, \ldots, K$ competing events and is given by

$$F_i(t) = 1 - S_i(t) = 1 - \exp\left(-\sum_{k=1}^{K} H_{ik}(t)\right) \quad (14)$$

where $H_{ik}(t) = \int_{s=0}^{t} h_{ik}(s)ds$. One is often also interested in the risk of individual i specifically experiencing event type k prior to time t, often called the cause-specific cumulative incidence (Koller et al., 2012). This is defined as

$$F_{ik}(t) = 1 - \int_0^t S_i(s)\, h_{ik}(s)ds \quad (15)$$

This shows that the cause-specific cumulative incidence for event type k at time t is a function of the cause-specific hazard rates for all $k = 1, \ldots, K$ competing events. This means that, for example, a covariate with a positive association with the hazard rate for event type k will not necessarily be associated with an increased risk (i.e. probability) of that event occurring; rather, it will also depend on the association between that covariate and the hazard rates for each of the $K - 1$ other event types.

A consequence of the lack of a one-to-one correspondence between the cause-specific hazard function (rate) and the cause-specific cumulative incidence (risk) is that inferences obtained from a competing risks time-to-event model might differ depending on which

quantity one chooses to examine. More specifically, the cause-specific hazard functions and cause-specific cumulative incidence functions relate to different aspects of the association between the covariate and the time-to-event outcome, and each is more (or less) appropriate for answering a given type of research question. Cause-specific hazard functions are generally considered more appropriate for research questions related to the potential for etiological associations, whereas the cause-specific cumulative incidence functions are generally considered more appropriate for prognostic modelling (i.e. risk prediction). These aspects of the competing risks model are discussed further in the paper presented in Chapter 4 of this thesis, as well as in Andersen et al. (2012) and Koller et al. (2012).

Competing risks have been considered within a joint modelling framework by a number of authors. Lu (2017) proposed a competing risks joint model which they used to analyse CD4 cell counts and their association with cause-specific mortality (human immunodeficiency virus (HIV) related deaths versus other deaths) in HIV patients. They included a number of extensions to the model to allow for longitudinal measurements that were both skewed and subject to a lower detection limit. Andrinopoulou et al. (2014) proposed a joint model for bivariate longitudinal data (i.e. two repeatedly measured biomarkers) and competing risks data. They used their model to analyse the associations between aortic valve performance and the risk or death or reoperation in patients who had undergone an aortic valve replacement. Other authors have also considered competing risks joint models (Huang et al., 2010; Williamson et al., 2008). In the application presented in Chapter 4 of this thesis, a competing risks joint model will be used to explore the association between longitudinal changes in BMI and the rates of two competing event types, kidney transplantation and death without transplantation, each occurring within a haemodialysis population.

With regard to user-friendly software implementations of the competing risks joint model there are only a small number of examples. The **JM** (Rizopoulos, 2010) R package accommodates competing risks under a shared parameter joint modelling framework, whilst the **lcmm** (Proust-Lima et al., 2017) R package accommodates competing risks under a latent class joint modelling framework (the latent class joint modelling framework is discussed in Section 2.4).

2.3.3.2 Recurrent events

Recurrent events are defined as events which can occur more than once for a single individual (Amorim and Cai, 2015). These event types have also been considered in a joint longitudinal and time-to-event modelling framework. An example of a recurrent event process is the occurrence of new tumour lesions in patients with non-small cell lung cancer (NSCLC). Under a standard joint modelling framework, one may be interested in the association between a repeatedly measured clinical biomarker, such as ctDNA, and the time from a defined baseline, such as initiation of treatment, until a terminating event such as death. However, for each patient we may also observe the time from baseline until the appearance of each new tumour lesion that occurs prior to death or censoring. Under an extended joint model we can allow for the associations between: (i) ctDNA and the risk of death; (ii) ctDNA and the rate at which new lesions appear; and (iii) the rate at which new lesions appear and the risk of death.

A joint model for a longitudinal biomarker, a recurrent event, and a terminating event can be formulated as follows. The observed timing of the r^{th} occurrence ($r = 1, \dots, R_i$) of the recurring event can be denoted $T_{ir} = \min(T_{ir}^*, T_i^*, C_i)$ where T_{ir}^* denotes the so-called "true" event time for the r^{th} occurrence of the recurring event, which may not be observed due to right-censoring (i.e. $C_i < T_{ir}^*$) or the prior occurrence of the terminating event (i.e. $T_i^* < T_{ir}^*$). An event indicator for the r^{th} occurrence of the recurrent event can be defined as $d_{ir} = I(T_{ir} = T_{ir}^*)$, whilst the event indicator for the terminating event remains the same as previously defined, i.e. $d_i = I(T_i^* \leq C_i)$.

The most common shared parameter joint modelling approach for accommodating the recurrent event process postulates a separate proportional hazards regression for each of the recurrent and terminating events (defined below). Moreover, to allow for a possible dependence between the terminating and recurrent event processes themselves, it is common to introduce a shared individual-specific parameter (i.e. random effect), u_i, commonly known as a shared frailty term. Therefore, to accommodate the recurrent event within a shared parameter joint modelling framework the event submodel can be redefined as the following set of proportional hazard regression equations

$$\begin{cases} h_{ir}(t) = h_{0r}(t)\exp\left(\boldsymbol{w}_{ir}'(t)\boldsymbol{\gamma}_r + \sum_{m=1}^{M}\alpha_{mr}\eta_{im}(t) + u_i\right) \\ h_i(t) = h_0(t)\exp\left(\boldsymbol{w}_i'(t)\boldsymbol{\gamma} + \sum_{m=1}^{M}\alpha_m\eta_{im}(t) + \delta u_i\right) \end{cases} \quad (16)$$

where $h_{ir}(t)$ is the hazard of the r^{th} occurrence of the recurrent event for individual i at time t, $h_{0r}(t)$ is the baseline hazard for the r^{th} occurrence of the recurrent event at time t, $\boldsymbol{w}_{ir}(t)$ is a vector of covariates with associated vector of population-level (i.e. fixed effects) parameters $\boldsymbol{\gamma}_r$, and α_{mr} is the population-level association parameter for the m^{th} biomarker. The vector of covariates included in the regression equation for the recurrent event process, $\boldsymbol{w}_{ir}(t)$, may or may not coincide with the vector of covariates included in the regression equation for the terminating event model, $\boldsymbol{w}_i(t)$.

The frailty parameters u_i are scaled in the terminating event regression by the population-level parameter δ to provide a flexible dependence between the recurrent and terminating event processes. For example, if $\delta = 0$ then there is no dependence between the hazard rates for the terminating and recurrent events, however, each can still be associated with the longitudinal biomarker(s). An additional consideration is the distribution given to u_i which might be, for example, normal or log-Gamma. Moreover, one may or may not wish to allow for correlation between u_i and the individual-specific parameters from the longitudinal submodel, $\boldsymbol{b}_i$. This choice is likely to relate to the context of the application and considerations related to computational complexity (including the amount of data available with which to estimate the parameters in the model).

The likelihood for the shared parameter joint model defined by equations (6), (7) and (16) is given by

$$L_i = \int\int \left(\prod_{m=1}^{M}\prod_{j=1}^{n_{im}} p\left(y_{ijm} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right)\right) p(T_i, d_i \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta}) \left(\prod_{r=1}^{R_i} p(T_{ir}, d_{ir} \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta})\right) p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, p(u_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i\, du_i \quad (17)$$

where the contribution from the r^{th} occurrence of the recurrent event is

$$p(T_{ir}, d_{ir} \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta}) = h_{ir}(T_{ir} \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta})^{d_{ir}} \exp\left(- \int_{T_{i(r-1)}}^{T_{ir}} h_{ir}(s \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta})\, ds \right) \quad (18)$$

with $T_{i(r-1)}$ set equal to 0 (i.e. baseline) when $r = 1$. This formulation assumes that the timescale for the recurrent event continues regardless of whether a recurrent event occurs or not, sometimes known as "calendar time". However, an alternative specification uses the so-called "gap time" approach in which the timescale for the recurrent event is reset to zero each time a recurrent event occurs. Under the gap time approach, the likelihood contribution from the r^{th} occurrence of the recurrent event is defined as

$$p(T_{ir}, d_{ir} \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta}) = h_{ir}\left(T_{ir} - T_{i(r-1)} \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta}\right)^{d_{ir}} \exp\left(- \int_{0}^{T_{ir} - T_{i(r-1)}} h_{ir}(s \mid \boldsymbol{b}_i, u_i, \boldsymbol{\theta})\, ds \right) \quad (19)$$

Joint models accommodating both recurrent and terminal events have been discussed by only a small number of authors. Krol et al. (2016) used such a model to identify the associations between tumour size, the appearance of new lesions, and the risk of death, in patients with colorectal cancer. In addition, their model allowed for left censoring in the longitudinal biomarker, since the tumour size measurements were subject to a lower detection limit. The authors also discussed how dynamic predictions can be obtained and assessed using their model. Liu and Huang (2009) used such a model to explore the associations between CD4 cell counts, the appearance of opportunistic diseases, and the risk of death, in HIV patients. However, their model included a time-fixed association structure between the longitudinal and each event process that was based on shared random effects, rather than a time-dependent association structure based on the current value of the biomarker. Kim et al. (2012) proposed a similar model, for a longitudinal outcome, recurrent event, and terminating event, but modelled the event outcomes using a flexible transformation of the cumulative intensities, rather than modelling them directly on the hazard scale. Some authors have also used a recurrent event submodel to relax the common assumption of an uninformative measurement schedule for the longitudinal outcome (this assumption is discussed in greater detail in Section 2.3.5) (Liu et al., 2008).

There are currently very few user-friendly software implementations of the recurrent event joint model. A notable exception is the **frailtypack** (Król et al., 2017; Rondeau et al., 2012)

R package, which allows the user to estimate a joint model consisting of a longitudinal biomarker, a recurrent event, and a terminating event. Although the implementation of the model in **frailtypack** is relatively flexible, it is also limited in some regards. For example, the user may choose between different options for the baseline hazard, or can choose between a gap time or calendar time scale. However, the vector $\boldsymbol{b}_i$ can be a maximum length of two, only a small number of association structures are accommodated, and one must assume the longitudinal biomarker is normally distributed.

2.3.3.3 Alternatives for proportional hazards

In a number of situations the assumption of proportional hazards may not be valid, thereby leading to potentially biased inferences. One approach to deal with such situations is to allow the relevant hazard ratios in the event submodel (i.e. those where the proportionality assumption is not valid) to be a function of time; this leads to what is commonly referred to as *non-proportional hazards* or *time-dependent effects*.

Although non-proportional hazards have been widely discussed in the survival analysis literature, they have not been widely discussed in the context of joint modelling. However, it is noteworthy that the most recent release of the **JMbayes** (Rizopoulos, 2016) R package (version 0.8-70) does allow for a time-dependent association parameter. Examples of this functionality are provided in the package vignette (Rizopoulos, 2017). Similarly, the **megenreg** (Crowther, 2017a, 2017c) Stata package allows the user to specify a joint model that incorporates a time-dependent association parameter. A tutorial showing this functionality is available on the **megenreg** author's website (Crowther, 2017b). These implementations provide increased flexibility for researchers by relaxing the assumption that the association between the biomarker and the hazard of the event must be constant over time.

However, another approach to dealing with non-proportional hazards is to specify an alternative formulation for the event submodel entirely, one that does not assume covariates have a multiplicative effect on the hazard of the event. Such alternatives include the use of additive hazards or AFT models.

Additive hazards regression models assume that covariates have an additive effect on the hazard, rather than the multiplicative effect seen in the proportional hazards regression

model. Accordingly, an additive hazards event submodel with a current value association structure can be specified as

$$h_i(t) = h_0(t) + w_i'(t)\gamma + \sum_{m=1}^{M} \alpha_m \eta_{im}(t) \qquad (20)$$

Here, the association parameters α_m $(m = 1, \dots, M)$ correspond to absolute differences on the hazard scale. In contrast, the association parameters under the proportional hazards event submodel in equation (3) correspond to absolute differences on the *log* hazard scale, which are more easily interpreted as relative differences on the hazard scale (i.e. hazard ratios, obtained by taking the exponential of α_m). In some settings the estimation of absolute (rather than relative) differences may be more appropriate and therefore an additive hazards model may be preferable. However, there are very few examples of joint modelling approaches that have incorporated an additive hazards formulation. One exception is Moreno-Betancur et al. (2017), who incorporated an additive hazards event submodel within an extended two-stage joint modelling approach based on multiple imputation.

AFT models are another alternative, whereby covariates have a multiplicative effect on the time-to-event outcome itself, rather than a multiplicative effect on the hazard of the event. Although several authors have considered an AFT event submodel under a joint modelling framework, they have not proven as popular as the proportional hazards modelling approach. A primary reason for this is likely to be that the introduction of time-varying covariates into the AFT model creates complications for interpretation of the parameters. It is worth noting, however, that a method for circumventing issues around the introduction of a time-varying covariate into the AFT event submodel is to use a time-independent association structure based on shared random effects; see, for example, Wu et al. (2010). Since AFT models are not considered any further as part of this thesis, the reader is referred to Tseng et al. (2005) and Rizopoulos (2012b) for a discussion of these issues.

2.3.4 Extensions to the association structure

2.3.4.1 Specification

As mentioned previously, the dependence between the longitudinal and event submodels is captured through the *association structure* of the joint model, which can be specified in a

number of ways. So far, the only association structure that has been explicitly formalised in this thesis is the so-called *current value association structure*. Recall that for a multivariate joint model with a current value association structure the event submodel can be specified as

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w}_i'(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \alpha_m \eta_{im}(t) \right) \tag{21}$$

Here, it is the $\sum_{m=1}^{M} \alpha_m \eta_{im}(t)$ term that is referred to as the *association structure*, since it is this term that captures the assumed dependence between the longitudinal and event outcomes. More specifically, under this formulation, dependence between the event and the m^{th} longitudinal outcome is assumed to be captured through a linear association between the log hazard of the event at time t and the current value of the linear predictor of the m^{th} longitudinal submodel at time t.

In a situation where the longitudinal submodel is a linear mixed model (i.e. an identity link function and normal error distribution) the current value association structure can be viewed as a method for including the underlying "true" value of the biomarker as a time-varying covariate in the event submodel. The term "true" is used here to refer to a value of the biomarker which is not subject to measurement error or discrete-time observation. Of course, for the expected value from the longitudinal submodel to be considered the so-called "true" underlying biomarker value, the longitudinal submodel would need to be correctly specified, which may not be the case in practice. However, it has been shown that a joint modelling approach with a current value association can provide relatively unbiased estimates of the association parameter even under (specific forms of) misspecification of the longitudinal submodel (Crowther et al., 2016).

Although the current value association structure is the most widely used association structure in the joint modelling literature to date, there are a variety of other association structures that can be specified. These other structures have been reviewed and discussed by other authors (Hickey et al., 2016; Rizopoulos and Ghosh, 2011; Lawrence Gould et al., 2015; Crowther, Lambert, et al., 2013; Mauff et al., 2017).

To introduce a general form for the association structure, the multivariate shared parameter joint model can be specified as

$$h_i(t) = h_0(t)\exp(w_i'(t)\gamma + \sum_{m=1}^{M}\sum_{q=1}^{Q_m}\alpha_{mq}f_{mq}(\boldsymbol{\beta}, \boldsymbol{b}_i; t, u)) \quad (22)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ is the combined vector of fixed effect parameters across all M longitudinal submodels, $f_{mq}(.)$ is a known function defining the q^{th} $(q = 1, \ldots, Q_m)$ association structure term related to the m^{th} $(m = 1, \ldots, M)$ longitudinal submodel, Q_m is the total number of association structure terms related to the m^{th} longitudinal submodel, and u is a potential lag time used in the association structure.

The functions $f_{mq}(.)$ determine the association structure for the multivariate joint model and must be specified. Three example association structures are defined as follows. First, suppose that $Q_m = 1$ for all m; that is, for each of the $1, \ldots, M$ longitudinal biomarkers there is a single term in the association structure linking that biomarker to the log hazard of the event. If $f_{m1}(\boldsymbol{\beta}, \boldsymbol{b}_i; t, u) = \eta_{im}(t)$ for all m then equation (22) reduces to the *current value* association structure that has been discussed earlier in this chapter (i.e. an association structure based on the current value of the linear predictor from the longitudinal submodels). Secondly, if $f_{m1}(\boldsymbol{\beta}, \boldsymbol{b}_i; t, u) = \eta_{im}(t-u)$ for all m then equation (22) reduces to a *lagged value* association structure; the log hazard of the event at time t is assumed to be linearly associated with the value of the linear predictor for each longitudinal submodel evaluated at time $t-u$. Thirdly, if $f_{m1}(\boldsymbol{\beta}, \boldsymbol{b}_i; t, u) = \frac{d\eta_{im}(t)}{dt}$ for all m then equation (22) reduces to a *current slope* association structure; the log hazard of the event at time t is assumed to be linearly associated with the rate of change (i.e. slope) of the linear predictor for each longitudinal submodel evaluated at time t.

Table 2 provides a relatively comprehensive list of possible association structures for shared parameter joint models that goes beyond the three examples given above. In addition to those association structures presented in Table 2, it is also possible to define a shared correlation structure for the individual-level random effects in the longitudinal submodel(s) and a frailty term in the event submodel. That is, suppose the event submodel takes the form

$$h_i(t) = h_0(t)\exp(\boldsymbol{w}_i'(t)\boldsymbol{\gamma} + u_i) \quad (23)$$

and then define $\begin{pmatrix} \boldsymbol{b}_i \\ u_i \end{pmatrix} \sim F_{\boldsymbol{\Sigma}}$ where $F_{\boldsymbol{\Sigma}}$ is a multivariate distribution allowing for dependence between the individual-level random effects from the longitudinal submodel(s), $\boldsymbol{b}_i$, and the individual-level frailty term in the event submodel, u_i, through some variance-covariance matrix $\boldsymbol{\Sigma}$.

Lastly, the association structure can be defined based on latent classes, whereby it is not *between-individual* differences in the biomarker that are associated with differences in event risk, but rather it is *between-class* differences in the biomarker that are associated with differences in event risk. Joint models with a latent class association structure will be discussed in detail in Section 2.4.

An extensive review of possible association structures is provided by Hickey et al. (2016). In addition to describing the aforementioned association structures in detail, they also cite studies that have used or discussed each of the association structures. Although the majority of these association structures have been discussed in various places throughout the literature, most joint modelling software packages only make available a limited set. The joint modelling functionality in the **rstanarm** (Brilleman et al., 2018; Stan Development Team, 2017a) R package, which will be introduced in Chapter 5 of this thesis, aims to implement the full set of association structures defined in Table 2, with the exception of weighted cumulative effects.

2.3.4.2 Comparison and selection

The choice of association structure should ideally be driven by clinical context and the study's main research question. In the joint modelling literature there are several examples of associations structures that were used to help answer specific clinical or epidemiological research questions. Crowther et al. (2013) showed how a joint modelling approach with a

Table 2. Forms for the association structure of a shared parameter joint model.

Association structure	Specification[1] for $f_{mq}(\boldsymbol{\beta}, \boldsymbol{b}_i; t, u)$	Additional comments
Current value	$\eta_{im}(t)$	
Lagged value	$\eta_{im}(t-u)_+$	For some specified lag time u, and with $(t-u)_+ = t-u$ when $t > u$ and 0 otherwise.
Current slope	$\frac{d^s \eta_{im}(t)}{dt^s}$	The derivative here is generalised through the index s (for some $s \geq 0$), however, in most instances $s = 1$. If $s = 0$, then the association structure simplifies to the current value.
Lagged slope	$\frac{d^s \eta_{im}(t-u)_+}{dt^s}$	For some specified lag time u, and with $(t-u)_+ = t-u$ when $t > u$ and 0 otherwise.
Cumulative effects	$\int_{s=0}^{t} \eta_{im}(s)ds$	
Lagged cumulative effects	$\int_{s=0}^{(t-u)_+} \eta_{im}(s)ds$	For some specified lag time u, and with $(t-u)_+ = t-u$ when $t > u$ and 0 otherwise.
Weighted cumulative effects	$\int_{s=0}^{t} \omega(t-s)_+ \eta_{im}(s)ds$	Here, $\omega(.)$ is some known weight function, with $(t-s)_+ = t-s$ when $t > s$ and 0 otherwise.
Shared random effects	b_{im}^s	Where b_{im}^s is the s^{th} element of the $\boldsymbol{b}_{im}$ vector, corresponding to the desired random effect that is to be used in the association structure (e.g. the random intercept parameter for the i^{th} individual).

Shared random and fixed effects	$b_{im}^{s} + g\left(\boldsymbol{\beta}_{m}^{(s)}\right)$	Where b_{im}^{s} is the same as defined for the shared random effects association structure, $\boldsymbol{\beta}_{m}^{(s)}$ is the subset of parameters from the vector of fixed effects, $\boldsymbol{\beta}_{m}$, that correspond to the random effect b_{im}^{s} in the association structure, and $g(.)$ is a specified function. In most cases, $\boldsymbol{\beta}_{m}^{(s)}$ will be a single parameter (e.g. the population-level intercept) and $g(.)$ will be the identity function; however, it is possible for $\boldsymbol{\beta}_{m}^{(s)}$ to be a vector with more than one element, in which case $g(.)$ is commonly a function returning the summation across the elements of $\boldsymbol{\beta}_{m}^{(s)}$.
Interactions with observed data	$\eta_{im}(t)c_i(t)$	Where $c_i(t)$ is an observed covariate, which may be either time-fixed or time-varying.[2]
Interactions between biomarkers	$\eta_{im}(t)\eta_{im'}(t)$	For some $m \neq m'$.

[1] In the specification of any of these association structures, one could replace the linear predictor for the m^{th} longitudinal submodel, $\eta_{im}(t)$, with the expected value $\mu_{im}(t)$.
[2] For example, suppose $c_i(t) = c_i$ takes the value 1 if individual i is female and 0 otherwise, and that the event submodel is specified as $h_i(t) = h_0(t)\exp(w_i'(t)\gamma + \alpha_{m1}\eta_{im}(t) + \alpha_{m2}\eta_{im}(t)c_i)$ then this would allow the effect of biomarker m on the hazard of the event to differ by gender.

shared random effects association structure can improve the use of baseline systolic blood pressure in prognostic models for cardiovascular risk. Their shared random effects association structure only included the individual-specific random intercept. This allowed the authors to use information from repeated SBP measurements to improve their estimate of an error-free baseline measure of SBP, thereby leading to a less biased estimate of the association between baseline SBP and cardiovascular risk. Mauff et al. (2017) described a novel association structure based on *recency-weighted cumulative effects* of the biomarker. They formulated a joint model in which cumulative HbA1c levels were associated with the risk of sight threatening retinopathy amongst type 2 diabetes patients. Through the use of a parametric weight function for the cumulative effects of the biomarker, they specified that more recent HbA1c levels would be more strongly associated (i.e. given a larger weighting) with the risk of the event. By estimating the parameters of the weight function they were able to describe the relative temporal importance of HbA1c levels.

In some situations, however, it is not possible to determine the most appropriate association structure based on clinical context alone. For example, if the aim of a study is to develop a risk prediction model, then researchers may wish to consider a variety of association structures and choose the best performing joint model(s) based on some objective criteria. In this setting, one may wish to compare models using an information criterion or a measure of predictive accuracy.

Zhang et al. (2014) proposed a decomposition of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for joint models. The benefit of their decomposition is that it allows researchers to compare several joint models based on the fit of either the longitudinal submodel component, the event submodel component, or both. However, if the study objective is to develop a prognostic model, then AIC and BIC are known to be suboptimal. Commenges et al. (2012) therefore proposed an information criterion based on expected prognostic cross-entropy, adapted to joint models, and aimed at assessing predictive accuracy.

Another alternative to using an information criterion is to instead directly assess the mean error in the event submodel predictions (i.e. calibration) using some form of scoring rule. In the context of joint models, measures for the mean error in predicted survival probabilities have been proposed by Henderson et al. (2002) and Schoop et al. (2008). These two authors differed primarily in the choice of loss function for the scoring rule, use

of an absolute error versus squared error, with the latter resulting in the so-called Brier score.

Another common approach to assessing the event submodel predictions from a joint model is to focus on discrimination. That is, the ability of the model to discriminate between those individuals who do, and do not, experience the event. The most common measure of discrimination is the area under the receiver operating characteristic curve (AUC), which has been adapted to the dynamic nature of survival predictions from joint models (Rizopoulos, 2011). In addition, some authors have extended such measures of predictive error and discrimination to non-standard joint models, such as those with competing risks (Blanche et al., 2015). It is also worth noting that the aforementioned measures of predictive accuracy can be applied in the context of cross-validation, using either approximate cross-validation measures (for example, approximate leave-one-out cross-validation for the expected prognostic cross-entropy measure discussed by Commenges et al. (2012)) or explicit K-fold cross validation.

Lastly, if it is not necessary to derive a single final model, but simply to derive a prognostic tool for generating dynamic survival predictions, then model averaging is also possible. For example, Rizopoulos et al. (2014) proposed a Bayesian model averaging approach for joint models with model weights based on the posterior probability of a given model $\mathcal{M}_k$ being the true model from a given set of models $\mathcal{M}_1, \dots, \mathcal{M}_K$ conditional on the observed data. This approach avoids the need to choose a single joint model with the most appropriate association structure, and instead averages across the survival predictions from two or more joint models with different association structures, with greater weight provided to those models with a larger posterior probability.

2.3.5 Assumptions

2.3.5.1 Conditional independence

Here we define a set of assumptions for the multivariate shared parameter joint model. The so-called *conditional independence* assumption of the shared parameter joint model postulates

$$y_{im}(t) \perp T_i^* \mid \boldsymbol{b}_i, \boldsymbol{\theta} \tag{24}$$

$$y_{im}(t) \perp y_{im}(t') \mid \boldsymbol{b}_i, \boldsymbol{\theta}$$

$$y_{im}(t) \perp y_{im'}(t) \mid \boldsymbol{b}_i, \boldsymbol{\theta}$$

for some $t \neq t'$ and $m \neq m'$, and where $\boldsymbol{\theta}$ is the combined vector of all remaining (fixed effect) parameters in the model. That is, conditional on the individual-specific parameters $\boldsymbol{b}_i$ and population-level parameters $\boldsymbol{\theta}$, the following are assumed: (i) any biomarker measurement for individual i is independent of that individual's true event time T_i^*; (ii) any two measurements of the m^{th} biomarker taken on the i^{th} individual at two distinct time points t and t' (i.e. longitudinal or repeated measurements) are independent of one another; and (iii) any two measurements of two different biomarkers, taken on the i^{th} individual at some time point t are independent of one another. It is worth noting however that the assumption of independence between longitudinal measurements taken at different time points conditional on $\boldsymbol{b}_i$, is also a standard assumption of linear mixed models and not unique to joint models.

For the univariate shared parameter joint model (i.e. one longitudinal outcome, such that $M = 1$) we obtain a slightly reduced set of conditional independence assumptions that do not involve the index m; that is

$$\begin{aligned} y_i(t) &\perp T_i^* \mid \boldsymbol{b}_i, \boldsymbol{\theta} \\ y_i(t) &\perp y_i(t') \mid \boldsymbol{b}_i, \boldsymbol{\theta} \end{aligned} \tag{25}$$

The key benefit of these conditional independence assumptions is that they entail a convenient factorisation of the likelihood function for the full joint model. That is, the joint distribution of the longitudinal and time-to-event data conditional on the individual-specific parameters is the product of the separate conditional distributions for each component. More specifically, with the conditional independence assumption in place, the likelihood for the i^{th} individual under the multivariate shared parameter joint model can be written as

$$L_i = \int \left(\prod_{m=1}^{M} \prod_{j=1}^{n_{im}} p\left(y_{ijm} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \right) p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i \tag{26}$$

Importantly, this factorisation of the full likelihood helps facilitate estimation of the joint model.

2.3.5.2 Censoring and visiting processes

Moreover, we require that, conditional on baseline covariates and observed longitudinal biomarker data for individual i, the

(i) censoring process for the event outcome, and

(ii) visiting process by which the observation times t_{ij} (for $j = 1, \ldots, n_i$) are determined

are both independent of the true event time T_i^* and all missing or future unobserved longitudinal biomarker measurements.

The assumption related to the *censoring process* is a relatively standard assumption required for most survival modelling approaches. Whilst the assumption related to the *visiting process* (i.e. the measurement schedule of the longitudinal outcome), has been relatively standard within the joint modelling literature. However, some authors have proposed methods to accommodate an informative visiting process; see for example Liu at el. (2008) and Han et al. (2014). Moreover, Rizopoulos et al. (2015) proposed an approach under which the joint model can be used to optimise the timing of future longitudinal measurements for a patient, conditional on their set of biomarker measurements observed prior to the current time.

2.3.6 Extensions to the clustering structure

A common feature of the shared parameter joint models described thus far has been that they consist of a two-level hierarchical structure. That is, the longitudinal response is assumed to be observed at time points (level 1) which are clustered within individuals (level 2). The data structure therefore consists of two levels which are defined based on a single clustering factor, the individual.

However, it is common in the health research setting for studies to give rise to observed data that contains more than one clustering factor. One example is a situation in which biomarker measurements are taken at time points (level 1) for patients (level 2) clustered within clinics (level 3). Another example is tumour size measurements taken at time points (level 1) for multiple tumour lesions (level 2) clustered within patients (level 3). Although these types of multilevel clustered data are common, they have not been discussed in the joint modelling literature.

One relevant exception is the meta-analysis of joint longitudinal and time-to-event data, which involves longitudinal measurements taken at time points (level 1) for patients (level 2) clustered within studies (level 3). In a meta-analysis of either individual patient data or aggregate data, one may wish to allow for these two clustering factors (i.e. the individual and the study). Sudell et al. (2017) describe analysis methods for this type of data structure. However, their method is based on a two-stage meta-analytic approach that pools the study-specific parameter estimates in the second stage, rather than directly specifying a single model for the three-level hierarchical structure of the individual patient data.

In Chapter 6 of this thesis, a methodological framework will be presented for the joint analysis of longitudinal and time-to-event data in the presence of more than one clustering factor. The methods are motivated by an application in clinical oncology, whereby repeated measurements are taken on tumour lesions clustered within non-small cell lung cancer patients. However, the methods that are proposed can be more widely applied to a range of clinical and epidemiological contexts.

2.3.7 Baseline hazards

One other point of consideration in joint models is the specification of the baseline hazard in the event submodel. The most common approach in standard survival analysis has been to use the Cox proportional hazards model, in which the baseline hazard $h_0(t)$ is left unspecified. The main advantage of leaving the baseline hazard unspecified is that the researcher is no longer required to assume and specify some parametric form for the baseline hazard which, in some cases, may be overly restrictive. For instance, the widely used Weibull proportional hazards model enforces a monotonicity constraint on the baseline hazard function, which may not be realistic in some settings.

However, there are two main disadvantages with leaving the baseline hazard unspecified in the context of joint modelling of longitudinal and time-to-event data. The first is that a common focus of joint modelling is to estimate future survival probabilities for individuals, whether they be individuals used in the estimation of the model (i.e. in sample) or new individuals for whom we may collect covariate and biomarker data in the future (i.e. out of sample). Leaving the baseline hazard unspecified means that survival probabilities cannot be directly calculated. Although approaches exist for deriving a non-parametric estimate of the baseline hazard, such approaches may not be desirable since they do not provide a smooth function for the baseline hazard and may be noisy in areas where there are few

events (e.g. the extremities of the observed time frame). The second reason that leaving the baseline hazard unspecified in joint modelling may be avoided is that it has been shown to result in an underestimate of the standard errors for the association parameter (Hsieh et al., 2006). A consequence of this is that an alternative method, such as bootstrapping, should ideally be used to obtain correct standard errors. However, such approaches can be computationally intensive.

An alternative approach, which avoids leaving the baseline hazard unspecified and also avoids the need to assume an overly simplistic parametric form for the baseline hazard, is to approximate the baseline hazard using a flexible non-linear function. A common choice has been to use some form of spline function, for example, restricted cubic splines (Crowther et al., 2012), B-splines (Rizopoulos, 2016), or M-splines (Proust-Lima et al., 2017). As long as the knot locations and/or degrees of freedom are chosen appropriately, then cubic splines provide a flexible and relatively parsimonious way to approximate the underlying baseline hazard function. In addition, some authors have added a penalty to the spline function to avoid overfitting and reduce sensitivity to the choice of knot locations or degrees of freedom (Rizopoulos, 2016). Nonetheless, it has been suggested that results from flexible parametric survival models are generally quite robust to the choice of knot locations or degrees of freedom (Rutherford et al., 2015).

Throughout this thesis, parametric forms will be used for the baseline hazard. This is because the software packages that will be used for fitting the models are all implemented with parametric forms for the baseline hazard. In some situations, this will be a simple parametric form, as for a Weibull distributed time-to-event outcome, and in other situations it will be a more flexible spline-based function.

2.4 Latent class joint models

In the previous section, the focus was on the so-called *shared parameter joint model*. Under a shared parameter joint modelling approach, it is assumed that there is a single homogeneous population. Within that population, differences in event risk are attributable to between-individual differences in some aspect(s) of the underlying longitudinal biomarker histories in addition to covariates.

A second, and alternative, modelling approach that has been proposed in the joint modelling literature is commonly known as a *latent class joint model* (Garre et al., 2007; Lin et al.,

2002; Proust-Lima et al., 2014). In contrast to the shared parameter joint model, the latent class joint model assumes that there are multiple (i.e. more than one) heterogeneous groups within our population and that it is differences in the marginal (average) longitudinal profiles within each of these groups that explain the observed differences in event risk. That is, differences in event risk are not explained by *between-individual* differences in the longitudinal biomarker histories, but rather they are explained by the *between-class* differences in the longitudinal biomarker histories. The heterogeneous groups that are believed to exist are not themselves directly observable (i.e. we do not know which group an individual belongs to). These groups are therefore referred to as "latent classes", to reflect the fact they are unobservable, and the probability of each individual i belonging to each of the G latent classes can be modelled. The latent class joint model therefore consists of at least three regression submodels which can be defined as follows.

Class membership submodel

It is assumed that individual i $(i = 1, \dots, N)$ belongs to one of the $g = 1, \dots, G$ latent classes. Let the random variable c_i denote the latent class to which individual i belongs. The probabilities of individual i belonging to each of the possible latent classes can be modelled through a multinomial logistic regression model

$$P(c_i = g) = \frac{\exp(v_i' \xi_g)}{\sum_g \exp(v_i' \xi_g)} \tag{27}$$

where $P(c_i = g)$ denotes the probability that individual i belongs to latent class g, and v_i is a vector of time-fixed covariates with an associated vector of class-specific population-level (i.e. fixed effects) parameters ξ_g. For identifiability a reference class must be specified. This can be achieved, for example, by setting $\xi_1 = 0$.

Longitudinal submodel

Similar to the definitions for the shared parameter joint model, suppose that $y_{ij} = y_i(t_{ij})$ corresponds to the value of a (longitudinal) biomarker observed at the j^{th} $(j = 1, \dots, n_i)$ time point, t_{ij}, for the i^{th} $(i = 1, \dots, N)$ individual. A class-specific linear mixed model that assumes the expected value of the longitudinal biomarker evolves in continuous time can be specified as

$$
\begin{aligned}
y_i(t)|_{c_i=g} &= \mu_{ig}(t) + \varepsilon_{ig}(t) \\
\mu_{ig}(t) &= \boldsymbol{x}_i'(t)\boldsymbol{\beta}_g + \boldsymbol{z}_i'(t)\boldsymbol{b}_{ig}
\end{aligned}
\tag{28}
$$

where $\mu_{ig}(t)$ is the expected value of the biomarker for individual i in class g at time t, $\varepsilon_{ig}(t)$ is a random (measurement) error term assumed to be drawn from a normal distribution with mean zero and variance σ_{y}^2, $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ are vectors of covariates, possibly time-dependent, with associated vectors of class-specific population-level (i.e. fixed effects) parameters $\boldsymbol{\beta}_g$ and class-specific individual-level (i.e. random effects) parameters $\boldsymbol{b}_{ig}$. The individual-specific parameters are assumed to be normally distributed, with $\boldsymbol{b}_{ig} \sim N(0, \boldsymbol{\Sigma}_{\mathrm{b}})$ for some unstructured variance-covariance matrix Σ_b.

In general, it is possible to extend the random effects structure by allowing the variance-covariance matrix to differ across the latent classes, that is, assume $\boldsymbol{b}_{ig} \sim N\left(0, \boldsymbol{\Sigma}_{\mathrm{bg}}\right)$. Here the subscript g denotes that the variance-covariance matrix is class-specific. Similarly, one could allow the error variance to be class-specific, for example, specifying σ_{yg}^2 instead of σ_{y}^2. However, in practical terms, such flexibility can lead to difficulties with estimation (Proust-Lima et al., 2014).

Event submodel

Let T_i^* denote the event time for individual i, C_i denote the right-censoring time, $T_i = \min(T_i^*, C_i)$ denote the observed event time, and $d_i = I(T_i^* \leq C_i)$ denote the event indicator. A class-specific proportional hazards model takes the form

$$
h_i(t)|_{c_i=g} = h_{0g}(t)\exp(\boldsymbol{w}_i'(t)\boldsymbol{\gamma}_g) \tag{29}
$$

where $h_{0g}(t)$ is the class-specific baseline hazard at time t, and $\boldsymbol{w}_i(t)$ is a vector of covariates, possibly time-dependent, with associated vector of class-specific population-level (i.e. fixed effects) parameters $\boldsymbol{\gamma}_g$. The population-level parameters are assumed to be class-specific for generality, but for simplicity one could assume that they were common across all latent classes; that is, assume $\boldsymbol{\gamma}_g = \boldsymbol{\gamma}$ for all g. Since the population-level parameters in each submodel are class-specific there is an association between the longitudinal and event outcomes that is induced through an individual's class membership. In the next subsection the assumptions related to this dependence are more explicitly defined.

Conditional independence

The latent class joint model assumes that within a given latent class g ($g = 1, \ldots, G$), and conditional on any other covariates in our event submodel, the expected event risk does not differ between individuals. More formally, a conditional independence assumption for the latent class joint model can be defined as

$$y_i(t) \perp T_i^* \mid c_i, \boldsymbol{\theta} \tag{30}$$

which defines that the longitudinal and event outcomes are independent, conditional on an individual's latent class, c_i, and the combined vector of all remaining (fixed effect) parameters in the model, $\boldsymbol{\theta}$. However, the individual-specific parameters are still required for the conditional independence between repeated measurements of the longitudinal outcome, that is

$$y_i(t) \perp y_i(t') \mid c_i, \boldsymbol{b}_{ig}, \boldsymbol{\theta} \tag{31}$$

Various extensions to the standard latent class joint model have also been proposed, for example, accommodating multiple longitudinal biomarkers (Proust-Lima et al., 2009), recurrent events (Han et al., 2007), competing risks (Proust-Lima et al., 2016), or interval-censored event times (Rouanet et al., 2016).

Although the latent class joint modelling approach has not been as widely used as the shared parameter joint modelling approach it can provide benefits in some settings. In particular, the latent class joint model can potentially allow for a more flexible relationship between the biomarker and the event since it does not enforce a specific parametric form for the association between the two processes (Proust-Lima et al., 2014). Moreover, if we believe that there are in fact latent heterogeneous subgroups within the population, and our research question directly relates to the identification of these subgroups, then a latent class joint modelling approach may be more appropriate. For example, in the paper presented in Chapter 4 of this thesis, the aim is to identify groups of haemodialysis patients who are similar with regard to their longitudinal changes in BMI. Moreover, we would like to explore how the BMI trajectories for those groups are related to their risk of kidney transplantation or death. A latent class joint modelling approach lends itself to this type of research question.

2.5 Estimation approaches

2.5.1 Classical methods of estimation

The models described in the previous sections can be estimated through maximisation of the full joint likelihood function. Recall from Section 2.3.1 that the full likelihood function for the standard shared parameter joint model can be specified as

$$L_i = \int \left(\prod_{j=1}^{n_i} p\left(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \right) p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i \qquad (32)$$

The most common approach for obtaining a maximum likelihood solution to equation (32) has been to use an expectation-maximisation (EM) algorithm (Gruttola and Tu, 1994; Henderson et al., 2000; Viviani et al., 2014). However, other approaches have also been used, for example, Newton-Rhapson (Crowther et al., 2012) or Marquardt (Thiébaut et al., 2005) algorithms.

Moreover, aside from the choice of maximisation algorithm, a major computational hurdle exists. Classical maximisation of the likelihood requires integration over the joint distribution of the individual-specific (i.e. random effects) parameters $\boldsymbol{b}_i$, as shown in equation (32). In practice, numerical integration must be used since this integral generally has no tractable solution. Numerical integration approaches used in the joint modelling literature to date have included adaptive Gauss-Hermite quadrature (Crowther et al., 2012), pseudo-adaptive Gaussian quadrature (Rizopoulos, 2012a), or Monte-Carlo simulation (Thiébaut et al., 2005). It is worth noting however, that the need for such numerical integration techniques is not unique to joint models. Indeed, with the exception of linear mixed models (i.e. identity link, normal error distribution), which do have a tractable solution, the estimation of generalised linear mixed models also requires numerical integration techniques.

Finally, another layer of numerical integration may be required in some settings where the association structure introduces time-dependency in the event submodel. Recall that under the standard shared parameter joint model, the contribution to the likelihood from the event submodel is

$$p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = h_i(T_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})^{d_i} \exp\left(-\int_0^{T_i} h_i(s \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, ds\right) \qquad (33)$$

If, for example, the common *current value association structure* is used in the definition of the event submodel, then this introduces a time-varying covariate into the log hazard function $h_i(t)$. In most situations, this will mean there is no closed-form solution to the integral that appears in the latter term of equation (33). Accordingly, one must use a numerical integration approach to evaluate the cumulative hazard $H_i(T_i) = \int_0^{T_i} h_i(s \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, ds$. A common approach in the joint modelling literature to date has been to use Gauss-Kronrod quadrature (Laurie, 1997), see for example Crowther et al. (2013) or Rizopoulos (2012a).

2.5.2 Bayesian estimation

An alternative to classical maximisation of the full joint likelihood is to obtain inferences from the joint model under a Bayesian approach. For the standard shared parameter joint model, one can specify a joint posterior distribution as

$$p(\boldsymbol{b}_i, \boldsymbol{\theta} \mid T_i, d_i, \boldsymbol{y}_i) \propto \left(\prod_{j=1}^{n_i} p\left(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right)\right) p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \qquad (34)$$

where $\boldsymbol{y}_i$ denotes the vector collecting the longitudinal biomarker measurements for individual i, and $p(\boldsymbol{\theta})$ denotes the likelihood for the joint prior distribution across all remaining unknown parameters in the model.

The posterior distribution in equation (34) does not require the individual-specific parameters $\boldsymbol{b}_i$ to be integrated out of the likelihood function, as was the case for the classical estimation methods. Instead, Markov chain Monte Carlo (MCMC) methods can be used to obtain draws from the posterior distribution. With a sufficient number of draws, inferences can be made about the joint distribution of all model parameters, including $\boldsymbol{b}_i$, conditional on the observed data.

Various MCMC algorithms exist and can be used to obtain draws from the joint posterior distribution in equation (34). Arguably the most popular MCMC algorithm in Bayesian statistics has been the Gibbs sampler and this is also true for the Bayesian estimation of joint models. A number of early publications on joint modelling used Gibbs sampling as

the method of estimation (see for example Faucett and Thomas (1996) or Wang and Taylor (2001)). Gibbs sampling has also been used to estimate increasingly complex joint models; for example, those with multiple longitudinal biomarkers (Rizopoulos and Ghosh, 2011), two-part longitudinal submodels (Hatfield et al., 2011, 2012), or complex association structures (Mauff et al., 2017). One of the major factors contributing to the widespread use of Gibbs sampling was that it was the main estimation algorithm used by the popular Bayesian software packages BUGS (Lunn et al., 2000) and JAGS (Plummer, 2016).

However, the ability to obtain valid inferences from a MCMC algorithm depends on two factors. First, that the algorithm converges from its initial location to the target posterior distribution. Second, that the correlation between subsequent draws in the MCMC chain (or those MCMC draws that are retained) is small enough such that the complete set of MCMC draws can be considered a "random" sample from the target posterior. However, when the target posterior distribution is high dimensional and/or has a complex geometry, MCMC algorithms such as Gibbs sampling can be slow to converge to the target distribution (Hoffman and Gelman, 2014). Therefore, more recently, there has been increasing interest in alternatives to Gibbs sampling.

In particular, the introduction of the Bayesian software package Stan (Carpenter et al., 2017) has meant that statistical models can now be more easily estimated using a variant of Hamiltonian Monte Carlo (HMC) (Hoffman and Gelman, 2014). In complex problems, the HMC algorithm can potentially avoid some of the inefficiencies associated with Gibbs sampling and other random walk MCMC algorithms. Specifically, under HMC, the random walk behaviour associated with Gibbs sampling and other random walk MCMC methods is replaced with a directed sampling path based on Hamiltonian dynamics (Neal, 2011). Therefore, some authors have recently used HMC for the estimation of joint models (see for example Desmee et al. (2017a) or Brilleman et al (2016)). In this thesis, Stan and its HMC implementation will be the primary method of estimation for the joint models described in Chapters 5 and 6. Further details on the estimation will be provided in those chapters.

2.5.3 Two-stage joint models

In general, it is computationally intensive to estimate joint models using the full joint likelihood function. For the most part, this is attributable to the fact there is a joint dependence of both the longitudinal and event outcomes on the individual-specific (i.e.

random effects) parameters. In many settings, the distribution of the individual-specific parameters may be of a relatively high dimension, for example, if one is allowing for flexible non-linear longitudinal trajectories through the use of splines or if the joint model includes multiple longitudinal biomarkers. In a classical estimation framework, one needs to integrate over the joint distribution of these individual-specific parameters, which can be a computationally demanding task. Whilst in a Bayesian framework, one needs to draw samples from the joint posterior distribution of all parameters, where the posterior distribution is likely to be high dimensional and have a complex geometry.

To circumvent these computational difficulties, one approach has been to estimate the joint model in two stages (Albert and Shih, 2010; Bycott and Taylor, 1998; Ye et al., 2008a, 2008b). In the first stage, the longitudinal submodel is estimated, whilst ignoring any information that is related to the event outcome. In the second stage, estimates from the longitudinal submodel (for example the expected value of the biomarker, or an estimate of the individual-specific parameters) are included in the estimation of the event submodel. This two-stage process allows separate estimation of each of the submodels, thereby decreasing the overall computational burden. However, separate estimation of the two submodels has two potential downsides. First, uncertainty in the estimates obtained for the longitudinal submodel in the first stage might not be carried through to the estimation of the event submodel in the second stage. Second, the time-to-event outcome (which is generally indicative of an informative dropout mechanism affecting the longitudinal process) is ignored entirely when the longitudinal submodel estimates are obtained in the first stage. For these reasons, several authors have identified that standard two-stage joint modelling approaches generally do not perform as well as joint modelling approaches based on the full joint likelihood (Sweeting and Thompson, 2011).

A method described by Murawska et al. (2012) attempted to deal with the first of these issues. That is, the authors used a Monte Carlo simulation approach to capture uncertainty in their estimates of the longitudinal submodel parameters when fitting the event submodel in the second stage. However, more recently, there has been work towards developing a two-stage joint modelling approach that appropriately deals with both of the issues previously raised. Specifically, Moreno-Betancur et al. (2017) proposed a two-stage joint modelling approach for multiple longitudinal biomarkers and a time-to-event outcome. They used a multiple imputation approach to propagate uncertainty in the estimates from the first stage when estimating the event submodel in the second stage. Moreover, by

including an estimate of the cumulative hazard and a modified event indicator in their imputation model they were able to correct their estimates in the first stage for the potentially informative missingness due to the event. The approach performed well under a variety of scenarios and, importantly, it was shown to be computationally fast relative to the full joint likelihood estimation approach.

2.6 Chapter summary

This chapter has provided an overview of methods for the joint modelling of longitudinal and time-to-event data. Moreover, it has provided a review of some of the most important contributions to the joint modelling literature. The review has focussed on literature and methodology that is most relevant to the subsequent chapters of this thesis. It is important to recognise that the joint modelling literature is now large and a number of topics have not been discussed here.

Chapter 3: Application of a shared parameter joint model: disaster exposure, disability and death in older Americans

3.1 Chapter introduction

Although the last two decades have seen a huge amount of interest in joint modelling methodology, the uptake of joint modelling methodology in the applied literature has been limited. In the applied project described in this chapter, a shared parameter joint modelling approach was used to better understand associations between a community-level exposure variable related to major disasters and individual-level changes in disability and risk of death. The primary outcome data used in this project came from the Health and Retirement Study (HRS), an ongoing longitudinal cohort study carried out in the United States. Specifically, repeated measurements of a disability score (based on activities of daily living) were observed for each individual in the HRS. In addition, mortality outcomes were obtained via linkage to the national death index.

The motivations for the study related to a hypothesis that major disaster events (for example earthquakes, hurricanes, or floods) would impact on the long-term disability trajectories of older individuals and/or their associated risk of death. A shared parameter joint modelling approach with a current value association structure allowed for the investigation of this hypothesis, whilst allowing for the implicit dependence between the disability trajectory and mortality. That is, the approach implicitly models deaths as informative dropouts in the longitudinal model, and in turns yields estimates in the survival model that correct for the discrete-time, imperfect observation of the disability function, which is truly an unobserved continuous process.

In this study, one relatively non-standard form for the joint model was that the longitudinal submodel was based on a generalised linear mixed model with a *negative binomial* error

distribution and log link function. The use of a non-normally distributed longitudinal response variable has been relatively less common in applied joint modelling papers. Given the use of a log link function in the longitudinal submodel, the association structure for the joint model assumed that the log hazard of death for a given individual at time t was linearly associated with their expected log disability score at time t.

The main content of this chapter is presented in the next section, in the form of an applied research paper that has been published in the journal *Social Science & Medicine*: The model described in the paper was estimated using the **JMbayes** (Rizopoulos, 2016) R package. The supplementary material for the paper is provided in Appendix A of this thesis.

3.2 Manuscript

This section herein contains the following applied research paper:

> Brilleman SL, Wolfe R, Moreno-Betancur M, Sales AE, Langa KM, Li Y, Daugherty Biddison EL, Rubinson L, Iwashyna TJ. Associations between community-level disaster exposure and individual-level changes in disability and risk of death for older Americans. *Social Science & Medicine*. 2017;173:118-125.

Social Science & Medicine 173 (2017) 118–125

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Associations between community-level disaster exposure and individual-level changes in disability and risk of death for older Americans

Samuel L. Brilleman [a, b, *], Rory Wolfe [a, b], Margarita Moreno-Betancur [a, b, c], Anne E. Sales [d, e], Kenneth M. Langa [d, e, f, g], Yun Li [h], Elizabeth L. Daugherty Biddison [i], Lewis Rubinson [j], Theodore J. Iwashyna [d, e]

[a] Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[b] Victorian Centre for Biostatistics (ViCBiostat), Melbourne, Australia
[c] Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne, Australia
[d] Department of Medicine, University of Michigan Medical School, Ann Arbor, MI, USA
[e] Center for Clinical Management Research, Veterans Affairs Ann Arbor Health System, Ann Arbor, MI, USA
[f] Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
[g] Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, USA
[h] Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
[i] Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA
[j] R Adams Cowley Shock Trauma Center, University of Maryland Medical School, Baltimore, MD, USA

ARTICLE INFO

Article history:
Received 23 August 2016
Received in revised form
2 December 2016
Accepted 5 December 2016
Available online 5 December 2016

Keywords:
Death
Disability
Disaster
Health and Retirement Study
Joint model
Shared parameter model
Survival

ABSTRACT

Disasters occur frequently in the United States (US) and their impact on acute morbidity, mortality and short-term increased health needs has been well described. However, barring mental health, little is known about the medium or longer-term health impacts of disasters. This study sought to determine if there is an association between community-level disaster exposure and individual-level changes in disability and/or the risk of death for older Americans. Using the US Federal Emergency Management Agency's database of disaster declarations, 602 disasters occurred between August 1998 and December 2010 and were characterized by their presence, intensity, duration and type. Repeated measurements of a disability score (based on activities of daily living) and dates of death were observed between January 2000 and November 2010 for 18,102 American individuals aged 50–89 years, who were participating in the national longitudinal Health and Retirement Study. Longitudinal (disability) and time-to-event (death) data were modelled simultaneously using a 'joint modelling' approach. There was no evidence of an association between community-level disaster exposure and individual-level changes in disability or the risk of death. Our results suggest that future research should focus on individual-level disaster exposures, moderate to severe disaster events, or higher-risk groups of individuals.

1. Introduction

World-wide, the United States (US) has been ranked amongst the top five countries most frequently experiencing a natural disaster (Guha-Sapir et al., 2015). In 2010, 738 US counties, which represent nearly one in four counties, experienced events devastating enough to qualify for a US Federal Emergency Management Agency (FEMA) disaster declaration. Most recently, FEMA reported a total of 79 major disaster declarations for 2015; during preceding years this figure reached as high as 242 (Federal Emergency Management Agency, 2016a).

Disasters have been defined as "a situation or event which overwhelms local capacity, necessitating a request to a national or international level for external assistance; an unforeseen and often sudden event that causes great damage, destruction and human

* Corresponding author. Level 6, Alfred Centre, 99 Commercial Road, Melbourne, VIC 3004, Australia.
E-mail address: sam.brilleman@monash.edu (S.L. Brilleman).

http://dx.doi.org/10.1016/j.socscimed.2016.12.007
0277-9536/

suffering" (Guha-Sapir et al., 2015). Although the psychological impacts of disasters have been widely studied (DiGrande et al., 2011; Galea et al., 2005; North and Pfefferbaum, 2013), few epidemiological studies have examined the medium or longer-term health consequences of disasters beyond the realm of mental health. Clearly, disasters may result in direct injury. We hypothesized that disasters could also result in medium-term adverse health impacts through two distinct *community-level* mechanisms related to the physical (Sampson et al., 2016) and social (Hikichi et al., 2016) environments of the community.

First, exposure to a disaster might worsen disability through the disruption of the adapted environment that individuals had crafted around themselves to mitigate physical risks for disability. Within the healthcare system, disasters may disrupt ongoing medical care for therapies ranging from daily insulin availability to longitudinal chemotherapy courses. More subtly still, disasters may strip an individual of the adaptations he or she uses to keep a physical activity limitation from becoming a social disability. For example, consider an individual with potentially limited mobility, but who through the use of assistive devices and a careful understanding of her local geography has mapped out routes without obstacles allowing her to go to the store or church. Certainly disaster debris will undo the effectiveness of these adaptations; but even after clean-up, reconstructions along the paths of her life space may present new and difficult barriers.

Second, exposure to a disaster might operate as a community-level exposure because of its wider-ranging disruption on interlocking social support networks and institutions. Because such networks are often informal, activated only upon a contingent need, or because they involve several chains of connection, we hypothesized that there might be multiple points of brittleness that are exposed to a disaster. We hypothesized that these disruptions might unmask the *social* adaptations that are the counterpart to the physical and environmental adaptations just discussed. Consider, for example, an individual who does not currently suffer from a physical activity limitation. A disaster may not physically interfere with her body or home, however, the disaster might still lead to disability if it disrupts those networks and institutions on whom she depends to maintain her good health.

Current understanding of disasters' medium or longer-term influence on disability or even death is scant and is largely based on case studies of particular types of disasters or specific to certain communities; thus generalizability is unclear (Aldrich et al., 2010; Hendrickson and Vogt, 1996; Sastry and Gregory, 2013; Wade et al., 2004). We therefore proposed to study the impact of a range of disasters across an extended time period and impacting many US communities to determine the potential association of disaster exposure and subsequent disability or death.

2. Methods

We matched community-level disaster events to individual outcomes for older Americans who were participating in a representative longitudinal panel study. Individual outcomes were death and a repeated measure of (instrumental) activities of daily living, the latter being specifically a measure of functional independence, but considered as a surrogate for disability in this study. Since these two outcomes are endogenous in older people, they must be modelled together (Marioni et al., 2014, 2015). Since we were directly interested in both outcomes we chose to use novel statistical methodology known as joint modelling (Henderson et al., 2000; Rizopoulos, 2012; Wulfsohn and Tsiatis, 1997), which allowed us to simultaneously model the longitudinal disability trajectory for each individual and their associated risk of death.

2.1. Data and sample

We utilized data from the Health and Retirement Study (HRS), a longitudinal panel survey that biannually followed a representative sample of US individuals over the age of 50 years and their spouses (Sonnega et al., 2014). We included 18,102 individuals, aged 50–89 years at baseline, who were enrolled in 1998 and had at least one follow-up survey between 1st January 2000 and 30th November 2010. Baseline, for each individual, was the date on which they completed their first eligible survey. Since individuals who complete a survey in the immediate aftermath of a disaster are likely to be an unrepresentative group, a survey was considered ineligible if it occurred during the 6 months immediately following a disaster (n = 9111 surveys, 12%). The 1998 data were only used to identify an individual's initial county of residence (prior to baseline); county of residence was modified, if changed at each wave, and modelled as if the change happened on the day of that survey.

Disasters were identified through the FEMA Public Assistance Program database (Federal Emergency Management Agency, 2016b), which contains all disaster declarations that qualified for federal relief funding. We obtained data which included 602 unique disaster declarations, starting on 411 unique dates between 22nd August 1998 and 26th December 2010. To test our hypotheses regarding the effects of disaster as a community-level exposure, we needed to operationalize "community" in a concrete way. Disaster events were matched to individuals based on their US county of residence. County was the finest-grained level at which national data on disaster exposure could be ascertained from this FEMA database.

2.2. Variable definitions

2.2.1. Outcome variables

Activities of Daily Living (ADLs) and Instrumental Activities of Daily Living (IADLs) were reported at each survey wave. We formulated a score based on the sum of ADLs and IADLs. The score took integer values between 0 and 11, inclusive, with higher values corresponding to lower levels of functional independence, which we consider a surrogate for higher levels of functional disability. We refer to this measure as a disability score throughout the paper, although we recognise that ADL/IADLs do not capture all the recommended dimensions of functioning and disability that are described in the World Health Organisation's (WHO) International Classification of Functioning, Disability and Health (ICF) framework (World Health Organization, 2001).

Time-to-death was defined as time from baseline to the date of death. An individual was censored if still alive on 30th November 2010.

2.2.2. Exposure variables

Different representations of the time-dependent disaster exposure were considered, based on disaster presence, intensity, duration and type. Disaster presence was the simplest form of exposure variable used, and indicated whether an individual had been exposed to a disaster during the two years prior to the current observation time t. For modelling the association between disaster exposure and disability, presence was operationalized as a binary exposure: no disaster within the previous two years; or disaster within the previous two years. For modelling the association between disaster exposure and death we separated out acute and medium-term effects by defining presence as a three-level categorical exposure: no disaster within the previous two years; disaster within the previous two years but not within the previous 21 days; or disaster within the previous 21 days. A period

of 21 days was deemed to be adequate for capturing the majority of deaths which would occur in the immediate aftermath of a disaster.

Disaster intensity was defined as the cumulative amount of FEMA funding for rebuilding — not upgrading — damaged public infrastructure (in millions of US dollars) received for disasters beginning within the previous two years. Each disaster event also had a duration (in days), as determined by FEMA. We created an exposure variable based on disaster duration, defined as the cumulative duration of disasters beginning within the previous two years. We categorised both disaster intensity and duration, with the non-zero part of the distribution divided into five equally-sized quantiles; cut-points for the categories are shown in Table 5.

Disaster type was a binary indicator taking value 1 if the individual experienced a disaster beginning sometime within the previous two years and that disaster was of a specific type. The types of disasters we considered, using FEMA's own categorization, were: earthquake, fire, flood, hurricane, tornado, storm, snow, or other.

2.2.3. Other covariates

Age at baseline was categorised as: ≥50, <60; ≥60, <65; ≥65, <70; ≥70, <75; ≥75, <80; ≥80, <85; and ≥85, <90 years. Race was categorised as: white/Caucasian; black or African American; or other. Wealth was defined at baseline for each individual using their total household (individual and spouse combined) wealth (excluding housing values) (Smith, 1995), and categorised into ten deciles ranging from decile 1 (most wealth) to decile 10 (least wealth).

2.3. Modelling

The joint model used in the analyses consisted of two distinct submodels: a longitudinal submodel for the disability score and a survival submodel for time-to-death. The dependency between the two outcomes is specified by allowing the survival submodel to depend on the current expected value of the log disability score as determined by the longitudinal submodel. This specification implicitly models deaths as informative drop-outs in the longitudinal model, and in turns yields estimates in the survival model that correct for the discrete-time, imperfect observation of the disability function, which is truly an unobserved continuous process. Effectively, this is achieved by the simultaneous estimation of all parameters. Alternative specifications for the association structure between the two outcomes are discussed in the Web Appendix.

2.3.1. Longitudinal submodel

Let $\boldsymbol{y_i} = \{y_{i1}, \ldots, y_{in_i}\}$ denote the vector of all observed disability measurements for individual i $(i = 1, \ldots, N)$ where $y_{ij} = y_i(t_{ij})$ denotes a single observed disability measurement at time point t_{ij} $(j = 1, \ldots, n_i)$. We specified a generalised linear mixed model $y_{ij} \sim NB(\mu_i(t_{ij}), \phi)$, where $NB(.)$ represents the negative binomial distribution with scale parameter ϕ, and linear predictor related to the mean via a log link function $\eta_i(t_{ij}) = \log(\mu_i(t_{ij}))$ with

$$\begin{aligned}\eta_i(t_{ij}) = {} & \beta_0 + \beta_1 t_{ij} + \sum_{\boldsymbol{a}} \beta_{2a} A_{ia} + \sum_{\boldsymbol{g}} \beta_{3g} G_{ig} + \sum_{\boldsymbol{r}} \beta_{4r} R_{ir} \\ & + \sum_{\boldsymbol{w}} \beta_{5w} W_{iw} + \sum_{\boldsymbol{a}} \beta_{6a} (A_{ia} t_{ij}) + \sum_{\boldsymbol{d}} \beta_{7d} D_{ijd} + b_{1i} \\ & + b_{2i} t_{ij} \end{aligned} \tag{1}$$

where A_{ia}, G_{ig} and R_{ir} are dummy variables for the baseline age, gender and race categories respectively, W_{iw} are dummy variables for the ten wealth deciles, D_{ijd} are dummy variables for the time-varying disaster exposure categories, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{7d})$ is a vector of fixed coefficients and $\boldsymbol{b_i} = (b_{1i}, b_{2i})$ is a vector of individual level random coefficients (intercept and slope). We assume that $\boldsymbol{b_i} \sim MVN(0, \boldsymbol{\Sigma})$.

2.3.2. Survival submodel

Let $h_i(t)$ denote the hazard of death for individual i at time t. We assumed a proportional hazards model of the form

$$\begin{aligned} h_i(t) = h_0(t) \exp\Bigg(& \sum_a \gamma_{1a} A_{ia} + \sum_g \gamma_{2g} G_{ig} + \sum_r \gamma_{3r} R_{ir} + \gamma_4 W_i^* \\ & + \sum_a \gamma_{5a} \left(A_{ia} W_i^* \right) + \sum_d \gamma_{6d} D_{ijd} + \alpha_1 \eta_i(t) \Bigg) \end{aligned} \tag{2}$$

where $h_0(t)$ is the baseline hazard (the hazard for an individual in the reference category of all covariates) evaluated at time t, W_i^* is a numeric variable taking integer values for each of the wealth categories from 0 (decile 1) to 9 (decile 10), $\boldsymbol{\gamma} = (\gamma_{1a}, \ldots, \gamma_{6d})$ is a vector of fixed coefficients, and α_1 is a fixed coefficient known as the association parameter. Using the numeric variable W_i^*, rather than the dummy variables W_{iw}, allowed us to fit a linear trend across wealth deciles, which was more parsimonious and resulted in little difference in model fit. The baseline hazard in the survival submodel was approximated using a parametric penalised splines-based method (Rizopoulos, 2016).

The coefficient α_1 provides a measure of strength of the association between the longitudinal and survival processes. Disaster, the exposure of interest, is present in both the longitudinal and survival submodels and hence the regression coefficient(s) γ_{6d} provide an estimate of the so-called direct effect of disaster on death, that is, the effect not mediated by the impact of disaster on disability.

2.3.3. Model estimation

We took a Bayesian approach to model estimation and used the JMbayes package (Rizopoulos, 2016) in R 3.2.2 (Core Team, 2015) to fit the model. JMbayes fits joint models using a Metropolis-based Markov chain Monte Carlo (MCMC) algorithm. The Web Appendix contains further details of the model estimation.

3. Results

3.1. Descriptive statistics

3.1.1. Baseline characteristics

Table 1 shows baseline characteristics of the 18,102 individuals in the study cohort. The sample was relatively balanced in terms of gender (57.8% female). The majority were white/Caucasian (83.1%), with a substantial minority being black or African American (13.4%). There was large variation in individual wealth, for example, median wealth in the poorest and richest deciles, respectively, was $400 and $1.3 million. The mean (SD) disability score at baseline was 0.7 (1.9), with this increasing by age. There were 543 (3%) individuals with missing baseline wealth data who were excluded from the regression modelling.

3.1.2. Outcome data

During the study period 67,135 disability score measurements were observed, with a mean (max) of 3.8 (Sampson et al., 2016) measurements per individual. Our joint model provides valid

Table 1
Baseline characteristics of the study cohort.

Characteristic	Estimate
Total number of individuals	18102
Age at baseline (in years), n (%)	
≥50, <60	4000 (22.1%)
≥60, <65	3580 (19.8%)
≥65, <70	3256 (18.0%)
≥70, <75	2526 (14.0%)
≥75, <80	2188 (12.1%)
≥80, <85	1610 (8.9%)
≥85, <90	942 (5.2%)
Gender, n (%)	
Female	10507 (58.0%)
Race, n (%)	
White/Caucasian	14933 (82.5%)
Black or African American	2518 (13.9%)
Other	651 (3.6%)
Wealth at baseline by decile (in USD thousands), median (min, max)[a]	
Decile 1 (most wealth)	1324.0 (857, 90708)
Decile 2	636.0 (495, 856)
Decile 3	398.6 (325, 494)
Decile 4	268.4 (223, 325)
Decile 5	187.0 (156, 223)
Decile 6	128.3 (105, 156)
Decile 7	83.6 (66, 105)
Decile 8	51.0 (36, 66)
Decile 9	20.0 (6, 36)
Decile 10 (least wealth)	0.4 (0, 6)
Disability score at baseline, mean (SD)	
Stratified by age at baseline (in years)	
≥50, <60	0.4 (1.3)
≥60, <65	0.5 (1.4)
≥65, <70	0.6 (1.6)
≥70, <75	0.6 (1.7)
≥75, <80	0.9 (2.1)
≥80, <85	1.5 (2.6)
≥85, <90	2.6 (3.4)
Overall	0.8 (1.9)

USD, United States dollars; SD, standard deviation.
[a] There were 543 (3%) individuals with missing wealth data at baseline.

estimates under the assumption that missing disability measurements at the survey level (3601 (5.1%) surveys) are missing at random (Little and Rubin, 2002).

Fig. 1 shows the observed disability score trajectories for all individuals, stratified by age category and whether the individual died or was censored. A LOWESS smoothed average curve is overlaid in each plot. On average, disability scores increased during follow-up with some non-linearity also evident. Older individuals had higher disability scores on average as well as having faster rates of increase in disability. Individuals who died had higher average disability scores than those who were censored, as well as faster rates of increase in disability over time. By the censoring date of 30th November 2010, 5304 (30%) individuals had died.

3.1.3. Exposure data

Of the total follow up time for all individuals (148,485 person-years), approximately 45% (67,210 person-years) was spent classified as exposed to a disaster within the previous 2 years. Of the 17,559 individuals included in the regression modelling, 6388 (37%) were exposed to a disaster within the previous 2 years at baseline. At the time of their terminating event (death or censoring) 6911 (39%) individuals were exposed to a disaster within the previous 2 years and 16,075 (92%) individuals had experienced at least one disaster some time during the study period. The mean (max) number of disasters experienced by each individual prior to death or censoring was 3.6 (Rizopoulos, 2012), however, the incidence of disaster exposure differed by disaster type (Table 2). Storm, hurricane, and snow were the most frequently experienced types of disaster event. In our discussion section we discuss why we may have observed such high disaster incidence rates in this study.

Disasters types were clustered within geographical regions; for example, hurricanes were experienced by 37% of individuals (25% of person-disaster exposure events), yet 71% of those exposures occurred within just two US states. The rates of disasters were associated with individual-level baseline characteristics, in particular age and wealth (Tables S1 and S2 of Web Appendix).

3.2. Modelling

3.2.1. Associations between disaster exposure and disability or death

There was no evidence that the presence of a disaster within the previous 2 years was associated with any increase in disability (disability score ratio = 0.98, 95% CrI: 0.93, 1.03) (Table 3). We also found very little evidence that the presence of a disaster within the previous 2 years (but not within the previous 21 days) was associated with any increase in the risk of death (HR = 1.03, 95% CrI: 0.96, 1.11) (Table 4). There was large uncertainty around the hazard ratio associated with disaster presence in the previous 21 days (HR = 0.96, 95% CrI: 0.75, 1.21), likely due to the small number of deaths which occurred within this narrow time frame (n = 114, just 2% of the total number of deaths). There was also no evidence that the mean disability score or risk of death increased in proportion to disaster intensity or duration (Table 5); even for an individual in the uppermost disaster intensity category (FEMA spending >$10 million) there was no evidence that the mean disability score was higher than an individual who had not been exposed to any disaster within the previous 2 years (HR = 0.99, 95% CrI: 0.90, 1.07).

3.2.2. Associations between baseline characteristics and disability or death

From the longitudinal submodel (Table 3) older age, less wealth, and non-white race were associated with higher levels of disability. The estimated annual increase in disability was also higher for individuals of older ages. Gender was not associated with the estimated disability score. From the survival submodel (Table 4) older age, being male, and less wealth were associated with a higher hazard of death. The magnitude of the association between wealth and the hazard of death diminished with increasing age.

3.2.3. Association between disability and death

There was evidence that the estimated disability score was strongly associated with the hazard of death (Table 4). A twofold increase in the estimated disability score for an individual (equivalent to a 0.693 unit increase in the log disability score, as $\exp(0.693) = 2$) was associated with a 36% (HR = 1.36, calculated as $\exp(0.693 \times \log(1.56))$, 95% CrI: 1.29, 1.41) increase in the hazard of death, for given fixed values of the baseline covariates and disaster exposure.

3.2.4. Exposure to specific disaster types

Table 6 shows disability score ratios and hazard ratios associated with exposure to specific disaster types, with the binary exposure variables for each disaster type all included in a single joint model. The largest posterior means were associated with exposure to a tornado (disability score ratio = 1.20, 95% CrI: 0.86, 1.67; HR = 1.66, 95% CrI: 1.12, 2.44), however, the statistical evidence to support these associations was relatively weak (wide credible intervals), potentially owing to the fact that tornados were relatively rare (662 person-disaster events, 1% of the total). The most prevalent disaster types (storms, hurricanes, snow) did not appear to be associated with increased disability or death.

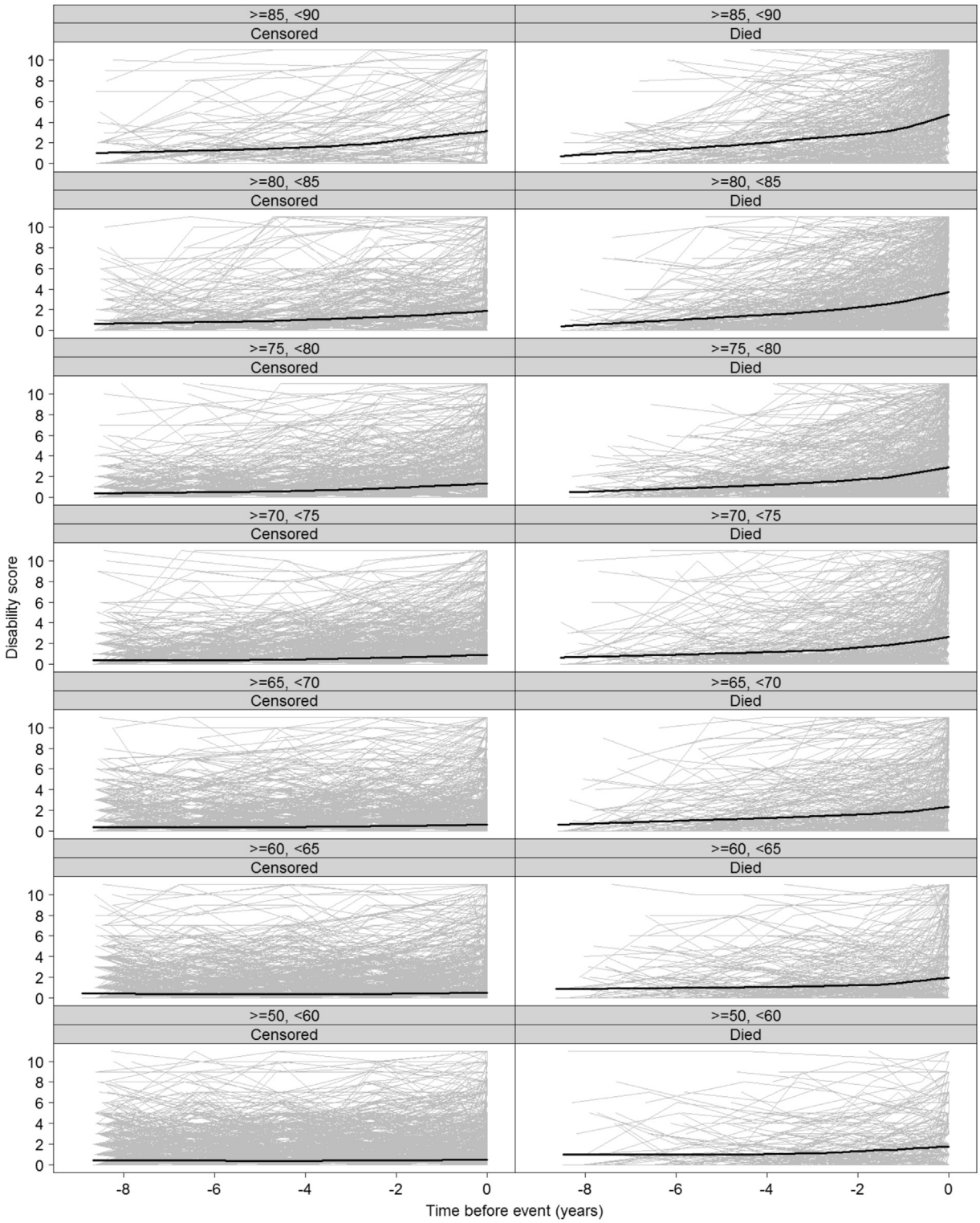


Fig. 1. Observed disability score trajectories for all individuals stratified by age category and whether the individual was censored or died. The red line overlaid in the plots is the LOWESS smoothed average.

3.2.5. Sensitivity analysis

In a sensitivity analysis (see Web Appendix) we refitted the disaster presence joint model to the subset of individuals with a baseline disability score of 0, 1 or 2 ("low" baseline disability). We found a slight change in the estimated disability score ratio for disaster exposure, such that it was positive, but the 95% credible interval still incorporated a value of 1 (disability score ratio = 1.04, 95% CI: 0.98 to 1.10). The estimates from the survival submodel remained almost unchanged.

Table 2
Number of individuals experiencing each disaster type at least once, as well as the total number of person-disaster events for each disaster type.

Disaster type	Number of individuals experiencing this disaster type at least once (%)	Number of person-disaster events (%)
Storm	12944 (74%)	28894 (45.2%)
Hurricane	6415 (37%)	16090 (25.2%)
Snow	5496 (31%)	10436 (16.3%)
Fire	3229 (18%)	4291 (6.7%)
Flood	1083 (6%)	1294 (2.0%)
Tornado	662 (4%)	662 (1.0%)
Earthquake	259 (1%)	259 (0.4%)
Other	1943 (11%)	1943 (3.0%)
All disasters	16075 (92%)	63869 (100%)

Notes. The 'storm' category includes severe storm, severe ice storm or coastal storm. The 'other' category includes dam/levee break, freezing, terrorist or not otherwise specified. The percentages shown are: % of total individuals (left column) and % of total person-disaster events (right column).

Table 3
Disability score ratios from longitudinal submodel of the fitted joint model for disaster presence. Estimates presented are the posterior means and 95% credible intervals.

	Disability score ratio	95% credible interval
Constant	0.02	0.02, 0.03
Time (years)	1.03	1.01, 1.04
Age category (ref: ≥50, <60y)		
≥60, <65y	0.92	0.82, 1.03
≥65, <70y	1.19	1.06, 1.33
≥70, <75y	1.72	1.51, 1.95
≥75, <80y	3.04	2.66, 3.48
≥80, <85y	5.70	4.96, 6.64
≥85, <90y	9.75	8.12, 11.79
Age category * time interaction		
≥60, <65y	1.05	1.03, 1.06
≥65, <70y	1.10	1.08, 1.11
≥70, <75y	1.18	1.16, 1.20
≥75, <80y	1.22	1.20, 1.24
≥80, <85y	1.27	1.25, 1.30
≥85, <90y	1.27	1.24, 1.30
Gender (ref: Male)		
Female	1.02	0.95, 1.09
Race (ref: White or Caucasian)		
Black or African American	1.30	1.17, 1.44
Other	1.15	0.95, 1.37
Wealth category (ref: Decile 1, most wealth)		
Decile 2	1.10	0.94, 1.30
Decile 3	1.27	1.08, 1.49
Decile 4	1.74	1.49, 2.05
Decile 5	1.86	1.61, 2.17
Decile 6	2.23	1.91, 2.60
Decile 7	3.06	2.60, 3.57
Decile 8	3.71	3.16, 4.32
Decile 9	5.28	4.46, 6.18
Decile 10, least wealth	9.52	8.12, 11.25
Disaster exposure		
Within previous 2 years	0.98	0.93, 1.03

ref, Reference category.

4. Discussion

This study investigated the health impacts of a temporally representative sample of disasters occurring in the US, rather than considering single disaster events as case studies. We matched community-level disaster exposures for a range of disaster types (for example hurricanes, earthquakes, fires, tornados) to individuals participating in a nationally representative longitudinal study of older Americans. We found no evidence of an association between community-level disaster exposure and individual-level changes in ADL/IADL outcome, the latter being a measure of functional independence and a surrogate of disability. Similar results were obtained even when considering several different representations (presence, intensity or duration) of disaster exposure.

Table 4
Hazard ratios from the survival submodel of the fitted joint model for disaster presence. Estimates presented are the posterior means and 95% credible intervals.

	Hazard ratio	95% credible interval
Age category (ref: ≥50, <60y)		
≥60, <65y	2.40	1.54, 3.62
≥65, <70y	3.58	2.52, 5.33
≥70, <75y	3.80	2.65, 5.59
≥75, <80y	5.81	4.12, 8.54
≥80, <85y	7.88	5.39, 11.25
≥85, <90y	9.88	6.65, 14.78
Gender (ref: Male)		
Female	0.60	0.56, 0.65
Race (ref: White or Caucasian)		
Black or African American	0.90	0.80, 0.99
Other	0.74	0.61, 0.92
Wealth trend across deciles		
Linear trend (0 = Decile 1; 9 = Decile 10)	1.13	1.07, 1.18
Age category * wealth trend interaction		
≥60, <65y	0.93	0.87, 0.99
≥65, <70y	0.91	0.86, 0.96
≥70, <75y	0.93	0.88, 0.98
≥75, <80y	0.91	0.86, 0.96
≥80, <85y	0.89	0.85, 0.94
≥85, <90y	0.88	0.83, 0.93
Disaster exposure		
Within previous 21 days	0.96	0.75, 1.21
Within previous 2 years, but not 21 days	1.03	0.96, 1.11
Association parameter		
Current value	1.56	1.45, 1.65

ref, Reference category.

Nonetheless, there are important limitations to our study that need to be recognised. These are discussed in greater detail below, but in brief, they include the inability of the community-level exposure variable to accurately reflect individual-level exposure, the potential for geographical or severity misclassification of disasters when using county-level FEMA declarations, as well as the potential insensitivity of the outcome measure to changes in disability.

It may be that, contrary to our hypotheses, the effects of disasters (with the magnitude observed during the study period) are predominantly direct physical injury, without impact on social cohesion and bonds. Alternatively, it may be that disasters do impact on health by disrupting social cohesion and bonds, but that this occurs at a level of granularity smaller than the county. Counties in the US vary dramatically in terms of both land area and population and, therefore, a disaster event is seldom large or pervasive enough to impact all individuals living in that county. In this sense, disasters may act as a community-level exposure on an individual's disability, but there was sufficient misclassification in our exposure variable such that we could not detect an effect.

There are also several possible mechanisms by which individuals or communities may find themselves resilient to the impacts of disasters (Institute of Medicine, 2015). Social cohesion, for example, may help to accelerate the recovery process following a disaster, or may prevent a community from becoming fragmented at the time of the event. Pre-disaster social support networks have been found to reduce the risk of adverse mental health outcomes following a disaster for individuals of low socioeconomic status (Chan et al., 2015) and the elderly (Hikichi et al., 2016). It is possible that social support networks are similarly protective against the medium-term physical health impacts of disasters for older people.

Table 5
Disability score ratios and hazard ratios (posterior means and 95% credible intervals) associated with disaster intensity or disaster duration. Separate joint models were fit for each of the exposure variables (i.e., 2 separate joint models). To save space, parameter estimates associated with the baseline covariates (age, gender, race, wealth) have been omitted but were similar to those contained in Tables 3 and 4.

Disaster exposure variable	Range of exposure category (min to max)	Longitudinal submodel:		Survival submodel:	
		Disability score ratios	95% credible interval	Hazard ratios	95% credible interval
Disaster spending within previous 2 years (ref: $0)					
>$0, Quintile 1	$892 to $295,828	0.97	0.90, 1.03	0.96	0.85, 1.08
>$0, Quintile 2	$295,877 to $1,198,329	0.97	0.90, 1.04	1.06	0.90, 1.22
>$0, Quintile 3	$1,203,047 to $3,405,042	0.96	0.88, 1.03	0.99	0.85, 1.16
>$0, Quintile 4	$3,432,852 to $9,906,982	1.03	0.95, 1.11	1.07	0.88, 1.27
>$0, Quintile 5	>$10 million	0.99	0.90, 1.07	1.07	0.93, 1.22
Total duration of disasters beginning within previous 2 years (ref: 0 days)					
>0 days, Quintile 1	1 to 6 days	0.98	0.91, 1.05	1.03	0.90, 1.17
>0 days, Quintile 2	7 to 18 days	0.97	0.90, 1.04	1.01	0.88, 1.16
>0 days, Quintile 3	19 to 34 days	0.96	0.89, 1.03	1.00	0.88, 1.12
>0 days, Quintile 4	35 to 75 days	1.00	0.91, 1.09	1.04	0.89, 1.21
>0 days, Quintile 5	80 to 289 days	1.02	0.93, 1.12	1.06	0.89, 1.26

ref, Reference category.

Table 6
Disability score ratios and hazard ratios (posterior means and 95% credible intervals) associated with different types of disasters; a single joint model was fit which included 7 dummy variables (one for each disaster type). To save space, parameter estimates associated with the baseline covariates (age, gender, race, wealth) have been omitted but were similar to those contained in Tables 3 and 4 The estimates for each disaster type are relative to a reference category of no disaster exposure (of that type) within the previous 2 years.

	Longitudinal submodel:		Survival submodel:	
	Disability score ratios	95% credible interval	Hazard ratios	95% credible interval
Disaster exposure within previous 2 years				
Storm	1.00	0.94, 1.05	1.03	0.95, 1.13
Hurricane	0.99	0.93, 1.05	1.03	0.92, 1.21
Snow	1.00	0.94, 1.07	1.05	0.91, 1.18
Fire	1.00	0.91, 1.09	0.99	0.85, 1.16
Flood	0.88	0.74, 1.05	0.73	0.43, 1.19
Tornado	1.20	0.86, 1.67	1.66	1.12, 2.44
Earthquake	0.98	0.72, 1.34	1.30	0.58, 2.53
Other	1.01	0.91, 1.11	0.95	0.74, 1.20

Notes. The 'storm' category includes severe storm, severe ice storm or coastal storm. The 'other' category includes dam/levee break, freezing, terrorist or not otherwise specified.

One of the strengths of this study was the disaster surveillance design. This led to a broadly defined exposure variable which incorporated a range of disaster types. In contrast with most studies in the literature, which consider single disaster events, our study design can provide novel insight into the wider impact of disasters on health outcomes. However, the generality of our exposure means it is difficult to compare our results directly with those from previous studies. Our exposure variable may have also been subject to misclassification. The incidence of disaster exposure was very high in this study. It may be that the FEMA definition of a disaster is too broad to be meaningful in this context. Less severe events, for example snow storms, may have attracted federal financial assistance but may not have caused the "great damage, destruction and human suffering" (Guha-Sapir et al., 2015) necessary to impact on the disablement process of individuals living in the affected community. This has important implications for the interpretation of our findings, since although our results showed no association between community-level disaster exposure and disability or death across a broad range of disaster exposures, it cannot be inferred that specific disaster events do not have long-term impacts on disability. It would be unreasonable to conclude that communities, or all subgroups within a community, are resilient to the long-term impacts of all disasters based on our study. Nonetheless, we found no evidence of an association between disaster exposure and disability even when considering a graded exposure variable, such as disaster intensity or duration.

Our analyses did not adjust for county-level clustering which may have led to standard errors which were slightly narrow — which would bias *away* from the null, reinforcing the lack of statistical association in our study. In a sensitivity analysis we reran the joint model for disaster presence also including dummy variables for each US state of residence. We found the width of the 95% credible intervals for the estimated effects of disaster exposure on either disability or death were almost unchanged. Our analyses also assumed that individuals remained in the same county of residence between surveys, potentially introducing some misclassification of disaster exposures. However, the number of misclassifications is likely to be small since, for example, 16,028 (91%) individuals in our analysis were residing in the same US state at all surveys (including the 1998 wave) and, furthermore, 93.5% of all community-dwelling respondents in the HRS lived in the same metropolitan statistical area as at their prior wave, suggesting participants in the HRS are a relatively non-transient population. Lastly, since we did not have data on clinical or self-reported diagnoses of comorbidities we were not able to adjust for individual-level comorbidities in the regression analyses.

Although the sum of ADLs and IADLs has been widely used as a measure of disability, there are several limitations to this outcome measure that are worth highlighting. First, it is recognised that the sum of ADLs and IADLs is likely to suffer from construct underrepresentation when used as a measure of disability (Buz and Cortes-Rodriguez, 2016). This refers to the fact that the ADL/IADL measure is likely to capture only part of the disability construct that we are truly interested in. That is, the sum of ADLs and IADLs is likely to provide only an imperfect measure of functional disability, and is more likely to provide a measure of functional independence. Second, a measure based on ADLs and IADLs is likely to exhibit differential item functioning, especially with regards to age (Buz and Cortes-Rodriguez, 2016; Fleishman et al., 2002; LaPlante, 2010). This suggests that the response probabilities for specific ADL and IADL items may be affected by age-related characteristics of the individual that are not related to the underlying disability level. Whether this leads to the sum of ADLs and IADLs being a biased measure of the severity of disability is however uncertain (Buz and Cortes-Rodriguez, 2016; LaPlante, 2010).

The analyses in this study were based on novel statistical methodology, known as joint modelling. Joint models have been widely discussed in the statistical literature in the last decade

(Crowther et al., 2012; Lawrence Gould et al., 2014; Sweeting and Thompson, 2011; Tsiatis and Davidian, 2004), yet their presence within the epidemiological literature remains limited. The slow uptake of joint modelling methods in the epidemiological literature is likely a consequence of their recent development, their higher degree of complexity compared with standard regression methods, and their only very recent availability in mainstream statistical software. Our study highlights the usefulness of these methods for epidemiological research. We anticipate that joint models will become more widely adopted by epidemiologists now that implementations are increasingly available for standard statistical software packages (Lawrence Gould et al., 2014).

In conclusion, this study found no evidence of an association between community-level disaster exposure and individual-level changes in either disability or the risk of death. Nonetheless, due to the limitations of the exposure variable, these findings should not be used as a basis for policy decisions regarding the long-term assistance provided to disaster-affected communities. Rather, our findings suggest that future research should focus on individual-level disaster exposures, moderate to severe events or disasters of a common type, and potentially consider a focus on higher-risk groups of individuals within the community.

Disclaimer

This work does not necessarily reflect the view of the US Government or the Department of Veterans Affairs.

Funding support

This work was supported by National Institutes of Health (NIH) grant R21AG044752. The Health and Retirement Study is funded by the National Institute on Aging (U01 AG009740), and performed at the Institute for Social Research, University of Michigan. SLB is funded by an Australian National Health and Medical Research Council (NHMRC) Postgraduate Scholarship (APP1093145).

Acknowledgements

We thank Vanessa Dickerman for her expert production of the analytic files from the HRS.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.socscimed.2016.12.007.

References

Aldrich, T.K., Gustave, J., Hall, C.B., et al., 2010. Lung function in rescue workers at the World Trade Center after 7 years. N. Engl. J. Med. 362 (14), 1263–1272.
Buz, J., Cortes-Rodriguez, M., 2016. Measurement of the severity of disability in community-dwelling adults and older adults: interval-level measures for accurate comparisons in large survey data sets. BMJ Open 6 (9), e011842.
Chan, C.S., Lowe, S.R., Weber, E., et al., 2015. The contribution of pre- and post-disaster social support to short- and long-term mental health after Hurricanes Katrina: a longitudinal study of low-income survivors. Soc. Sci. Med. 138, 38–43.
Core Team, R., 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Crowther, M.J., Abrams, K.R., Lambert, P.C., 2012. Flexible parametric joint modelling of longitudinal and survival data. Stat. Med. 31 (30), 4456–4471.
DiGrande, L., Neria, Y., Brackbill, R.M., et al., 2011. Long-term posttraumatic stress symptoms among 3,271 civilian survivors of the september 11, 2001, terrorist attacks on the World trade center. Am. J. Epidemiol. 173 (3), 271–281.
Federal Emergency Management Agency. Disaster Declarations by Year. US Department of Homeland Security. https://www.fema.gov/disasters/grid/year. Accessed 11 January 2016a.
Federal Emergency Management Agency, 2016. FEMA Public Assistance Funded Projects Detail - Open Government Initiative. US Department of Homeland Security. https://www.fema.gov/media-library/assets/documents/28331. Accessed 10 March 2016b.
Fleishman, J.A., Spector, W.D., Altman, B.M., 2002. Impact of differential item functioning on age and gender differences in functional disability. J. Gerontology Ser. B-Psychological Sci. Soc. Sci. 57 (5), S275–S284.
Galea, S., Nandi, A., Vlahov, D., 2005. The epidemiology of post-traumatic stress disorder after disasters. Epidemiol. Rev. 27, 78–91.
Guha-Sapir, D., Hoyois, P., Below, R., 2015. Annual Disaster Statistical Review 2014: the Numbers and Trends. WHO Centre for Research on the Epidemiology of Disasters (CRED), Brussels, Belgium.
Henderson, R., Diggle, P., Dobson, A., 2000. Joint modelling of longitudinal measurements and event time data. Biostatistics 1 (4), 465–480.
Hendrickson, L.A., Vogt, R.L., 1996. Mortality of kauai residents in the 12-month period following hurricane iniki. Am. J. Epidemiol. 144 (2), 188–191.
Hikichi, H., Aida, J., Tsuboya, T., et al., 2016. Can community social cohesion prevent posttraumatic stress disorder in the aftermath of a disaster? A natural experiment from the 2011 tohoku earthquake and tsunami. Am. J. Epidemiol. 183, 902–910.
Institute of Medicine, 2015. Healthy, Resilient, and Sustainable Communities after Disasters: Strategies, Opportunities, and Planning for Recovery. The National Academies Press, Washington, DC.
LaPlante, M.P., 2010. The classic measure of disability in activities of daily living is biased by age but an expanded IADL/ADL measure is not. J. Gerontology Ser. B-Psychological Sci. Soc. Sci. 65 (6), 720–732.
Lawrence Gould, A., Boye, M.E., Crowther, M.J., et al., 2014. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat. Med. 34, 2181–2195.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, second ed. Wiley, Hoboken, N.J.
Marioni, R.E., Proust-Lima, C., Amieva, H., et al., 2014. Cognitive lifestyle jointly predicts longitudinal cognitive decline and mortality risk. Eur. J. Epidemiol. 29 (3), 211–219.
Marioni, R.E., Proust-Lima, C., Amieva, H., et al., 2015. Social activity, cognitive decline and dementia risk: a 20-year prospective cohort study. BMC Public Health 15, 1089.
North, C.S., Pfefferbaum, B., 2013. Mental health response to community disasters: a systematic review. JAMA 310 (5), 507–518.
Rizopoulos, D., 2012. Joint Models for Longitudinal and Time-to-event Data: with Applications in R. CRC Press.
Rizopoulos, D., 2016. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. J. Stat. Softw. 72 (7), 1–46.
Sampson, L., Lowe, S.R., Gruebner, O., et al., 2016. Perceived service need after hurricane sandy in a representative sample of survivors: the roles of community-level damage and individual-level stressors. Disaster Med. Public Health Prep. 10, 428–435.
Sastry, N., Gregory, J., 2013. The effect of Hurricane Katrina on the prevalence of health impairments and disability among adults in New Orleans: differences by age, race, and sex. Soc. Sci. Med. 80, 121–129.
Smith, J.P., 1995. Racial and ethnic differences in wealth in the health and retirement study. J. Hum. Resour. 30, S158–S183.
Sonnega, A., Faul, J.D., Ofstedal, M.B., et al., 2014. Cohort profile: the health and retirement study (HRS). Int. J. Epidemiol. 43 (2), 576–585.
Sweeting, M.J., Thompson, S.G., 2011. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. Biom J. 53 (5), 750–763.
Tsiatis, A.A., Davidian, M., 2004. Joint modeling of longitudinal and time-to-event data: an overview. Stat. Sin. 14 (3), 809–834.
Wade, T.J., Sandhu, S.K., Levy, D., et al., 2004. Did a severe flood in the midwest cause an increase in the incidence of gastrointestinal symptoms? Am. J. Epidemiol. 159 (4), 398–405.
World Health Organization, 2001. International Classification of Functioning, Disability and Health : ICF. World Health Organization, Geneva, Switzerland.
Wulfsohn, M.S., Tsiatis, A.A., 1997. A joint model for survival and longitudinal data measured with error. Biometrics 53 (1), 330–339.

Chapter 4: Application of a latent class joint model: BMI trajectories and risk of death or transplant in haemodialysis

4.1 Chapter introduction

In end-stage kidney disease (ESKD) patients are at high-risk of mortality. However, within ESKD patients undergoing dialysis treatment, it has been commonly observed that overweight and obese patients seem to have better survival outcomes relative to those patients who are normal or underweight (Fleischmann et al., 1999; Kalantar-Zadeh et al., 2005; Park et al., 2014). Since this relationship between weight and survival is the reverse of what is observed in the general population, it has been termed the "obesity paradox" (Sperrin et al., 2016). While a number of possible explanations have been proposed for the obesity paradox, there still has not been any clear consensus on what is its primary explanation (Schmidt and Salahudeen, 2007).

One potential explanation for the obesity paradox is that previous studies have predominantly considered baseline or time-fixed measures of weight and have not investigated how longitudinal changes in body mass index (BMI) may be associated with mortality. In the study presented in this chapter, a latent class joint modelling approach was used to explore in patients on kidney dialysis the association between longitudinal changes in BMI and the rates of two competing events: kidney transplantation or death without transplantation.

The study was conducted using a large registry-based population that included all ESKD patients initiating haemodialysis in Australia and New Zealand over a ten-year period. It was hypothesised that a latent class joint modelling approach would allow identification of underlying heterogeneous groups of haemodialysis patients that differed in terms of both

their marginal longitudinal BMI trajectories and their associated rates of the competing event outcomes.

The main content of this chapter is presented in the next section, in the form of an applied research paper that has been recently submitted and is currently under review: The model described in the paper was estimated using the **lcmm** (Proust-Lima et al., 2017) R package. The supplementary material for the paper is provided in Appendix B of this thesis.

4.2 Manuscript

This section herein contains the following applied research paper:

> Brilleman SL, Moreno-Betancur M, Polkinghorne KR, McDonald SP, Crowther MJ, Thomson J, Wolfe R. Longitudinal changes in body mass index and the competing outcomes of death and transplant in patients undergoing hemodialysis: a joint latent class mixed model approach. *Submitted for publication*.

Longitudinal changes in body mass index and the competing outcomes of death and transplant in patients undergoing hemodialysis

Samuel L. Brilleman[1,2], Margarita Moreno-Betancur[1,2,3,4], Kevan R. Polkinghorne[1,5,6], Stephen P. McDonald[7,8], Michael J. Crowther[9], Jim Thomson[10], Rory Wolfe[1,2]

Author affiliations: [1]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; [2]Victorian Centre for Biostatistics (ViCBiostat), Melbourne, Australia; [3]Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne, Australia; [4]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia; [5]Department of Nephrology, Monash Medical Centre, Melbourne, Australia; [6]Department of Medicine, Monash University, Melbourne, Australia; [7]ANZDATA Registry, SA Health and Medical Research Institute, Adelaide, Australia; [8]Department of Medicine, University of Adelaide, Adelaide, Australia; [9]Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, UK; [10]Arthur Rylah Institute for Environmental Research, Department of Environmental, Land, Water and Planning, Victoria, Australia

***Corresponding author:** sam.brilleman@monash.edu

Abstract

Background: The relationship between body mass index (BMI) and patient survival in end-stage kidney disease in not well understood and has been the subject of much debate over recent years.

Methods: This study used a latent class joint modelling approach to identify latent groups that underpinned associations between patterns of change in BMI during hemodialysis and two competing events: transplant and death without transplant. We included all adult patients who initiated chronic hemodialysis treatment in Australia or New Zealand between 2005 and 2014.

Results: There were 16,414 patients included in the analyses; 2,365 (14%) received a transplant, 5,639 (34%) died prior to transplant, and 8,410 (51%) were administratively censored. Our final model characterised patients based on five broad patterns of weight change (BMI trajectories): "late BMI decline" (about two years after commencing hemodialysis); "rapid BMI decline" (immediately after commencing hemodialysis); "stable and normal/overweight BMI"; "stable and morbidly obese BMI"; or "increasing BMI". Mortality rates were highest amongst classes with declining BMI, and the timing of weight loss coincided with the timing of increases in mortality. Within the two stable BMI classes, death rates were slightly lower amongst the morbidly obese.

Conclusions: Our findings provide some evidence to support the paradoxical protective effect of obesity against mortality. However, they also suggest that the shape of the BMI trajectory is important, with stable BMI trajectories being beneficial. Future research should be aimed at understanding the causes of weight changes during dialysis, to determine whether there could be strategies to improve patient survival.

Introduction

The relationship between body mass index (BMI) and patient survival in end-stage kidney disease (ESKD) has been the subject of intense interest and debate over recent years [1–3]. This has predominantly occurred as the result of consistent findings suggesting that, amongst patients undergoing hemodialysis treatment, those who are overweight and obese have superior survival outcomes than those who are normal or underweight [1,4]. This apparent protective effect of obesity in the hemodialysis population has been labelled the "obesity paradox" or more generally "reverse epidemiology", owing to the fact that it is the opposite of what is seen in the general population where obesity is associated with increased mortality [5,6]. Interestingly, similar

associations have also been reported for other chronic disease populations, for example, patients with cardiovascular disease or heart failure [7,8].

In the majority of studies involving patients with ESKD, body mass index (BMI) has been used as a measure of obesity. Moreover, early studies have predominantly considered baseline measures of BMI when examining the association between BMI and subsequent mortality risk [3]. However, an understanding of the relationship between BMI and mortality should ideally incorporate information about changes in BMI over the course of dialysis. It is likely that the mortality risk of a patient undergoing dialysis is related not only to their BMI at the time of initiating treatment, but also related to changes in their BMI leading up to their death or otherwise. Moreover, considering longitudinal changes in BMI and their association with mortality helps direct us towards a more meaningful clinical question, since it provides insight as to whether we might still be able to intervene (in the BMI pathway) even after the commencement of dialysis treatment [9]. Whether intervening on the BMI pathway is in fact able to provide a survival benefit for patients depends on the causal mechanisms underpinning any associations between weight change and mortality. Nonetheless, a first step is gaining an understanding of whether such associations (causal or otherwise) exist.

Recent studies assessing the effect of early weight change after commencement of hemodialysis suggest that weight change is associated with survival (improved survival with weight gain, reduced survival with weight loss) [10–12]. In another study weight gain provided a survival benefit to patients with normal or low BMI, but not to obese subjects; however, this was assessed amongst prevalent hemodialysis patients rather than from hemodialysis commencement [13].

A rapidly evolving field of statistical methodology, known as joint modelling, allows for the joint analysis of longitudinal changes in an outcome (here a clinical biomarker such as BMI) and the time to an event of interest [14–17]. By jointly modelling the longitudinal and event outcomes, one can estimate the association between the two processes, as well as – under certain parametric modelling assumptions – the expected longitudinal process accounting for non-random dropout due to the event. More specifically, a latent class joint modelling approach is one in which underlying "latent" classes are assumed to capture the association between longitudinal changes in the biomarker and the risk of the event [18]. A latent class can be thought of as an underlying class (or group) of patients who are similar in their longitudinal marker trajectories (BMI over the course of hemodialysis) and event-risk profiles. Several latent classes are believed to exist, and although an individual's class membership is not observable, it can be modelled. Latent class joint modelling approaches have been extended to model multiple competing events simultaneously; for example, receipt of a kidney transplant and death without transplant [19].

In this study, we used a latent class joint modelling approach to characterise hemodialysis patients by BMI trajectories over the course of treatment and their association with the rates of transplant and death without transplant. We thus identified latent classes that underpin the associations between longitudinal BMI trajectories and the event rates. We used registry data that included all patients in Australia and New Zealand who initiated chronic hemodialysis treatment over a ten year period.

Materials and methods

Data and sample

The Australia and New Zealand Dialysis and Transplant Registry (ANZDATA) collects data from all Australian and New Zealand renal units on a wide range of characteristics related to patients undergoing renal replacement therapy (RRT). Comprehensive demographic and comorbidity data is collected on each patient at the initiation of RRT. Longitudinal data is collected ongoingly, for example, dates on which a patient's transplant status, survival status, or dialysis modalities are changed. In addition ANZDATA conducts annual surveys in which clinicians provide clinical information on each patient in the registry, for example, their most recent weight or treatment measurements [20].

This study included all patients aged 18 years and above who initiated dialysis in

Australia and New Zealand between 1/1/2005 and 31/12/2014 and who, at 90 days after initiating dialysis (or at the time of death, transplant or censoring if this occurred earlier than 90 days), were recorded as receiving hemodialysis. We focus on hemodialysis since it is the more common modality worldwide (compared with peritoneal dialysis) and the majority of studies reporting the obesity paradox were based on hemodialysis patients. Each patient was followed up from the initiation of their dialysis treatment until the earliest of: (i) transplant; (ii) death without transplant; or (iii) administrative censoring (occurring on the earliest date of either 31/12/2014 or 5 years after initiating dialysis). The ANZDATA cohort is generally considered to be non-transient (due to age, health status, and nature of the treatment) and so we did not consider loss to follow up. Moreover, since only 3049 patients (18% of the analysis sample) remained at risk beyond 5 years, it was deemed an appropriate maximum follow up time, to ensure we did not draw inferences based on time periods with much fewer events.

Measurements of BMI during the follow up period were derived for each patient using their baseline height measured at the initiation of dialysis and longitudinal dry weight measurements reported as part of each ANZDATA survey. The weights reported in the survey are recorded by nursing staff. Clinically, they are used as the endpoint for dialysis treatment. Patients are weighed prior to treatment and the difference between that weight and their intended post-treatment dry weight ("ideal body weight") is used to determine the fluid removal during dialysis.

Diabetes was coded as yes (type I or II) or no, and each of the remaining comorbidities as yes (including suspected) or no. As the registry's intent is to collect data on long-term chronic dialysis use for ESKD, our analyses excluded those patients who were recorded as having recovered kidney function after initiating dialysis. In addition, we excluded patients with missing comorbidity status, an unspecified race, a missing or extreme (<130 cm) baseline height measurement, or an extreme (≥45, <17.5 kg/m^2) BMI measurement. Figure S1 in the Supplementary Materials illustrates derivation of the analysis sample.

Modelling

Our latent class joint modelling approach aimed to capture the association between an individual's unobserved "true" longitudinal BMI trajectory whilst undergoing hemodialysis treatment and their corresponding time to occurrence of either of the competing events: (i) transplant; or (ii) death without transplant. The modelling approach consisted of the following regression submodels.

Class membership submodel

We assumed that each individual i ($i = 1, \ldots, N$) belonged to one of a set of possible latent classes which differentiate types of BMI trajectory as well as being linked to the transplant and death outcomes. Since the true latent class membership for individual i is unknown, the random variable c_i, the latent class to which individual i belongs, remains unobserved and we model the probabilities of individual i belonging to each of the possible latent classes through a multinomial distribution where $P(c_i = g) = \pi_{ig}$ for latent classes $g = 1, \ldots, G$. This model specification assumes that an individual's latent class membership does not depend on any of their observed characteristics.

Longitudinal submodel for BMI trajectories

The longitudinal BMI trajectory for individual i is modelled using a class-specific linear mixed effects model. We let BMI_{ij} be the observed BMI measurement for individual i at time t_{ij} ($j = 1, \ldots, n_i$). To improve numerical stability and aid convergence of the model estimation procedure we model $y_{ij} = \frac{BMI_{ij} - 25}{10}$ where the standardizing of each BMI measurement around 25 kg/m^2 and per 10 kg/m^2 merely scales the magnitude of all covariate effects in the subsequent model without affecting their relative values (in the results section we back-transform to enable interpretation on the BMI scale). We assume that outcomes in class g, $y_{ij}|_{c_i=g}$, follow:

$$y_{ij}|_{c_i=g} = X_{ij}{}' \beta_g + u_{ig} + \varepsilon_{ij}$$

where X_{ij} denotes a vector of covariates evaluated at time t_{ij} with associated class-specific fixed effects β_g, u_{ig} denotes a class-

specific random intercept term for individual i where we assume $u_{ig} \sim N(0, \sigma_u^2)$, and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ represents residual error. We allow for flexibility in the longitudinal trajectories through the use of cubic splines. Specifically, the vector X_{ij} contains both an intercept term and the time-dependent basis terms for natural cubic splines with two degrees of freedom and accordingly, the vector of class-specific parameters β_g allows each latent class g to have its own class-specific non-linear BMI trajectory. The individual-level random effects u_{ig} allow for the correlation between repeated measurements on the same patient. In general the individual-level random effects can be extended beyond a random intercept, however we elected not to do so since the number of BMI measurements per individual was modest. Moreover, we did not include patient characteristics in the longitudinal submodel for the BMI trajectory. Some additional discussion around the choice of model structure is contained in the Supplementary Materials.

Competing risks time-to-event submodel for transplant or death without transplant

For the competing events of transplant and death without transplant we assumed class-specific Weibull proportional cause-specific hazard models. The class-specific hazard of death without transplant occurring at time t for individual i was modelled as

$$h_{i1}(t)|_{c_i=g} = h_{01g}(t) exp\,(Z_i{}'\delta_1)$$

where Z_i denotes the vector of baseline covariates for individual i with associated vector of coefficients δ_1, and $h_{01g}(t)$ denotes the class-specific baseline hazard of death without transplant evaluated at time t, assumed to have the form of the hazard of a Weibull-distributed time-to-event outcome. Similarly, the class-specific hazard of transplant occurring at time t for individual i is

$$h_{i2}(t)|_{c_i=g} = h_{02g}(t) exp\,(Z_i{}'\delta_2)$$

where δ_2 and $h_{02g}(t)$ are the corresponding vector of coefficients and class-specific baseline hazard for the competing event of transplant. The magnitude of the covariate effects, captured by the coefficient vectors δ_1 and δ_2, were assumed to be constant across latent classes. In both event submodels the vector Z_i includes covariates for age, gender, race (Caucasian, Aboriginal/Torres Strait Islander, Asian, Maori/Pacific Islander), calendar period for initiating RRT (2005-09, 2010-14; to capture temporal changes in risk that may have resulted from improvements in technologies and treatment over time), primary cause of renal disease (diabetes, glomerulonephritis, hypertension, other/uncertain), and comorbidity indicators (chronic lung disease, coronary artery disease, cerebrovascular disease, peripheral vascular disease, diabetes). Age was included in the model as a linear term (on the log hazard scale) for the number of years greater than age 50 at the start of RRT; this provided a parsimonious way to accommodate the fact that age showed almost no association with the event outcomes below age 50, but a strong association with the event outcomes at older ages.

Model estimation and comparison

Our interpretation of the joint model focuses on the association between the "typical" longitudinal BMI trajectory for latent class g and the corresponding hazard functions for each of the competing events for latent class g. An understanding of the association is most easily obtained by plotting the predicted trajectories and hazard functions for each latent class.

An important aspect of the modelling is determining the appropriate number of latent classes. The number of latent classes must be specified explicitly, however, the true number of latent classes is unknown. We compared models with varying numbers of latent classes using the Bayesian information criterion (BIC). BIC indicates the relative *marginal* likelihood of competing models, where the marginal likelihood penalizes free parameters. The model with the smallest BIC has the greatest support in the data [21]. A further consideration is meaningfulness of the latent classes. A sufficient number of patients in each latent class would lend clinical usefulness. Moreover, the model should provide classes that, from a clinical perspective, are sufficiently differentiated from one another to be informative. Although the model does not allocate each patient to a single latent class (rather it estimates their probability of being included in each latent class), we can allocate each patient to the class they have the highest estimated probability of being in. Our choice of

final model can then be informed by: (i) the number of patients allocated to each latent class in this way, and (ii) the magnitude and clinical importance of differences between the latent classes (with regard to the predicted longitudinal trajectories and hazard functions).

We used the lcmm package in R [22,23]. Further computational details are contained in the Supplementary Materials. They include the model code, a description of the initial values used, computation time, a description of alternative model structures that we considered (e.g. a more complex individual-level random effects structure), and an assessment of goodness of fit for the final model. A GRoLTS (Guidelines for Reporting on Latent Trajectory Studies) checklist is also included in the Supplementary Materials [24].

Results

Descriptive statistics

A total of 19,264 patients initiated hemodialysis during the study period. Following exclusions (see Supplementary Materials) there were 16,414 patients with at least one BMI measurement recorded prior to death, transplant or censoring who were included in the main analyses. Median (IQR) follow-up time was 2.3 (1.0, 4.3) years. The percentage of patients with 1, 2, 3, 4, and 5 BMI measurements respectively (prior to death, transplant or censoring) was 18%, 21%, 17%, 14% and 30%. Not all patients had their first BMI measurement taken at baseline (i.e. at the time of their initiating dialysis); the median (IQR) time to the first BMI measurement was 0.45 (0.22, 0.71) years. Figure S2 (Supplementary Materials) shows observed BMI trajectories for a random sample of patients.

Of the 16,414 patients, 2,365 (14%) received a transplant, 5,639 (34%) died, and 8,410 (51%) were censored. Figure 1 shows crude cumulative incidence curves for the two competing events. An estimated 18% of patients receive a kidney transplant by 5 years while 44% die (without transplant) within the same period.

Determining the number of latent classes

Table 1 shows a comparison of the models for a varying number of latent classes. The Bayesian information criterion (BIC) values suggested that higher numbers of latent classes consistently resulted in better fitting models (i.e. smaller BIC). This reflects the fact that in large datasets more complex models can easily appear justified based on statistical criterion alone. After examining the predicted BMI trajectories and hazard functions, we concluded that greater than five latent classes resulted in less useful and clinically meaningful characterisations (see Supplementary Materials).

Our chosen model had five classes. The estimated mean probability of an individual being in their specified class (that is, the class for which they had the highest probability) ranged between 0.73 and 0.89 (Table S1, Supplementary Materials) and the relative entropy for the model was 0.745 (Table 1), suggesting that this model provided good discrimination. With six or more latent classes the relative entropy decreased slightly (Table 1).

Modelling

The left panel of Figure 2 shows the predicted longitudinal trajectories for each of the five latent classes. These are population average predictions, since the longitudinal submodel did not adjust for demographic or clinical characteristics.

The first class contained the majority (75%) of patients, and was characterised by relatively "stable and normal/overweight BMI" over the course of hemodialysis. The second class contained 14% of patients and was categorised by "stable and morbidly obese BMI". The next two classes were both characterised by declining BMI. One contained around 6% of patients and was characterised by slower and delayed decline in BMI (during the first two years of hemodialysis BMI remained relatively stable) whilst the other, which contained only 2% of patients, was characterised by rapid decline in BMI from the initiation of hemodialysis. We refer to these classes as "late BMI decline" and "rapid BMI decline". The fifth class, which contained 4% of patients, was the only class characterised by "increasing BMI", and the average increase was

approximately 5 kg/m^2 over the first 3 years of hemodialysis.

Table 2 shows the characteristics of the patients, overall and stratified by their latent class membership. Overall, almost two thirds of the cohort were male, the majority were Caucasian, and comorbidities were common. Patients in the "increasing BMI" and "stable and morbidly obese BMI" classes were less likely to be in the oldest age groups (>70 years), were more likely to be Maori/Pacific Islander, have diabetes at initiation of RRT, and have diabetes as their primary cause of kidney disease. The characteristics of the "rapid BMI decline" class did not appear to differ from those of the "late BMI decline" class. Patients in the "stable and normal/overweight BMI" class were more likely to be in the older age groups, more likely to be Asian or Caucasian, and less likely to be diabetic. There were no large differences in the rates of current, former, and never smokers across the five latent classes. Patients with a single BMI measurement were almost exclusively allocated to the stable BMI classes.

In Figure 2, we present class-specific estimates of the cause-specific hazard functions for transplant and death without transplant in the right and centre panels, respectively. These show the instantaneous rate (i.e. "hazard") of each event at a given time *t*, amongst those individuals who are still at risk of the event. The estimates shown are for a Caucasian male aged ≤50 years, with diabetic nephropathy, cerebrovascular disease and coronary artery disease, initiating dialysis between 2005-09. However, since our competing risks event submodels assumed proportional hazards, the relative comparison between latent classes will be similar regardless of a patient's covariate profile.

The highest rate of death without transplant was observed in the "rapid BMI decline" class, with a dramatic increase from the initiation of dialysis concurrent with the drop in BMI. This class also had an increasing rate of transplant receipt, beginning from the initiation of dialysis. The "late BMI decline" class had a relatively low rate of death without transplant during early treatment, however, once weight loss started to occur the rate of death increased.

The two stable BMI classes had similar rates of death without transplant, but with the morbidly obese having slightly better survival than the normal/overweight. These two classes had different rates of transplant. The "stable and normal/overweight BMI" class had a relatively high incidence of transplant. Conversely, and as might be expected due to the relative contraindication of obesity to kidney transplantation, the "stable and morbidly obese BMI" class had a lower rate of transplant. The "increasing BMI" class generally had the lowest estimated rate of death without transplant and was the only group that showed a decreasing rate of transplant over the duration of dialysis.

Table 3 shows the estimated hazard ratios quantifying the associations between the baseline covariates and the rate of each competing event. Through the latent class joint modelling approach, the hazard ratios are also adjusted for the BMI trajectories. There was strong evidence that comorbidities increased the rate of death without transplant and decreased the rate of transplant. Initiating dialysis in more recent years decreased the rate of death without transplant, and increased the rate of transplant. Maori/Pacific Islanders or Aboriginal/Torres Strait Islanders had much lower rates of transplant receipt compared with Caucasians, however, they had only a slightly higher rate of death without transplant.

In the Supplementary Materials we present the cumulative incidence functions for each event. In contrast to the cause-specific hazard functions, the cumulative incidence functions show the cumulative risk (i.e. probability) of the event having occurred at any point up to time *t*. Although cumulative incidence functions can be useful, particularly for understanding patient prognosis, they are less aligned with understanding potential etiological associations [25].

Discussion

There is currently a limited understanding of the paradoxical relationship observed between BMI and mortality risk in ESKD patients undergoing hemodialysis when compared with the general population. This study investigated the association between longitudinal changes in BMI over the course of hemodialysis treatment

and the risk of kidney transplant and death without transplant using competing risks latent class joint models. This is one of only a few studies that have considered longitudinal changes in BMI rather than BMI values at the initiation of treatment alone. Similarly, it is one of the few studies to consider the competing risks of both death and transplant (treated here as distinct but competing events), which is particularly important given the marked effect of BMI on probability of receiving a kidney transplant, rather than considering BMI and its association with mortality only.

We identified five broad patterns of weight change (BMI trajectories) following commencement of hemodialysis with an assessment of survival. Two classes were characterised by a decline in BMI, either "late" (about two years after commencing hemodialysis) or immediately from the start of dialysis. In both groups, the decline in BMI was associated with increasing mortality rates and risks. In the late BMI decline group, the temporal increase in the risk of death matched the decline in BMI seen at approximately two years post dialysis initiation. This suggests that the weight loss is indicative of an underlying illness leading to increased mortality rates. Both groups exhibited the greatest rates of death by the end of study period.

Two of the remaining classes demonstrated stable weight during the study period and differed only by the starting weight; being either normal/overweight or in the (morbidly) obese range. Compared to the classes characterised by weight loss, death (without transplant) rates were lower amongst the stable BMI classes. Within the two stable BMI classes the death rates were slightly lower amongst the morbidly obese (potentially supporting the obesity paradox). However, these two classes with stable BMI differed quite markedly in their rates of transplantation. As might be expected from the relative contraindication of obesity to access to transplant programmes in Australia and New Zealand, obese patients appeared to have a much smaller chance of receiving a transplant.

The final class was characterised by a progressive increase in weight following dialysis initiation, increasing further into the morbidly obese range over the course of dialysis. This group had relatively low rates of death and a relatively high rate of transplant, although the former increased during dialysis while the latter decreased.

Amongst the baseline characteristics associated with the hazard of death, we found patients initiating dialysis more recently had better survival. This finding is consistent with the reduction in mortality seen in dialysis patients globally. Although there is no clear cause, likely contributors include improved cardiovascular disease treatment.

Previous studies report that greater BMI or obesity at dialysis initiation appears protective against mortality in ESKD [1–4]. More recent work suggested that early weight gain in the first year after dialysis initiation was associated with improved survival with risk similar when stratified across BMI groups [12]. Another study suggested that any benefit in weight gain was seen only in patients with normal or low BMI at dialysis initiation and not in obese subjects [13]. However these studies did not treat or consider kidney transplantation as a competing risk. This study therefore adds to previous work aimed at explaining or disentangling the apparent paradoxical relationship between increased BMI and survival in dialysis patients. Here we show that it is the longitudinal weight change that is associated with differences in patient mortality rates and risk, with a beneficial effect of stable BMI that was consistent irrespective of the BMI at dialysis commencement (normal weight or morbidly obese).

Although our final analysis focuses on the five-class model, it must be recognised that the choice of final model involves some subjectivity. Moreover, our final model included only a small percentage of patients in each of the classes with a non-stable BMI trajectory. Regardless of the chosen number of classes, we found that the majority of patients belonged to stable BMI trajectories: either a single stable BMI class containing the majority of patients (as in the two- through four-class models) or they were separated into two stable BMI classes that differed based on baseline BMI (normal/overweight versus morbidly obese, as in the models with more than five classes).Our study used a large registry-based cohort of all adult patients who initiated hemodialysis in Australia or New Zealand over

a ten-year period. We consider the findings are generalizable to the wider hemodialysis population in countries with similar population BMI profiles and dialysis care. However, it is likely that our results do not generalise to patients on peritoneal dialysis. Due to the glucose present in peritoneal dialysis solutions, we expect that the BMI trajectories may differ across the two modalities, especially during the early phase of treatment.

Further strengths of this study include the use of an inception cohort (with very few exclusions due to missing data or loss to follow-up), assessment of longitudinal changes in BMI, and treating transplant as a competing risk. However, it is important to recognise that the observational data used in this study only allows us to infer associations but not causal relationships. Even though this study considered longitudinal changes in BMI, and these changes in BMI preceded the event(s) of interest, our results do not imply that, for example, decreases in BMI caused mortality events. Rather we simply found that mortality events appeared to occur at a higher rate amongst those individuals who had decreases in their BMI, even after adjusting for the effects of demographics and comorbidities. For example, we cannot conclude whether deliberate weight loss among those who are overweight or obese is associated with the same increase in mortality as we observed in our cohort.

Some limitations of this study need acknowledgment. The cohort included those individuals receiving hemodialysis 90 days after initiating any RRT for the first time (or at the point of death or transplant if that occurred earlier than 90 days), however, modality switching means that some individuals in our analysis did not remain on hemodialysis over the entire course of RRT. Specifically, there were 1,442 (9%) patients who received peritoneal dialysis at some time after 90 days. This is in line with previous studies which have shown that when modality switches occur, they predominantly occur earlier in the treatment period [26]. Another important aspect is the potential for bias due to selecting patients based on their modality type post-baseline (sometimes known as an immortal time bias). However, in our study a patient was still able to have an event during the first 90 days, so we would only expect an immortal time bias if their modality in the first 90 days is related to their mortality risk during that same period. Moreover, the number of events during the first 90 days was not large (562 deaths, 179 transplants) and, therefore, we would expect any bias due to our conditioning on 90-day modality status to be small.

One additional consideration is that patients required at least one BMI measurement to be included in the analysis. Excluding individuals without any BMI measurement (e.g. if they died before a BMI measurement was obtained) may have induced a form of immortal time bias, however, we expect any such bias to be small since there were only 171 patients who did not have any BMI measurement. Lastly, our ability to accurately capture patient-specific longitudinal trajectories may have been improved by more frequent measurements of BMI.

This study provides evidence to support the paradoxical protective effect of obesity against mortality. Amongst hemodialysis patients with a stable BMI trajectory, morbidly obese patients appeared to have slightly better survival rates than normal or overweight patients. However, our findings also suggest that the shape of the BMI trajectory is important, with stable BMI trajectories being beneficial. Progressive weight loss either following dialysis initiation or after two years of dialysis treatment, was strongly associated with poorer patient survival. However, the mechanism or effects leading to the weight loss cannot be determined by this study. The processes that lead to weight loss during dialysis are likely to be important in determining whether targeted interventions aimed at maintaining weight would improve probability of survival. Therefore future research should be aimed at understanding the causal mechanisms that lead to significant weight changes during dialysis treatment, to help determine whether targeted interventions could be successful in improving patient survival.

Conflicts of interest

The authors declare that they have no conflict of interest.

Availability of data

Availability of data: The Australia and New Zealand Dialysis and Transplant Registry (ANZDATA) dataset used in this study is not publicly available. ANZDATA data is available for research purposes, but an application must be submitted to, and approved by, the executive committee.

Funding support

SLB is funded by an Australian National Health and Medical Research Council (NHMRC) Postgraduate Scholarship (ref: APP1093145), with additional support from an NHMRC Centre of Research Excellence grant (ref: 1035261) awarded to the Victorian Centre for Biostatistics (ViCBiostat). MJC is partly funded by a UK Medical Research Council (MRC) New Investigator Research Grant (MR/P015433/1). The ANZDATA Registry is funded by the Australian Organ and Tissue Donation and Transplantation Authority, the New Zealand Ministry of Health and Kidney Health Australia.

References

1. Schmidt DS, Salahudeen AK. Cardiovascular and Survival Paradoxes in Dialysis Patients: Obesity-Survival Paradox-Still a Controversy? Semin Dial. 2007;20:486–92.
2. Kalantar-Zadeh K, Abbott KC, Salahudeen AK, Kilpatrick RD, Horwich TB. Survival advantages of obesity in dialysis patients. Am J Clin Nutr. 2005;81:543–54.
3. Park J, Ahmadi S-F, Streja E, Molnar MZ, Flegal KM, Gillen D, et al. Obesity Paradox in End-Stage Kidney Disease Patients. Prog Cardiovasc Dis. 2014;56:415–25.
4. Fleischmann E, Teal N, Dudley J, May W, Bower JD, Salahudeen AK. Influence of excess weight on mortality and hospital stay in 1346 hemodialysis patients. Kidney Int. 1999;55:1560–7.
5. Sperrin M, Candlish J, Badrick E, Renehan A, Buchan I. Collider Bias Is Only a Partial Explanation for the Obesity Paradox: Epidemiology. 2016;27:525–30.
6. Kalantar-Zadeh K, Block G, Humphreys MH, Kopple JD. Reverse epidemiology of cardiovascular risk factors in maintenance dialysis patients. Kidney Int. 2003;63:793–808.
7. Banack HR, Kaufman JS. The obesity paradox: Understanding the effect of obesity on mortality among individuals with cardiovascular disease. Prev Med. 2014;62:96–102.
8. Oga EA, Eseyin OR. The Obesity Paradox and Heart Failure: A Systematic Review of a Decade of Evidence. J Obes. 2016;2016:1–9.
9. Vansteelandt S. Asking too much of epidemiologic studies: the problem of collider bias and the obesity paradox. Epidemiology. 2017;1.
10. Kalantar-Zadeh K, Streja E, Molnar MZ, Lukowsky LR, Krishnan M, Kovesdy CP, et al. Mortality Prediction by Surrogates of Body Composition: An Examination of the Obesity Paradox in Hemodialysis Patients Using Composite Ranking Score Analysis. Am J Epidemiol. 2012;175:793–803.
11. Kalantar-Zadeh K, Streja E, Kovesdy CP, Oreopoulos A, Noori N, Jing J, et al. The Obesity Paradox and Mortality Associated With Surrogates of Body Size and Muscle Mass in Patients Receiving Hemodialysis. Mayo Clin Proc. 2010;85:991–1001.
12. Chang TI, Ngo V, Streja E, Chou JA, Tortorici AR, Kim TH, et al. Association of body weight changes with mortality in incident hemodialysis patients. Nephrol Dial Transplant. 2017;32:1549–58.
13. Cabezas-Rodriguez I, Carrero JJ, Zoccali C, Qureshi AR, Ketteler M, Floege J, et al. Influence of Body Mass Index on the Association of Weight Changes with Mortality in Hemodialysis Patients. Clin J Am Soc Nephrol. 2013;8:1725–33.
14. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53:330–9.
15. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. Stat Sin. 2004;14:809–34.
16. Rizopoulos D. Joint models for longitudinal and time-to-event data: with applications in R. Boca Raton: CRC Press; 2012.
17. Lawrence Gould A, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat Med. 2015;34:2181–95.
18. Proust-Lima C, Séne M, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. Stat Methods Med Res. 2014;23:74–90.
19. Proust-Lima C, Dartigues J-F, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. Stat Med. 2016;35:382–98.
20. Australia and New Zealand Dialysis and Transplant Registry. ANZDATA Data Forms [Internet]. 2015 [cited 2017 May 17]. Available from: http://www.anzdata.org.au/v1/data_forms.html
21. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978;6:461–4.

22. Proust-Lima C, Philipps V, Diakite A, Liquet B. lcmm: Extended mixed models using latent classes and latent processes. R package version: 1.7.7. 2017; Available from: https://cran.r-project.org/package=lcmm
23. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2017; Available from: https://www.R-project.org/
24. van de Schoot R, Sijbrandij M, Winter SD, Depaoli S, Vermunt JK. The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies. Struct Equ Model Multidiscip J. 2017;24:451–67.
25. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Stat Med. 2012;31:1089–97.
26. Kasza J, Wolfe R, McDonald SP, Marshall MR, Polkinghorne KR. Dialysis modality, vascular access and mortality in end-stage kidney disease: A bi-national registry-based cohort study. Nephrology. 2016;21:878–86.

Table 1. Model comparison for varying numbers of latent classes.

Number of latent classes	Log-likelihood	Number of parameters	BIC	Relative entropy[b]	Percentage of patients allocated to each latent class[a]							
					Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
2	-29395.1	47	59246	0.741	91.94	8.06						
3	-28570.3	56	57684	0.750	88.57	6.50	4.93					
4	-27974.2	65	56579	0.757	85.73	5.56	5.43	3.28				
5	-27434.3	74	55587	0.745	74.72	13.76	6.18	3.86	1.47			
6	-27036.3	83	54878	0.695	71.10	14.20	6.63	3.50	3.08	1.50		
7	-26804.3	92	54502	0.716	70.90	12.93	5.83	5.47	2.98	1.36	0.53	
8	-26546.2	101	54073	0.698	69.50	13.19	6.56	5.23	2.10	1.60	1.34	0.49

Abbreviations: BIC: Bayesian information criterion.

[a] Patients are allocated to the latent class that they have the highest posterior probability of being in.

[b] Relative entropy is calculated as $1 + \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\hat{\pi}_{ig}\log(\hat{\pi}_{ig})}{N \log G}$ where $\hat{\pi}_{ig}$ is the estimated posterior probability of individual i $(i = 1, \ldots, N)$ being in latent class g $(g = 1, \ldots, G)$.

Table 2. Patient characteristics, overall and stratified by latent class for the five-class model. Figures are number (and %) of patients, unless stated otherwise.

	Class description					
Characteristic	Stable and normal/overweight BMI	Stable and morbidly obese BMI	Late BMI decline	Rapid BMI decline	Increasing BMI	Total
Number of patients	12265 (74.7%)	2259 (13.8%)	1015 (6.2%)	241 (1.5%)	634 (3.9%)	16414
Number of BMI measurements per patient						
1	2546 (20.8%)	457 (20.2%)	7 (0.7%)	0 (0.0%)	6 (0.9%)	3016 (18.4%)
2	2904 (23.7%)	390 (17.3%)	54 (5.3%)	76 (31.5%)	29 (4.6%)	3453 (21.0%)
3	2146 (17.5%)	350 (15.5%)	157 (15.5%)	102 (42.3%)	93 (14.7%)	2848 (17.4%)
4	1544 (12.6%)	262 (11.6%)	278 (27.4%)	38 (15.8%)	115 (18.1%)	2237 (13.6%)
5	3125 (25.5%)	800 (35.4%)	519 (51.1%)	25 (10.4%)	391 (61.7%)	4860 (29.6%)
Baseline BMI						
Median (IQR)	25 (23, 28)	36 (34, 39)	31 (27, 34)	31 (26, 35)	32 (29, 35)	27 (23, 31)
Baseline BMI, categories[a]						
Underweight (less than 18.5 kg/m^2)	241 (2.0%)	0 (0.0%)	5 (0.5%)	2 (0.8%)	0 (0.0%)	248 (1.5%)
Normal weight (18.5 to 24.9 kg/m^2)	5476 (44.6%)	1 (0.0%)	149 (14.7%)	42 (17.4%)	60 (9.5%)	5728 (34.9%)
Overweight (25 to 29.9 kg/m^2)	4845 (39.5%)	9 (0.4%)	257 (25.3%)	61 (25.3%)	176 (27.8%)	5348 (32.6%)
Obese (30 kg/m^2 and over)	1703 (13.9%)	2249 (99.6%)	604 (59.5%)	136 (56.4%)	398 (62.8%)	5090 (31.0%)
Year of initiating dialysis						
2005 to 2009	6108 (49.8%)	972 (43.0%)	635 (62.6%)	108 (44.8%)	381 (60.1%)	8204 (50.0%)
2010 to 2014	6157 (50.2%)	1287 (57.0%)	380 (37.4%)	133 (55.2%)	253 (39.9%)	8210 (50.0%)
Age (years)	65 (52, 75)	60 (50, 68)	63 (52, 72)	63 (49, 71)	59 (50, 68)	63 (52, 73)
Median (IQR)						
Age (years), categories						
<30	478 (3.9%)	39 (1.7%)	29 (2.9%)	12 (5.0%)	19 (3.0%)	577 (3.5%)
31 to 40	734 (6.0%)	137 (6.1%)	58 (5.7%)	21 (8.7%)	47 (7.4%)	997 (6.1%)
41 to 50	1478 (12.1%)	395 (17.5%)	147 (14.5%)	32 (13.3%)	106 (16.7%)	2158 (13.1%)
51 to 60	2297 (18.7%)	637 (28.2%)	206 (20.3%)	47 (19.5%)	169 (26.7%)	3356 (20.4%)
61 to 70	2852 (23.3%)	633 (28.0%)	270 (26.6%)	68 (28.2%)	182 (28.7%)	4005 (24.4%)

71 to 80	3193 (26.0%)	367 (16.2%)	254 (25.0%)	42 (17.4%)	90 (14.2%)	3946 (24.0%)
>80	1233 (10.1%)	51 (2.3%)	51 (5.0%)	19 (7.9%)	21 (3.3%)	1375 (8.4%)
Gender						
Male	8127 (66.3%)	1274 (56.4%)	578 (56.9%)	127 (52.7%)	338 (53.3%)	10444 (63.6%)
Female	4138 (33.7%)	985 (43.6%)	437 (43.1%)	114 (47.3%)	296 (46.7%)	5970 (36.4%)
Race categories						
Caucasian	9336 (76.1%)	1393 (61.7%)	764 (75.3%)	173 (71.8%)	391 (61.7%)	12057 (73.5%)
Aboriginal/Torres Strait Islander	1145 (9.3%)	244 (10.8%)	93 (9.2%)	25 (10.4%)	97 (15.3%)	1604 (9.8%)
Asian	888 (7.2%)	57 (2.5%)	28 (2.8%)	3 (1.2%)	25 (3.9%)	1001 (6.1%)
Maori/Pacific Islander	896 (7.3%)	565 (25.0%)	130 (12.8%)	40 (16.6%)	121 (19.1%)	1752 (10.7%)
Chronic lung disease						
No	10135 (82.6%)	1823 (80.7%)	827 (81.5%)	181 (75.1%)	519 (81.9%)	13485 (82.2%)
Yes	2130 (17.4%)	436 (19.3%)	188 (18.5%)	60 (24.9%)	115 (18.1%)	2929 (17.8%)
Coronary artery disease						
No	7028 (57.3%)	1258 (55.7%)	576 (56.7%)	127 (52.7%)	350 (55.2%)	9339 (56.9%)
Yes	5237 (42.7%)	1001 (44.3%)	439 (43.3%)	114 (47.3%)	284 (44.8%)	7075 (43.1%)
Cerebrovascular disease						
No	10306 (84.0%)	1961 (86.8%)	874 (86.1%)	199 (82.6%)	539 (85.0%)	13879 (84.6%)
Yes	1959 (16.0%)	298 (13.2%)	141 (13.9%)	42 (17.4%)	95 (15.0%)	2535 (15.4%)
Peripheral vascular disease						
No	9165 (74.7%)	1583 (70.1%)	745 (73.4%)	166 (68.9%)	421 (66.4%)	12080 (73.6%)
Yes	3100 (25.3%)	676 (29.9%)	270 (26.6%)	75 (31.1%)	213 (33.6%)	4334 (26.4%)
Diabetes						
No	6825 (55.6%)	604 (26.7%)	426 (42.0%)	93 (38.6%)	220 (34.7%)	8168 (49.8%)
Yes	5440 (44.4%)	1655 (73.3%)	589 (58.0%)	148 (61.4%)	414 (65.3%)	8246 (50.2%)
Smoking status[b]						
Current	1582 (12.9%)	245 (10.8%)	129 (12.7%)	42 (17.4%)	92 (14.5%)	2090 (12.7%)
Former	5217 (42.5%)	1030 (45.6%)	450 (44.3%)	96 (39.8%)	246 (38.8%)	7039 (42.9%)
Never	5412 (44.1%)	972 (43.0%)	434 (42.8%)	102 (42.3%)	296 (46.7%)	7216 (44.0%)
Primary cause of renal disease						
Diabetes	4085 (33.3%)	1388 (61.4%)	433 (42.7%)	121 (50.2%)	346 (54.6%)	6373 (38.8%)

Glomerulonephritis	2579 (21.0%)	355 (15.7%)	186 (18.3%)	40 (16.6%)	103 (16.2%)	3263 (19.9%)
Hypertension	1933 (15.8%)	187 (8.3%)	132 (13.0%)	25 (10.4%)	57 (9.0%)	2334 (14.2%)
Other/Uncertain	3668 (29.9%)	329 (14.6%)	264 (26.0%)	55 (22.8%)	128 (20.2%)	4444 (27.1%)

Abbreviations: BMI: body mass index; IQR: interquartile range.

[a] Categorisation based on the World Health Organisation (WHO) International Classification of adult underweight, overweight and obesity according to BMI.

[b] There were 69 patients with unknown or missing smoking status.

Table 3. Hazard ratios (and 95% confidence limits) for the effects of baseline covariates on the hazard of transplant or death without transplant, as estimated under the five-class joint model and assuming constant effects over latent classes.

Covariate	Death without transplant	Transplant
Year of initiating dialysis (ref: 2005 to 2009)		
2010 to 2014	0.87 (0.82 to 0.92)	1.24 (1.14 to 1.36)
Gender (ref: Male)		
Female	1.05 (0.99 to 1.11)	0.89 (0.81 to 0.97)
Chronic lung disease (ref: No)		
Yes	1.28 (1.21 to 1.37)	0.53 (0.44 to 0.63)
Coronary artery disease (ref: No)		
Yes	1.33 (1.25 to 1.41)	0.69 (0.61 to 0.78)
Cerebrovascular disease (ref: No)		
Yes	1.28 (1.20 to 1.37)	0.63 (0.51 to 0.76)
Peripheral vascular disease (ref: No)		
Yes	1.27 (1.19 to 1.35)	0.76 (0.65 to 0.89)
Diabetes (ref: None)		
Yes	1.24 (1.14 to 1.35)	0.64 (0.53 to 0.78)
Primary cause of renal disease (ref: Diabetes)		
Glomerulonephritis	0.75 (0.68 to 0.83)	1.22 (0.98 to 1.51)
Hypertension	0.96 (0.87 to 1.06)	0.86 (0.67 to 1.11)
Other/Uncertain	1.14 (1.04 to 1.25)	1.06 (0.86 to 1.32)
Race categories (ref: Caucasian)		
Aboriginal/Torres Strait Islander	1.01 (0.91 to 1.12)	0.15 (0.12 to 0.19)
Asian	0.62 (0.54 to 0.71)	0.63 (0.54 to 0.74)
Maori/Pacific Islander	1.13 (1.02 to 1.25)	0.24 (0.20 to 0.30)
Number of years above age 50 at initiation of dialysis[1]	1.04 (1.03 to 1.04)	0.90 (0.89 to 0.90)

[1] "Number of years above age 50 at initiation of dialysis" is set equal to zero for ages below 50 years. For example age 45 would correspond to a covariate value of 0, whilst age 60 would correspond to a covariate value of 10.

Figure 1. Crude cumulative incidence curves (and 95% confidence limits) for the two competing events: transplant, and death without transplant.

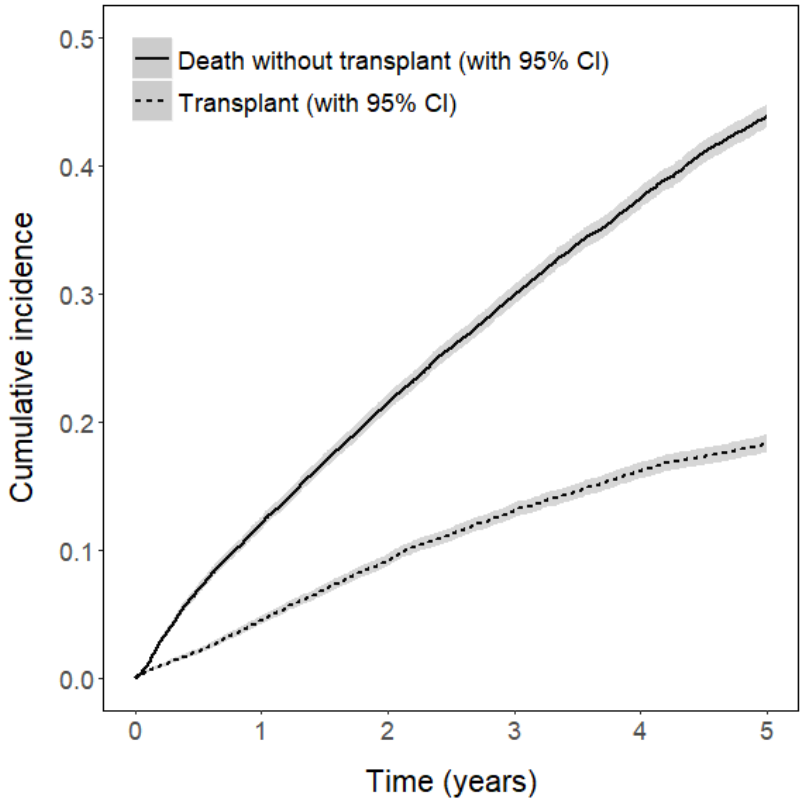
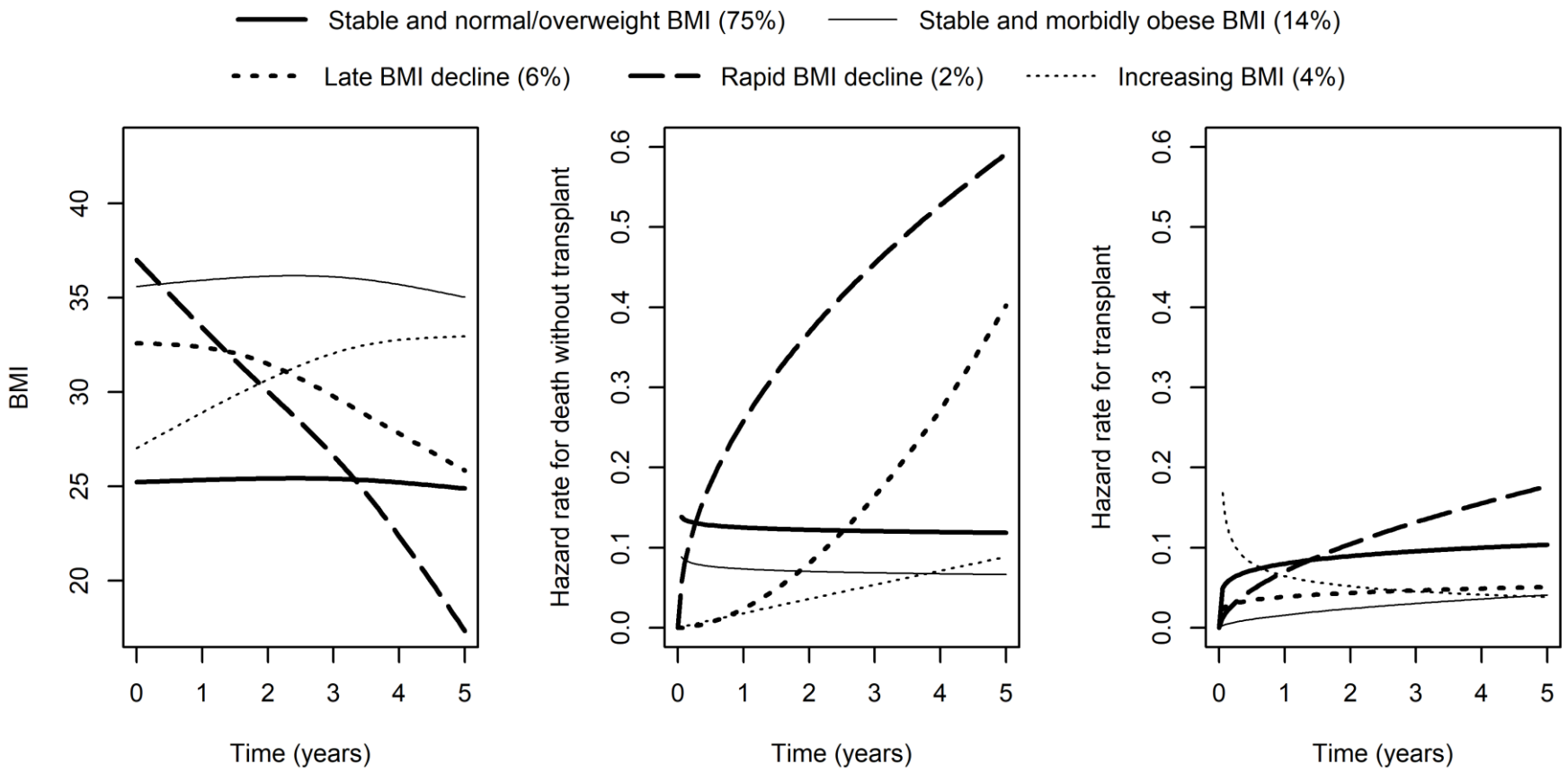


Figure 2. Predicted longitudinal BMI trajectories (left panel) and cause-specific hazard functions for death without transplant (middle panel) and transplant (right panel) from the five-class model. The predictions are shown for each of the five possible latent classes. The BMI predictions are on average (since no covariates were included in the BMI submodel), whilst the event outcome predictions are for a Caucasian male, aged ≤50 years, initiating RRT between 2005-09 with diabetic nephropathy, cerebrovascular disease and coronary artery disease. The percentage of patients in each latent class is shown in parentheses in the plot legend.



Chapter 5: Bayesian estimation of multivariate joint models

5.1 Chapter introduction

Until recently there was a lack of general purpose software available for fitting multivariate joint models (for example, see Hickey et al. (2016) for a recent review). However, since 2017, there have been several developments in this area.

The **joineRML** (Hickey et al., 2017) R package was officially released on the Comprehensive R Archive Network (CRAN). In addition, a development version of the **survtd** (Moreno-Betancur et al., 2017) R package was made publically available on GitHub. Both the aforementioned packages can be used to estimate a shared parameter joint model with multiple continuous (i.e. normally distributed) longitudinal biomarkers. The former uses a full likelihood-based joint modelling approach, whilst the latter uses an extended two-stage joint modelling approach. However, both are currently limited to a current value association structure.

A version of the **JMbayes** (Rizopoulos, 2016) R package was also released on CRAN that incorporates functionality for fitting multivariate joint models. The package currently allows for normal, Poisson, or Bernoulli distributed biomarkers, and provides significant flexibility in the definition of the association structure. Estimation of the model is performed under a Bayesian approach implemented using Gibbs sampling.

Lastly, the work of this PhD led to the release of a version (2.17.0 and thereafter) of the **rstanarm** (Brilleman et al., 2018; Stan Development Team, 2017a) R package on CRAN that incorporates functionality for fitting multivariate joint models. The package allows for different types of longitudinal outcomes through a range of link functions and error distributions. Moreover, a variety of association structures are accommodated. The package also allows for joint model data with multiple clustering factors; these methods are

discussed in the next chapter, however, it is noteworthy that none the aforementioned packages currently accommodate such data structures.

The remainder of this chapter will focus on describing the methodology, features, and performance of the joint modelling functionality in **rstanarm**. In Section 5.2, a peer-reviewed conference paper is presented. The paper introduces the formulation of the multivariate joint model used in **rstanarm**, a description of how the model is estimated using the Bayesian software Stan, and an example application showing usage of the package. However, one area that is not covered in the paper is the Hamiltonian Monte Carlo (HMC) theory and methodology that underpins the estimation of models in Stan itself. A review of HMC is outside the scope of this thesis. Instead, the reader is referred to two publications that aim to provide a conceptual explanation of the theory of HMC, namely Neal (2011) and Betancourt (2017), as well as Hoffman and Gelman (2014) for technical details on the Stan implementation of the method.

In Section 5.3, a simulation study is presented with the aim of evaluating the performance of the joint modelling functionality in **rstanarm**. As part of the simulation study, two additional R packages are described: **simsurv** (Brilleman, 2018b) and **simjm** (Brilleman, 2018a). The former has been developed for simulating complex time-to-event data using the method of Crowther and Lambert (2013) and the latter for simulating joint longitudinal and time-to-event data. The motivation for including these packages in this thesis is that they help facilitate the simulation study. However, the **simsurv** package in particular is far more general, providing users with the opportunity to simulate time-to-event data from standard parametric distributions, two-component mixture distributions, hazard functions that incorporate time-dependent effects (i.e. non-proportional hazards), or flexible user-defined hazard functions. Finally, in Section 5.4, a qualitative comparison is made between **rstanarm** and alternative software packages that are available for fitting multivariate joint models.

5.2 Conference paper

This section herein contains the following peer-reviewed conference paper:

> Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J, Wolfe R. Joint longitudinal and time-to-event models via Stan. In: *Proceedings of StanCon 2018*. 10-12 Jan 2018. Pacific Grove, California, USA. https://github.com/stan-dev/stancon_talks

Note that some parts of the “Introduction” and “Model formulation” sections from the conference paper were presented in Chapter 2 of the thesis.

Joint longitudinal and time-to-event models via Stan

Sam Brilleman, Michael Crowther, Margarita Moreno-Betancur, Jacqueline Buros Novik, Rory Wolfe*

Abstract

The joint modelling of longitudinal and time-to-event data has received much attention in the biostatistical literature in recent years. In this notebook, we describe the implementation of a shared parameter joint model for longitudinal and time-to-event data in Stan. The methods described in the notebook are a simplified version of those underpinning the `stan_jm` modelling function that has recently been contributed to the **rstanarm** R package. This notebook will proceed as follows. In Section 1 we provide an introduction to the joint modelling of longitudinal and time-to-event data, including briefly describing the potential motivations for using such an approach. In Section 2 we describe the formulation of a multivariate shared parameter joint model and introduce it's log likelihood function. In Section 3 we describe some of the more important features of the Stan code required to implement the model. In Section 4 we present a short applied example to demonstrate estimation of the model and the type of inferences that can be obtained. In Section 5 we close with a discussion.

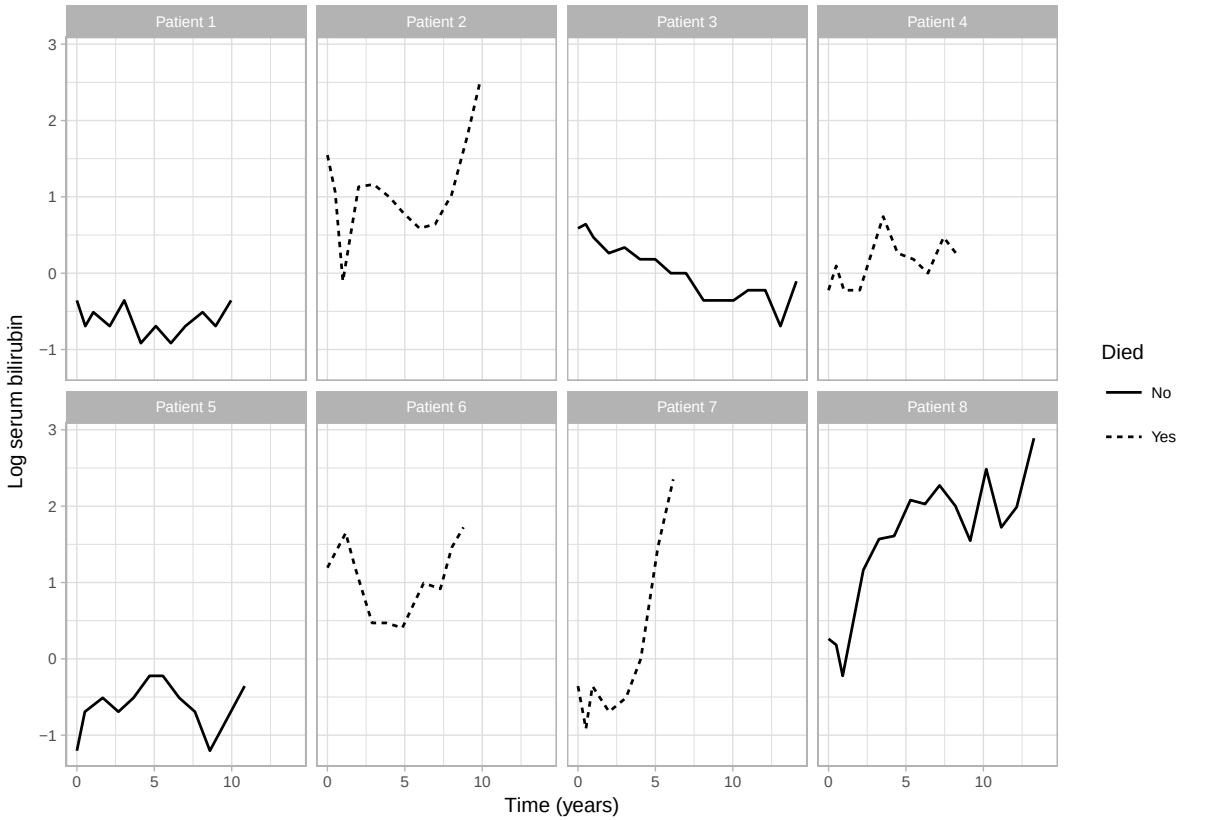
Date this notebook was compiled: 25 June 2018.

1 Introduction

Joint modelling can be broadly defined as the simultaneous estimation of two or more statistical models which traditionally would have been separately estimated. When we refer to a shared parameter joint model for longitudinal and time-to-event data, we generally mean the joint estimation of: 1) a longitudinal mixed effects model which analyses patterns of change in an outcome variable that has been measured repeatedly over time (for example, a clinical biomarker) and 2) a survival or time-to-event model which analyses the time until an event of interest occurs (for example, death or disease progression). Joint estimation of these so-called "submodels" is achieved by assuming they are correlated via individual-specific parameters (i.e. individual-level random effects).

Over the last two decades the joint modelling of longitudinal and time-to-event data has received a significant amount of attention [1-5]. Methodological developments in the area have been motivated by a growing awareness of the benefits that a joint modelling approach can provide. In clinical or epidemiological research it is common for a clinical biomarker to be repeatedly measured over time on a given patient. In addition, it is common for time-to-event data, such as the patient-specific time from a defined origin (e.g. time of diagnosis of a disease) until a terminating clinical event such as death or disease progression to also be collected. The figure below shows observed longitudinal measurements (i.e. observed "trajectories") of log serum bilirubin for a small sample of patients with primary biliary cirrhosis. Solid lines are used for those patients who were still alive at the end of follow up, while dashed lines are used for those patients who died. From the plots, we can observe between-patient variation in the longitudinal trajectories for log serum bilirubin, with some patients showing an increase in the biomarker over time, others decreasing, and some remaining stable. Moreover, there is variation between patients in terms of the frequency and timing of the longitudinal measurements.

*Corresponding author: sam.brilleman@monash.edu.



From the perspective of clinical risk prediction, we may be interested in asking whether the between-patient variation in the log serum bilirubin trajectories provides meaningful prognostic information that can help us differentiate patients with regard to some clinical event of interest, such as death. Alternatively, from an epidemiological perspective we may wish to explore the potential for etiological associations between changes in log serum bilirubin and mortality. Joint modelling approaches provide us with a framework under which we can begin to answer these types of clinical and epidemiological questions.

More formally, the motivations for undertaking a joint modelling analysis of longitudinal and time-to-event data might include one or more of the following:

- One may be interested in how *underlying changes in the biomarker influence the occurrence of the event*. However, including the observed biomarker measurements directly into a time-to-event model as time-varying covariates poses several problems. For example, if the widely used Cox proportional hazards model is assumed for the time-to-event model then biomarker measurements need to be available for all patients at all failure times, which is unlikely to be the case [3]. If simple methods of imputation are used, such as the "last observation carried forward" method, then these are likely to induce bias [6]. Furthermore, the observed biomarker measurements may be subject to measurement error and therefore their inclusion as time-varying covariates may result in biased and inefficient estimates. In most cases, the measurement error will result in parameter estimates which are shrunk towards the null [7]. On the other hand, joint modelling approaches allow us to estimate the association between the biomarker (or some function of the biomarker trajectory, such as rate of change in the biomarker) and the risk of the event, whilst allowing for both the discrete time and measurement-error aspects of the observed biomarker.
- One may be interested primarily in the evolution of the clinical biomarker but *may wish to account for what is known as informative dropout*. If the value of future (unobserved) biomarker measurements are related to the occurrence of the terminating event, then those unobserved biomarker measurements will be "missing not at random" [8,9]. In other words, biomarker measurements for patients who have an event will differ from those who do not have an event. Under these circumstances, inference based solely on observed measurements of the biomarker will be subject to bias. A joint modelling approach can help to adjust for informative dropout and has been shown to reduce bias in the estimated parameters associated with longitudinal changes in the biomarker [1,9,10].

- Joint models are naturally suited to the task of *dynamic risk prediction*. For example, joint modelling approaches have been used to develop prognostic models where predictions of event risk can be updated as new longitudinal biomarker measurements become available. Taylor et al. [11] jointly modelled longitudinal measurements of the prostate specific antigen (PSA) and time to clinical recurrence of prostate cancer. The joint model was then used to develop a web-based calculator which could provide real-time predictions of the probability of recurrence based on a patient's up to date PSA measurements.

In this notebook, we describe the implementation of a shared parameter joint model for longitudinal and time-to-event data in Stan. In Section 2 we describe the formulation for a multivariate joint model, that is, a joint model for multiple (i.e. more than one) longitudinal biomarkers and the time to a terminating event. In Section 3 we describe the important features of the Stan code required to fit the model. In Section 4 we present a brief applied example to demonstrate estimation of the model and the type of inferences that can be obtained. Note that the methods and code described in this paper are a simplified version of the `stan_jm` modelling function that is being contributed to the **rstanarm** R package [12,13], see https://github.com/stan-dev/rstanarm or https://github.com/sambrilleman/rstanarm.

2 Model formulation

A shared parameter joint model consists of related submodels which are specified separately for each of the longitudinal and time-to-event outcomes. These are therefore commonly referred to as the *longitudinal submodel(s)* and the *event submodel*. The longitudinal and event submodels are linked using shared individual-specific parameters, which can be parameterised in a number of ways. We describe each of these submodels below.

2.1 Longitudinal submodel(s)

We assume $y_{ijm}(t) = y_{im}(t_{ij})$ corresponds to the observed value of the m^{th} $(m = 1, ..., M)$ biomarker for individual i $(i = 1, ..., N)$ taken at time point t_{ij}, $j = 1, ..., n_i$. We specify a (multivariate) generalised linear mixed model that assumes $y_{ijm}(t)$ follows a distribution in the exponential family with mean $\mu_{ijm}(t)$ and linear predictor

$$\eta_{ijm}(t) = g_m(\mu_{ijm}(t)) = \boldsymbol{x}_{ijm}^T(t)\boldsymbol{\beta}_m + \boldsymbol{z}_{ijm}^T(t)\boldsymbol{b}_{im} \tag{1}$$

where $\boldsymbol{x}_{ijm}^T(t)$ and $\boldsymbol{z}_{ijm}^T(t)$ are both row-vectors of covariates (which likely include some function of time, for example a linear slope, cubic splines, or polynomial terms) with associated vectors of fixed and individual-specific parameters $\boldsymbol{\beta}_m$ and $\boldsymbol{b}_{im}$, respectively, and g_m is some known link function.

The distribution and link function are allowed to differ over the M longitudinal submodels. We assume that the dependence across the different longitudinal submodel (i.e. the correlation between the different longitudinal biomarkers) is captured through a shared multivariate normal distribution for the individual-specific parameters; that is, we assume

$$\begin{pmatrix} \boldsymbol{b}_{i1} \\ \vdots \\ \boldsymbol{b}_{iM} \end{pmatrix} = \boldsymbol{b}_i \sim \mathsf{Normal}\left(0, \boldsymbol{\Sigma}\right) \tag{2}$$

for some unstructured variance-covariance matrix $\boldsymbol{\Sigma}$.

2.2 Event submodel

We assume that we also observe an event time $T_i = \mathsf{min}\left(T_i^*, C_i\right)$ where T_i^* denotes the so-called "true" event time for individual i (potentially unobserved) and C_i denotes the censoring time. We define an

event indicator $d_i = I(T_i^* \leq C_i)$. We then model the hazard of the event using a parametric proportional hazards regression model of the form

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{w}_i^T(t)\boldsymbol{\gamma} + \sum_{m=1}^{M}\sum_{q=1}^{Q_m}\alpha_{mq}f_{mq}(\boldsymbol{\beta}_m, \boldsymbol{b}_{im}; t)\right) \tag{3}$$

where $h_i(t)$ is the hazard of the event for individual i at time t, $h_0(t)$ is the baseline hazard at time t, $\boldsymbol{w}_i^T(t)$ is a row-vector of individual-specific covariates (possibly time-dependent) with an associated vector of regression coefficients $\boldsymbol{\gamma}$ (log hazard ratios), and the α_{mq} are also coefficients (log hazard ratios).

The longitudinal and event submodels are assumed to related via an "association structure" based on shared individual-specific parameters and captured via the $\sum_{m=1}^{M}\sum_{q=1}^{Q_m}\alpha_{mq}f_{mq}(\boldsymbol{\beta}_m, \boldsymbol{b}_{im}; t)$ term in the linear predictor of equation (3). The coefficients α_{mq} are referred to as the "association parameters" since they quantify the strength of the association between the longitudinal and event processes, while the $f_{mq}(\boldsymbol{\beta}_m, \boldsymbol{b}_{im}; t)$ (for some functions $f_{mq}(.)$) can be referred to as the "association terms" and can be specified in a variety of ways which we describe in the next section.

We assume that the baseline hazard $h_0(t)$ is modelled parametrically. For simplicity, in the formulation of the joint model presented in this notebook we will restrict ourselves to modelling the log baseline hazard using B-splines. Note however that in the **rstanarm** package's `stan_jm` modelling function the baseline hazard can be specified as either an approximation using B-splines (the default), a Weibull distribution, or a piecewise constant baseline hazard (sometimes referred to as piecewise exponential). In the case of the piecewise constant or B-splines baseline hazard, the user can control the flexibility by explicitly specifying the knot points or degrees of freedom.

2.3 Association structures

As mentioned in the previous section, the dependence between the longitudinal and event submodels is captured through the association structure, which can be specified in a number of ways. In this notebook, we focus on the simplest association structure

$$f_{mq}(\boldsymbol{\beta}_m, \boldsymbol{b}_{im}; t) = \eta_{im}(t) \tag{4}$$

This is often referred to as a *current value* association structure since it assumes that the log hazard of the event at time t is linearly associated with the value of the longitudinal submodel's linear predictor also evaluated at time t. This is the most common association structure used in the joint modelling literature to date. In the situation where the longitudinal submodel is based on an identity link function and normal error distribution (i.e. a linear mixed model) the *current value* association structure can be viewed as a method for including the underlying "true" value of the biomarker as a time-varying covariate in the event submodel.[1]

However, there are a variety of other association structures that could be specified. For example, we could assume the log hazard of the event is linearly associated with the *current slope* (i.e. rate of change) of the longitudinal submodel's linear predictor, that is

$$f_{mq}(\boldsymbol{\beta}_m, \boldsymbol{b}_{im}; t) = \frac{d\eta_{im}(t)}{dt} \tag{5}$$

Moreover, there are a whole range of possible association structures, many of which have been discussed in the literature [14-16]. Also note that the full set of association structures that are accommodated in the **rstanarm** package's `stan_jm` modelling function are not described here but are discussed in the documentation for the `stan_jm` function itself.

[1] By "true" value of the biomarker, we mean the value of the biomarker which is not subject to measurement error or discrete time observation. Of course, for the expected value from the longitudinal submodel to be considered the so-called "true" underlying biomarker value, we would need to have specified the longitudinal submodel appropriately!

2.4 Conditional independence assumption

A key assumption of the multivariate shared parameter joint model is that the observed longitudinal measurements are independent of one another (both across the M biomarkers and across the n_i time points), as well as independent of the event time, conditional on the individual-specific parameters $\boldsymbol{b}_i$. That is, we assume

$$\text{Cov}\Big(y_{im}(t), y_{im'}(t)|\boldsymbol{b}_i\Big) = 0 \tag{6}$$

$$\text{Cov}\Big(y_{im}(t), y_{im}(t')|\boldsymbol{b}_i\Big) = 0 \tag{7}$$

$$\text{Cov}\Big(y_{im}(t), T_i|\boldsymbol{b}_i\Big) = 0 \tag{8}$$

for some $m \neq m'$ and $t \neq t'$.

Although this may be considered a strong assumption, it is useful in that it allows the full likelihood for joint model to be factorised into the likelihoods for each of the component parts (i.e. the likelihoods for each of the submodels and the likelihood for the distribution of the individual-specific parameters).

2.5 Log posterior distribution

Under the conditional independence assumption, the log posterior for the i^{th} individual can be specified as

$$p(\boldsymbol{\theta}, \boldsymbol{b}_i|\boldsymbol{y}_i, T_i, d_i) \propto \log\left[\left(\prod_{m=1}^{M}\prod_{j=1}^{n_i} p(y_{ijm}|\boldsymbol{b}_i, \boldsymbol{\theta})\right) p(T_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})\right] \tag{9}$$

which we can rewrite as

$$p(\boldsymbol{\theta}, \boldsymbol{b}_i|\boldsymbol{y}_i, T_i, d_i) \propto \left(\sum_{m=1}^{M}\sum_{j=1}^{n_i} \log p(y_{ijm}|\boldsymbol{b}_i, \boldsymbol{\theta})\right) + \log p(T_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta}) + \log p(\boldsymbol{b}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{10}$$

where $\sum_{j=1}^{n_i} \log p(y_{ijm}|\boldsymbol{b}_i, \boldsymbol{\theta})$ is the log likelihood for the m^{th} longitudinal submodel, $\log p(T_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ is the log likelihood for the event submodel, $\log p(\boldsymbol{b}_i|\boldsymbol{\theta})$ is the log likelihood for the distribution of the group-specific parameters (i.e. random effects), and $\log p(\boldsymbol{\theta})$ represents the log likelihood for the joint prior distribution across all remaining unknown parameters.[2]

We can rewrite the log likelihood for the event submodel as

$$\log p(T_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta}) = d_i * \log h_i(T_i) - \int_0^{T_i} h_i(s)ds \tag{11}$$

[2]In this notebook we assume normal prior distributions for all unbounded parameters (e.g. regression coefficients) and half-normals for all lower-bounded parameters (e.g. standard deviations). However, in the **rstanarm** package there is a variety of prior distributions available to the user. For the prior distribution for the variance-covariance matrix of the group-specific parameters (i.e. the variance-covariance matrix for the individual-level random effects) we use a decomposition of the variance-covariance matrix into a vector of standard deviations and a correlation matrix. We then place a half-Cauchy prior distribution on each of the standard deviations, and use the LKJ correlation matrix distribution (parameterised in terms of it's Cholesky factor) as the prior for the correlation matrix of the group-specific parameters. Further details of this prior distribution are described in the Stan User Manual and the implementation can be seen in the `jm.stan` file included with this notebook. Importantly however, when using the `stan_jm` modelling function in the **rstanarm** package, the user can control the hyperparameters related to this prior distribution. Moreover, the user can instead choose to place a prior on a further decomposed version of the variance-covariance matrix, whereby the vector of standard deviations are further decomposed into a trace and a simplex vector. This latter option is taken directly from the prior distribution described for variance-covariance matrices in the **rstanarm** package's `stan_glmer` modelling, and we refer the reader to the documentation of that package for further details.

and then use Gauss-Kronrod quadrature with Q nodes to approximate $\int_0^{T_i} h_i(s)ds$, such that

$$\int_0^{T_i} h_i(s)ds \approx \frac{T_i}{2} \sum_{q=1}^{Q} w_q h_i\left(\frac{T_i(1+s_q)}{2}\right) \tag{12}$$

where w_q and s_q, respectively, are the standardised weights and locations ("abscissa") for quadrature node q ($q = 1, ..., Q$) [17]. In this notebook we choose to use $Q = 15$ quadrature nodes.[3]

Therefore, once we have an individual's event time T_i we can evaluate the design matrices for the event submodel and longitudinal submodels at the $Q + 1$ necessary time points (which are the event time T_i and the quadrature points $\frac{T_i(1+s_q)}{2}$ for $q = 1, ..., Q$) and then pass these to Stan's data block. We can then evaluate the log likelihood for the event submodel by simply calculating the hazard $h_i(t)$ at those $Q + 1$ time points and summing the quantities appropriately. This calculation will need to be performed each time we iterate through Stan's model block. The Stan code required to evaluate this log posterior is described in the next section.

3 Stan code

Here we describe the most important features of the Stan code used to estimate the joint model. The full Stan code is provided in the separate `jm.stan` file supplied with this notebook. To simplify things for the reader, we have limited ourselves to the situation in which $M = 2$ (i.e. we have two longitudinal biomarkers) and each of those longitudinal outcomes is modelled using a linear mixed model (i.e. identity link, normal distribution).

3.1 Data and transformed data blocks

The data block includes dimensions of the model, outcome vectors (e.g. observed biomarker values and event times), design matrices for the different submodels, and hyperparameters for the prior distributions. We do not discuss the data or transformed data blocks here in any detail.

3.2 Parameters block

Most of the parameters defined in the parameters block are "primitive" or "unscaled", meaning that they will be given a prior distribution with mean 0 and scale 1 and then converted into the actual parameters used in the regression equation via the transformed parameters block. Our parameters block therefore includes:

- `y1_gamma`, `y2_gamma`: the intercept for each of the longitudinal submodels. These intercept parameters are unbounded, given that each longitudinal submodel in our application consists of a linear mixed model (i.e. in our application we assume an identity link function and normal error distribution for each longitudinal biomarker).
- `y1_z_beta`, `y2_z_beta`: the primitive coefficients for each of the longitudinal submodels.
- `e_z_beta`, `a_z_beta`: the primitive coefficients and primitive association parameters for the event submodel.
- `y1_aux_unscaled`, `y2_aux_unscaled`: the unscaled standard deviations (SD) of the residual errors for each of the longitudinal submodels, combined into a single vector.
- `e_aux_unscaled`: the unscaled coefficients for the B-spline terms used in the baseline hazard.

The parameters block also includes the unscaled group-specific parameters (i.e. unscaled individual-level random effects). We specify these as a matrix, with the number of rows in the matrix equal to the total number of group-specific terms in the model, and the number of columns in the matrix equal to the total number of patients in the data (i.e. the total number of "groups"). We

[3] The `stan_jm` modelling function in the **rstanarm** package allows the user to choose between $Q = 15$ (the default), 11, or 7 quadrature nodes.

declare a parameter vector that contains the standard deviations for each of the group-specific parameters and a lower triangular matrix that corresponds to the Cholesky factor of the correlation matrix for the group-specific terms. The latter is declared using Stan's `cholesky_factor_corr` data type.

```
parameters {
  real y1_gamma; // intercepts in long. submodels
  real y2_gamma;
  vector[y_K[1]] y1_z_beta; // primitive coefs in long. submodels
  vector[y_K[2]] y2_z_beta;
  vector[e_K] e_z_beta; // primitive coefs in event submodel (log hazard ratios)
  vector[a_K] a_z_beta; // primitive assoc params (log hazard ratios)
  real<lower=0> y1_aux_unscaled; // unscaled residual error SDs
  real<lower=0> y2_aux_unscaled;
  vector[basehaz_df] e_aux_unscaled; // unscaled coefs for baseline hazard

  // group level params
  vector<lower=0>[b_K] b_sd; // group level sds
  matrix[b_K,b_N] z_b_mat;   // unscaled group level params
  cholesky_factor_corr[b_K > 1 ? b_K : 0]
    b_cholesky;              // cholesky factor of corr matrix
}
```

3.3 Transformed parameters block

The transformed parameters block includes code to alter the location and scale of the "primitive" or "unscaled" parameters, in order to obtain the actual parameters used in the regression submodels.

Note that in the code below `b_K` is the number of group-specific parameters in the model, so if `b_K > 1` then we will be estimating a correlation matrix for the group-specific parameters and, hence, we must transform the primitive group-specific parameters using `b_cholesky` and `b_sd`, rather than `b_sd` alone. If there was only one group-specific parameter in the model then there would be no correlation matrix (i.e. no `b_cholesky` parameter). Also note that for any *multivariate* joint model (i.e. more than one longitudinal outcome) we will have `b_K > 1`.

```
transformed parameters {
  ...
  // coefs for long. submodels
  y1_beta = y1_z_beta .* y1_prior_scale + y1_prior_mean;
  y2_beta = y2_z_beta .* y2_prior_scale + y2_prior_mean;

  // coefs for event submodel (incl. association parameters)
  e_beta = e_z_beta .* e_prior_scale + e_prior_mean;
  a_beta = a_z_beta .* a_prior_scale + a_prior_mean;

  // residual error SDs for long. submodels
  y1_aux = y1_aux_unscaled * y_prior_scale_for_aux[1] + y_prior_mean_for_aux[1];
  y2_aux = y2_aux_unscaled * y_prior_scale_for_aux[2] + y_prior_mean_for_aux[2];

  // b-spline coefs for baseline hazard
  e_aux = e_aux_unscaled .* e_prior_scale_for_aux + e_prior_mean_for_aux;

  // group level params
  if (b_K == 1)
    b_mat = (b_sd[1] * z_b_mat)';
  else if (b_K > 1)
```

```
    b_mat = (diag_pre_multiply(b_sd, b_cholesky) * z_b_mat)';
}
```

3.4 Model block

The model block consists of several distinct parts. We describe each of these separately.

In the first part of the model block, we evaluate the linear predictor for each of the M longitudinal submodels at the respective observation times. We then increment the target with the resulting likelihood. To evaluate the linear predictor we call a user-defined function which is defined in the `functions {}` block at the start of the `jm.stan` file. This function takes the form:

```
  /**
  * Evaluate the linear predictor for the glmer submodel
  *
  * @param X Design matrix for fe
  * @param Z Design matrix for re, for a single grouping factor
  * @param Z_id Group indexing for Z
  * @param gamma The intercept parameter
  * @param beta Vector of population level parameters
  * @param bMat Matrix of group level params
  * @param shift Number of columns in bMat
  *   that correpond to group level params from prior glmer submodels
  * @return A vector containing the linear predictor for the glmer submodel
  */
  vector evaluate_eta(matrix X, vector[] Z, int[] Z_id, real gamma,
                      vector beta, matrix bMat, int shift) {
    int N = rows(X);    // num rows in design matrix
    int K = rows(beta); // num predictors
    int p = size(Z);    // num group level params
    vector[N] eta;

    if (K > 0) eta = X * beta;
    else eta = rep_vector(0.0, N);

    for (k in 1:p)
      for (n in 1:N)
        eta[n] = eta[n] + (bMat[Z_id[n], k + shift]) * Z[k,n];

    return eta;
  }
```

Such that the code in our model block is the following:

```
model {
  //---- Log-lik for longitudinal submodels
  {
    // declare linear predictors
    vector[y_N[1]] y1_eta;
    vector[y_N[2]] y2_eta;

    // evaluate linear predictor for each long. submodel
    y1_eta = evaluate_eta(y1_X, y1_Z, y1_Z_id, y1_gamma, y1_beta, b_mat, 0);
    y2_eta = evaluate_eta(y2_X, y2_Z, y2_Z_id, y2_gamma, y2_beta, b_mat, b_KM[1]);

    // increment the target with the log-lik
```

```
  target += normal_lpdf(y1 | y1_eta, y1_aux);
  target += normal_lpdf(y2 | y2_eta, y2_aux);
}
...
```

To evaluate the event submodel likelihood we must evaluate $h_i(T_i)$ for individuals who experienced the event (i.e. $d_i = 1$) (i.e. the hazard at their event time) as well as the cumulative hazard $\int_0^{T_i} h_i(s)ds$ for all individuals. Since we are going to evaluate the cumulative hazard using Gauss-Kronrod quadrature, this means calculating the hazard $h_i(t)$ at 15 quadrature points between 0 and T_i for each individual i. To do this, we have constructed the design matrices in R evaulated at the necessary times; these are passed to Stan's data block (not shown here) as `e_Xq`, `y1_Xq`, `y2_Xq` etc. In the code below there are several steps:

- In **Step 1** we use the event submodel design matrices to evaluate the $\boldsymbol{w}_i^T(t)\boldsymbol{\gamma}$ part of the event submodel's linear predictor at the observed event times and the 15 quadrature points between 0 and T_i.
- The remainder of the event submodel's linear predictor consists of the term corresponding to the association structure: $\sum_{m=1}^{M} \alpha_m \eta_{im}(t)$. This involves the current value of the longitudinal submodel's linear predictor, so we must also evaluate the longitudinal submodel's linear predictor at the event times and the 15 quadrature points between 0 and T_i. This is shown in **Step 2** of the code below.
- In **Step 3** we evaluate the log baseline hazard at the event times and the 15 quadrature points between 0 and T_i.
- In **Step 4** we combine the log baseline hazard with the event submodel linear predictor, that is, we evaluate
$$\log h_i(t) = \log h_0(t) + \left(\boldsymbol{w}_i^T(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \alpha_m \eta_{im}(t) \right)$$
- In **Steps 5** and **6** the hazard evaluated at the event times is separated out from the hazard evaluated at each of the quadrature points. The latter will be used in **Step 7** to evaluate the approximate cumulative hazard at the event time via the Gauss-Kronrod quadrature rule described in equation (12).
- In **Step 7** we evaluate the log likelihood for the event submodel as
$$\log p(T_i, d_i | \boldsymbol{b}_i, \boldsymbol{\theta}) = d_i * \log h_i(T_i) - \int_0^{T_i} h_i(s)ds$$
The first term in **Step 7** is the log hazard contribution to the log likelihood for the event submodel. The second term is the log survival contribution to the log likelihood for the event submodel. The latter is obtained by summing over the quadrature points to get the approximate integral (i.e. cumulative hazard). Note that the `qwts` vector already incorporates the necessary scaling such that the integral is evaluated over limits $(0, T_i)$ rather than $(-1, +1)$. We increment the target with the resulting log likelihood.

```
//----- Log-lik for event submodel (Gauss-Kronrod quadrature)
{
  vector[nrow_y_Xq[1]] y1_eta_q;
  vector[nrow_y_Xq[2]] y2_eta_q;
  vector[nrow_e_Xq] e_eta_q;
  vector[nrow_e_Xq] log_basehaz;
  vector[nrow_e_Xq] ll_haz_q;
  vector[Nevents] log_haz_etimes;
  vector[Npat_times_qnodes] log_haz_qtimes;

  // Step 1: event submodel linear predictor at event time and quadrature points
  e_eta_q = e_Xq * e_beta;

  // Step 2: long. submodel linear predictor at event time and quadrature points
  y1_eta_q = evaluate_eta(y1_Xq, y1_Zq, y1_Zq_id, y1_gamma, y1_beta, b_mat, 0);
  y2_eta_q = evaluate_eta(y2_Xq, y2_Zq, y2_Zq_id, y2_gamma, y2_beta, b_mat, b_KM[1]);
```

```
    // Step 2 (continued): add on contribution from association structure to
    // the event submodel linear predictor at event time and quadrature points
    e_eta_q = e_eta_q + a_beta[1] * y1_eta_q + a_beta[2] * y2_eta_q;

    // Step 3: log baseline hazard at event time and quadrature points
    log_basehaz = basehaz_X * e_aux;

    // Step 4: log hazard at event time and quadrature points
    ll_haz_q = log_basehaz + e_eta_q;

    // Step 5: log hazard at event times only
    // (i.e. log hazard contribution to the likelihood)
    log_haz_etimes = head(log_haz_q, Nevents);

    // Step 6: log hazard at quadrature points only
    log_haz_qtimes = tail(log_haz_q, Npat_times_qnodes);

    // Step 7: log likelihood for event submodel
    target += sum(log_haz_etimes) - dot_product(qwts, exp(log_haz_qtimes));
  }
```

We then increment the target with the log priors for each of the intercepts, coefficients, auxiliary parameters (including coefficients for the B-splines baseline hazard), and group-specific terms (i.e. individual-level random effects):

```
  //----- Log-priors

    // intercepts for long. submodels
    target += normal_lpdf(y1_gamma |
      y_prior_mean_for_intercept[1], y_prior_scale_for_intercept[1]);
    target += normal_lpdf(y2_gamma |
      y_prior_mean_for_intercept[2], y_prior_scale_for_intercept[2]);

    // coefficients for long. submodels
    target += normal_lpdf(y1_z_beta | 0, 1);
    target += normal_lpdf(y2_z_beta | 0, 1);

    // coefficients for event submodel
    target += normal_lpdf(e_z_beta | 0, 1);
    target += normal_lpdf(a_z_beta | 0, 1);

    // residual error SDs for long. submodels
    target += normal_lpdf(y1_aux_unscaled | 0, 1);
    target += normal_lpdf(y2_aux_unscaled | 0, 1);

    // b-spline coefs for baseline hazard
    target += normal_lpdf(e_aux_unscaled | 0, 1);

    // group level terms
      // sds
      target += student_t_lpdf(b_sd | b_prior_df, 0, b_prior_scale);
      // primitive coefs
      target += normal_lpdf(to_vector(z_b_mat) | 0, 1);
      // corr matrix
      if (b_K > 1)
        target += lkj_corr_cholesky_lpdf(b_cholesky | b_prior_regularization);
}
```

4 Application

4.1 Data

In order to make this notebook freely available we use a motivating example based on a publically accessible dataset. The Mayo Clinic's widely used primary biliary cirrhosis (PBC) data contains 312 individuals with primary biliary cirrhosis, who participated in a randomised placebo controlled trial of D-penicillamine conducted at the Mayo Clinic between 1974 and 1984 [18]. In our secondary analysis of this trial data, our primary research is *not* concerned with the efficacy of the randomised treatment but rather understanding how the clinical biomarker histories for these patients are associated with their overall survival. Specifically, we focus on the associations between two repeatedly measured clinical biomarkers, log serum bilirubin and serum albumin, and the risk of death. Given that the joint modelling methods are computationally intensive we restrict our analyses to a small random subset of just 40 patients from the PBC dataset. This ensures that the computation time for the joint models described in later sections are kept to a minimum and therefore this notebook can be compiled in a relatively short time. However, this also means that the clinical findings from this analysis should not to be overinterpreted. Rather, this notebook aims to simply demonstrate the joint modelling framework and describe how these models can be estimated using Stan.

The PBC data are contained in two separate data frames, each saved as an RDS object. The first data frame (saved as "Data/pbcLong.rds"), contains multiple-row per patient longitudinal biomarker information, as shown in

```
head(pbcLong)
```

```
##   id      age sex trt      year     logBili albumin platelet
## 1  1 58.76523   f   1 0.0000000  2.67414865    2.60      190
## 2  1 58.76523   f   1 0.5256674  3.05870707    2.94      183
## 3  2 56.44627   f   1 0.0000000  0.09531018    4.14      221
## 4  2 56.44627   f   1 0.4982888 -0.22314355    3.60      188
## 5  2 56.44627   f   1 0.9993155  0.00000000    3.55      161
## 6  2 56.44627   f   1 2.1026694  0.64185389    3.92      122
```

while the second data frame (saved as "Data/pbcSurv.rds"), contains single-row per patient survival information, as shown in

```
head(pbcSurv)
```

```
##    id      age sex trt futimeYears status death
## 1   1 58.76523   f   1    1.095140      2     1
## 3   2 56.44627   f   1   14.151951      0     0
## 12  3 70.07255   m   1    2.770705      2     1
## 16  4 54.74059   f   1    5.270363      2     1
## 23  5 38.10541   f   0    4.120465      1     0
## 29  6 66.25873   f   0    6.852841      2     1
```

The variables included across the two datasets can be defined as follows:

- `age` in years
- `albumin` serum albumin (g/dl)
- `logBili` logarithm of serum bilirubin
- `death` indicator of death at endpoint
- `futimeYears` time (in years) between baseline and the earliest of death, transplantion or censoring
- `id` numeric ID unique to each individual
- `platelet` platelet count
- `sex` gender (m = male, f = female)
- `status` status at endpoint (0 = censored, 1 = transplant, 2 = dead)
- `trt` binary treatment code (0 = placebo, 1 = D-penicillamine)
- `year` time (in years) of the longitudinal measurements, taken as time since baseline)

4.2 Estimation using the simplified jm.stan file

We fit a multivariate joint model to the two longitudinal biomarkers, log serum bilirubin and serum albumin, and time-to-death. Note that patients are censored if they had a transplant prior to death (here we ignore the fact that this is likely to be informative censoring). We fit a linear mixed model (identity link, normal distribution) for each biomarker with a patient-specific intercept and linear slope but no other covariates. In the event submodel we include gender (`sex`) and treatment (`trt`) as baseline covariates. Each biomarker is assumed to be associated with the log hazard of death at time t via it's expected value at time t (i.e. a *current value* association structure).

To save needing to carry out any data manipulation steps we instead used the `stan_jm` modelling function in **rstanarm** to generate the R list for passing to **rstan**. This data is saved as an RDS object and supplied with the notebook ("Stan/standata.rds"). In addition, a function to generate a list of initial values has also been supplied as an RDS object with the notebook ("Stan/staninit.rds"). Of course, the stan file containing the model is also supplied ("Stan/jm.stan"). We can therefore estimate this model using the **rstan** package:

```
standata <- readRDS("Stan/standata.rds")
staninit <- readRDS("Stan/staninit.rds")
mod1 <- with_filecache(
  stan(
    file = "Stan/jm.stan",
    data = standata,
    init = function() staninit,
    chains = 2, seed = 12345),
  filename = "mod1.rds")
```

Since our primary interest is in the association between the current value of each of the biomarkers (log serum bilirubin and serum albumin) and the hazard of death, we focus on the estimated association parameters. The summary of the posterior distribution for each of the association parameters follows:

```
print(mod1, pars = "a_beta")
```

```
## Inference for Stan model: jm.
## 2 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=2000.
##
##            mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## a_beta[1]  0.74    0.01 0.29  0.20  0.54  0.73  0.93  1.36  2000 1.00
## a_beta[2] -3.21    0.02 0.72 -4.79 -3.63 -3.17 -2.72 -1.92   838 1.01
##
## Samples were drawn using NUTS(diag_e) at Fri Mar 09 12:25:30 2018.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

We see that a one unit increase in log serum bilirubin is associated with a 0.72 (95% CrI: 0.17 to 1.29) unit increase in the log hazard of death, equivalent to a 2.05-fold (95% CrI: 1.19 to 3.63) *increase* in the hazard of death. Similarly, a one unit increase in serum albumin is associated with a 3.23 (95% CrI: 1.90 to 4.77) unit *decrease* in the log hazard of death. These estimates are broadly in line with what we would expect from a clinical perspective; that is, that higher serum bilirubin is associated with *worse* patient outcomes (i.e. higher risk of mortality), whilst higher serum albumin is associated with *better* patient outcomes (i.e. lower risk of mortality). However, recall that we have estimated this model with a very small dataset only used for demonstration purposes. Moreover, the number of mortality events ($N = 29$) is even less than the number of patients since some patients are censored.

4.3 Estimation using the joint modelling functionality in rstanarm

The `jm.stan` file provided with this notebook is a simplified version of the Stan code underlying the `stan_jm` modelling function in the **rstanarm** package. However, estimating the model using the **rstanarm** provides us with much nicer output (for example, meaningful variable names!) as well as a broad range of post-estimation functionality, including model diagnostics, posterior predictions, dynamic predictions and more.

To see this, we can use the development version of **rstanarm** with joint modelling functionality to refit our model, this time using `stan_jm` with the customary R formula syntax and data frames:

```
mod2 <- with_filecache(
  stan_jm(
    formulaLong = list(
        logBili ~ year + (year | id),
        albumin ~ year + (year | id)),
    formulaEvent = survival::Surv(futimeYears, death) ~ sex + trt,
    dataLong = pbcLong, dataEvent = pbcSurv,
    time_var = "year", assoc = "etavalue", basehaz = "bs",
    chains = 2, seed = 12345),
  filename = "mod2.rds")
```

We can now examine the output from the fitted model, for example

```
print(mod2)
```

```
## stan_jm
##  formula (Long1): logBili ~ year + (year | id)
##  family  (Long1): gaussian [identity]
##  formula (Long2): albumin ~ year + (year | id)
##  family  (Long2): gaussian [identity]
##  formula (Event): survival::Surv(futimeYears, death) ~ sex + trt
##  baseline hazard: bs
##  assoc:           etavalue (Long1), etavalue (Long2)
## ------
##
## Longitudinal submodel 1: logBili
##             Median MAD_SD
## (Intercept) 0.665  0.184
## year        0.229  0.041
## sigma       0.354  0.017
##
## Longitudinal submodel 2: albumin
##             Median MAD_SD
## (Intercept)  3.521  0.078
## year        -0.160  0.024
## sigma        0.291  0.013
##
## Event submodel:
##                  Median  MAD_SD  exp(Median)
## (Intercept)       6.821   2.892 916.606
## sexf             -0.145   0.637   0.865
## trt              -0.492   0.488   0.611
## Long1|etavalue    0.788   0.283   2.198
## Long2|etavalue   -3.105   0.905   0.045
## b-splines-coef1  -0.923   1.155      NA
## b-splines-coef2   0.606   0.912      NA
## b-splines-coef3  -2.622   1.307      NA
```

```
## b-splines-coef4  -0.469   1.810      NA
## b-splines-coef5  -1.166   1.789      NA
## b-splines-coef6  -2.515   1.864      NA
##
## Group-level error terms:
##  Groups Name              Std.Dev. Corr
##  id     Long1|(Intercept) 1.2405
##         Long1|year        0.1925    0.51
##         Long2|(Intercept) 0.5013   -0.64 -0.52
##         Long2|year        0.1015   -0.60 -0.82  0.47
## Num. levels: id 40
##
## Sample avg. posterior predictive distribution
## of longitudinal outcomes:
##                Median MAD_SD
## Long1|mean_PPD 0.588  0.029
## Long2|mean_PPD 3.344  0.024
##
## ------
## For info on the priors used see help('prior_summary.stanreg').
```

or we can examine the summary output for the association parameters alone:

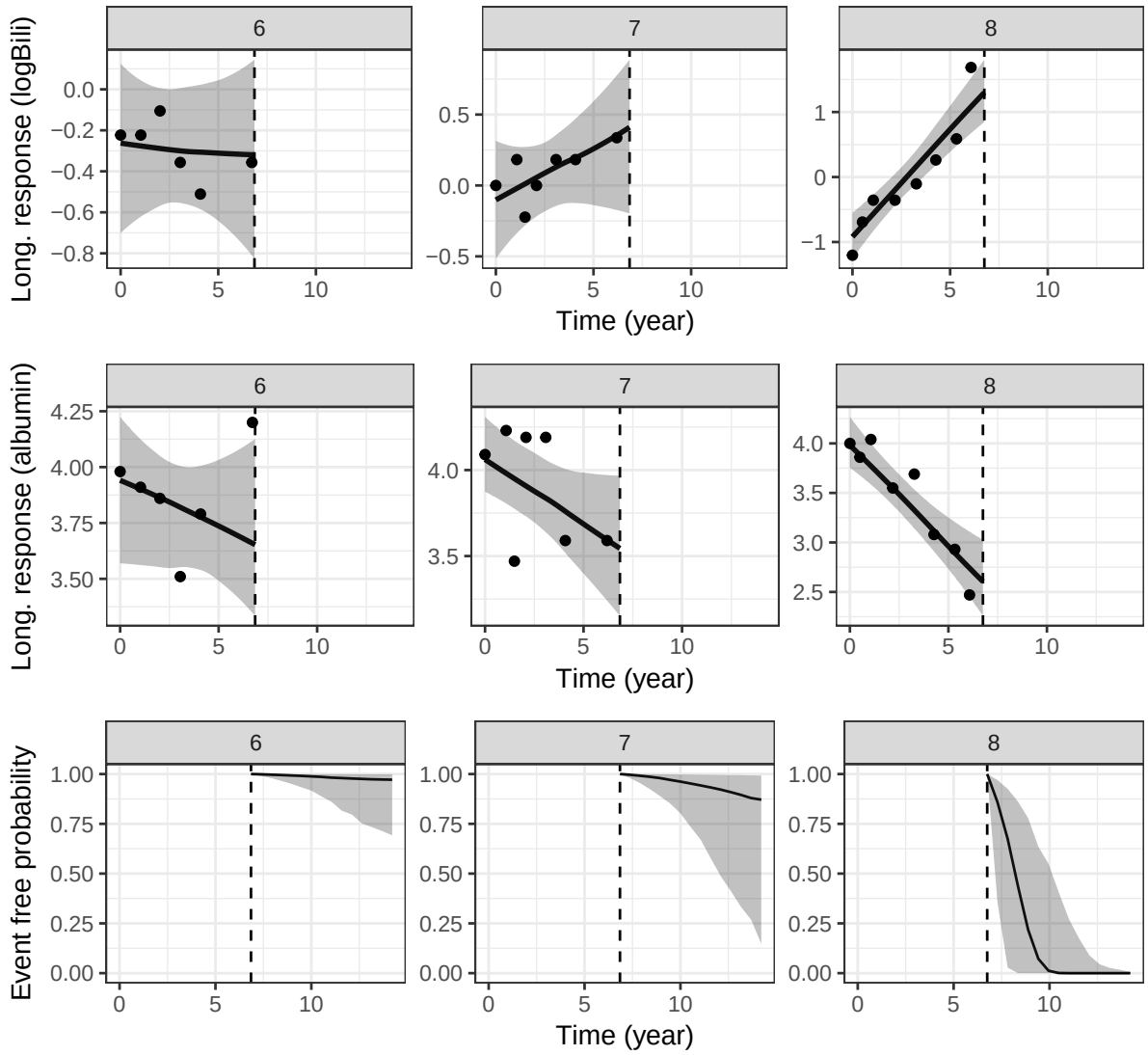
```
summary(mod2, pars = "assoc")
```

```
##
## Model Info:
##
##  function:        stan_jm
##  formula (Long1): logBili ~ year + (year | id)
##  family  (Long1): gaussian [identity]
##  formula (Long2): albumin ~ year + (year | id)
##  family  (Long2): gaussian [identity]
##  formula (Event): survival::Surv(futimeYears, death) ~ sex + trt
##  baseline hazard: bs
##  assoc:           etavalue (Long1), etavalue (Long2)
##  algorithm:       sampling
##  priors:          see help('prior_summary')
##  sample:          2000 (posterior sample size)
##  num obs:         304 (Long1), 304 (Long2)
##  num subjects:    40
##  num events:      29 (72.5%)
##  groups:          id (40)
##  runtime:         1.3 mins
##
## Estimates:
##                         mean   sd     2.5%   25%    50%    75%    97.5%
## Assoc|Long1|etavalue  0.797  0.296  0.242  0.599  0.788  0.984  1.404
## Assoc|Long2|etavalue -3.149  0.931 -5.176 -3.728 -3.105 -2.503 -1.511
##
## Diagnostics:
##                       mcse  Rhat  n_eff
## Assoc|Long1|etavalue 0.007 0.999 2000
## Assoc|Long2|etavalue 0.029 1.001 1000
##
```

We can see that the estimated association parameters are similar to those obtained from the model in the previous section. However, we can now also access a range of post-estimation functions

(described in the `stan_jm` and related help documentation; see for example `help(posterior_traj)` or `help(posterior_survfit)`). As an example, let's plot the predicted trajectories for each biomarker and the predicted survival function for three selected individuals in the dataset using `stan_jm` post-estimation functions:

```
p1 <- posterior_traj(mod2, m = 1, ids = 6:8)
p2 <- posterior_traj(mod2, m = 2, ids = 6:8)
p3 <- posterior_survfit(mod2, ids = 6:8, draws = 200)
pp1 <- plot(p1, plot_observed = TRUE, vline = TRUE)
pp2 <- plot(p2, plot_observed = TRUE, vline = TRUE)
plot_stack_jm(yplot = list(pp1, pp2), survplot = plot(p3))
```



Here we can see the strong relationship between the underlying values of the biomarkers and mortality. Patient `8` who, relative to patients `6` and `7`, has a higher underlying value for log serum bilirubin and a lower underlying value for serum albumin at the end of their follow up has a far worse predicted probability of survival.

5 Discussion

In this notebook we have introduced the formulation of a shared parameter joint model for longitudinal and time-to-event data. The formulation of the joint model can allow for multiple longitudinal biomarkers

along with a terminating event. The association between the longitudinal and event processes can be parameterised in a variety of ways, but here we have focussed on the so-called *current value* association structure which serves as the simplest and natural starting point.

The aim of this notebook was to introduce some of the ideas underpinning the estimation of these joint models in Stan. One key feature of the Stan code that we have tried to describe in detail is the implementation of the Gauss-Kronrod quadrature rule. The Gauss-Kronrod quadrature rule is required to approximate the cumulative hazard in the likelihood of the event submodel. This aspect makes evaluating the log likelihood for the event submodel more computationally intensive than if there were a closed-form solution to the integral. In addition, the models are computationally intensive due to the relatively large number of group-specific parameters that often need to be estimated. Nonetheless, estimating joint models under a Bayesian framework can provide a number of benefits. The specification of complex association structures can be made much easier. Furthermore, a Bayesian approach can lead more naturally to dynamic predictions. For these, and other reasons, we believe it is of interest to try and optimise the estimation of these models in Stan. The hope is that by describing the Stan code in some detail as part of this notebook, those reading it will have the opportunity to provide guidance on how increases in speed, efficiency, or numerical stability might be achieved.

6 Acknowledgements

Much of the joint modelling functionality that has been contributed to the **rstanarm** package has been built upon code that was already included in that package, and that code was written primarily by Ben Goodrich and Jonah Gabry [12]. We are also grateful to them for their ongoing support in helping to get the joint modelling functionality up and running in **rstanarm**.

7 References

1. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000;**1**(4):465-80.
2. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997;**53**(1):330-9.
3. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. *Stat Sinica* 2004;**14**(3):809-34.
4. Gould AL, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat Med.* 2015;**34**(14):2181-95.
5. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R* CRC Press; 2012.
6. Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J Biopharm Stat* 2002;**12**(2):207-26.
7. Prentice RL. Covariate Measurement Errors and Parameter-Estimation in a Failure Time Regression-Model. *Biometrika* 1982;**69**(2):331-42.
8. Baraldi AN, Enders CK. An introduction to modern missing data analyses. *J Sch Psychol* 2010;**48**(1):5-37.
9. Philipson PM, Ho WK, Henderson R. Comparative review of methods for handling drop-out in longitudinal studies. *Stat Med* 2008;**27**(30):6276-98.
10. Pantazis N, Touloumi G. Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Applied Statistics* 2005;**54**:405-23.
11. Taylor JM, Park Y, Ankerst DP, et al. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 2013;**69**(1):206-13.
12. Stan Development Team. *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.14.1. http://mc-stan.org/. 2016.
13. R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2015.

14. Crowther MJ, Lambert PC, Abrams KR. Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: a joint modelling approach. *BMC Med Res Methodol* 2013;**13**.
15. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Med Res Methodol* 2016;**16**(1):117.
16. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat Med.* 2011;**30**(12):1366-80.
17. Laurie DP. Calculation of Gauss-Kronrod quadrature rules. *Math Comput* 1997;**66**(219):1133-45.
18. Therneau T, Grambsch P. *Modeling Survival Data: Extending the Cox Model* Springer-Verlag, New York; 2000. ISBN: 0-387-98784-3

5.3 Evaluating the performance of rstanarm

This section describes a simulation study used to evaluate the performance of the joint modelling functionality in **rstanarm**. Subsection 5.3.1 describes the methods required to simulate event times under a complex data generating process, such as a joint longitudinal and time-to-event model. This includes background to the required methodology, followed by a description of the development of two new R packages: one for simulating simple or complex time-to-event data, and one specifically for simulating joint longitudinal and time-to-event data. Subsection 5.3.2 describes the simulation study itself, including the data generating models, the inferential quantities used, and the results.

5.3.1 Simulating time-to-event data

5.3.1.1 Background

To simulate time-to-event data, a common approach has been to make simplifying parametric assumptions about the distribution of event times. For example, many authors limit their simulation studies to settings in which the simulated event times are drawn from exponential or Weibull distributions. However, the exponential and Weibull distributions may be unrealistic in many settings. This is because the former assumes a constant hazard function, whilst the latter assumes a monotonically increasing or decreasing hazard function. These limiting forms for the hazard function are likely to be implausible in many health related applications where the underlying hazard function may have one or more turning points.

A more general method, described throughout this section, is referred to herein as the *cumulative hazard inversion method* (Bender et al., 2005). This method allows one to simulate event times under a proportional hazards model data generating process with any parametric formulation for the baseline hazard. This includes as special cases the exponential, Weibull and Gompertz distributions. However, it also provides a much more general framework that can be extended beyond those standard parametric distributions to allow for more flexible baseline hazard functions.

Using the cumulative hazard inversion method one can easily simulate an event time T_i^s for individual i by drawing a uniform random variable $U_i \sim U(0,1)$ and evaluating

$$T_i^s = H_0^{-1}(-\log(U_i)\exp(-\boldsymbol{w}_i'\boldsymbol{\gamma})) \quad (35)$$

where $H_0^{-1}(.)$ corresponds to the inverted cumulative baseline hazard, and $\boldsymbol{w}_i$ is a vector of covariates with associated population-level (i.e. fixed effect) parameters $\boldsymbol{\gamma}$. The parameters $\boldsymbol{\gamma}$ are the so-called "true" log hazard ratios used to simulate the data.

Therefore, all that is required to generate simulated event times under the cumulative hazard inversion method is an invertible cumulative baseline hazard function and access to independent draws of the random uniform variable U_i. The latter is easily obtainable, since standard statistical software packages have built in functions to simulate uniform random variables. However, access to the former will depend on how complex the definition of the baseline hazard is. If an analytical form for the inverted cumulative baseline hazard can be obtained, then the cumulative hazard inversion method is simple and computationally efficient. Moreover, even in situations where one cannot obtain an analytical form for the inverted cumulative baseline hazard, numerical approximations can be used. The following sections step through the intuition behind the cumulative hazard inversion method, as well as describing a number of related extensions.

5.3.1.2 Simulating survival probabilities for known parametric distributions

Recall that the survival function for individual i is the probability that their "true" event time T_i^* is greater than the current time t. That is, the survival function can be defined as

$$S_i(t) = P(T_i^* > t) \quad (36)$$

Moreover, the corresponding probability of having experienced the event at or before time t is the complement to the survival function. That is, the probability of failure is defined as

$$F_i(t) = P(T_i^* \leq t) = 1 - S_i(t) \quad (37)$$

If the event time T_i^* is known to be drawn from some parametric distribution, then it also holds that the definition of the probability of failure in equation (37) is equivalent to the definition of the cumulative distribution function (CDF) for the distribution of event times.

The probability integral transformation tells us that transforming a continuous random variable by its own CDF leads a new random variable that must follow a uniform distribution on the range 0 to 1 (Mood et al., 1974). That is, $F_X(X) \sim U(0,1)$ where $F_X(.)$ denotes the CDF for the continuous random variable X. Similarly, a new random variable

obtained by taking the complement of X transformed by its own CDF must also follow a uniform distribution on the range 0 to 1, that is, $1 - F_X(X) \sim U(0,1)$. This result therefore allows one to conclude that under a standard parametric distributional assumption for the event times T_i^* $(i = 1, \dots, N)$, the survival probability for individual i at their true event time will be a uniform random variable on the range 0 to 1. That is,

$$S_i(T_i^*) = U_i \sim U(0,1) \tag{38}$$

5.3.1.3 Extending to the proportional hazards data generating process

It is then possible to extend these results to the setting of a proportional hazards model. Under a proportional hazards model the survival probability for individual i at their event time T_i^* can be written as

$$S_i(T_i^*) = \exp(-H_0(T_i^*)\exp(\boldsymbol{w}_i'\boldsymbol{\gamma})) \tag{39}$$

where $H_0(t) = \int_0^t h_0(s)\, ds$ is the cumulative baseline hazard evaluated at time t, and again $\boldsymbol{w}_i$ is a vector of covariates with associated population-level (i.e. fixed effect) parameters $\boldsymbol{\gamma}$. This is because the proportional hazards assumption also implies proportional cumulative hazards. Of course, using equation (39), the survival probability $S_i(T_i^*)$ can be replaced by the uniform random variable U_i. Moreover, since the objective is to simulate a new event time for individual i $(i = 1, \dots, N)$, rather than to evaluate the survival probability at the known true event time, one can replace T_i^* with the simulated event time for individual i, T_i^s. This leads to

$$U_i = \exp(-H_0(T_i^s)\exp(\boldsymbol{w}_i'\boldsymbol{\gamma})) \tag{40}$$

To obtain the formula for the cumulative hazard inversion method, one simply needs to rearrange equation (40) to solve for the event time. That is,

$$T_i^s = H_0^{-1}(-\log(U_i)\exp(-\boldsymbol{w}_i'\boldsymbol{\gamma})) \tag{41}$$

which is the formula that was proposed by Bender et al. (2005).

5.3.1.4 Extending to complex data generating processes

If one can obtain an algebraic closed-form solution for the inverse cumulative baseline hazard, $H_0^{-1}(.)$, then a major benefit of the cumulative hazard inversion method is that it is

simple and computationally efficient. Moreover, it can be used to generate event times for a variety of parametric baseline hazards, including standard choices such as the exponential, Weibull or Gompertz distributions. However, two potential hurdles can be encountered when applying the method to complex data generating processes. First, one may not be able to obtain a closed-form solution to the cumulative baseline hazard $H_0(t)$. Second, the cumulative baseline hazard may not be invertible.

Crowther and Lambert (2013) therefore proposed an extension to overcome these two difficulties. Their extension incorporates numerical root finding and/or numerical quadrature. The root finding is used to numerically solve for T_i^s in situations where the cumulative baseline hazard function cannot be inverted analytically. A convenient choice of algorithm is Brent's univariate root finder (Brent, 1973), which can effectively find a solution to the equation

$$S_i(T_i^s) - U_i = 0 \tag{42}$$

or equivalently

$$\exp\left(-H_i(T_i^s)\right) - U_i = 0 \tag{43}$$

by treating T_i^s as the single unknown. The quadrature is used to numerically evaluate the cumulative hazard function in settings where it does not have a tractable-form. A standard choice of algorithm is either Gauss-Legendre or Gauss-Kronrod quadrature, whereby the cumulative hazard $H_i(T_i^s) = \int_{s=0}^{T_i^s} h_i(s)ds$ can be approximated by

$$H_i(T_i^s) \approx \frac{T_i^s}{2}\sum_{q=1}^{Q} v_q h_i\left(\frac{T_i^s\left(1+z_q\right)}{2}\right) \tag{44}$$

where v_q and z_q are, respectively, the standardised weights and locations ("abscissa") for the $q = 1, \ldots, Q$ quadrature nodes. In some situations, when the form of the baseline hazard function is complex, both the root finding and quadrature steps may be required, and so their approach involves iterating between the two until an appropriate solution to the root finding equation is obtained.

Simulating event times under a joint longitudinal and time-to-event model is one example of a situation in which the method of Crowther and Lambert can be applied. In general, with a current value association structure, or any other time-dependent association

structure, there is difficulty in obtaining a closed-form solution to the cumulative hazard function of the event submodel. This is because the general form of the time-dependency in the hazard function (which depends on the longitudinal submodel specification) often leads to an intractable integral for the cumulative hazard. Moreover, even if a closed-form solution for the cumulative hazard can be obtained, one would need to do this for each formulation of the joint model, thereby decreasing the wider applicability of such an approach (Sweeting and Thompson, 2011). The standard cumulative hazard inversion method is therefore not suitable for simulating event times under a variety of joint model formulations. Rather, incorporating the extension proposed by Crowther and Lambert provides a far more general solution. The use of numerical quadrature to evaluate the cumulative hazard ensures that one only needs to specify the form of the hazard or log hazard function, which is typically known, since it is usually explicitly defined in the model specification.

A unique set of circumstances warrant special mention. They correspond to the situation in which an individual is effectively cured (i.e. is no longer at risk of the event). This can occur, for example, if the simulated biomarker trajectory for an individual leads to a monotonically decreasing log hazard that effectively reaches negative infinity at some post-baseline time $t > 0$. When the log hazard reaches negative infinity, the hazard of the event is effectively zero, and the cumulative hazard and survival curve for that individual both become bounded (i.e. they reach an asymptote). If such a setting occurs, then – depending on the value drawn for the uniform random variable U_i – equation (43) may not have a solution; that is, any univariate root finding algorithm will not be able to find a solution for T_i^s on any positive finite interval. Although the event time in this setting is undefined, one may effectively consider the event time to be infinity for an individual who is cured. Therefore, a simple solution is to specify a maximum follow up time T_{max} and, if an individual has a simulated event time that is known to be greater than the maximum follow up time (including the situation where the simulated event time is after the time at which cure occurs), then that individual is given a censoring time equal to T_{max}. This is the approach adopted in the **simjm** package described in Section 5.3.1.6.

5.3.1.5 simsurv: An R package for simulating simple and complex time-to-event data

The method described by Crowther and Lambert (2013) is implemented as part of the **survsim** (Crowther and Lambert, 2012) package in the Stata software. The package allows

users to simulate event times from standard distributions (exponential, Weibull, Gompertz), 2-component mixture distributions, or any user-defined baseline hazard. The effects of covariates can be incorporated under a proportional hazards assumption or, alternatively, under non-proportional hazards by interacting covariates with a function of time.

However, until recently, users of R software did not have access to an equivalent package that allowed simulation of event times under a time-to-event model with any user-defined hazard function. Therefore, as part of the work of this PhD, the **simsurv** (Brilleman, 2018b) package was written in R. The **simsurv** package is publically available on CRAN (https://cran.r-project.org/package=simsurv). The package has been modelled on the **survsim** package from Stata. Although the native syntax of the two packages differ slightly, the respective underlying methodologies and the packages' functionality are intended to coincide.

Although the **simsurv** package has functionality for simulating from standard parametric proportional hazards models, 2-component mixture distributions, or under non-proportional hazards, these features will not be described here. Rather, in this section the only usage example provided will show how the **simsurv** package can be used to simulate event times under a shared parameter joint model. For further examples or technical details related to the package, the reader is referred to the vignettes provided in Appendix D.

Example: Simulating event times under a joint model using simsurv

This example shows how the **simsurv** package can be used to simulate event times under a shared parameter joint model. The simulated event times will be generated under the following random intercept and random slope model formulation for the longitudinal submodel

$$
\begin{gathered}
y_i(t) = \mu_i(t) + \varepsilon_i(t) \\
\mu_i(t) = \beta_{0i} + \beta_{1i}t + \beta_2 x_{1i} + \beta_3 x_{2i} \\
\text{where } \beta_{0i} = \beta_{00} + b_{0i}, \;\; \beta_{1i} = \beta_{10} + b_{1i}, \\
(b_{0i}, b_{1i})^T \sim N(0, \boldsymbol{\Sigma}) \text{ and } \varepsilon_i(t) \sim N\left(0, \sigma_{\mathrm{y}}^2\right)
\end{gathered}
\tag{45}
$$

where the random error terms $\varepsilon_i(t)$ are assumed to be independent of the individual-level random effects b_{0i} and b_{1i}, and the event submodel

$$h_i(t) = \delta t^{\delta-1} \exp(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \alpha \mu_i(t)) \quad (46)$$

where x_{1i} is an indicator variable for a binary covariate, x_{2i} is a continuous covariate, b_{0i} and b_{1i} are individual-level parameters (i.e. random effects) for the intercept and slope for individual i, the β and γ terms are population-level parameters (i.e. fixed effects), and δ is the shape parameter for the Weibull baseline hazard.

This specification allows for an individual-specific linear trajectory for the longitudinal submodel, a Weibull baseline hazard in the event submodel, a current value association structure, and the effects of a binary and a continuous covariate in both the longitudinal and event submodels. Although the binary and continuous covariates used in this example are common across the longitudinal and event submodels, this restriction is not required in general. In this example, we will use the following distributions for generating the covariates: $x_{1i} \sim \text{Bernoulli}(0.45)$ and $x_{2i} \sim N(44, 8.5)$. The latter could, for example, mimic the age distribution in a study of adults.

The R code required to return the hazard function for this joint model is:

```
# First define the hazard function to pass to simsurv
haz <- function(t, x, betas, ...) {
  betas[["delta"]] * (t ^ (betas[["delta"]] - 1)) * exp(
    betas[["gamma_0"]] +
    betas[["gamma_1"]] * x[["x1"]] +
    betas[["gamma_2"]] * x[["x2"]] +
    betas[["alpha"]] * (
      betas[["beta_0i"]] +
      betas[["beta_1i"]] * t +
      betas[["beta_2"]]  * x[["x1"]] +
      betas[["beta_3"]]  * x[["x2"]]
    )
  )
}
```

The next step is to define the "true" parameter values and the covariate data for each individual study subject. This is achieved by specifying two data frames: one for the parameter values, and one for the covariate data. Each row of the data frame will correspond to a different individual. The R code required to achieve this is:

```
# Then construct data frames with the true parameter
# values and the covariate data for each individual
set.seed(5454) # set seed before simulating data
N <- 200       # number of individuals

# Population-level (fixed effect) parameters
betas <- data.frame(
  delta   = rep(2,    N),
  gamma_0 = rep(-11.9,N),
  gamma_1 = rep(0.6,  N),
  gamma_2 = rep(0.08, N),
  alpha   = rep(0.03, N),
  beta_0  = rep(90,   N),
  beta_1  = rep(2.5,  N),
  beta_2  = rep(-1.5, N),
  beta_3  = rep(1,    N)
)

# Individual-level (random effect) parameters
b_corrmat <- matrix(c(1, 0.5, 0.5, 1), 2, 2)
b_sds     <- c(20, 3)
b_means   <- rep(0, 2)
b_z       <- MASS::mvrnorm(n = N, mu = b_means, Sigma = b_corrmat)
b         <- sapply(1:length(b_sds),
                    FUN = function(x) b_sds[x] * b_z[,x])
betas$beta_0i <- betas$beta_0 + b[,1]
betas$beta_1i <- betas$beta_1 + b[,2]

# Covariate data
covdat <- data.frame(
  x1 = stats::rbinom(N, 1, 0.45), # the binary covariate
  x2 = stats::rnorm(N, 44, 8.5)   # the continuous covariate
)
```

The final step is to generate the simulated event times using a call to the ‘simsurv’ function. The only arguments that need to be specified are the user-defined hazard function, the true parameter values, and the covariate data. In this example, a maximum follow up time of ten units will also be used (for example, ten years), after which individuals will be censored if they have not yet experienced the event. The R code required to generate the simulated event times and then display the first few rows of the resulting data frame is as follows:

```
# Set seed for simulations
set.seed(546546)
```

```
# Then simulate the event times based on the user-defined
# hazard function, covariates data, and true parameter values
times <- simsurv(hazard = haz, x = covdat, betas = betas, maxt = 10)
head(times)
## id eventtime status
##  1  4.813339      1
##  2  9.763900      1
##  3  5.913436      1
##  4  2.823562      1
##  5  2.315488      1
##  6 10.000000      0
```

5.3.1.6 simjm: An R package for simulating joint longitudinal and time-to-event data

In the previous section, it was demonstrated how the **simsurv** R package can be used to simulate event times under a shared parameter joint model. The end-user of the package required relatively little knowledge about the underlying statistical methodology, since the package only required a user-defined hazard function to be provided and this was readily available as the event submodel specified in the data generating model. However, generating event times under a joint model using the **simsurv** package requires numerous lines of computing code. It also requires a new definition of the hazard function for each joint model association structure one may wish to use.

To perform the simulation study in the following section of this thesis, it is necessary to simulate event times under several joint model specifications, for example, varying numbers of longitudinal outcomes and different types of association structures. To facilitate this process, the **simjm** (Brilleman, 2018a) R package was developed as part of this PhD. The **simjm** package acts as a wrapper to the **simsurv** package, and provides a user-friendly interface specifically for simulating joint longitudinal and time-to-event data. In addition to simulating event times under the joint model formulation, it also returns the simulated longitudinal data. The longitudinal data, for example, repeated biomarker measurements, are generated as random draws from the appropriate conditional distribution specified in the data generating model, thereby incorporating measurement error, and evaluated at either scheduled visit times or random visit times depending on the choice made by the user. The

use of the **simjm** package is demonstrated next, by simulating data under the same joint model as used in the **simsurv** example in the previous section.

Example: Simulating joint longitudinal and time-to-event data using simjm

The code below shows how to simulate longitudinal and time-to-event data for 200 individuals under the joint model that was specified in equations (45) and (46):

```
# Use simjm to generate joint longitudinal and event data
simdat <- simjm(
  M = 1, n = 200,
  max_fuptime          = 10,
  betaEvent_aux        = 2,
  betaEvent_intercept  = -11.9,
  betaEvent_binary     = 0.6,
  betaEvent_continuous = 0.08,
  betaEvent_assoc      = 0.03,
  betaLong_intercept   = 90,
  betaLong_slope       = 2.5,
  betaLong_binary      = -1.5,
  betaLong_continuous  = 1,
  b_sd    = c(20, 3),
  b_rho   = 0.5,
  family  = gaussian(),
  prob_Z1 = 0.45,
  mean_Z2 = 44,
  sd_Z2   = 8.5)
```

Arguments are provided in the call to the 'simjm' function for specifying the number of longitudinal biomarkers, the number of individuals, the maximum follow up time, the type of distribution for each longitudinal biomarker (i.e. family), the distributions of the covariates, and the "true" values for each of the parameters. Other arguments that are not shown in the code above (since they were set equal to their defaults) include the type of longitudinal trajectory (linear or quadratic), the choice of association structure, and more. The user is not required to explicitly specify the hazard or log hazard function (as they would have to if using **simsurv** directly), since this is determined based on the user inputs to the arguments of the 'simjm' function.

The simulated data are returned in a separate data frame for each submodel, that is, one for each longitudinal submodel (i.e. each biomarker, measured with error) and one for the event

submodel (i.e. event times and event indicator). We can examine the first few rows of each of these data frames using the following code:

```
# Examine the first few rows of the longitudinal data
head(simdat$Long1)
##  id Z1      Z2 eventtime status        tij    Yij_1
##   1  0 34.2582  7.408827      1 0.00000000 140.4578
##   1  0 34.2582  7.408827      1 2.99705843 158.9575
##   1  0 34.2582  7.408827      1 4.39900056 166.8687
##   1  0 34.2582  7.408827      1 4.49202040 166.3300
##   1  0 34.2582  7.408827      1 6.44825677 176.7721
##   1  0 34.2582  7.408827      1 6.75222545 179.8334

# Examine the first few rows of the event data
head(simdat$Event)
##  id Z1       Z2 eventtime status
##   1  0 34.25820  7.408827      1
##   3  0 54.39843  5.548837      1
##   4  0 45.73779  1.256674      1
##   5  1 47.81405  1.578142      1
##   6  0 68.17889  1.104480      1
```

The first data frame shows the multiple-row per-individual biomarker data, with the first individual having at least six biomarker measurements, taken at at baseline and random post-baseline measurement times. The second data frame shows the single-row per-individual event data with the first six individuals all experiencing the event during follow up. The **simjm** package will be used to generate data for the **s**imulation study described in the following section.

5.3.2 Simulation study

This section describes a simulation study evaluating the performance of the joint modelling functionality in the **rstanarm** R package. The objective of the simulation study was to assess whether **rstanarm** was able to recover the true value for each of the parameters used in the data generating models. The data generating models, the inferential quantities, and the results of the simulation study are described in the following sections.

5.3.2.1 Data generating models

Two sets of simulations were performed in this simulation study. The first set of simulations was based on a univariate joint model (i.e. one longitudinal outcome) for the data

generating process, and considered several different types of association structure. The second set of simulations was based on a multivariate joint model (i.e. more than one longitudinal outcome) and considered a current value association structure. The simulation study using the multivariate joint model took significantly longer computing time to complete and therefore an extensive simulation study incorporating a wide range of association structures was not feasible.

The data in the first set of simulations were generated under the following univariate joint model. The longitudinal submodel took the form

$$\begin{gathered} y_i(t) = \mu_i(t) + \varepsilon_i(t) \\ \mu_i(t) = \beta_{0i} + \beta_{1i}t + \beta_2 x_{1i} + \beta_3 x_{2i} \\ \text{where } \beta_{0i} = \beta_{00} + b_{0i}, \ \ \beta_{1i} = \beta_{10} + b_{1i}, \\ (b_{0i}, b_{1i})^T \sim N(0, \boldsymbol{\Sigma}) \text{ and } \varepsilon_i(t) \sim N\left(0, \sigma_y^2\right) \end{gathered} \tag{47}$$

where the random error terms $\varepsilon_i(t)$ are assumed to be independent of the individual-level random effects b_{0i} and b_{1i}. For the event submodel, three types of association structures were considered. These were as follows

$$\text{Model 1:} \quad h_i(t) = \delta t^{\delta-1} \exp(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \alpha_1 \mu_i(t)) \tag{48}$$

$$\text{Model 2:} \quad h_i(t) = \delta t^{\delta-1} \exp(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \alpha_2 \tfrac{d\mu_i(t)}{dt}) \tag{49}$$

$$\text{Model 3:} \quad h_i(t) = \delta t^{\delta-1} \exp(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \alpha_3 \textstyle\int_0^t \mu_i(u)\, du) \tag{50}$$

These models each have the same baseline hazard (Weibull) and time-fixed covariates (one binary and one continuous), but differ in the assumed association structure (current value, current slope, and cumulative effects respectively).

The covariate values for each individual were simulated according to the following distributions, $x_{1i} \sim Bern(0.5)$ and $x_{2i} \sim N(0,1)$. The true parameter values that were used for the data generating models are shown in Table 3. For the variance-covariance matrix of the individual-specific parameters, $\boldsymbol{\Sigma}$, consider the decomposition $\boldsymbol{\Sigma} = \mathbf{V}\,\mathbf{R}\,\mathbf{V}$ where $\mathbf{R}$ is the correlation matrix for the individual-specific parameters and $\mathbf{V} = \text{diag}(\boldsymbol{\sigma}_\mathbf{b})$ is a square diagonal matrix with the diagonal elements being the values from a vector of standard

deviations for the individual-specific parameters, $\boldsymbol{\sigma}_{\mathbf{b}}$. For the univariate joint models, the true standard deviations were taken to be $\boldsymbol{\sigma}_{\mathbf{b}} = (2, 1)$ and the true correlation matrix was $\mathbf{R} = \begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix}$.

The data in the second set of simulations were generated under the following multivariate joint model. The longitudinal submodel for the m^{th} $(m = 1,2,3)$ longitudinal outcome takes the form

$$\begin{gathered} y_i^{(m)}(t) = \mu_i^{(m)}(t) + \varepsilon_i^{(m)}(t) \\ \mu_i^{(m)}(t) = \beta_{0i}^{(m)} + \beta_{1i}^{(m)}(t) + \beta_2^{(m)} x_{1i} + \beta_3^{(m)} x_{2i} \\ \text{where } \beta_{0i}^{(m)} = \beta_{00}^{(m)} + b_{0i}^{(m)}, \;\; \beta_{1i}^{(m)} = \beta_{10}^{(m)} + b_{1i}^{(m)}, \\ \left(b_{0i}^{(1)}, b_{1i}^{(1)}, b_{0i}^{(2)}, b_{1i}^{(2)}, b_{0i}^{(3)}, b_{1i}^{(3)}\right)^T \sim N(0, \boldsymbol{\Sigma}) \text{ and } \varepsilon_i^{(m)}(t) \sim N\left(0, \sigma_{\mathrm{y}}^{(m)}\right) \end{gathered} \tag{51}$$

where each of the three random error terms are assumed to be independent of the six individual-level random effect terms. For the event submodel, a current value association structure is used for linking each expected biomarker value to the log hazard of the event. Therefore, the event submodel takes the form

$$h_i(t) = \delta t^{\delta - 1} \exp(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \sum_{m=1}^{3} \alpha_m \mu_i^{(m)}(t)) \tag{52}$$

Again, the covariate values were simulated according to $x_{1i} \sim Bern(0.5)$ and $x_{2i} \sim N(0,1)$ distributions. The "true" parameter values that were used for the data generating model are shown in Table 4. For the multivariate joint model, the true standard deviations used for the individual-specific parameters were $\boldsymbol{\sigma}_{\mathbf{b}} = (2, 1, 2, 1, 2, 1)$ and the correlation matrix used was

$$\mathbf{R} = \begin{bmatrix} 1 & 0.2 & 0.5 & 0 & -0.3 & 0 \\ 0.2 & 1 & 0 & 0.2 & 0 & 0 \\ 0.5 & 0 & 1 & 0.3 & 0.1 & 0 \\ 0 & 0.2 & 0.3 & 1 & 0 & 0 \\ -0.3 & 0 & 0.1 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0 & 0.2 & 1 \end{bmatrix} \tag{53}$$

5.3.2.2 Inferential quantities

For each data generating model (i.e. three univariate joint models, and one multivariate joint model) there were $D = 200$ simulated datasets generated, with each dataset containing $N = 200$ individuals. An analysis model (intended to coincide with the data generating model) was fit to each simulated dataset using a single chain of 2000 MCMC iterations, which included a warm-up phase of 1000 iterations that were not used for inference. Convergence for each simulated dataset was addressed by ensuring that each parameter had an estimated R-hat statistic less than 1.1 (Stan Development Team, 2017b).

The ability to recover the true parameter values was assessed using the following approach. For the model that was fit to the d^{th} $(d = 1, \ldots, D)$ simulated dataset the following estimates were calculated:

- The mean of the posterior distribution for parameter k, denoted $\hat{\theta}_k^{(d)}$
- The bias for parameter k, defined as $\hat{B}_k^{(d)} = \hat{\theta}_k^{(d)} - \theta_k$ where θ_k denotes the true value of parameter k that was used to simulate the data

Table 3. True parameter values used for simulation study (univariate joint models)

Parameter	Value	Description of parameter
β_{00}	0	Population-level intercept in longitudinal submodel
β_{10}	1	Population-level linear slope in longitudinal submodel
β_2	1	Population-level coefficient for binary covariate in longitudinal submodel
β_3	1	Population-level coefficient for continuous covariate in longitudinal submodel
γ_0	-4	Population-level intercept in event submodel
γ_1	1	Population-level coefficient for binary covariate in event submodel
γ_2	0	Population-level coefficient for continuous covariate in event submodel
α_1	0.2	Association parameter (current value association structure, under Model 1)
α_2	0.2	Association parameter (current slope association structure, under Model 2)
α_3	0.2	Association parameter (cumulative effects association structure, under Model 3)
δ	1.2	Shape parameter for Weibull baseline hazard

Table 4. True parameter values used for simulation study (multivariate joint model)

Parameter	Value	Description of parameter
$\beta_{00}^{(1)}, \beta_{00}^{(2)}, \beta_{00}^{(3)}$	1, -1, 0	Population-level intercept in longitudinal submodels
$\beta_{10}^{(1)}, \beta_{10}^{(2)}, \beta_{10}^{(3)}$	1, 2, -1	Population-level linear slope in longitudinal submodels
$\beta_{2}^{(1)}, \beta_{2}^{(2)}, \beta_{2}^{(3)}$	1, 1, 1	Population-level coefficient for binary covariate in longitudinal submodels
$\beta_{3}^{(1)}, \beta_{3}^{(2)}, \beta_{3}^{(3)}$	1, 1, 1	Population-level coefficient for continuous covariate in longitudinal submodels
γ_0	-4	Population-level intercept in event submodel
γ_1	1	Population-level coefficient for binary covariate in event submodel
γ_2	0	Population-level coefficient for continuous covariate in event submodel
α_1	0.1	Association parameter (current value association structure for biomarker 1)
α_2	0.2	Association parameter (current value association structure for biomarker 2)
α_3	-0.1	Association parameter (current value association structure for biomarker 3)
δ	1.2	Shape parameter for Weibull baseline hazard

- The relative bias for parameter k, defined as $\hat{R}_k^{(d)} = \theta_k^{-1}\left(\hat{\theta}_k^{(d)} - \theta_k\right)$ where θ_k denotes the true value of parameter k that was used to simulate the data
- The standard deviation of the posterior distribution (i.e. estimated standard error) for parameter k, denoted $\hat{S}_k^{(d)}$

For inference in the simulation study, the following quantities and plots were then calculated using the estimates obtained across the D datasets:

- The mean bias for parameter k, defined as $\bar{B}_k = \frac{1}{D}\sum_{\mathrm{d}=1}^{D} \hat{B}_k^{(d)}$ ("mean bias")
- The mean relative bias for parameter k, defined as $\bar{R}_k = \frac{1}{D}\sum_{\mathrm{d}=1}^{D} \hat{R}_k^{(d)}$ ("mean relative bias")
- The mean standard deviation of the posterior distribution for parameter k, defined as $\bar{S}_k = \frac{1}{D}\sum_{\mathrm{d}=1}^{D} \hat{S}_k^{(d)}$ ("mean estimated standard error")
- The standard error of the posterior mean for parameter k, defined as the standard deviation of the estimates $\left\{\hat{\theta}_k{}^{(d)}; d = 1, \ldots, D\right\}$ ("empirical standard error").
- Density plots of the posterior mean estimates $\left\{\hat{\theta}_k{}^{(d)}; d = 1, \ldots, D\right\}$ for parameter k; these were overlaid with a dashed line showing the true parameter value θ_k that was used to simulate the data ("empirical sampling distribution of the posterior mean").

5.3.2.3 Results

Figure 2 through 5 show, for each of the data generating models (i.e. three univariate joint models and one multivariate joint model), density plots of the posterior mean estimates for each parameter (i.e. the empirical distribution of $\hat{\theta}_k$). The dashed lines in the plots show the location of the true parameter value used to simulate the data (i.e. the location of θ_k). Tables 5 through 8 show, for each of the data generating models, the estimated mean bias, mean relative bias, mean estimated standard error, and empirical standard error for each of the parameters.

Overall, the results from the simulation study suggest that **rstanarm** was able to recover the true parameter values used in the data generating models. The true values for the parameters (the dashed lines in Figures 2 through 5) are located close to the centre of the sampling distribution for the posterior mean. Moreover, the tables demonstrate that

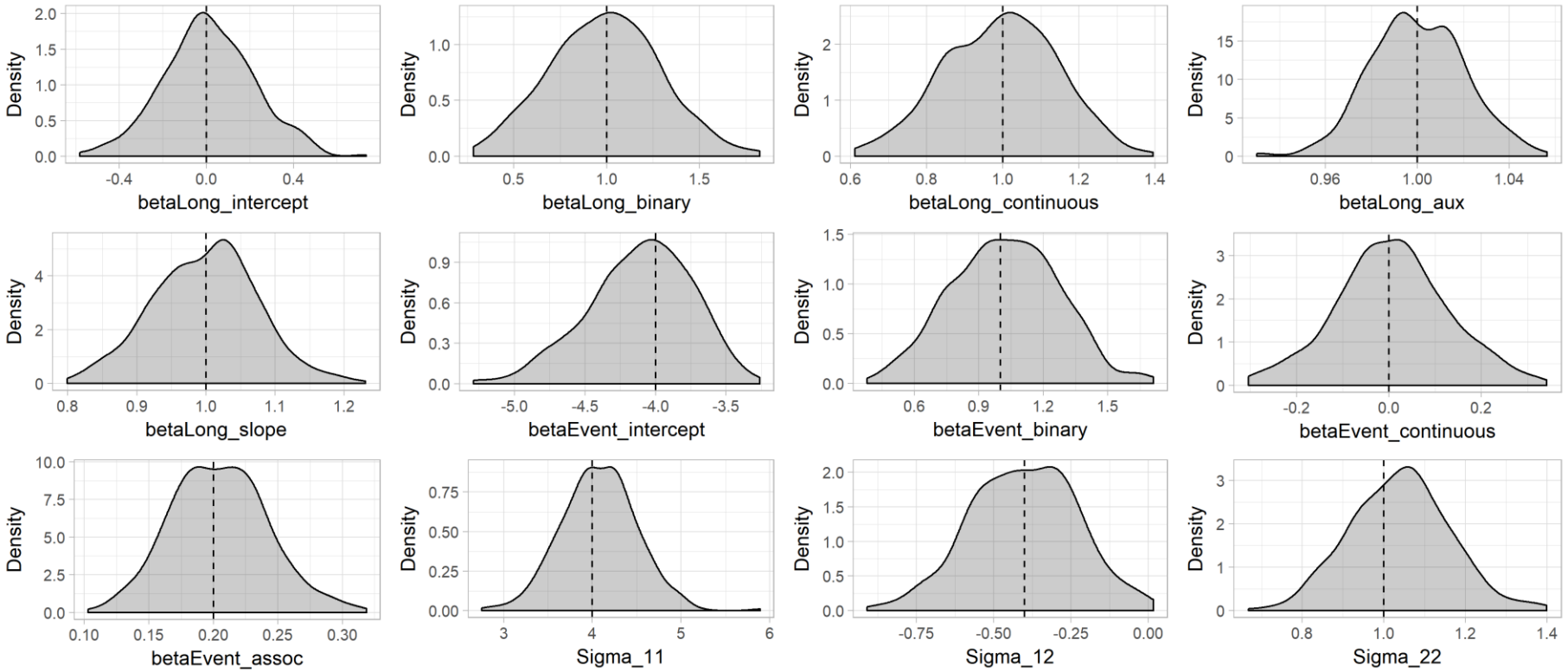


Figure 2. Simulation study results for univariate joint model 1; kernel density plots showing the distribution (across the 200 simulated datasets) of the estimated posterior mean for each parameter. The dashed line shows the true parameter value used in the data generating model.

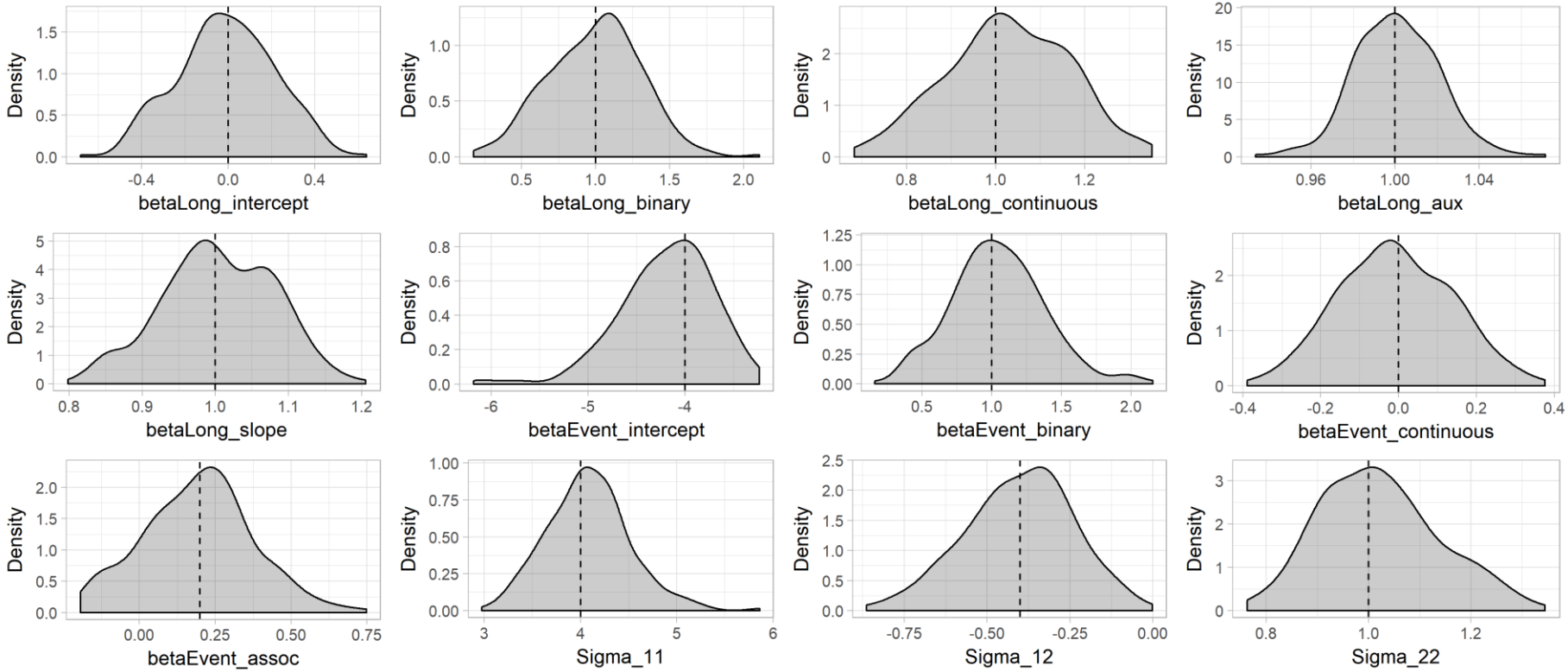


Figure 3. Simulation study results for univariate joint model 2; kernel density plots showing the distribution (across the 200 simulated datasets) of the estimated posterior mean for each parameter. The dashed line shows the true parameter value used in the data generating model.

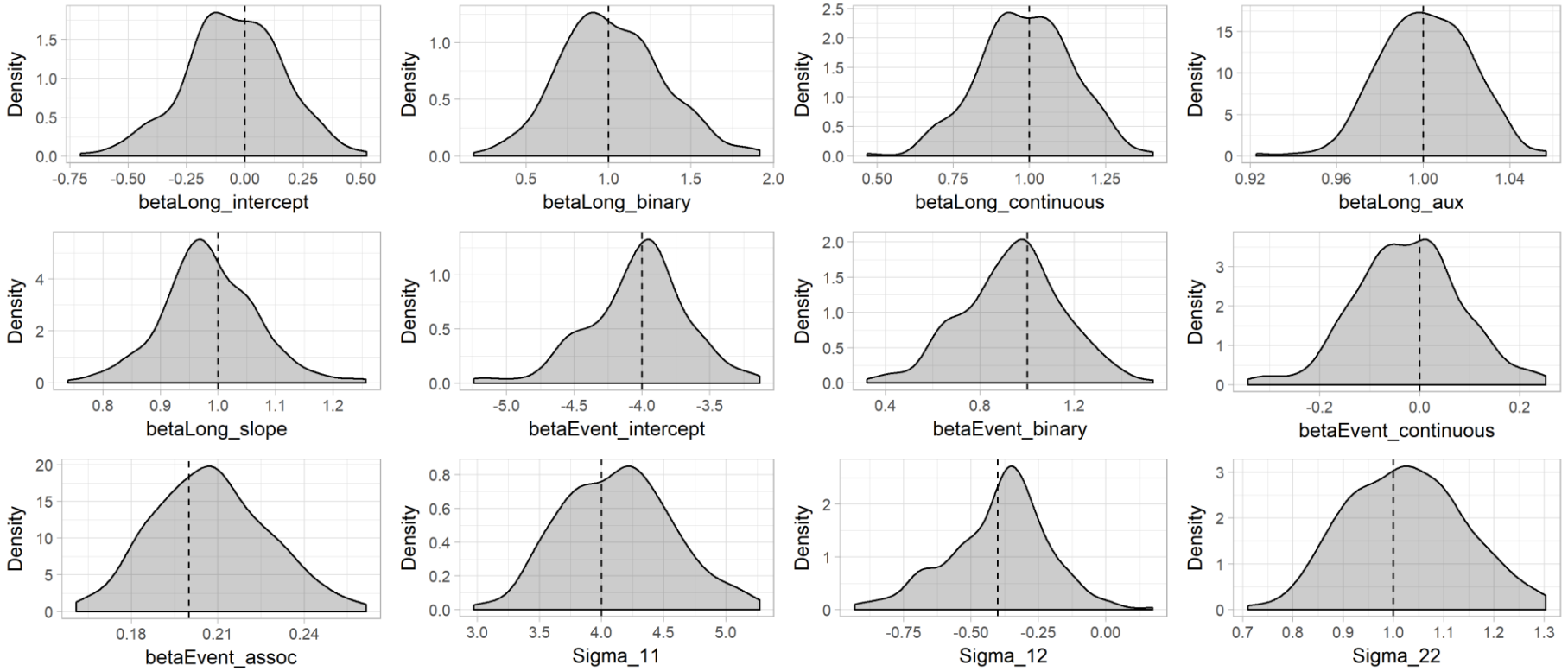


Figure 4. Simulation study results for univariate joint model 3; kernel density plots showing the distribution (across the 200 simulated datasets) of the estimated posterior mean for each parameter. The dashed line shows the true parameter value used in the data generating model.

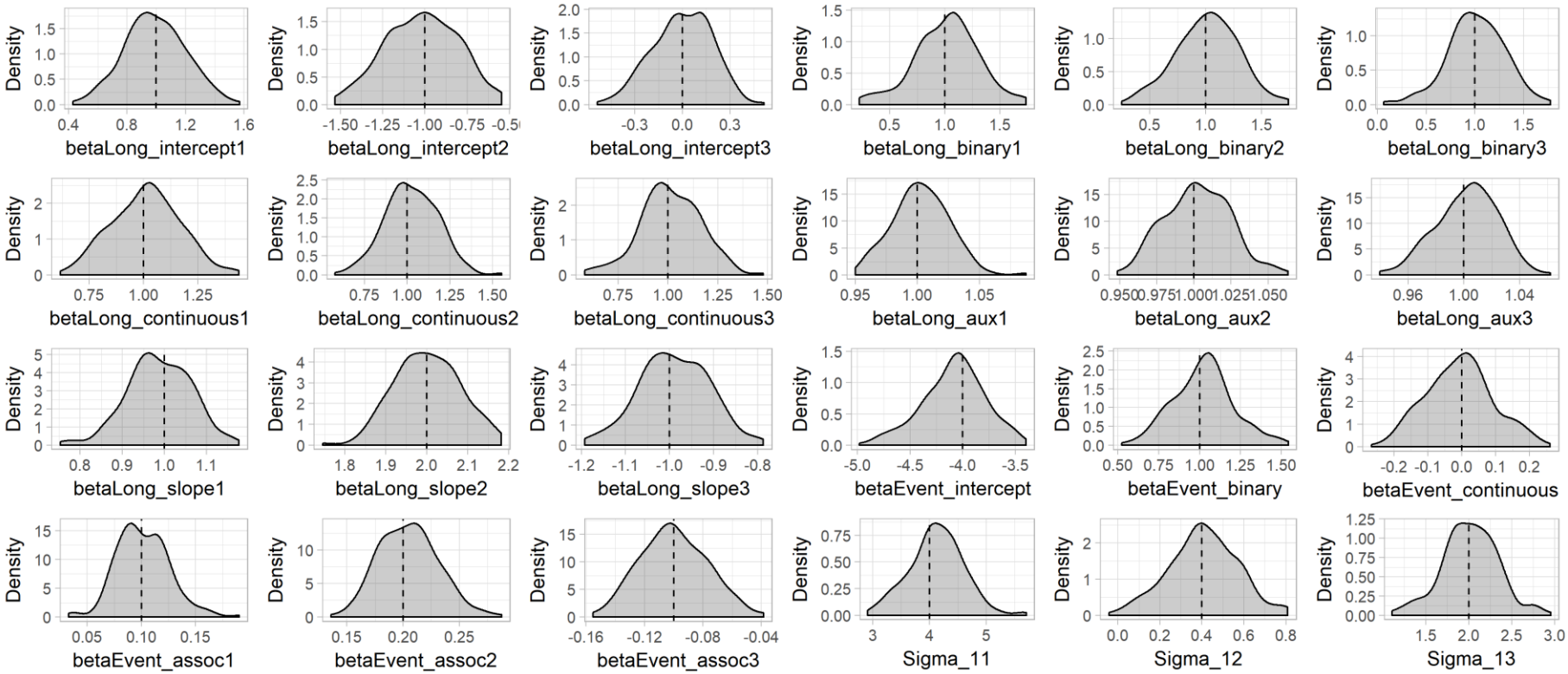


Figure 5. Simulation study results for the multivariate joint model; kernel density plots showing the distribution (across the 200 simulated datasets) of the estimated posterior mean for each parameter. The dashed line shows the true parameter value used in the data generating model.

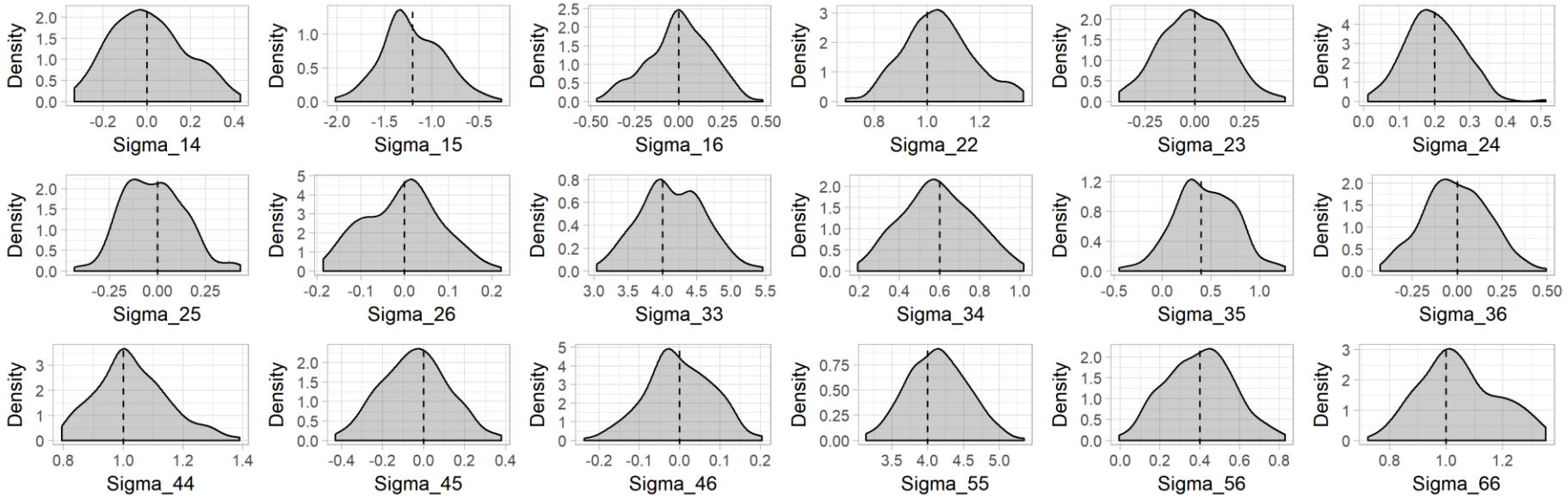


Figure 5 (cont'd). Simulation study results for the multivariate joint model; kernel density plots showing the distribution (across the 200 simulated datasets) of the estimated posterior mean for each parameter. The dashed line shows the true parameter value used in the data generating model.

Table 5. Simulation study results for univariate joint model 1. Estimated mean bias, mean relative bias, mean estimated standard error, and empirical standard error for each of the parameters.

Parameter	True	$\bar{B}_k$	$\bar{R}_k$	$\bar{S}_k$	$sd(\hat{\theta}_k)$
β_{00}	0	0.011	n/a	0.208	0.209
β_{10}	1	-0.008	-0.812	0.289	0.295
β_2	1	-0.006	-0.616	0.146	0.147
β_3	1	-0.001	-0.119	0.078	0.075
σ_y	1	0.000	0.002	0.020	0.020
γ_0	-4	-0.109	2.715	0.346	0.365
γ_1	1	0.022	2.208	0.251	0.248
γ_2	0	0.007	n/a	0.120	0.121
α_1	0.2	0.005	2.310	0.036	0.037
$\mathbf{\Sigma}_{[1,1]}$	4	0.075	1.877	0.456	0.423
$\mathbf{\Sigma}_{[1,2]}$	0.4	-0.003	0.692	0.172	0.172
$\mathbf{\Sigma}_{[2,2]}$	1	0.039	3.940	0.122	0.122

Abbreviations. $\bar{B}_k$: mean bias for parameter k; $\bar{R}_k$: mean relative bias (%) for parameter k; $\bar{S}_k$: mean estimated standard error for parameter k; $sd(\hat{\theta}_k)$: empirical standard error of the posterior mean for parameter k; n/a: not applicable.

Table 6. Simulation study results for univariate joint model 2. Estimated mean bias, mean relative bias, mean estimated standard error, and empirical standard error for each of the parameters.

Parameter	True	$\bar{B}_k$	$\bar{R}_k$	$\bar{S}_k$	$sd(\hat{\theta}_k)$
β_{00}	0	-0.011	n/a	0.207	0.223
β_{10}	1	0.004	0.403	0.078	0.076
β_2	1	-0.005	-0.524	0.289	0.313
β_3	1	0.027	2.729	0.145	0.138
σ_y	1	0.001	0.071	0.019	0.020
γ_0	-4	-0.191	4.763	0.454	0.487
γ_1	1	0.046	4.638	0.326	0.336
γ_2	0	-0.009	n/a	0.146	0.145
α_2	0.2	-0.006	-2.862	0.187	0.179
$\mathbf{\Sigma}_{[1,1]}$	4	0.080	2.012	0.454	0.434
$\mathbf{\Sigma}_{[1,2]}$	0.4	-0.003	0.765	0.166	0.163
$\mathbf{\Sigma}_{[2,2]}$	1	0.021	2.128	0.116	0.116

Abbreviations. $\bar{B}_k$: mean bias for parameter k; $\bar{R}_k$: mean relative bias (%) for parameter k; $\bar{S}_k$: mean estimated standard error for parameter k; $sd(\hat{\theta}_k)$: empirical standard error of the posterior mean for parameter k; n/a: not applicable.

Table 7. Simulation study results for univariate joint model 3. Estimated mean bias, mean relative bias, mean estimated standard error, and empirical standard error for each of the parameters.

Parameter	True	$\bar{B}_k$	$\bar{R}_k$	$\bar{S}_k$	$sd(\hat{\theta}_k)$
β_{00}	0	-0.047	n/a	0.210	0.208
β_{10}	1	-0.016	-1.600	0.078	0.082
β_2	1	0.019	1.857	0.293	0.308
β_3	1	-0.013	-1.279	0.147	0.153
σ_y	1	0.001	0.118	0.021	0.021
γ_0	-4	-0.027	0.680	0.339	0.360
γ_1	1	-0.062	-6.169	0.203	0.211
γ_2	0	-0.024	n/a	0.103	0.105
α_3	0.2	0.008	3.790	0.020	0.020
$\mathbf{\Sigma}_{[1,1]}$	4	0.127	3.163	0.466	0.438
$\mathbf{\Sigma}_{[1,2]}$	0.4	0.003	-0.798	0.176	0.184
$\mathbf{\Sigma}_{[2,2]}$	1	0.024	2.406	0.122	0.116

Abbreviations. $\bar{B}_k$: mean bias for parameter k; $\bar{R}_k$: mean relative bias (%) for parameter k; $\bar{S}_k$: mean estimated standard error for parameter k; $sd(\hat{\theta}_k)$: empirical standard error of the posterior mean for parameter k; n/a: not applicable.

Table 8. Simulation study results for the multivariate joint model. Estimated mean bias, mean relative bias, mean estimated standard error, and empirical standard error for each of the parameters.

Parameter	True	$\bar{B}_k$	$\bar{R}_k$	$\bar{S}_k$	$sd(\hat{\theta}_k)$
$\beta_{00}^{(1)}$	1	-0.008	-0.813	0.211	0.209
$\beta_{00}^{(2)}$	-1	-0.016	1.623	0.211	0.214
$\beta_{00}^{(3)}$	0	-0.002	n/a	0.212	0.186
$\beta_{10}^{(1)}$	1	-0.013	-1.276	0.082	0.075
$\beta_{10}^{(2)}$	2	0.002	0.095	0.082	0.080
$\beta_{10}^{(3)}$	-1	0.011	-1.060	0.082	0.079
$\beta_{2}^{(1)}$	1	0.015	1.486	0.299	0.292
$\beta_{2}^{(2)}$	1	0.015	1.503	0.296	0.280
$\beta_{2}^{(3)}$	1	0.013	1.251	0.300	0.293
$\beta_{3}^{(1)}$	1	0.012	1.193	0.150	0.157
$\beta_{3}^{(2)}$	1	0.024	2.419	0.149	0.155
$\beta_{3}^{(3)}$	1	0.013	1.269	0.150	0.151
$\sigma_{y}^{(1)}$	1	0.003	0.275	0.022	0.022
$\sigma_{\mathrm{y}}^{(2)}$	1	0.002	0.226	0.022	0.022
$\sigma_{\mathrm{y}}^{(3)}$	1	0.002	0.230	0.021	0.022
γ_0	-4	-0.085	2.135	0.307	0.297
γ_1	1	0.023	2.290	0.190	0.184
γ_2	0	-0.006	n/a	0.096	0.099
α_1	0.1	0.002	1.516	0.025	0.024
α_2	0.2	0.005	2.510	0.026	0.026

α_3	-0.1	0.000	-0.261	0.024	0.023
$\boldsymbol{\Sigma}_{[1,1]}$	4	0.124	3.100	0.468	0.481
$\boldsymbol{\Sigma}_{[1,2]}$	0.4	0.008	1.974	0.175	0.164
$\boldsymbol{\Sigma}_{[1,3]}$	2	0.015	0.752	0.356	0.324
$\boldsymbol{\Sigma}_{[1,4]}$	0	0.009	n/a	0.174	0.170
$\boldsymbol{\Sigma}_{[1,5]}$	-1.2	0.000	0.020	0.336	0.319
$\boldsymbol{\Sigma}_{[1,6]}$	0	0.004	n/a	0.177	0.174
$\boldsymbol{\Sigma}_{[2,2]}$	1	0.044	4.372	0.130	0.132
$\boldsymbol{\Sigma}_{[2,3]}$	0	-0.002	n/a	0.176	0.165
$\boldsymbol{\Sigma}_{[2,4]}$	0.2	0.000	-0.139	0.092	0.080
$\boldsymbol{\Sigma}_{[2,5]}$	0	-0.019	n/a	0.176	0.159
$\boldsymbol{\Sigma}_{[2,6]}$	0	-0.004	n/a	0.091	0.085
$\boldsymbol{\Sigma}_{[3,3]}$	4	0.143	3.580	0.468	0.472
$\boldsymbol{\Sigma}_{[3,4]}$	0.6	-0.007	-1.204	0.177	0.177
$\boldsymbol{\Sigma}_{[3,5]}$	0.4	0.036	9.003	0.326	0.307
$\boldsymbol{\Sigma}_{[3,6]}$	0	-0.007	n/a	0.177	0.176
$\boldsymbol{\Sigma}_{[4,4]}$	1	0.029	2.936	0.130	0.120
$\boldsymbol{\Sigma}_{[4,5]}$	0	-0.042	n/a	0.173	0.161
$\boldsymbol{\Sigma}_{[4,6]}$	0	-0.001	n/a	0.091	0.082
$\boldsymbol{\Sigma}_{[5,5]}$	4	0.136	3.408	0.470	0.415
$\boldsymbol{\Sigma}_{[6,6]}$	0.4	-0.002	-0.418	0.176	0.168
$\boldsymbol{\Sigma}_{[6,6]}$	1	0.047	4.652	0.131	0.137

Abbreviations. $\bar{B}_k$: mean bias for parameter k; $\bar{R}_k$: mean relative bias (%) for parameter k; $\bar{S}_k$: mean estimated standard error for parameter k; $sd(\hat{\theta}_k)$: empirical standard error of the posterior mean for parameter k; n/a: not applicable.

the mean estimated standard error (i.e. the mean standard deviation for the posterior distribution) was close to the empirical standard error for all parameters.

5.4 Comparison of software for multivariate joint models

This section provides a qualitative comparison of the software packages currently available for fitting multivariate joint models for longitudinal and time-to-event data. *Multivariate* is defined here as meaning joint models that can accommodate more than one longitudinal outcome.

In 2016, Hickey et al. reviewed the methods for joint modelling of multivariate longitudinal and time-to-event data, and summarised the software that was available for fitting such models at that time. They highlighted that the software available for fitting multivariate joint models was extremely limited. However, since 2016 there have been a number of developments in this area. In particular, during 2017 several new or updated software packages were released that provide functionality for estimating multivariate shared parameter joint models. Here, we focus on the general purpose software packages that are now available. Specifically, we include the following five packages in our comparison: **rstanarm**, **JMbayes**, **joineRML**, **survtd**, and **megenreg**. A brief overview of each package is provided below.

5.4.1 The rstanarm R package

The **rstanarm** (Brilleman et al., 2018; Stan Development Team, 2017a) R package estimates shared parameter joint models under a Bayesian approach via the software Stan. The joint modelling functionality in **rstanarm** has been developed as part of the work for this PhD. An introduction to this package was provided in the paper presented in Section 5.2. Nonetheless, the key details related to the package will be described again in brief here.

The joint model can be univariate (i.e. one longitudinal outcome) or multivariate (i.e. more than one longitudinal outcome). The longitudinal outcomes can be continuous, binary, or counts and each longitudinal outcome can be of a different type. Clustering factors beyond that of the individual (e.g. patients clustered within clinics, or multiple tumour lesions clustered within patients) are allowed. The event submodel is assumed to be a parametric proportional hazards model, and several options are allowed for the baseline hazard including a Weibull distribution, a piecewise constant baseline hazard, or more flexible shapes by using B-splines for modelling the log baseline hazard.

The user specifies the joint model using customary R formula syntax and data frames. The main modelling function ('stan_jm') returns a model object for which there are a large number of post-estimation functions available, for example, diagnostic plots or dynamic predictions. The back-end estimation of the model is carried out using the Bayesian software Stan, a C++ library for full Bayesian inference based on an implementation of Hamiltonian Monte Carlo (although approximate Bayesian inference and optimisation algorithms are also available).

5.4.2 The JMbayes R package

The **JMbayes** (Rizopoulos, 2016, 2017) R package estimates shared parameter joint models under a Bayesian framework using a Metropolis algorithm. The **JMbayes** package has been publically available on CRAN since 2012, however, in 2017 it was extended to handle multivariate joint models. The package currently allows for either continuous, binary or count biomarker data through a linear, logistic, or Poisson mixed effects regression submodel for the longitudinal outcome. The event submodel is specified as a parametric proportional hazards model with the baseline hazard estimated using penalised cubic splines. There is significant flexibility in the specification of the association structure for the joint model, since a user-specified definition of the association structure is possible. A range of post-estimation functions are available for the fitted joint model, including dynamic predictions.

5.4.3 The joineRML R package

The **joineRML** (Hickey et al., 2017) R package estimates multivariate shared parameter joint models under an extension of the joint model proposed by Henderson et al. (2000). The longitudinal outcomes are assumed to be normally distributed continuous biomarkers, specified using a (multivariate) linear mixed effects regression submodel. The event submodel is specified as a semi-parametric proportional hazards model with the baseline hazard left unspecified. Estimation of the model is performed under an EM algorithm. Approximate standard errors are provided by default, although the user can also elect to obtain standard errors by bootstrapping (however, the latter can be relatively more time consuming). The association structure for the joint model is based on a current value association structure. A range of post-estimation functions are available for the fitted joint model, including dynamic predictions.

5.4.4 The survtd R package

The **survtd** (Moreno-Betancur, 2017) R package fits multivariate joint models using a two-stage multiple imputation for joint modelling (MIJM) approach. The methods are described in Moreno-Betancur et al. (2017) and were briefly discussed in Section 2.5.3 of this thesis. Although the paper is not part of this thesis, it was included in the *"Research outcomes during enrolment"* section at the start of this thesis, owing to my contribution to the work and software, and co-authorship of the resulting publication. In brief, the approach uses the following two stages:

(i) the so-called "true" underlying values of each longitudinal biomarker are estimated via multiple imputation by chained equations (MICE), with a novel adjustment in the imputation model such that it corrects for informative dropout in the longitudinal process due to occurrence of the event, and

(ii) the multiple imputations are included in a semi-parametric Cox or additive hazards model, and parameter estimates are obtained via Rubin's rules (Rubin, 1987).

Although the approach is not based on the full likelihood for the joint model, it has been shown to perform well in a range of simulations. The software is currently limited to longitudinal outcomes that are assumed to be continuous and normally distributed. The association structure for the joint model is based on a current value association structure. Because of the two-stage process, the software provides flexibility in the formulation of the event submodel. Specifically, the user can choose between either a proportional hazards submodel or additive hazards submodel. The latter may be desirable in settings where one wishes to estimate absolute rate differences rather than relative rate differences (i.e. hazard differences not hazard ratios). The baseline hazard is left unspecified in both the proportional and additive hazard formulations.

5.4.5 The megenreg Stata package

The **megenreg** (Crowther, 2017a, 2017c) Stata package provides an estimation framework for fitting extended multivariate generalised linear and non-linear mixed effects models. Joint models for longitudinal and time-to-event data are included within the scope of the package. However, rather than being designed for estimating a specific class of models, the package provides a general framework for estimating models with any number of correlated outcomes and/or multiple levels of clustering. With regard to shared parameter joint models

for longitudinal and time-to-event data, both univariate and multivariate models can be estimated. Moreover, several association structures are accommodated through a series of built-in functions that generate current value, current slope, or cumulative effects of the biomarker. Additional association structures can also be accommodated through the ability to specify a user-defined likelihood function. The package is currently only available as a development version. Moreover, post-estimation functionality is currently limited, with predictions from the fitted model not yet available. For these latter reasons, the **megenreg** package is excluded from the comparison table shown in the following section.

5.4.6 Summary of features

Table 9 provides a qualitative summary of the features of each of the aforementioned packages for fitting multivariate joint models.

Table 9. Summary of the features available in each of the multivariate joint modelling packages.

	rstanarm	JMbayes	joineRML	survtd
Estimation algorithm				
MCMC	✓	✓		
MC Expectation Maximization			✓	
MICE, Partial likelihood				✓
Longitudinal submodel(s)				
Distributional families				
Gaussian	✓	✓	✓	✓
Bernoulli	✓	✓		
Poisson	✓	✓		
Negative-binomial	✓			
Gamma	✓			
Inverse-Gaussian	✓			
Clustering beyond patient-level	✓			
Event submodel				
Hazard effect types				
Proportional hazards	✓	✓	✓	✓
Additive hazards				✓
Non-proportional hazards[1]		✓		
Baseline hazard types				
Unspecified			✓	✓
Penalised splines		✓		
B-splines	✓			
Weibull	✓			
Piecewise constant	✓			
Association structures				
Current value[2]	✓	✓	✓	✓
Current slope	✓	✓		
AUC	✓	✓		
Weighted AUC		✓		
Shared random effects	✓	✓		
Lagged associations	✓	✓		
Interaction terms with observed covariate data	✓	✓		
Interactions between longitudinal outcomes	✓			
Post-estimation functions				
Dynamic predictions	✓	✓	✓	

Abbreviations: MCMC: Markov chain Monte-Carlo. MC: Monte-Carlo. MICE: multiple imputation by chained equations. AUC: area under the curve.

1 In other words, time-dependent hazard ratios.

2 For joineRML this is using the approach of Henderson et al. (2000).

Chapter 6: Joint models for multilevel hierarchical data

6.1 Chapter introduction

Clustered data arise in many settings in applied health research. Longitudinal measures are one such example whereby observation times are clustered within the observational units, often individual study participants. Longitudinal data therefore represent a two-level hierarchical structure; observation times are at level 1 of the hierarchy and individuals are at level 2. Moreover, we generally refer to the individual as being the single *clustering factor*.

However, there are also many situations in which there are two or more clustering factors. In oncology, for example, it is possible to track changes in tumour size over time following the initiation of treatment. Since patients can have multiple tumour lesions, the resulting longitudinal data structure may contain repeated measurements taken at observation times (level 1) clustered within lesions (level 2) within patients (level 3). Such data would have a three-level hierarchical structure with two clustering factors, the lesion and the patient. In this example, the patient (or individual) represents the highest level of the hierarchy.

An alternative situation is one in which the additional clustering factor (that is, the clustering factor that is not the individual) occurs higher in the data hierarchy than the individual. One example is longitudinal measurements taken at observation times (level 1) on individuals (level 2) clustered within clinics (level 3). A second example is an individual patient data (IPD) meta-analysis in which there are repeated measurements for each individual, such that observations (level 1) are clustered within individuals (level 2) who are clustered within studies (level 3). These two examples illustrate situations that have a three-level hierarchical structure with two clustering factors, the individual and either the clinic or study. However, in these examples the individual is no longer at the highest level

of the data hierarchy. This feature of the data structure will have implications for the model formulations described in this chapter.

As discussed in Chapter 2, there have been numerous extensions to standard joint modelling approaches. Nonetheless, one feature of all joint modelling approaches proposed in the literature to date has been that they are limited to a two-level hierarchical structure, whereby the individual is the only clustering factor. In this chapter, novel methods are described for the joint modelling of longitudinal and time-to-event data in the presence of more than one clustering factor. The paper in the following section describes the methodology. The main application presented in the paper is related to patients with non-small cell lung cancer, however, several other motivating examples are also discussed. The methods described in the paper have been implemented as part of the **rstanarm** (Brilleman et al., 2018; Stan Development Team, 2017a) R package, which was introduced in Chapter 5. This software option makes these methods readily available to other researchers who wish to apply them in their own studies.

6.2 Manuscript

This section herein contains the following methodological research paper:

> Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J, Dunyak J, Al-Huniti N, Fox R, Hammerbacher J, Wolfe R. Joint longitudinal and time-to-event models for multilevel hierarchical data. *Submitted for publication.*

Joint longitudinal and time-to-event models for multilevel hierarchical data

Samuel L. Brilleman[1,2], Michael J Crowther[3], Margarita Moreno-Betancur[2,4,5], Jacqueline Buros Novik[6], James Dunyak[7], Nidal Al-Huniti[7], Robert Fox[7], Jeff Hammerbacher[6,8], Rory Wolfe[1,2]

Author affiliations: [1]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; [2]Victorian Centre for Biostatistics (ViCBiostat), Melbourne, Australia; [3]Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, UK; [4]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Australia; [5]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia; [6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [7]Quantitative Clinical Pharmacology, IMED Biotech Unit, AstraZeneca, Waltham, MA, USA; [8]Department of Microbiology and Immunology, Medical University of South Carolina, Charleston, SC, USA

***Corresponding author:** sam.brilleman@monash.edu

Abstract

Joint modelling of longitudinal and time-to-event data has received much attention recently. Increasingly, extensions to standard joint modelling approaches are being proposed to handle complex data structures commonly encountered in applied research. In this paper we propose a joint model for hierarchical longitudinal and time-to-event data. Our motivating application explores the association between tumor burden and progression-free survival in non-small cell lung cancer patients. We define tumor burden as a function of the sizes of target lesions clustered within a patient. Since a patient may have more than one lesion, and each lesion is tracked over time, the data have a three-level hierarchical structure: repeated measurements taken at time points (level 1) clustered within lesions (level 2) within patients (level 3). We jointly model the lesion-specific longitudinal trajectories and patient-specific risk of death or disease progression by specifying novel association structures that combine information across lower level clusters (e.g. lesions) into patient-level summaries (e.g. tumor burden). We provide user-friendly software for fitting the model under a Bayesian framework. Lastly, we discuss alternative situations in which additional clustering factor(s) occur at a level higher in the hierarchy than the patient-level, since this has implications for the model formulation.

1. Introduction

In clinical or epidemiological research studies, longitudinal data may be in the form of a clinical biomarker that is repeatedly measured over time on a given patient, whilst time-to-event data may refer to the patient-specific time from a defined origin (e.g. time of diagnosis of a disease) until a clinical event of interest such as death or disease progression. A common motivation for collecting such data is to explore how changes in the biomarker are associated with the occurrence of the event. A rapidly evolving field of statistical methodology, known as "joint modelling", aims to model both the longitudinal and time-to-event data simultaneously providing several potential benefits over more traditional approaches [1–3]. Compared with using the observed biomarker measurements as covariates in a time-to-event model, a joint modelling approach can protect against bias due to missing data or measurement error in estimating the association between the value of the biomarker and the risk of occurrence of the event [1,4]. Moreover, we can explore the associations between more complex aspects of the biomarker trajectory (such as the rate of change) and the occurrence of the event. Lastly, we might wish to use the longitudinal biomarker data in the development of a "dynamic" risk prediction model, and joint

modelling approaches lend themselves to this purpose [5,6].

The so-called "shared parameter" joint modelling approach consists of two regression submodels, one for the longitudinal biomarker measurements (the "longitudinal submodel") and one for the time-to-event outcome (the "event submodel"). Dependence between the two submodels is allowed for by assuming that the model for the time-to-event outcome includes as predictor some functional form of the patient-specific parameters from the longitudinal submodel, commonly referred to as an association structure. In the joint modelling literature to date, primary focus has been on a situation in which there is a single normally-distributed biomarker measured repeatedly over time for each patient and a unique, possibly right-censored, time to a terminating event of interest [4,7]. However, a number of extensions have been proposed for the standard shared parameter joint model, such as competing risks [8], interval censored event times [9], non-normally distributed biomarkers [10], and multiple biomarkers [11].

Nonetheless, a common aspect of the proposed methodology has been that the longitudinal data have a two-level hierarchical structure; longitudinal measurements of the biomarker are observed at time points (level 1 of the hierarchy) which are clustered within patients (level 2 of the hierarchy). The patient is therefore considered to be the only clustering factor. An example of this data structure is shown in Figure 1a. However, there exist many situations in clinical and epidemiological research in which we wish to analyse longitudinal and time-to-event data where the longitudinal data component (and potentially also the time-to-event component) has a hierarchical structure with clustering factors beyond just that of the patient.

In this paper we describe a joint modelling approach that can be applied to longitudinal and time-to-event data with more than one clustering factor. In Section 2 we introduce several motivating examples which describe the types of data structures our joint modelling approach is intended for. In Sections 3 and 4 we describe the formulation and estimation of a joint model that is suitable when an additional clustering factor occurs at a level lower in the hierarchy than the patient-level. In Section 5 we describe an application in which we use this joint model to explore the association between tumor burden and risk of death or disease progression in non-small cell lung cancer (NSCLC) patients undergoing treatment. In Section 6 we describe the formulation of the joint model under alternative scenarios in which the additional clustering factor occurs at a level higher in the hierarchy than the patient-level. In Section 7 we close with a discussion.

2. Motivating examples

2.1 Tumor burden and progression-free survival in non-small cell lung cancer

In our primary motivating example interest lies in exploring the relationship between tumor burden and the risk of death or disease progression in patients with non-small cell lung cancer (NSCLC). After a patient initiates treatment the size of each tumor lesion is measured repeatedly over time in order to assess the effectiveness of treatment and aid clinical decision making. Accordingly, for a given patient, we can define the tumor burden as some patient-level summary of the sizes of their individual tumor lesions. Given that a patient may have more than one lesion, our data consists of a hierarchy in which the longitudinal measurements are observed at time points (level 1) which are clustered within a specific lesion (level 2) for a given patient (level 3), as represented in Figure 1b.

Consideration of the multilevel structure of the data is important for several reasons. Firstly, the underlying growth trajectories may vary across different lesions, even when those lesions are clustered within the same patient. We can allow for between-lesion variation in the growth trajectories through the use of lesion-specific, as well as patient-specific, parameters in the longitudinal submodel. Equivalently, the introduction of lesion-specific parameters in the longitudinal submodel allows us to account for the within-cluster correlation of longitudinal measurements made on the same lesion and therefore appropriately estimate standard errors. Secondly, the hierarchical structure of the data is a key aspect to consider when specifying the form of the association between the longitudinal

and event processes, something we discuss further in Section 3.3.

2.2 Visual field progression in glaucoma

Our second motivating example comes from research on eye disease. In ophthalmology it is of interest to use repeated measurements of eye-specific biomarkers to help predict the occurrence of disease-specific events. For example, in glaucoma research we may be interested in the association between optic nerve head surface depth (ONHSD) and visual field progression. Previous studies [12] have used joint models to explore this association by treating each eye as independent and modelling the association between the eye-specific longitudinal trajectory for ONHSD and the eye-specific event endpoint (visual field progression). However, this approach ignores the dependence between measurements taken on the two eyes clustered within a patient. Arguably, a more appropriate analysis approach would model the correlation between measurements taken on a person's two eyes. Hence, consider a joint modelling approach in which we assume the ONHSD measurements are observed at time points (level 1) which are clustered within a specific eye (level 2) for a given patient (level 3). We could then explore the association between the longitudinal trajectory for ONHSD and a patient-specific endpoint for the time to visual field progression.

2.3 Patients within clinics or the meta-analysis of joint model data

Our final two motivating examples relate to an alternative data structure in which the additional clustering factor occurs at a level which is higher in the hierarchy than the patient. One example is where repeated observation times (level 1) exist for patients (level 2) and those patients are clustered within clinics (level 3). Another example is an individual patient data (IPD) meta-analysis where observation times (level 1) are for patients (level 2) clustered within randomised clinical trials (level 3) [13]. In both of these examples, we wish to include the additional clustering factor (i.e. the clinic or the trial) in our joint modelling approach, so that we appropriately allow for the correlation structure. However, because the additional clustering factor occurs at a level higher than the patient-level, there are different implications for the specification of the joint model association structure compared with our previous motivating examples. For this reason we describe a formulation of the joint model for this type of data structure separately; in Section 6 of the paper.

3. Model formulation

3.1 Longitudinal submodel

Consider the situation in which we have a three-level hierarchical structure for our longitudinal data, where the patient represents the highest level of the hierarchy (in Section 6 we discuss the situation in which the patient does not represent the highest level of the hierarchy). We assume our longitudinal outcome measurements $y_{ijk} = y_{ij}(t_{ijk})$ are obtained at a set of time points $k = 1, \dots, K_{ij}$ which are assumed to be nested within unit j $(j = 1, \dots, J_i)$ of the level 2 clustering factor which in turn is nested within patient i $(i = 1, \dots, N)$, the level 3 clustering factor. We model the longitudinal outcome in continuous time using a generalised linear mixed effects model where we assume $Y_{ij}(t)$ is governed by a distribution in the exponential family with expected value $\mu_{ij}(t) = g^{-1}(\eta_{ij}(t))$ for some known link function $g(.)$. Specific choices of family and link function lead to, for example, linear, logistic or Poisson regression. We specify a three-level hierarchical model for the linear predictor

$$\eta_{ij}(t) = x'_{ij}(t)\beta + z'_{ij}(t)b_i + w'_{ij}(t)u_{ij} \qquad (1)$$

where $x_{ij}(t)$, $z_{ij}(t)$, and $w_{ij}(t)$ are vectors of covariates, possibly time-dependent. The vector β contains fixed-effect parameters, and u_{ij} and b_i are vectors of level 2 (cluster-specific) and level 3 (patient-specific) parameters, each assumed to be normally distributed with mean zero and unstructured variance-covariance matrix, that is $u_{ij} \sim N(0, \Sigma_u)$ and $b_i \sim N(0, \Sigma_b)$. We assume that u_{ij} and b_i are uncorrelated.

3.2 Event submodel

We observe an event time $T_i = min\{T_i^*, C_i\}$, where T_i^* denotes the true event time for patient i and C_i denotes the right-censoring time, and define an indicator of observed event

occurrence $d_i = I(T_i^* \leq C_i)$. We model the hazard of the event using a proportional hazards regression model

$$h_i(t) = h_0(t)exp\,(v_i'(t)\gamma + \sum_{q=1}^{Q} \alpha_q f_q(\Theta_{ij}(t);\, j = 1, \ldots, J_i)) \quad (2)$$

where $h_i(t)$ is the hazard of the event for patient i at time t, $h_0(t)$ is the baseline hazard at time t, $v_i(t)$ is a vector of covariates with an associated vector of fixed-effect parameters γ, and $\sum_{q=1}^{Q} \alpha_q f_q(\Theta_{ij}(t);\, j = 1, \ldots, J_i)$ forms the "association structure" for the joint model which consists of some specified set of functions $f_q(.)$ applied to the full set of (possibly time-varying) parameters from the longitudinal submodel $\Theta_{ij}(t) = \{\beta, b_i, u_{ij}, \mu_{ij}(t), \eta_{ij}(t)\}$ with associated fixed effects α_q $(q = 1, \ldots, Q)$. The functions $f_q(.)$ might each correspond to a functional of the longitudinal submodel parameters for a given patient $i$and cluster j, for example, the expected value or rate of change in the longitudinal biomarker. Alternatively, they might be functions of the longitudinal submodel parameters for a given patient i across all J_i clusters, representing different methods for combining the level 2 clusters into a patient-level summary (as described in the next section). We refer to the fixed effects α_q as "association parameters" since they quantify the magnitude of the association between aspects of the longitudinal process and the event process. In the next section we describe the variety of ways in which the association structure for the joint model can be specified.

3.3 Association structures for patient-level summaries

Given that the event time T_i is measured at the patient-level, the patient represents the level of the hierarchy at which our primary interest lies for understanding the association between the longitudinal and event processes. Accordingly, we wish to formulate a model that captures the association between the longitudinal and event processes at any given time t in a meaningful way at the patient-level. A decision is required about how information from the level 2 clustering factor (that is, the clustering factor between the patient-level and the observation-level) is used in the formulation of the association structure.

Since the number of level 2 units may differ for each patient (i.e. it isn't necessarily the case that $J_i = J_{i'}$ for all $i \neq i'$) we must combine the information in the level 2 units into some patient-level time-specific summary. Obvious choices for a patient-level summary measure are likely to be the summation, average, maximum or minimum taken across the level 2 units within patient i. That is

$$f_q(\Theta_{ij}(t);\, j = 1, \ldots, J_i) = \sum_{j=1}^{J_i} \mu_{ij}(t) \quad (3)$$

$$f_q\big(\Theta_{ij}(t);\, j = 1, \ldots, J_i\big) = J_i^{-1} \sum_{j=1}^{J_i} \mu_{ij}(t) \quad (4)$$

$$f_q\big(\Theta_{ij}(t);\, j = 1, \ldots, J_i\big) = max\,(\mu_{ij}(t);\; j = 1, \ldots, J_i) \quad (5)$$

$$f_q\big(\Theta_{ij}(t);\, j = 1, \ldots, J_i\big) = min\,(\mu_{ij}(t);\; j = 1, \ldots, J_i) \quad (6)$$

The association structure resulting from equation (3) assumes that the hazard of the event for patient i at time t is associated with the sum of the expected values (at time t) for each of the level 2 units clustered within that patient. In contrast, the J_i^{-1}term in equation (4) provides us with the average of the level 2 cluster-specific expected values within patient i rather than their summation alone. Lastly, equations (5) and (6), respectively, assume that the hazard of the event for patient i at time t is associated with the level 2 cluster (within patient i) that has the largest or smallest expected value at time t. It is possible for more than one of these summary functions to be included in a single model (i.e. $Q > 1$). Moreover, other summary functions are possible but are not described here.

The most appropriate summary function(s) may be determined based on clinical context, or by choosing the association structure that provides the best model performance based on some criterion. For instance, returning to the first motivating example introduced in Section 2, we may believe that risk of death or disease progression for a patient with NSCLC is driven by treatment-failure occurring at a single lesion. We may therefore assume the hazard of the event for patient i at time t is associated with

the maximum (i.e. largest) of the lesion-specific expected values, since this would represent the lesion with the most advanced disease, for example due to it having the worst treatment response.

Moreover, we could easily replace the $\mu_{ij}(t)$ in equations (3) through (6) with some other function of the longitudinal submodel parameters, such as the level 2 cluster-specific rate of change in the marker at time t (i.e. $\frac{d\mu_{ij}(t)}{dt}$) or the area under the level 2 cluster-specific marker trajectory up to time t (i.e. $\int_0^t \mu_{ij}(u)\,du$). For instance, we may assume that the lesion with largest growth rate may be most informative of treatment failure. Such extensions follow naturally from association structures that have been proposed elsewhere for shared parameter joint models [11,14].

The specifications in equations (3) and (4) both assume a constant magnitude of association between the expected value of each level 2 unit and the hazard of the event; that is, there is an implicit assumption that the level 2 units within a patient are exchangeable since their expectations are each multiplied by the same fixed effect association parameter α_q. It is worth noting that this is in contrast to the situation in which we have multiple longitudinal biomarkers (for example lesion size and circulating DNA) each measured repeatedly over time. In this situation the multiple biomarkers within a patient are *not* exchangeable and therefore each biomarker would have a different coefficient quantifying its association with the hazard of the event. Methods for the joint modelling of multiple longitudinal biomarkers and time-to-event data, where the multiple biomarkers are not exchangeable, have been described elsewhere [11,15,16]. Although it is outside the scope of this paper, the methodology described here could be extended to a situation in which we have multiple longitudinal biomarkers, some of which may or may not have additional levels of clustering. This type of data structure is therefore represented in Figure 1c.

4. Model estimation

The joint model proposed in Section 3 can be estimated using the 'stan_jm' modelling function within the rstanarm R package [17]. The association structure can be based on the expected value and/or slope of the longitudinal biomarker, however, currently only a single patient-level summary function (summation, average, maximum, or minimum) can be chosen by the user (this restriction may be relaxed in a future release). Estimation of the model requires a full Bayesian specification with prior distributions on all unknown parameters. We provide further details of the estimation (for example prior distributions and computation) in the Supplementary Materials.

5. Application

We demonstrate the use of our modelling approach by exploring the association between tumor burden and the hazard of death or disease progression amongst NSCLC patients undergoing treatment.

5.1 Data

The Iressa Pan-Asia Study (IPASS) was an open label, phase 3 trial of 1,217 untreated NSCLC patients in East Asia randomized to: (i) gefitinib (250mg per day), or (ii) carboplatin (dose calculated to provide 5-6 mg per milliliter per minute) plus paclitaxel (200 mg per square meter of body-surface area) [18]. The primary endpoint was progression-free survival, however, the study was extended to track overall survival in the longer term. We restricted our analyses to the 430 (35%) patients with an available test result for epidermal growth factor receptor (EGFR) mutation since this has been shown to be associated with both tumor dynamics and treatment response [19]. We thereby defined a group covariate corresponding to either: (i) EGFR+, (ii) EGFR- and receiving gefitinib; or (iii) EGFR- and receiving carboplatin plus paclitaxel.

5.2 Model specification

5.2.1 Longitudinal submodel

We modelled repeated measurements of the longest diameter (in millimetres) of each lesion using a linear mixed effects model (identity link, normal distribution) with a linear predictor as in equation (1) where: the observation-level covariates with fixed (population-average) effects, $x_{ij}(t)$, included an intercept, the 3-category EGFR-group covariate, linear and quadratic terms for time, and an interaction

between group and each of the linear and quadratic terms for time; the patient-level vector $z_{ij}(t)$ included an intercept only; and the lesion-level vector $w_{ij}(t)$included an intercept, and linear and quadratic terms for time. This specification allowed for lesion-specific nonlinear (quadratic) evolutions of the longitudinal trajectory, while also allowing the average (i.e. population-level) estimate of the nonlinear longitudinal trajectory to differ between the three groups (through the group by time interaction terms).

5.2.2 Event submodel

We modelled the hazard of death or disease progression using the proportional hazards model in equation (2). We approximated the log baseline hazard using B-splines with 6 degrees of freedom and included a 3 category physical functioning measure (normal activity; restricted activity; in bed >50% of the time) [20] as a baseline covariate in $v_i(t)$. We considered several models which each differed in terms of their association structure. Specifically we considered the following: (i) no association structure (i.e. no biomarker information in the event submodel), (ii) association structures based on the sum, average, maximum or minimum of the lesion-specific expected values (i.e. the association structures defined in equations (3) through (6)), and (iii) association structures based on both the lesion-specific expected value *and* slope, that is an association structure of the form

$$\sum_{q=1}^{Q} \alpha_q f_q\left(\Theta_{ij}(t)\right) = \alpha_1 f\left(\mu_{ij}(t);\, j = 1, \ldots, J_i\right) + \alpha_2 f\left(\frac{d\mu_{ij}(t)}{dt};\, j = 1, \ldots, J_i\right) \quad (7)$$

where the function $f(.)$ was taken to be either the sum, average, maximum or minimum, $\mu_{ij}(t)$ is the size and $\frac{d\mu_{ij}(t)}{dt}$ is the rate of change in the size of lesion j in patient i at time t, J_i is the total number of target lesions identified for patient i at baseline, and α_1 and α_2 are association parameters.

5.3 Model comparison

An ideal feature of our model would be that it is able to inform clinical decision making by accurately predicting a patient's future risk of death or disease progression in the clinical setting. We therefore compared different possible association structures for our proposed joint model using a measure of predictive accuracy for the event outcome. Specifically, we used the estimated area under the (time-dependent) receiver operating characteristic curve (AUC) to assess how well each of the models discriminated between those patients who did and did not have the event [3].

To do this we first used the fitted joint model to generate conditional survival probabilities for each patient at some time horizon $t_L + \Delta t$, conditional on: (i) their still being at risk at some landmark time t_L, and (ii) their longitudinal biomarker data up to the landmark time t_L (following the methods described in [3]). These survival probabilities were then used in combination with the observed event times and censoring indicators for each patient, taken over the interval $(t_L,\ t_L + \Delta t)$, to calculate the time-dependent AUC measure.

5.4 Results

In our analysis, 360 (84%) of the 430 patients progressed or died prior to censoring. The overall Kaplan-Meier curve is shown in Figure 2. There were 1209 lesions across the 430 patients, and 138 (32%), 101 (23%), 71 (17%) and 120 (28%) patients with 1, 2, 3 or 4+ lesions, respectively. A total of 6132 size measurements were observed, corresponding to a median number of 5 (IQR: 3 to 7; range: 1 to 17) measurements per lesion.

Table 1 shows the estimated AUC values for the fitted models. The results are shown for a landmark time of t = 5 months and a horizon time of $t + \Delta t$ = 10 months. For an association structure based on the expected value (i.e. diameter of the lesion) only, a summary function based on the sum or maximum of the lesions showed better discriminatory performance compared with using the mean or minimum. We found that also including the slope (i.e. rate of change in the diameter of the lesion) in the association structure improved the predictive performance. When both the expected value and slope were used in the association structure then summaries based on the sum, mean or maximum of the lesions all performed similarly. The summary based on the minimum (i.e. size of the smallest lesion, and rate of change in the slowest growing lesion, at

time t) was the worst in terms of predicting the risk of death or disease progression. These results are in line with what we would expect from a clinical perspective, that is, those summaries that incorporate information on the largest and/or fastest growing lesion at time t are likely to provide better predictive performance. This is because they capture information about the most aggressive tumor, which may have escaped treatment and is therefore likely to impact most severely on the risk of death disease progression and death.

Table 2 shows the parameter estimates from the model with the best performance based on the AUC measure (with an association structure based on the maximum of the lesion-specific expected values and slopes). The estimated hazard ratio corresponding to the first association parameter (i.e. $exp(\alpha_1)$) was 1.011 (95% credible interval (CrI): 1.004 to 1.017), suggesting that a one millimetre increase in the diameter of a patient's largest lesion was associated with a 1.1% (95% CrI: 0.4 to 1.7%) increase in their hazard of death or disease progression (conditional on the other covariates in the model). Similarly, a one millimetre per month increase in the rate of change of their fastest growing lesion was associated with a 56% (95% CrI: 42 to 75%) increase in their hazard. Figure 3 shows the fitted lesion-specific longitudinal trajectories and observed measurements for a selection of patients under the fitted model.

6. Alternative data structure: clustering *above* the patient-level

In our analysis of the IPASS data, patient represented the top level of the data hierarchy and the additional clustering factor – "lesion" – occurred at a level which was *lower* in the hierarchy than patient; that is, lesions were clustered within patients rather than patients being clustered within lesions. An alternative situation is that in which the additional clustering factor(s) occur at a level which is *higher* in the hierarchy than the patient-level. An example is where repeated observation times (level 1) exist for patients (level 2) and the patients are clustered within clinics (level 3). Another example is an individual patient data (IPD) meta-analysis with repeated observation times (level 1) for patients (level 2) clustered within randomised clinical trials (level 3) [13].

Recall however that the event time T_i is measured at the patient-level and, therefore, the patient represents the level of the hierarchy at which our primary interest lies for understanding the association between the longitudinal and event processes. For this reason, the relative locations within the hierarchy of the patient and the additional clustering factor have implications for specifying the association structure of the joint model.

6.1 Model formulations based on a *patient*-level association structure

In Section 3.3 we proposed association structures based on a patient-level time-specific summary of the J_i level 2 units clustered within patient i. However, with clustering above the patient-level, there is no need to construct such a patient-level summary.

Suppose that the longitudinal outcome $y_{lik} = y_{li}(t_{lik})$ is measured at time point k $(k = 1, \dots, K_{li})$ which is nested within unit i $(i = 1, \dots, N_l)$ of the level 2 clustering factor (the patient) which in turn is nested within unit l $(l = 1, \dots, L)$ of a level 3 clustering factor (clinic, say, for example). If we again model the longitudinal outcome in continuous time using a generalised linear mixed effects model where $Y_{li}(t)$ is governed by a distribution in the exponential family with expected value $\mu_{li}(t) = g^{-1}(\eta_{li}(t))$ we might, for example, consider a specification for the longitudinal submodel of the form

$$\eta_{li}(t) = x'_{li}(t)\beta + z'_{li}(t)b_{li} + q'_{li}(t)c_l \quad (8)$$

where $x_{li}(t)$, $z_{li}(t)$ and $q_{li}(t)$ are vectors of covariates, possibly time-dependent, b_{li} still represents the vector of patient-specific parameters (but now patient i is nested within the level 3 cluster l), and c_l represents the vector of level 3 parameters such that $c_l \sim N(0, \Sigma_c)$. The corresponding specification of the event submodel may take the form

$$h_{li}(t) = h_0(t) exp\left(v'_{li}(t)\gamma + \alpha\, \mu_{li}(t)\right) \quad (9)$$

Because the additional clustering occurs at a level in the hierarchy that is higher than the patient we can simply use an association structure based on the patient-level expected value of the longitudinal outcome, without any

need to derive a summary quantity based on lower-level units. The specification in (9) would assume that the hazard of the event for patient i at time t is associated with the patient-specific expected value of the longitudinal marker at time t, incorporating the effects of any higher level clustering. Note that the specification in (9) could be easily extended to any other patient-level function of the longitudinal submodel parameters, such as the patient-specific rate of change in the marker (i.e. slope) at time t or the area under the patient-specific marker trajectory (i.e. integral) up to time t.

A possible extension would be to include a shared frailty term in the event submodel

$$h_{li}(t) = h_0(t)exp\left(v'_{li}(t)\gamma + \alpha\,\mu_{li}(t) + \delta_l\right) \quad (10)$$

where δ_l is, for example, assumed to follow a normal or log-Gamma distribution. The inclusion of the shared frailty term does not induce an association with the longitudinal submodel, but it does allow for correlation in the event times of patients within a level 3 cluster. Note that if the variance of the δ_l parameters is close to zero, then this would suggest there is little within-cluster correlation in the event times and the shared frailty term could be dropped from the model. Moreover, if the number of level 3 groups was small then another alternative would be to include the level 3 group as a fixed effect covariate in the event submodel or as a stratification factor for the baseline hazard. The benefit of these latter models is that they may be computationally simpler than specifying a shared frailty term as a random effect.

6.2 Model formulation based on a *higher*-level association structure

An alternative possibility is that the hazard of the event for patient i need only be related to the higher-level cluster's deviation from the average. That is, we can consider a shared random effects joint model of the form

$$h_{li}(t) = h_0(t)exp\left(v'_{li}(t)\gamma + \alpha\, c_l\right) \quad (11)$$

where c_l might, for example, represent the clinic-level random intercept. In this case, we would have a model in which we assume that the hazard of the event for patient i is associated with the way in which their clinic's biomarker measurements deviate from the average clinic, but not with any time-varying characteristics of the patient themselves. Here, the random effect c_l serves two purposes in the event submodel. First, it allows for within-cluster correlation in the event times (as previously described for the shared frailty). Second, it allows for dependence between the event and longitudinal processes through a shared parameter at the level of clustering factor l.

7. Discussion

Increasingly complex data structures are being accommodated under a joint longitudinal and time-to-event modelling framework. In this paper we have described a new joint modelling approach that allows for multilevel hierarchical data, where the data structure includes clustering factors beyond that of the individual. Such data structures commonly appear in clinical and epidemiological research, however, they have not previously been incorporated into a joint modelling framework. Standard joint modelling approaches aim to model patient-level measurements of a clinical biomarker, however, greater flexibility can be achieved by incorporating both patient-specific and cluster-specific effects in the longitudinal submodel when those levels of clustering are present in the underlying data structure. Moreover, it allows an additional set of association structures to be used for modelling the association between the longitudinal biomarker and the patient-level risk of the event. We proposed a set of possible association structures that could be used in most settings, however, the most appropriate choice of association structure is likely to depend on the primary research question and data structure that is relevant to the application at hand. By incorporating the multilevel structure into our joint modelling approach, we are able to formulate a model that answers the research question appropriately. For instance, in our application, patient-level summaries of the lesion-specific trajectories are likely to be meaningful in a way that quantities obtained by ignoring the lesion level would not be.

A potential limitation of the modelling in our application is that the observed event times were subject to interval censoring. This interval censoring is evident from the "steps" that can be seen in the Kaplan-Meier curve in Figure 2. This is due to clinicians in the IPASS trial declaring

disease progression at the scheduled clinic visits. In our application we ignored this interval censoring and so an avenue for future work will be to accommodate this interval censoring within our proposed joint modelling framework. Moreover, in future work, we would like to separately assess the competing event outcomes of death and disease progression by considering cause-specific competing risks event submodels. In this way, we will be able to separate out the cause-specific associations between tumor burden and each of the competing events.

A significant strength of this paper is that our proposed model, described in Section 3, has been implemented as part of the rstanarm R package. A benefit of having implemented this model as part of that package is that researchers can easily fit the model to their data, via a user-friendly interface with customary R formula syntax and data frames. The back-end estimation of the model is carried out under a full Bayesian specification with priors on all unknown parameters. A variety of prior distributions are available to the user, as well as a variety of exponential family and link function options for the longitudinal outcome, thereby providing significant flexibility. In addition, the package allows users to estimate a joint model with multiple longitudinal outcomes (i.e. a multivariate joint model) of which one or more can have the multilevel structure described in Section 3. We hope that by providing user-friendly software and example code (Supplementary Materials) for fitting the proposed model, we will help to facilitate its use in a wide variety of applications.

Conflicts of interest

None.

Authors' contributions

Research idea and study design: all authors; data acquisition: SLB, KRP, SPM, MMB, RW; statistical analysis: SLB, JT; statistical supervision: RW, MMB, KRP, MJC; first written draft of manuscript: SLB; comments and input to final manuscript: all authors. Each author contributed important intellectual content during manuscript drafting or revision and accepts accountability for the overall work by ensuring that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved.

Acknowledgements

SLB is funded by an Australian National Health and Medical Research Council (NHMRC) Postgraduate Scholarship (ref: APP1093145), with additional support from an NHMRC Centre of Research Excellence grant (ref: 1035261) awarded to the Victorian Centre for Biostatistics (ViCBiostat). MJC is partly funded by a UK Medical Research Council (MRC) New Investigator Research Grant (ref: MR/P015433/1). We thank Eric Novik at Generable (http://www.generable.com/) for his support and management of the collaboration between SLB, JBN and the team at AstraZeneca (NAH, RH, JD).

Supplementary materials

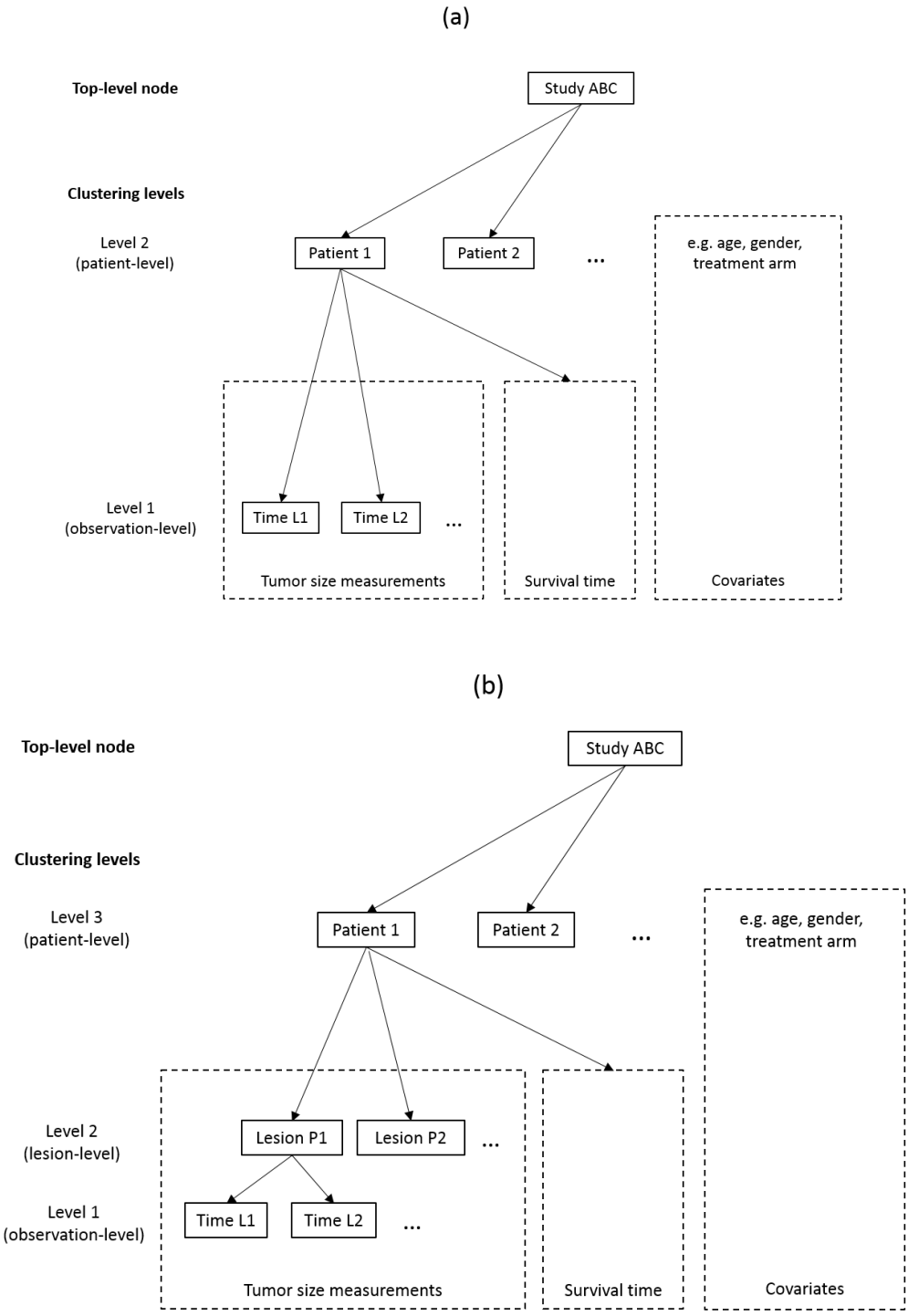
We have developed user-friendly software for fitting the model described in the paper. The software is publically available as part of the rstanarm package, downloadable from the Comprehensive R Archive Network (https://cran.r-project.org/). The supplementary materials include an example of the code required to fit the model and additional details about the model estimation. However, the Iressa Pan-Asia Study (IPASS) dataset used in our application is not publicly available.

References

1. Tsiatis AA, Davidian M. JOINT MODELING OF LONGITUDINAL AND TIME-TO-EVENT DATA: AN OVERVIEW. Stat Sin. Institute of Statistical Science, Academia Sinica; 2004;14: 809–834.
2. Lawrence Gould A, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat Med. 2015;34: 2181–2195.
3. Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. CRC Press; 2012.
4. Wulfsohn MS, Tsiatis AA. A joint model

for survival and longitudinal data measured with error. Biometrics. 1997;53: 330–339.
5. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics. 2011;67: 819–829.
6. Taylor JMG, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, et al. Real-time individual predictions of prostate cancer recurrence using joint models. Biometrics. 2013;69: 206–213.
7. Henderson R. Joint modelling of longitudinal measurements and event time data. Biostatistics. 2000;1: 465–480.
8. Williamson PR, Kolamunnage-Dona R, Philipson P, Marson AG. Joint modelling of longitudinal and competing risks data. Stat Med. 2008;27: 6426–6438.
9. Gueorguieva R, Rosenheck R, Lin H. Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. J R Stat Soc Ser A Stat Soc. 2011;175: 417–433.
10. Brilleman SL, Crowther MJ, May MT, Gompels M, Abrams KR. Joint longitudinal hurdle and time-to-event models: an application related to viral load and duration of the first treatment regimen in patients with HIV initiating therapy. Stat Med. 2016;35: 3583–3594.
11. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Stat Med. 2011;30: 1366–1380.
12. Wu Z, Lin C, Crowther M, Mak H, Yu M, Leung CK-S. Impact of Rates of Change of Lamina Cribrosa and Optic Nerve Head Surface Depths on Visual Field Progression in Glaucoma. Invest Ophthalmol Vis Sci. 2017;58: 1825–1833.
13. Sudell M, Kolamunnage-Dona R, Tudur-Smith C. Joint models for longitudinal and time-to-event data: a review of reporting quality with a view to meta-analysis. BMC Med Res Methodol. 2016;16: 168.
14. Andrinopoulou E-R, Rizopoulos D. Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. Stat Med. 2016;35: 4813–4823.
15. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. BMC Med Res Methodol. 2016;16: 117.
16. Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). Biostatistics. 2017 (Epub ahead of print) doi:10.1093/biostatistics/kxx046
17. Stan Development Team. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.15.4. http://mc-stan.org/. 2016;
18. Mok TS, Wu Y-L, Thongprasert S, Yang C-H, Chu D-T, Saijo N, et al. Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma. N Engl J Med. 2009;361: 947–957.
19. Fukuoka M, Wu Y-L, Thongprasert S, Sunpaweravong P, Leong S-S, Sriuranpong V, et al. Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non-small-cell lung cancer in Asia (IPASS). J Clin Oncol. 2011;29: 2866–2874.
20. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol. 1982;5: 649–656.

Figure 1. Example of the hierarchical structure of joint model data under three possible scenarios: (a) one longitudinal biomarker (tumor size) where the patient is the only clustering factor; (b) one longitudinal biomarker (tumor size) where there are two clustering factors (lesions clustered within patients); (c) two longitudinal biomarkers (tumor size and circulating DNA), one of which has one clustering factor (the patient), and one of which has two clustering factors (lesions clustered within patients).



(c)

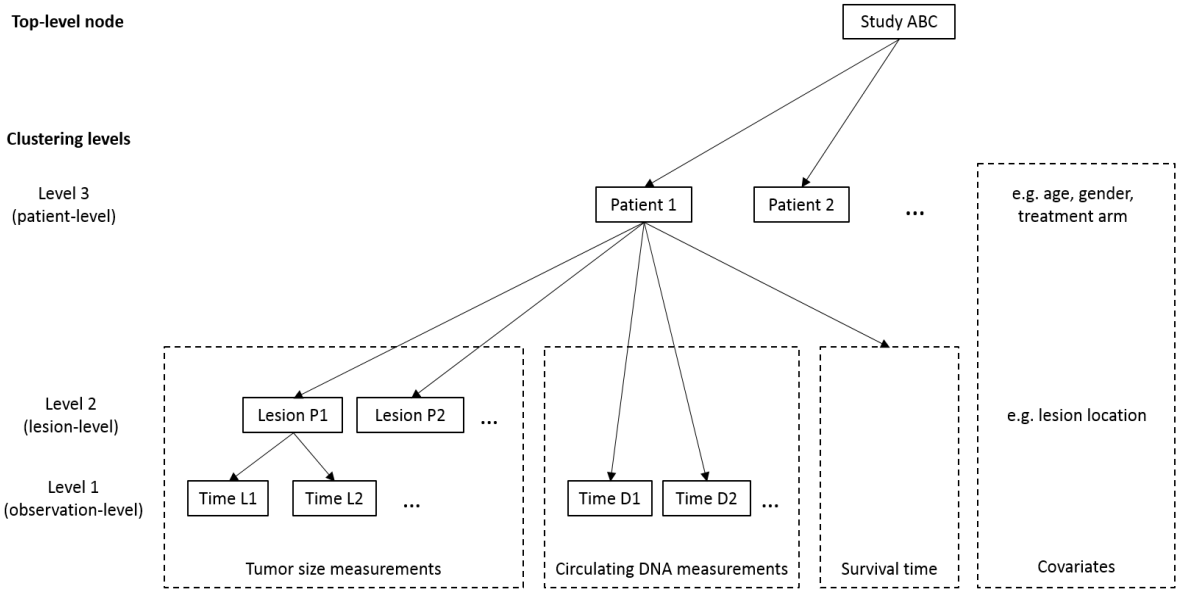

Top-level node
Study ABC
Clustering levels
Level 3
(patient-level)
Patient 1
Patient 2
...
e.g. age, gender,
treatment arm
Level 2
(lesion-level)
Lesion P1
Lesion P2
...
e.g. lesion location
Level 1
(observation-level)
Time L1
Time L2
...
Time D1
Time D2
...
Tumor size measurements
Circulating DNA measurements
Survival time
Covariates

Figure 2. Overall Kaplan-Meier curve for progression-free survival. The values provided at the top of the plot are the numbers of patients still at risk for the event.

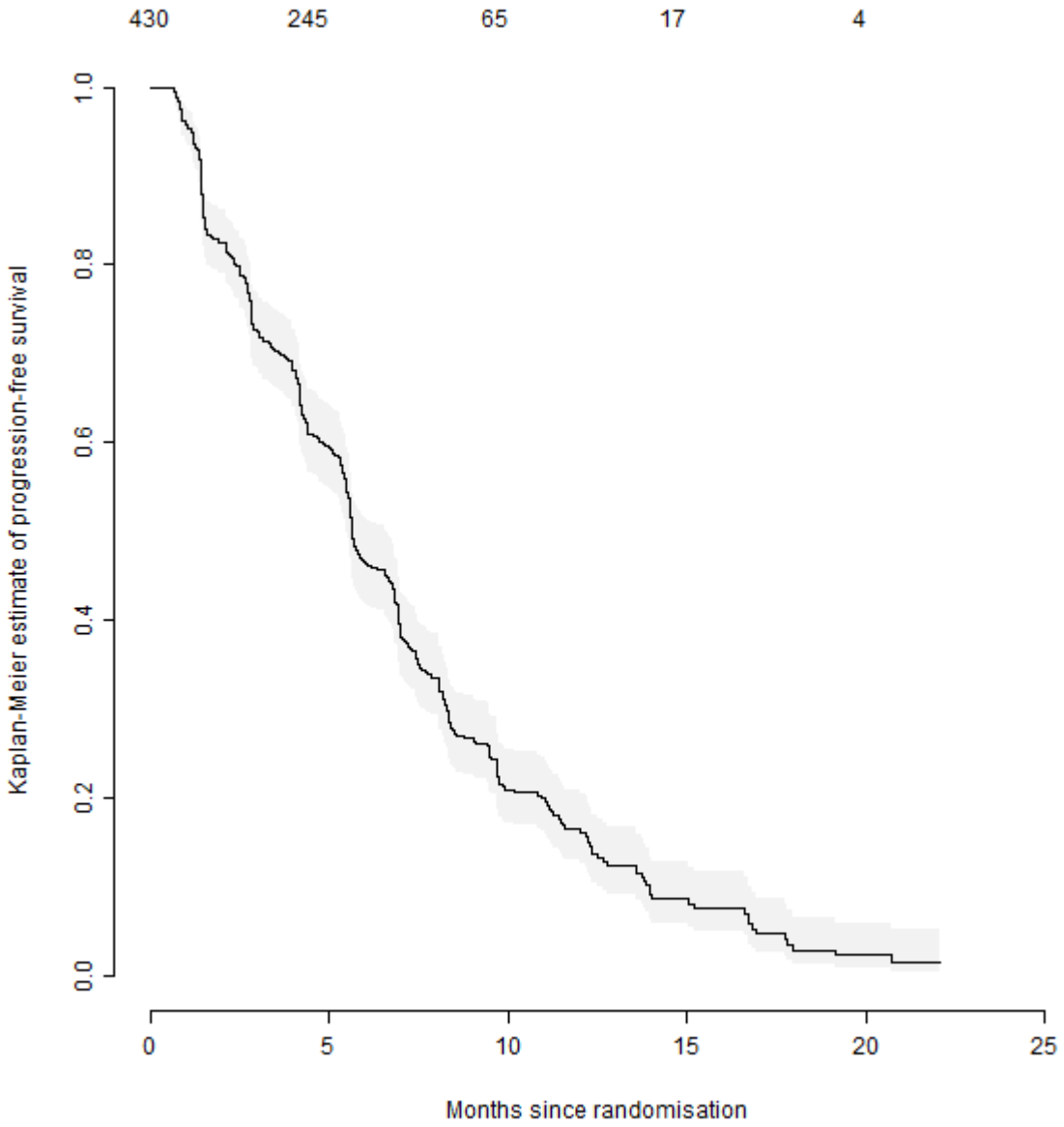


Figure 3. Observed longitudinal biomarker measurements (longest diameter of the lesion) and the fitted lesion-specific longitudinal trajectories (with 95% prediction intervals) under the joint model, for a selection of patients. Each panel of the figure shows a different patient, with some patients having multiple lesions. The dashed vertical line shows each patient's event or censoring time.

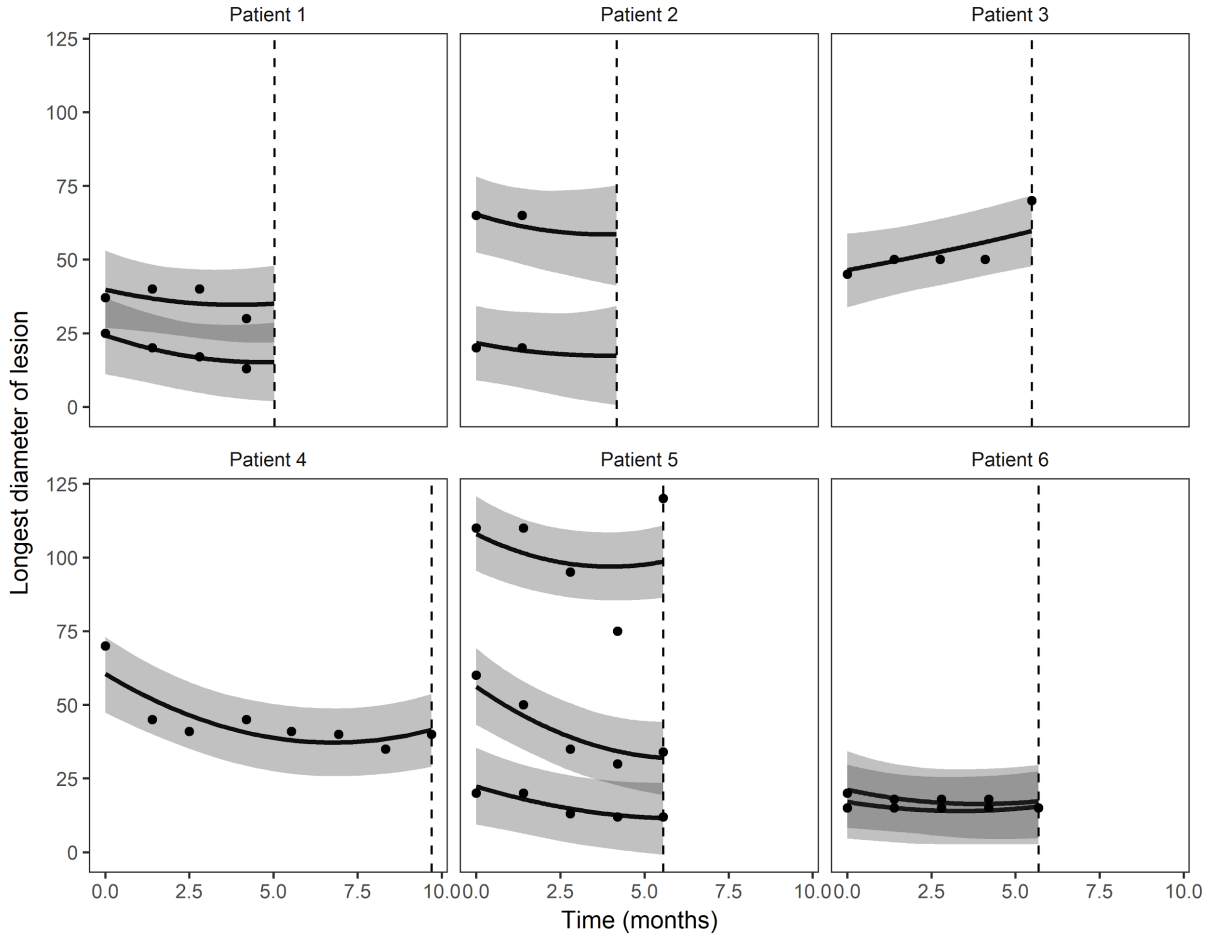


Table 1. Estimated time-dependent AUC for the proposed joint model using various association structures. The AUC is calculated using a landmark time of $t = 5$ months and horizon time of $t = 10$ months.

Association structure	Time-dependent AUC
No biomarker data (i.e. no association structure)	0.50
Lesion-specific value	
Sum	0.62
Average	0.56
Maximum	0.61
Minimum	0.55
Lesion-specific value & slope	
Sum	0.65
Average	0.64
Maximum	0.66
Minimum	0.59

Abbreviations. AUC: area under the (receiver operating characteristic) curve.

Table 2. Fixed effect parameter estimates (posterior means and 95% credible interval limits) from the joint model. The estimates for the event submodel are hazard ratios and the coefficients for the B-splines baseline hazard have been omitted.

Parameter	Estimate	Lower	Upper
Longitudinal submodel			
Intercept	23.0	21.3	24.7
Group (ref: EGFR+)			
EGFR-, carboplatin plus paclitaxel	4.0	0.8	7.1
EGFR-, gefitinib	16.9	13.2	20.4
Time effects			
Linear term (orthogonalised)	-0.1	-73.3	76.7
Quadratic term (orthogonalised)	450.3	391.6	512.5
Group * Linear interaction			
EGFR-, carboplatin plus paclitaxel * Linear	315.2	195.1	438.4
EGFR-, gefitinib * Linear	390.0	127.5	660.4
Group * Quadratic interaction			
EGFR-, carboplatin plus paclitaxel * Quadratic	23.7	-74.3	123.4
EGFR-, gefitinib * Quadratic	-524.8	-697.0	-351.1
Event submodel			
Physical functioning (ref: in bed >50% of the time)			
Normal activity	0.6	0.4	1.0
Restricted activity	0.6	0.4	1.0
Association parameters (exponentiated)			
Value (diameter of largest lesion at time t)	1.011	1.004	1.017
Slope (rate of change in fastest growing lesion at time t)	1.56	1.42	1.75

Abbreviations. ref: reference category; EGFR: epidermal growth factor receptor (mutation status).

Chapter 7: Discussion

7.1 Summary of the findings and contributions

The research presented in this thesis has covered three key areas related to the joint modelling of longitudinal and time-to-event data. First, the application of the methods in order to help answer research questions of importance to health. Second, the development of a joint modelling framework that accommodates multilevel hierarchical data structures encountered in a number of clinical and epidemiological research studies. Third, the development of software for estimating both univariate and multivariate joint models under a Bayesian approach.

Chapter 3 and 4 of the thesis each described an applied research project. The first project, in Chapter 3, involved modelling longitudinal changes in disability and the risk of death in a cohort of older Americans. Specifically, the analysis was based on a shared parameter joint modelling approach with a current value association structure. The primary research question in the study was to examine the effect that community-level disaster exposure may have on an individual's disability trajectory or their risk of death. The joint modelling approach allowed this research question to be answered, whilst simultaneously accounting for: (i) missingness in the longitudinal disability information that is attributable to informative dropouts due to death; (ii) adjustment for the effect of the underlying (error-free) disability measure on the risk of death; and (iii) the possible joint dependence of both disability and risk of death on the primary exposure variable in the study (i.e. disaster exposure). From the findings of the study there was no evidence to suggest that community-level disaster exposure affected an individual's longitudinal disability trajectory or risk of death. However, it remains unclear whether this was attributable to a relatively weak measure of disaster exposure, the broad definition of a "disaster" event, or the true absence of an underlying association between disaster exposure and disability or death.

The second applied project, in Chapter 4, involved patients with end-stage kidney disease (ESKD) who were undergoing haemodialysis. The project investigated the associations between longitudinal changes in BMI and the competing risks of either transplant or death without transplant in those patients. The study was based on a hypothesis that haemodialysis patients could be meaningfully separated into a set of underlying ("latent") groups that are related to longitudinal changes in BMI. That is, the underlying latent groups are distinguishable with regard to their longitudinal BMI trajectories. Moreover, it was hypothesised that these differences in the underlying longitudinal BMI trajectories would be associated with differences in the rates of either: receiving a kidney transplant, or dying without having received a transplant. A latent class joint modelling approach was well suited to this project, since it helped to identify latent classes that were based on differences in both the marginal longitudinal BMI trajectory and the cause-specific hazard functions for death and transplant. The findings from the study suggested that there were five latent classes, and that the longitudinal BMI trajectories were associated with differences in the rates of death or transplant. This adds insight to the so-called "obesity paradox" in dialysis, since previous studies had identified that rates of death are associated with cross-sectional measures of BMI, but few studies had considered the importance of longitudinal changes in BMI. The study findings suggested that, although BMI at the initiation of haemodialysis is important, it is *changes* in BMI over the course of haemodialysis that are more strongly associated with differences in the rates of death.

In the latter parts of this thesis, namely Chapters 5 and 6, the focus shifted from applied work using joint modelling approaches to methodological and software developments. Chapter 5 focussed on describing the development of software for joint modelling under a Bayesian framework. This was implemented as an R package, with the back-end estimation of the joint models achieved using the Bayesian software Stan. Initially, the motivations for developing the software were that, in 2016 at the time of commencing the work, several joint model formulations were not widely accommodated by existing software. In particular, software for estimating joint models with multiple longitudinal outcomes was not readily available. However, given that joint modelling is a fertile area of research, it is unsurprising that during 2017 there were several developments in this area, as discussed in Chapter 5. Nonetheless, the software package described in Chapter 5 makes several important and novel contributions. It can be used to estimate joint models with multiple longitudinal outcomes, multilevel clustering structures, non-normally distributed

longitudinal outcomes (e.g. binary or counts), a range of association structures, interaction effects in the association structure, and more. Several of these features are not readily available in most other packages. In addition, estimating the model under a Bayesian framework can provide other advantages. For instance, users of the package have a choice of prior distributions available for each of the parameters. This includes, for example, shrinkage priors (Piironen and Vehtari, 2017). They also have the opportunity to incorporate informative or semi-informative prior information into the model. Moreover, a Bayesian approach also allows one to explore the full posterior distribution of each parameter, to make direct probability statements with regard to the parameters, and to easily calculate uncertainty bounds on complex functions of the parameters such as dynamic predictions of future event risk.

The importance of having freely available, general purpose, user-friendly software is becoming increasingly evident in applied biostatistics. Moreover, one needs to be able to communicate the use of that software in order to facilitate its uptake in applied work. For example, the Journal of Statistical Software (JSS) provides one of the few high quality peer-reviewed outlets for publishing papers describing software implementations of statistical methodology. Between 2003 and 2016, the JSS grew from being ranked amongst the lowest quartile of journals in probability and statistics, to being the fifth highest ranked journal in probability and statistics (SCImago, 2007). Similarly, the high-profile International Journal of Epidemiology (IJE) recently established a "Software Application Profile (SAP)" section to communicate practical, freely available, software applications to epidemiological researchers (Forgetta and Richards, 2016). Moreover, one can observe the growth in popularity of vignettes for R packages, which serve as an ideal alternative that circumvents many of the long delays associated with the peer-review and publishing process of traditional journals (Björk and Solomon, 2013; Wickham, 2015). However, the downside for vignettes is that a lack of peer-review may lead to compromised quality of the final output, in some cases. Nonetheless, these examples, to some degree, show the increasing value and importance that is being placed on translating methodological developments in statistics into usable, general-purpose, software implementations.

In Chapter 6, a methodological framework was developed for the joint modelling of longitudinal and time-to-event data, when the underlying data structure contains additional clustering factors beyond just that of the individual. The primary motivating application was related to non-small cell lung cancer. However, the methodological framework can be

applied in a variety of other settings; for instance, ophthalmology and meta-analysis as have been discussed in the chapter. Importantly, the methods were also implemented in the user-friendly software described in Chapter 5.

7.2 Strengths, limitations, and avenues for future work

7.2.1 Definitions of disaster exposure

The applied study in Chapter 3 is the first study to consider a comprehensive range of disaster events (for example, floods, fires, earthquakes, tornados, etc.) and their impact on disability and death, rather than focussing on a single disaster event (or type of disaster event). Although the definition of a disaster was based on the US Federal Emergency Management Agency's (FEMA) definition of a major declaration, it is possible that this definition was too broad for our epidemiological-related purposes. Therefore, with regard to the applied project in Chapter 3, future work should involve identifying opportunities for constructing an improved measure of disaster exposure. An improved measure of disaster exposure, whether defined at the community-level or individual-level, could then be incorporated into the shared parameter joint modelling framework discussed in Chapter 3.

7.2.2 Dialysis modalities

The applied study in Chapter 4 is one of very few studies that have considered longitudinal changes in BMI during haemodialysis and their association with rates of death and transplant. The study used a large, registry-based, cohort of all patients undergoing chronic haemodialysis treatment in Australia and New Zealand over a ten-year period. The findings of the study can therefore be considered generalizable to countries with similar clinical practices and haemodialysis patient populations as Australia. However, the effects of haemodialysis on weight gain or loss are potentially quite different to those of peritoneal dialysis. During peritoneal dialysis, a patient's blood is purified internally (i.e. inside the body) through the injection of a fluid that often contains glucose. Therefore, calories in the glucose can potentially lead to weight gain in the patient. Conversely, under haemodialysis, an individual's blood is purified externally (i.e. outside of the body) and therefore no calories are absorbed by the patient during the dialysis process. These differences in the clinical procedures could lead to significant differences in the underlying BMI trajectories for patients undergoing treatment and, potentially, be associated with differences in the

underlying rates of transplant or death. Therefore, the main avenue identified for future work in this area is to perform a similar study in a peritoneal dialysis patient population.

7.2.3 Limitations of registry-based data

A major limitation of the applied study in Chapter 4 was the infrequent measurement schedule for BMI. This likely limited the ability of the analysis to capture shorter-term fluctuations in weight, potentially leading to less diversity across the BMI trajectories for the different latent classes. The study was based on data from a nationwide registry, for which BMI measurements are only recorded approximately once per year for each patient. Therefore, although more frequent measurements of BMI would have been desirable for this study, such data would likely need to come from a new, or additional, data source.

7.2.4 Uninformative visiting process

As discussed in Section 2.3.5, a common assumption in joint modelling is that the timing of the measurements of the longitudinal biomarker are uninformative. This assumption is likely to be reasonable in the three main applications discussed in this thesis. In the first project, in Chapter 3, the longitudinal outcome was a disability score. The disability score measurements were derived from data obtained as part of a nationally representative longitudinal cohort study for which the interview schedule was likely to be pre-specified, or at least unrelated to the participants' responses. In the second project, in Chapter 4, the longitudinal biomarker measurements were of BMI. These measurements were obtained from a national registry that collected each patient's clinical data as part of an annual survey. The timing of the measurements for a given patient are therefore unlikely to be associated with the patient's BMI value. Lastly, in the methodological project discussed in Chapter 6, the primary motivating application was related to patients with non-small cell lung cancer. Tumour size measurements for each patient were collected as part of the IPASS clinical trial. The measurement schedule for the clinical trial was likely pre-specified and, therefore, it is unlikely that the timing of the biomarker measurements were related to a patient's disease state.

Although assumption of an uninformative visiting process was likely to be reasonable for the applications presented in this thesis, the same will not always be true in practice. For instance, an application that was explored but not pursued during this PhD candidature was related to patients with multiple sclerosis. The data for the project came from a clinical

database of multiple sclerosis patients, in which data was recorded at clinic visits. However, clinic visits for these patients tend to occur when the disease is in an exacerbation stage, and are less likely to occur when the disease is under control. Therefore, any biomarker measurement related to the disease state is likely to be subject to an informative measurement schedule. In this type of situation, one must model the informative visiting process alongside the longitudinal biomarker and event processes. Such methods have been described by several authors (Han et al., 2014; Liu et al., 2008), however, this in an ongoing area for methodological development.

7.2.5 Dynamic predictions under multilevel joint models

The methodological project in Chapter 6 described the development of a framework for joint modelling in the presence of multiple clustering factors. The paper in Section 6.2 focussed primarily on the formulation of the multilevel hierarchical joint model, association structures, and estimation. However, in the application, several joint models with different association structures were compared using a time-dependent measure of discrimination. The time-dependent measure of discrimination was based on the area under the receiver operating characteristic curve (AUC), defined by taking a landmark time of 5 months, and a horizon time of 10 months.

In the absence of multiple clustering factors (that is, if the individual is the only clustering factor), the calculation of a time-dependent AUC measure proceeds in the following manner. For some individual i who is still at risk at a landmark time t_L, longitudinal biomarker data observed between times 0 and t_L can be used in the calculation of a conditional survival probability

$$S_i(t_L + \Delta t \mid \widetilde{\boldsymbol{b}}_i, \boldsymbol{\theta}, T_i^* > t_L) = P(T_i^* > t_L + \Delta t \mid \widetilde{\boldsymbol{b}}_i, \boldsymbol{\theta}, T_i^* > t_L) \qquad (54)$$

where $t_L + \Delta t$ denotes a horizon time for evaluating the survival probability. Conditional survival probabilities of this form can then be used in combination with the observed event times and censoring indicators for each individual, taken over the interval $(t_L, t_L + \Delta t)$, to calculate a time-dependent AUC measure (Rizopoulos, 2012b).

The tilde on the individual-level parameters $\widetilde{\boldsymbol{b}}_i$ is used to denote that these parameters differ from those obtained from fitting the model to a sample of training data $\mathcal{D} = \{T_i, d_i, \boldsymbol{y}_i; i = 1, \ldots, N\}$. Specifically, new individual-level parameters $\widetilde{\boldsymbol{b}}_i$ need to be drawn from a

distribution of the form $p(\widetilde{\boldsymbol{b}}_i \mid \widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta}, T_i^* > t_L)$ where $\widetilde{\boldsymbol{y}}_i = \{y_i(t_{ij}); j = 1, \ldots, n_i, 0 \leq t_{ij} \leq t_L\}$ denotes a collection of "new" longitudinal biomarker measurements for some individual i taken prior to the landmark time t_L. The population-level parameters $\boldsymbol{\theta}$ are obtained from the joint posterior distribution conditional on the training data, that is, $p(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D})$. Predictions that require these new individual-specific parameters $\widetilde{\boldsymbol{b}}_i$ are encountered in two situations. First, when one wishes to predict for a new individual who was not included in our training data at all. Second, when one wishes to predict for an individual who was in our original training, but wants to condition on a restricted subset of their longitudinal biomarker data (for example, the biomarker data observed prior to the landmark time t_L).

In the joint modelling literature, predictions obtained conditional on these new individual-level parameters $\widetilde{\boldsymbol{b}}_i$ are often referred to as *dynamic predictions*. This is because they are conditional on a set of "new" data for individual i and can therefore be dynamically updated as additional biomarker measurements become available or, in the context of assessing predictive accuracy, if we were to change our landmark time t_L. Dynamic predictions have been discussed by several authors (Taylor et al., 2013; Rizopoulos, 2011, 2012b; Desmée et al., 2017a; Blanche et al., 2015; Proust-Lima and Taylor, 2009; Proust-Lima et al., 2014). A method similar to that discussed by Rizopoulos (2011, 2012b) was implemented in the **rstanarm** package discussed in Chapter 5.

The framework for dynamic predictions under joint models can be extended to the situation in which there are multiple clustering factors. In the paper in Chapter 6, dynamic predictions were generated under a joint model with multiple clustering factors and used for calculating the time-dependent AUC measure. The framework for those predictions was not formally described in the paper, but will be described in detail as part of a future publication. In brief, dynamic predictions under a joint model with multiple clustering factors can be generated as follows. Suppose longitudinal biomarker measurements $y_{ijk} = y_{ij}(t_{ijk})$ are observed at a set of time points $k = 1, \ldots, K_{ij}$ for unit j $(j = 1, \ldots, J_i)$ clustered within individual i $(i = 1, \ldots, N)$. Note that this is the same notation used in the paper presented in Chapter 6. Given a sample of training data $\mathcal{D} = \{T_i, d_i, \boldsymbol{y}_{ijk}; i = 1, \ldots, N, j = 1, \ldots, J_i, k = 1, \ldots, K_{ij}\}$, a joint model can be estimated using the framework described in Chapter 6 and, accordingly, estimates of the parameters can be obtained from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{b}_i, \boldsymbol{u}_{ij} \mid \mathcal{D})$ where $\boldsymbol{\theta}$ denotes the population-level parameters, $\boldsymbol{b}_i$

denotes the individual-specific parameters for individual i, and $\boldsymbol{u}_{ij}$ denotes the cluster-specific parameters for unit j clustered within individual i. With regard to dynamic predictions of the time-to-event outcome, our goal is to estimate a survival probability for individual i, that is

$$S_i(t_L + \Delta t \mid \widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{u}}_{ij}, \boldsymbol{\theta}, T_i^* > t_L) = P(T_i^* > t_L + \Delta t \mid \widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{u}}_{ij}, \boldsymbol{\theta}, T_i^* > t_L) \qquad (55)$$

where $p(\widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{u}}_{ij} \mid \widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta}, T_i^* > t_L; j = 1, \ldots, J_i)$ denotes the joint distribution of the new individual-specific and cluster-specific parameters for individual i conditional on a set of "new" data for that individual, denoted $\widetilde{\boldsymbol{y}}_i = \{y_{ij}(t_{ijk}); j = 1, \ldots, J_i, k = 1, \ldots, K_{ij}, 0 \leq t_{ijk} \leq t_L\}$.

Under an assumption that $\widetilde{\boldsymbol{b}}_i$ and $\widetilde{\boldsymbol{u}}_{ij}$ are uncorrelated, it is straightforward to obtain estimates for these new parameters using the approach of Taylor et al. (2013) or Rizopoulos (2011, 2012b). Moreover, it is evident that $\widetilde{\boldsymbol{y}}_i$ contains sufficient information about individual i and the lower-level clusters within that individual to draw the estimates of both $\widetilde{\boldsymbol{b}}_i$ and $\widetilde{\boldsymbol{u}}_{ij}$. However, suppose that the additional clustering factor (that is, the clustering factor that is not the individual) occurs at a level above the individual; for example, if individuals were clustered within hospitals. In that setting, there may not be sufficient information to draw both the new individual-specific and cluster-specific parameters conditional on new data observed for individual i alone. For instance, suppose we observe new data for an individual who was not in our training data and who comes from a new hospital that was not in our training data either. The data for this new individual will not provide sufficient information on its own to obtain estimates for the hospital-specific parameters, which are required to generate dynamic predictions for the new individual. Consequently, a reasonable approach may be to draw the individual-specific parameters for the new individual, conditional on their data, but to marginalise over the distribution of hospital-specific parameters. These ideas will be explored as part of future work.

7.2.6 Limitations and future developments for rstanarm

7.2.6.1 Delayed entry

The various formulations for the joint models described in Chapters 2 through 6 all assumed that individuals are at risk from baseline, i.e. $t = 0$. However, consider the situation in which some individuals do not have any observed biomarker measurements. That is, they

are included in our sample but they experience the event (or are censored) prior to any longitudinal biomarker measurements being observed. If we exclude these individuals from the likelihood function for the joint model (i.e. they are not used in the model estimation), then we are effectively enforcing a constraint that individuals are not at-risk of the event until the time of their first longitudinal biomarker measurement.

Under the standard joint model likelihood, excluding individuals without any biomarker information induces a negative bias (i.e. an underestimate) in the estimated baseline hazard. This is because the subset of individuals used in the model estimation are a so-called "survivor" subset of the full sample of all individuals; the longer an individual remains event-free, then the greater their probability of having at least one longitudinal measurement observed and therefore being included in the joint likelihood. The magnitude of the bias will likely depend on the frequency of the longitudinal biomarker measurements; in situations with sparse measurements of the biomarker, the bias in the estimated baseline hazard will be greater. Importantly, this survival advantage will occur even if the visiting process (i.e. the timing of the biomarker measurements) is unrelated to the value of the unobserved biomarker measurements or the event time, an assumption discussed in Section 2.3.5.2. However, it is noteworthy that this bias will not occur in the common situation whereby all individuals have a baseline biomarker measurement, since all individuals will be used in the estimation (assuming there are no other reasons for exclusions).

The issue described here is, effectively, one of delayed entry. Delayed entry can be defined as the situation in which an individual is not at-risk of the event until some post-baseline time, i.e. $t > 0$. Although individuals may in fact be at risk of the event from time zero, our exclusion criteria when fitting the model effectively enforces the presence of delayed entry, whereby the entry time for an individual is the time of their first biomarker measurement. Crowther et al. (2016) described the formulation and estimation of a shared parameter joint model incorporating delayed entry. The likelihood function for individual i can be written as

$$L_i = \frac{\int \left(\prod_{j=1}^{n_i} p\left(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \right) p(T_i, d_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i}{\int S_i(T_{0i} \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta})\, d\boldsymbol{b}_i} \tag{56}$$

where

$$S_i(T_{0i} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\left(-\int_0^{T_{0i}} h_i(s \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, ds\right) \quad (57)$$

is the survival probability at the entry time for individual i (i.e. the time at which individual i became at-risk). This likelihood function correctly adjusts for the fact that individual i is not at-risk of the event between times 0 and T_{0i}.

It is common for most, if not all, joint modelling software to require individuals to have at least one biomarker measurement in order to be included in the model estimation. This is also true for the **rstanarm** package described in Chapter 5. In practice, this means the user must exclude any individuals who experience the event before a biomarker measurement is observed. It is therefore necessary for the software to incorporate delayed entry in the likelihood function underpinning the estimation of the joint model, otherwise, the resulting estimates of the baseline hazard will be biased in situations where not all individuals have a baseline biomarker measurement. The **rstanarm** package does not currently support delayed entry. This became evident during the simulation study discussed in Chapter 5. In early simulations, a negative bias in the estimated baseline hazard was observed. The joint longitudinal and time-to-event data being used in the simulation study was initially generated using biomarker measurement times that were uniformly distributed between zero and the maximum follow-up time. This meant that some individuals who were in the simulated data, but who experienced the event prior to the timing of their first longitudinal measurement, were excluded from the model estimation. This led to the bias discussed above. In the simulation study, this was resolved by ensuring that every individual in the simulated data had at least a baseline biomarker measurement. However, in practice, this is not always the case. For example, in a study using a registry-based dataset, baseline might be defined as the date of diagnosis of a disease, but the timing of the earliest biomarker measurement may not coincide with the date of diagnosis. Therefore, a more general solution is required. Specifically, in a future release of **rstanarm** it will be necessary to incorporate delayed entry in the likelihood function using the form specified in equation (56).

7.2.6.2 Recurrent events

Another important future development for the joint modelling functionality in **rstanarm** is to incorporate recurrent events. There is currently limited software for joint longitudinal and time-to-event models that incorporate recurrent event processes (the notable exception

being the **frailtypack** (Król et al., 2017; Rondeau et al., 2012) R package). Nonetheless, there is significant scope for the wider uptake of these models. For example, shared frailty models for terminating and recurrent events have been identified as a suitable alternative to the suboptimal analysis approaches currently used in clinical trials in intensive care (Colantuoni et al., 2016). Therefore, incorporating shared frailty models for terminating and recurrent events is an important avenue for future development in **rstanarm**.

7.2.6.3 Model diagnostics

The other feature that will be incorporated in the near future will be a wider range of functions for assessing the accuracy of predictions of event risk. Currently, the fit of the joint model in **rstanarm** can be assessed using approximate leave-one-out cross-validation measures (Vehtari et al., 2017). Under leave-one-out cross-validation it must be determined what a unit of observation is (i.e. what "one" are we leaving out?). Currently, under a joint model, the **rstanarm** package treats an individual as the unit of observation when calculating the approximate leave-one-out cross-validation measures to assess model fit. However, these provide a measure of the overall fit of the joint model based on the predictive density for both the longitudinal and time-to-event outcomes. In the context of risk prediction, one may be more interested in assessing the accuracy of the event submodel predictions in particular (i.e. focus on the model performance with regard to the time-to-event outcome alone). Therefore, in a future release of **rstanarm**, measures of predictive accuracy, for example, time-dependent measures of model discrimination (Rizopoulos, 2011) will also be added.

7.2.6.4 Simulation study to compare joint modelling software

In Section 5.4 of the thesis a qualitative comparison of software for fitting multivariate joint models was presented. This included a summary table outlining the features of each of the software packages. However, a quantitative comparison was not made. Therefore, as part of future work, a simulation study comparing the various software packages will be undertaken. The simulation study will assess estimation properties such as bias and coverage, as well as investigating computational issues such as success at convergence and computation time.

7.3 Conclusions

This thesis provides an important contribution to several areas of joint modelling research, including their development, implementation and application. The publications presented in Chapters 3 and 4 demonstrate how novel joint modelling approaches can be applied to research studies in epidemiological, public health, or clinical medicine settings. The publications presented in Chapters 5 and 6 describe new methods and software that can help facilitate the wider uptake of joint modelling approaches in health research. Moreover, Chapters 5 and 6 provide a basis for ongoing developments and contributions to research in the area of joint modelling.

References

Albert PS and Shih JH (2010) On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. *Biometrics* 66(3): 983–987. DOI: 10.1111/j.1541-0420.2009.01324_1.x.

Amorim LD and Cai J (2015) Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology* 44(1): 324–333. DOI: 10.1093/ije/dyu222.

Andersen PK and Liestøl K (2003) Attenuation caused by infrequently updated covariates in survival analysis. *Biostatistics* 4(4): 633–649. DOI: 10.1093/biostatistics/4.4.633.

Andersen PK, Geskus RB, de Witte T, et al. (2012) Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology* 41(3): 861–870. DOI: 10.1093/ije/dyr213.

Andrinopoulou E-R, Rizopoulos D, Takkenberg JJM, et al. (2014) Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine* 33(18): 3167–3178. DOI: 10.1002/sim.6158.

Baraldi AN and Enders CK (2010) An introduction to modern missing data analyses. *Journal of School Psychology* 48(1): 5–37. DOI: 10.1016/j.jsp.2009.10.001.

Bender R, Augustin T and Blettner M (2005) Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 24(11): 1713–1723. DOI: 10.1002/sim.2059.

Betancourt M (2017) A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint:170102434*.

Björk B-C and Solomon D (2013) The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics* 7(4): 914–923. DOI: 10.1016/j.joi.2013.09.001.

Blanche P, Proust-Lima C, Loubère L, et al. (2015) Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* 71(1): 102–113. DOI: 10.1111/biom.12232.

Brent RP (1973) Chapter 4: An Algorithm with Guaranteed Convergence for Finding a Zero of a Function. In: *Algorithms for Minimization without Derivatives*. Englewood Cliffs, N.J.: Prentice-Hall.

Brilleman SL (2018a) simjm: Simulate Joint Longitudinal and Survival Data. R package version: 0.0.0. Available at: https://github.com/sambrilleman/simjm.

Brilleman SL (2018b) simsurv: Simulate Survival Data. R package version: 0.2.0. Available at: https://CRAN.R-project.org/package=simsurv.

Brilleman SL, Crowther MJ, May MT, et al. (2016) Joint longitudinal hurdle and time-to-event models: an application related to viral load and duration of the first treatment regimen in patients with HIV initiating therapy. *Statistics in Medicine* 35(20): 3583–3594. DOI: 10.1002/sim.6948.

Brilleman SL, Crowther MJ, Moreno-Betancur M, et al. (2018) Joint longitudinal and time-to-event models via Stan. In: *Proceedings of StanCon 2018*, Pacific Grove, CA, USA, 10 January 2018. Available at: https://github.com/stan-dev/stancon_talks.

Bycott P and Taylor J (1998) A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* 17(18): 2061–2077.

Carpenter B, Gelman A, Hoffman MD, et al. (2017) Stan : A Probabilistic Programming Language. *Journal of Statistical Software* 76(1). DOI: 10.18637/jss.v076.i01.

Colantuoni E, Dinglas VD, Ely EW, et al. (2016) Statistical methods for evaluating delirium in the ICU. *The Lancet Respiratory Medicine* 4(7): 534–536. DOI: 10.1016/S2213-2600(16)30138-2.

Commenges D, Liquet B and Proust-Lima C (2012) Choice of Prognostic Estimators in Joint Models by Estimating Differences of Expected Conditional Kullback-Leibler Risks. *Biometrics* 68(2): 380–387. DOI: 10.1111/j.1541-0420.2012.01753.x.

Cooper NJ, Lambert PC, Abrams KR, et al. (2007) Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis. *Health Economics* 16(1): 37–56. DOI: 10.1002/hec.1141.

Crowther MJ (2017a) Extended multivariate generalised linear and non-linear mixed effects models. Available at: https://arxiv.org/abs/1710.02223.

Crowther MJ (2017b) Joint longitudinal-survival model with time-dependent effects. Available at: https://www.mjcrowther.co.uk/software/megenreg/jointmodels/jointmodel_tde/ (accessed 4 March 2018).

Crowther MJ (2017c) Mixed Effects GENeralised REGression models. Available at: https://www.mjcrowther.co.uk/software/megenreg (accessed 4 March 2018).

Crowther MJ and Lambert PC (2012) Simulating complex survival data. *The Stata Journal* 12(4): 674–687.

Crowther MJ and Lambert PC (2013) Simulating biologically plausible complex survival data. *Statistics in Medicine* 32(23): 4118–4134. DOI: 10.1002/sim.5823.

Crowther MJ, Abrams KR and Lambert PC (2012) Flexible parametric joint modelling of longitudinal and survival data. *Statistics in Medicine* 31(30): 4456–4471. DOI: 10.1002/sim.5644.

Crowther MJ, Lambert PC and Abrams KR (2013) Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: a joint modelling approach. *BMC Medical Research Methodology* 13(1). DOI: 10.1186/1471-2288-13-146.

Crowther MJ, Abrams KR and Lambert PC (2013) Joint modelling of longitudinal and survival data. *Stata Journal* 13(1): 165–184.

Crowther MJ, Andersson TM-L, Lambert PC, et al. (2016) Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine* 35(7): 1193–1209. DOI: 10.1002/sim.6779.

Dantan E, Joly P, Dartigues J-F, et al. (2011) Joint model with latent state for longitudinal and multistate data. *Biostatistics* 12(4): 723–736. DOI: 10.1093/biostatistics/kxr003.

Desmée S, Mentré F, Veyrat-Follet C, et al. (2017a) Nonlinear joint models for individual dynamic prediction of risk of death using Hamiltonian Monte Carlo: application to metastatic prostate cancer. *BMC Medical Research Methodology* 17(1). DOI: 10.1186/s12874-017-0382-9.

Desmée S, Mentré F, Veyrat-Follet C, et al. (2017b) Using the SAEM algorithm for mechanistic joint models characterizing the relationship between nonlinear PSA kinetics and survival in prostate cancer patients: Joint Model for Nonlinear Kinetics and Survival Data. *Biometrics* 73(1): 305–312. DOI: 10.1111/biom.12537.

Faucett CL and Thomas DC (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 15(15): 1663–1685. DOI: 10.1002/(SICI)1097-0258(19960815)15:15<1663::AID-SIM294>3.0.CO;2-1.

Faucett CL, Schenker N and Elashoff RM (1998) Analysis of Censored Survival Data with Intermittently Observed Time-Dependent Binary Covariates. *Journal of the American Statistical Association* 93(442): 427–437. DOI: 10.1080/01621459.1998.10473692.

Ferrer L, Rondeau V, Dignam J, et al. (2016) Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine* 35(22): 3933–3948. DOI: 10.1002/sim.6972.

Fleischmann E, Teal N, Dudley J, et al. (1999) Influence of excess weight on mortality and hospital stay in 1346 hemodialysis patients. *Kidney International* 55(4): 1560–1567. DOI: 10.1046/j.1523-1755.1999.00389.x.

Forgetta V and Richards JB (2016) Software Application Profiles: useful and novel software for epidemiological data analysis. *International Journal of Epidemiology* 45(2): 309–310. DOI: 10.1093/ije/dyw064.

Garre FG, Zwinderman AH, Geskus RB, et al. (2007) A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(1): 299–308. DOI: 10.1111/j.1467-985X.2007.00514.x.

Gruttola VD and Tu XM (1994) Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time. *Biometrics* 50(4): 1003. DOI: 10.2307/2533439.

Han J, Slate EH and Peña EA (2007) Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in Medicine* 26(29): 5285–5302. DOI: 10.1002/sim.2915.

Han M, Song X and Sun L (2014) Joint modeling of longitudinal data with informative observation times and dropouts. *Statistica Sinica*. DOI: 10.5705/ss.2013.063.

Hatfield LA, Boye ME and Carlin BP (2011) Joint Modeling of Multiple Longitudinal Patient-Reported Outcomes and Survival. *Journal of Biopharmaceutical Statistics* 21(5): 971–991. DOI: 10.1080/10543406.2011.590922.

Hatfield LA, Boye ME, Hackshaw MD, et al. (2012) Multilevel Bayesian Models for Survival Times and Longitudinal Patient-Reported Outcomes With Many Zeros. *Journal of the American Statistical Association* 107(499): 875–885. DOI: 10.1080/01621459.2012.664517.

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

Henderson R, Diggle P and Dobson A (2002) Identification and efficacy of longitudinal markers for survival. *Biostatistics* 3(1): 33–50. DOI: 10.1093/biostatistics/3.1.33.

Hickey GL, Philipson P, Jorgensen A, et al. (2016) Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology* 16(1). DOI: 10.1186/s12874-016-0212-5.

Hickey GL, Philipson P, Jorgensen A, et al. (2017) joineRML: Joint Modelling of Multivariate Longitudinal Data and Time-to-Event Outcomes. R package version 0.4.0. Available at: https://CRAN.R-project.org/package=joineRML.

Hoffman MD and Gelman A (2014) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15: 1593–1623.

Hsieh F, Tseng Y-K and Wang J-L (2006) Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited. *Biometrics* 62(4): 1037–1043. DOI: 10.1111/j.1541-0420.2006.00570.x.

Huang X, Li G and Elashoff RM (2010) A joint model of longitudinal and competing risks survival data with heterogeneous random effects and outlying longitudinal measurements. *Statistics and Its Interface* 3(2): 185–195.

Kalantar-Zadeh K, Abbott KC, Salahudeen AK, et al. (2005) Survival advantages of obesity in dialysis patients. *The American Journal of Clinical Nutrition* 81(3): 543–554.

Kim S, Zeng D, Chambless L, et al. (2012) Joint Models of Longitudinal Data and Recurrent Events with Informative Terminal Event. *Statistics in Biosciences* 4(2): 262–281. DOI: 10.1007/s12561-012-9061-x.

Koller MT, Raatz H, Steyerberg EW, et al. (2012) Competing risks and the clinical community: irrelevance or ignorance? *Statistics in Medicine* 31(11–12): 1089–1097. DOI: 10.1002/sim.4384.

Król A, Ferrer L, Pignon J-P, et al. (2016) Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000-05 trial. *Biometrics* 72(3): 907–916. DOI: 10.1111/biom.12490.

Król A, Mauguen A, Mazroui Y, et al. (2017) Tutorial in Joint Modeling and Prediction: A Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event. *Journal of Statistical Software* 81(3). DOI: 10.18637/jss.v081.i03.

Lambert P and Vandenhende F (2002) A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine* 21(21): 3197–3217. DOI: 10.1002/sim.1249.

Laurie DP (1997) Calculation of Gauss-Kronrod quadrature rules. *Mathematics of Computation* 66(219): 1133–1146. DOI: 10.1090/S0025-5718-97-00861-2.

Lawrence Gould A, Boye ME, Crowther MJ, et al. (2015) Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* 34(14): 2181–2195. DOI: 10.1002/sim.6141.

Li N, Elashoff RM, Li G, et al. (2009) Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Statistics in Medicine*: n/a-n/a. DOI: 10.1002/sim.3798.

Lin H, Turnbull BW, McCulloch CE, et al. (2002) Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer. *Journal of the American Statistical Association* 97(457): 53–65. DOI: 10.1198/016214502753479220.

Liu G and Gould AL (2002) Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics* 12(2): 207–226.

Liu L (2009) Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine* 28(6): 972–986. DOI: 10.1002/sim.3497.

Liu L and Huang X (2009) Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired

immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58(1): 65–81. DOI: 10.1111/j.1467-9876.2008.00641.x.

Liu L, Huang X and O'Quigley J (2008) Analysis of Longitudinal Data in the Presence of Informative Observational Times and a Dependent Terminal Event, with Application to Medical Cost Data. *Biometrics* 64(3): 950–958. DOI: 10.1111/j.1541-0420.2007.00954.x.

Lixoft SAS (2018) *Monolix version 2018R1*. Antony, France. Available at: http://lixoft.com/products/monolix/.

Lu T (2017) Bayesian nonparametric mixed-effects joint model for longitudinal-competing risks data analysis in presence of multiple data features. *Statistical Methods in Medical Research* 26(5): 2407–2423. DOI: 10.1177/0962280215597939.

Lunn DJ, Thomas A, Best N, et al. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 10: 325–337.

Mauff K, Steyerberg EW, Nijpels G, et al. (2017) Extension of the association structure in joint models to include weighted cumulative effects. *Statistics in Medicine* 36(23): 3746–3759. DOI: 10.1002/sim.7385.

Mbogning C, Bleakley K and Lavielle M (2015) Joint modelling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the stochastic approximation expectation–maximization algorithm. *Journal of Statistical Computation and Simulation* 85(8): 1512–1528. DOI: 10.1080/00949655.2013.878938.

Mood A, Graybill F and Boes D (1974) *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Moreno-Betancur M (2017) survtd: Survival analysis with time-dependent covariates. R package version: 0.0.1. Available at: https://github.com/moreno-betancur/survtd.

Moreno-Betancur M, Carlin JB, Brilleman SL, et al. (2017) Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics (Oxford, England)*. DOI: 10.1093/biostatistics/kxx046.

Murawska M, Rizopoulos D and Lesaffre E (2012) A Two-Stage Joint Model for Nonlinear Longitudinal Response and a Time-to-Event with Application in Transplantation Studies. *Journal of Probability and Statistics* 2012: 1–18. DOI: 10.1155/2012/194194.

Neal RM (2011) Chapter 5: MCMC using Hamiltonian dynamics. In: *Handbook of Markov Chain Monte Carlo*. CRC Press.

Pantazis N, Touloumi G, Walker AS, et al. (2005) Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(2): 405–423. DOI: 10.1111/j.1467-9876.2005.00491.x.

Park J, Ahmadi S-F, Streja E, et al. (2014) Obesity Paradox in End-Stage Kidney Disease Patients. *Progress in Cardiovascular Diseases* 56(4): 415–425. DOI: 10.1016/j.pcad.2013.10.005.

Philipson PM, Ho WK and Henderson R (2008) Comparative review of methods for handling drop-out in longitudinal studies. *Statistics in Medicine* 27(30): 6276–6298. DOI: 10.1002/sim.3450.

Piironen J and Vehtari A (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* 11(2): 5018–5051. DOI: 10.1214/17-EJS1337SI.

Plummer M (2016) rjags: Bayesian Graphical Models using MCMC. R package version 4-6. Available at: https://CRAN.R-project.org/package=rjags.

Prentice RL (1982) Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69(2): 331–342. DOI: 10.1093/biomet/69.2.331.

Proust-Lima C and Taylor JMG (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* 10(3): 535–549. DOI: 10.1093/biostatistics/kxp009.

Proust-Lima C, Joly P, Dartigues J-F, et al. (2009) Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis* 53(4): 1142–1154. DOI: 10.1016/j.csda.2008.10.017.

Proust-Lima C, Séne M, Taylor JMG, et al. (2014) Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* 23(1): 74–90. DOI: 10.1177/0962280212445839.

Proust-Lima C, Dartigues J-F and Jacqmin-Gadda H (2016) Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in Medicine* 35(3): 382–398. DOI: 10.1002/sim.6731.

Proust-Lima C, Philipps V and Liquet B (2017) Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software* 78(2). DOI: 10.18637/jss.v078.i02.

Rizopoulos D (2010) JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software* 35(9). 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): 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): 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*. Chapman & Hall/CRC biostatistics series 6. Boca Raton: CRC Press.

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(7). DOI: 10.18637/jss.v072.i07.

Rizopoulos D (2017) Multivariate Joint Models. Available at: http://www.drizopoulos.com/vignettes/multivariate%20joint%20models (accessed 4 March 2018).

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): 1366–1380. DOI: 10.1002/sim.4205.

Rizopoulos D, Verbeke G, Lesaffre E, et al. (2008) A Two-Part Joint Model for the Analysis of Survival and Longitudinal Binary Data with Excess Zeros. *Biometrics* 64(2): 611–619. DOI: 10.1111/j.1541-0420.2007.00894.x.

Rizopoulos D, Hatfield LA, Carlin BP, et al. (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): 1385–1397. DOI: 10.1080/01621459.2014.931236.

Rizopoulos D, Taylor JMG, Van Rosmalen J, et al. (2015) Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*: kxv031. DOI: 10.1093/biostatistics/kxv031.

Rodríguez G (2007) Chapter 7: Survival Models. Lecture Notes on Generalized Linear Models. Available at: http://data.princeton.edu/wws509/notes/ (accessed 2 March 2018).

Rondeau V, Mazroui Y and Gonzalez JR (2012) frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *Journal of Statistical Software* 47(4). DOI: 10.18637/jss.v047.i04.

Rouanet A, Joly P, Dartigues J-F, et al. (2016) Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics* 72(4): 1123–1135. DOI: 10.1111/biom.12530.

Rubin D (1987) *Multiple Imputation for Nonresponse in Surveys*. NewYork: Wiley.

Rutherford MJ, Crowther MJ and Lambert PC (2015) The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation* 85(4): 777–793. DOI: 10.1080/00949655.2013.845890.

Schmidt DS and Salahudeen AK (2007) Cardiovascular and Survival Paradoxes in Dialysis Patients: Obesity-Survival Paradox-Still a Controversy? *Seminars in Dialysis* 20(6): 486–492. DOI: 10.1111/j.1525-139X.2007.00349.x.

Schoop R, Graf E and Schumacher M (2008) Quantifying the Predictive Performance of Prognostic Models for Censored Survival Data with Time-Dependent Covariates. *Biometrics* 64(2): 603–610. DOI: 10.1111/j.1541-0420.2007.00889.x.

SCImago (2007) SJR — SCImago Journal & Country Rank: Journal of Statistical Software. Available at: http://www.scimagojr.com/journalsearch.php?q=12137&tip=sid (accessed 4 March 2018).

Sperrin M, Candlish J, Badrick E, et al. (2016) Collider Bias Is Only a Partial Explanation for the Obesity Paradox: *Epidemiology* 27(4): 525–530. DOI: 10.1097/EDE.0000000000000493.

Spiegelhalter DJ (2002) Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(1): 115–133. DOI: 10.1111/1467-9876.00101.

Stan Development Team (2017a) rstanarm: Bayesian applied regression modeling via Stan. R package version 2.17.1. Available at: http://mc-stan.org/rstanarm.

Stan Development Team (2017b) Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. Available at: http://mc-stan.org.

Sudell M, Tudur Smith C, Gueyffier F, et al. (2017) Investigation of 2-stage meta-analysis methods for joint longitudinal and time-to-event data through simulation and real data application. *Statistics in Medicine*. DOI: 10.1002/sim.7585.

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

Taylor JMG, Park Y, Ankerst DP, et al. (2013) Real-Time Individual Predictions of Prostate Cancer Recurrence Using Joint Models. *Biometrics* 69(1): 206–213. DOI: 10.1111/j.1541-0420.2012.01823.x.

Thiébaut R, Jacqmin-Gadda H, Babiker A, et al. (2005) Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statistics in Medicine* 24(1): 65–82. DOI: 10.1002/sim.1923.

Tseng Y-K, Hsieh F and Wang J-L (2005) Joint modelling of accelerated failure time and longitudinal data. *Biometrika* 92(3): 587–603. DOI: 10.1093/biomet/92.3.587.

Tsiatis AA and Davidian M (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 14(3): 809–834.

Vehtari A, Gelman A and Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5): 1413–1432. DOI: 10.1007/s11222-016-9696-4.

Verbeke G, Fieuws S, Molenberghs G, et al. (2014) The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research* 23(1): 42–59. DOI: 10.1177/0962280212445834.

Viviani S, Alfó M and Rizopoulos D (2014) Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing* 24(3): 417–427. DOI: 10.1007/s11222-013-9378-4.

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): 895–905. DOI: 10.1198/016214501753208591.

Wickham H (2015) Vignettes: long-form documentation. In: *R Packages: Organize, Test, Document, and Share Your Code*. 1st ed. O'Reilly Media. Available at: http://r-pkgs.had.co.nz/vignettes.html (accessed 27 February 2018).

Williamson PR, Kolamunnage-Dona R, Philipson P, et al. (2008) Joint modelling of longitudinal and competing risks data. *Statistics in Medicine* 27(30): 6426–6438. DOI: 10.1002/sim.3451.

Wu L, Liu W and Hu XJ (2010) Joint Inference on HIV Viral Dynamics and Immune Suppression in Presence of Measurement Errors. *Biometrics* 66(2): 327–335. DOI: 10.1111/j.1541-0420.2009.01308.x.

Wulfsohn MS and Tsiatis AA (1997) A joint model for survival and longitudinal data measured with error. *Biometrics* 53(1): 330–339.

Ye W, Lin X and Taylor JMG (2008a) A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Its Interface* 1(1): 33–45. DOI: 10.4310/SII.2008.v1.n1.a4.

Ye W, Lin X and Taylor JMG (2008b) Semiparametric Modeling of Longitudinal Measurements and Time-to-Event Data-A Two-Stage Regression Calibration Approach. *Biometrics* 64(4): 1238–1246. DOI: 10.1111/j.1541-0420.2007.00983.x.

Zhang D, Chen M-H, Ibrahim JG, et al. (2014) Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. *Statistics in Medicine* 33(27): 4715–4733. DOI: 10.1002/sim.6269.

Appendix A. Supplementary materials for Chapter 3 paper

This appendix herein contains the supplementary materials for the following paper that was presented in Chapter 3:

Brilleman SL, Wolfe R, Moreno-Betancur M, Sales AE, Langa KM, Li Y, Daugherty Biddison EL, Rubinson L, Iwashyna TJ. Associations between community-level disaster exposure and individual-level changes in disability and risk of death for older Americans. *Social Science & Medicine*. 2017;173:118-125.

Supplementary material for: "Associations between community-level disaster exposure and individual-level changes in disability and risk of death for older Americans"

Samuel L Brilleman, Rory Wolfe, Margarita Moreno-Betancur, Anne Sales, Kenneth M Langa, Yun Li, Elizabeth Daugherty Biddison, Lewis Rubinson, Theodore J Iwashyna

1. Model estimation

Software. For fitting the joint models we used the JMbayes package (1) in R version 3.2.2 (2). The JMbayes package fits joint models under a Bayesian framework using a Metropolis-based Markov chain Monte Carlo (MCMC) algorithm. Since the dataset we were working with was relatively large (N = 17,559 individuals) the amount of computer RAM required to fit the models exceeded what was available on any local machine we had access to (approximately 12GB of RAM was required to fit each joint model). We chose, therefore, to fit the joint models using a UNIX-based computer cluster. The main joint model used in the analysis (with a current value association structure) took approximately 12 hours to fit based on a draw of 26,000 MCMC iterations (which included the adaptation and burn-in iterations).

Prior distributions. For each of the fixed regression coefficients ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$) and the association parameter (α_1), vague univariate normal prior distributions were used. For the variance-covariance matrix of the random effects ($\boldsymbol{\Sigma}$), an inverse-Wishart prior distribution was used. For the scale parameter for the negative binomial distribution (ϕ) an inverse Gamma distribution was used. These were the default prior distributions provided by the JMbayes package (1).

Sampling. For each joint model we obtained 26,000 MCMC iterations, drawn from one "chain" using a single set of initial values. The initial values are obtained by fitting separate longitudinal and survival models prior to fitting the joint model. The 26,000 MCMC iterations included 3000 iterations for each of the adaptation and "burn-in" phases. Furthermore, to reduce autocorrelation between parameter estimates obtained from subsequent MCMC iterations the final set of 20,000 MCMC iterations is thinned down to just 2,000 by retaining every 10th iteration. This "thinning" process helps to ensure the retained set of MCMC samples (i.e., those which are used for inference) represent a random sample from the joint posterior distribution for the model.

Density for the longitudinal submodel. The JMbayes package requires the user to explicitly specify the density for the longitudinal submodel. As described in our main manuscript, the longitudinal submodel in this study consisted of a generalised linear mixed model with log link function and negative binomial error distribution. The density for the longitudinal submodel, for a single observation y_{ij}, is specified as

$$p(y_{ij} | \mu_{ij}, \phi) = \binom{y_{ij} + \phi - 1}{y_{ij}} \left(\frac{\mu_{ij}}{\mu_{ij} + \phi}\right)^{y_{ij}} \left(\frac{\phi}{\mu_{ij} + \phi}\right)^{\phi}$$

where $\mu_{ij} = \mu_i(t_{ij})$ and ϕ is the negative binomial scale parameter (both parameters are specified in equation (1) of the main manuscript).

2. Characteristics associated with disaster rates

Of the 17,559 individuals included in the analysis there were 16,075 (92%) individuals who experienced at least one disaster during the study period. Table S1 shows the number of individuals experiencing each type of disaster at least once, as well as the total number of person-disaster events for each disaster type. Storm, hurricane and snow were the most widely experienced disaster types. There appeared to be reasonable correspondence between the percentage of individuals experiencing a specific type of disaster and the prevalence of that disaster type overall. In other words, there did not appear to be any disaster type which was frequently and repeatedly experienced by only a small number of individuals.

Poisson regression models were also used to assess which individual-level baseline characteristics were associated with individual-level disaster rates. We considered both the rates of overall disasters (i.e., any disaster type) as well as fitting a separate regression model for the rates of each disaster type. The rate ratios from the estimated models are shown in Tables S2 and S3.

3. Results from joint models with alternative association structures

In the main manuscript we provided the results from a joint model for longitudinal disability score measurements and time-to-death. The shared parameter joint model defined in the main manuscript was based on a "current value" association structure. In equation (3) in the main manuscript we defined the survival submodel as

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{w_i'}(t)\boldsymbol{\gamma} + \alpha_1\eta_i(t)\right)$$

where the fixed coefficient α_1 is referred to as the "current value" association parameter, since it measures the strength of association between the current expected value of the log disability score at time t and the log hazard of death at time t.

In this section we present the results for joint models based on two alternative association structures. The first is a "current value and slope" association structure, where the survival submodel takes the form

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{w_i'}(t)\boldsymbol{\gamma} + \alpha_2\eta_i(t) + \alpha_3{\eta_i}'(t)\right)$$

Under this joint model the hazard of death at time t is assumed to be associated with the current expected value of the log disability score at time t as well as the current rate of change (slope) in the log disability score. The second is a current value association structure with an current value by age category interaction term, where the survival submodel takes the form

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{w_i'}(t)\boldsymbol{\gamma} + \alpha_4\eta_i(t) + \sum\nolimits_{\boldsymbol{a}}\alpha_a\left(A_{ia}\eta_i(t)\right)\right)$$

where the A_{ia} are the dummy variables for the age categories (defined in the main manuscript). The specification of the longitudinal submodel, as well as the specification of $h_0(t)$ and $\boldsymbol{w_i}(t)$, are the same as described for the joint model in the main manuscript.

Tables S4 and S5, respectively, show the estimated disability score ratios and hazard ratios from the fitted joint models; for ease of comparison the joint model from the main manuscript is also

included. From the joint model with the current value by age category interaction we observed that the magnitude of the association between disability and death decreased slightly with age. For example, a doubling in the estimated disability score was associated with a 39%[1] increase in the risk of death for the youngest age category but only a 28%[2] increase in the risk of death for the oldest age category.

There was also some evidence that the rate of change (slope) in the log disability score was associated with the risk of death. From the joint model with the current value and slope parameterisation (Table S2) we observed that an individual's estimated risk of death is 40% (HR = 1.40[3], 95% CrI: 0.95 to 2.04) higher if their rate of increase in disability is twice as fast, for given fixed values of the level of disability, baseline covariates and disaster exposure. However, the credible interval surrounding this estimate was relatively wide which may suggest that the current level of disability is in fact a more important predictor of the risk of death rather than the rate of change in disability.

4. Sensitivity analysis only including individuals with "low" baseline disability

Following a reviewer's suggestion, we conducted a sensitivity analysis in which we only included those individuals with a baseline disability score of 0, 1 or 2. The parameter estimates from the model fitted to this subgroup are shown in Tables S5 and S6. The point estimate for the association between disaster exposure and disability was positive in this subgroup (that is, it suggested that the presence of a disaster within the previous 2 years was associated with slightly higher mean levels of disability), however, the 95% credible interval for this estimate included a value of 1 corresponding to no association (disability score ratio = 1.04, 95% CI: 0.98 to 1.10). The hazard ratio estimate for the association between disaster exposure and risk of mortality remained almost unchanged; that is it was similar in both this subgroup and the overall sample.

References

1. Rizopoulos D. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software* [in press].

2. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2015.

[1] Calculated as $\exp\{0.693 \times \log(1.61)\}$
[2] Calculated as $\exp\{0.693 \times (\log(1.61) + \log(0.89))\}$
[3] Calculated as $\exp\{0.693 \times \log(1.62)\}$

Table S1. Disaster rate ratios (and 95% confidence intervals) from Poisson regression models for the number of disasters experienced by each individual during follow up. The models are adjusted for the total follow up time (in years) for each individual. Separate models were fit for the overall disaster rate (i.e., any disaster) and for each disaster type. The table presented here includes: any disaster; storm; hurricane; snow; and fire.

	Any disaster type	Storm	Hurricane	Snow	Fire
Constant	0.44 (0.42 to 0.45)	0.21 (0.20 to 0.22)	0.09 (0.09 to 0.10)	0.08 (0.08 to 0.09)	0.03 (0.02 to 0.03)
Age category (ref: ≥50, <60y)					
≥60, <65y	1.02 (0.99 to 1.04)	1.02 (0.98 to 1.05)	1.04 (0.99 to 1.09)	0.98 (0.92 to 1.04)	1.05 (0.96 to 1.15)
≥65, <70y	1.05 (1.02 to 1.07)	1.06 (1.02 to 1.09)	1.08 (1.03 to 1.13)	1.00 (0.94 to 1.06)	1.02 (0.93 to 1.12)
≥70, <75y	1.02 (0.99 to 1.04)	0.96 (0.93 to 1.00)	1.04 (0.99 to 1.10)	1.10 (1.03 to 1.17)	1.14 (1.03 to 1.27)
≥75, <80y	1.06 (1.03 to 1.09)	1.01 (0.97 to 1.05)	1.12 (1.06 to 1.19)	1.09 (1.01 to 1.17)	1.20 (1.07 to 1.33)
≥80, <85y	1.15 (1.11 to 1.18)	1.04 (0.99 to 1.09)	1.20 (1.12 to 1.28)	1.31 (1.21 to 1.42)	1.40 (1.24 to 1.58)
≥85, <90y	1.21 (1.16 to 1.26)	1.03 (0.96 to 1.10)	1.23 (1.14 to 1.34)	1.61 (1.46 to 1.77)	1.45 (1.24 to 1.69)
Gender (ref: Male)					
Female	0.98 (0.96 to 0.99)	0.99 (0.97 to 1.01)	0.96 (0.93 to 0.99)	0.98 (0.94 to 1.02)	0.96 (0.90 to 1.02)
Race (ref: White, Caucasian)					
Black, African American	0.90 (0.88 to 0.92)	0.80 (0.77 to 0.83)	0.99 (0.95 to 1.04)	0.95 (0.89 to 1.01)	0.85 (0.77 to 0.94)
Other	0.90 (0.86 to 0.94)	0.78 (0.72 to 0.83)	0.90 (0.83 to 0.98)	0.77 (0.68 to 0.87)	2.02 (1.79 to 2.27)
Wealth category (ref: Decile 1, most wealth)					
Decile 2	1.00 (0.97 to 1.04)	1.03 (0.98 to 1.08)	0.98 (0.91 to 1.06)	0.99 (0.92 to 1.07)	0.99 (0.87 to 1.13)
Decile 3	1.03 (1.00 to 1.07)	1.08 (1.03 to 1.13)	1.02 (0.94 to 1.09)	0.99 (0.91 to 1.07)	1.11 (0.98 to 1.27)
Decile 4	1.00 (0.97 to 1.04)	0.99 (0.94 to 1.04)	1.06 (0.98 to 1.14)	1.04 (0.97 to 1.13)	1.01 (0.88 to 1.15)
Decile 5	1.00 (0.97 to 1.04)	1.02 (0.97 to 1.07)	1.08 (1.01 to 1.16)	0.96 (0.89 to 1.04)	1.00 (0.88 to 1.15)
Decile 6	1.01 (0.97 to 1.04)	1.03 (0.98 to 1.08)	1.10 (1.03 to 1.19)	0.90 (0.83 to 0.98)	1.00 (0.87 to 1.14)
Decile 7	1.02 (0.98 to 1.05)	1.00 (0.95 to 1.05)	1.39 (1.30 to 1.49)	0.73 (0.67 to 0.80)	0.88 (0.77 to 1.01)
Decile 8	0.98 (0.95 to 1.02)	0.93 (0.88 to 0.98)	1.49 (1.39 to 1.60)	0.65 (0.59 to 0.71)	0.86 (0.75 to 0.99)
Decile 9	1.01 (0.98 to 1.05)	0.98 (0.93 to 1.03)	1.47 (1.37 to 1.58)	0.67 (0.61 to 0.73)	1.02 (0.89 to 1.17)
Decile 10, least wealth	1.02 (0.98 to 1.06)	0.90 (0.85 to 0.95)	1.38 (1.28 to 1.49)	0.75 (0.68 to 0.82)	1.43 (1.26 to 1.63)

Abbreviations. “ref”: reference category.

Table S2. Disaster rate ratios (and 95% confidence intervals) from Poisson regression models for the number of disasters experienced by each individual during follow up. The models are adjusted for the total follow up time (in years) for each individual. Separate models were fit for the overall disaster rate (i.e., any disaster) and for each disaster type. The table presented here includes: flood; tornado; earthquake; and other.

	Flood	Tornado	Earthquake	Other
Constant	0.01 (0.01 to 0.02)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.01)	0.01 (0.01 to 0.01)
Age category (ref: ≥50, <60y)				
≥60, <65y	1.11 (0.95 to 1.30)	0.89 (0.71 to 1.11)	0.82 (0.56 to 1.19)	0.99 (0.87 to 1.14)
≥65, <70y	1.09 (0.92 to 1.28)	1.16 (0.93 to 1.44)	0.75 (0.51 to 1.12)	0.98 (0.85 to 1.13)
≥70, <75y	0.91 (0.76 to 1.10)	0.93 (0.72 to 1.20)	0.86 (0.57 to 1.31)	1.04 (0.89 to 1.22)
≥75, <80y	0.87 (0.71 to 1.07)	0.89 (0.67 to 1.19)	1.44 (0.98 to 2.11)	1.21 (1.03 to 1.42)
≥80, <85y	0.88 (0.69 to 1.13)	0.69 (0.48 to 0.99)	1.12 (0.68 to 1.85)	1.38 (1.16 to 1.65)
≥85, <90y	0.92 (0.65 to 1.29)	0.65 (0.39 to 1.09)	1.41 (0.74 to 2.69)	1.64 (1.33 to 2.04)
Gender (ref: Male)				
Female	0.92 (0.82 to 1.02)	1.00 (0.85 to 1.17)	0.97 (0.76 to 1.24)	0.99 (0.90 to 1.08)
Race (ref: White or Caucasian)				
Black or African American	0.21 (0.15 to 0.30)	1.72 (1.42 to 2.10)	0.23 (0.10 to 0.53)	1.77 (1.57 to 1.99)
Other	0.45 (0.29 to 0.69)	0.71 (0.44 to 1.16)	0.37 (0.12 to 1.16)	1.31 (1.04 to 1.63)
Wealth category (ref: Decile 1, most wealth)				
Decile 2	0.97 (0.79 to 1.19)	0.92 (0.63 to 1.35)	0.84 (0.57 to 1.23)	0.93 (0.76 to 1.14)
Decile 3	0.87 (0.71 to 1.08)	1.08 (0.75 to 1.57)	0.51 (0.33 to 0.80)	0.78 (0.64 to 0.97)
Decile 4	0.78 (0.63 to 0.98)	0.95 (0.65 to 1.40)	0.62 (0.41 to 0.95)	0.87 (0.71 to 1.06)
Decile 5	0.71 (0.56 to 0.89)	1.28 (0.89 to 1.83)	0.46 (0.29 to 0.74)	0.89 (0.73 to 1.09)
Decile 6	0.85 (0.68 to 1.06)	1.41 (0.99 to 2.00)	0.28 (0.15 to 0.50)	0.92 (0.75 to 1.13)
Decile 7	0.80 (0.64 to 1.01)	1.43 (1.01 to 2.03)	0.27 (0.15 to 0.49)	0.95 (0.78 to 1.16)
Decile 8	0.80 (0.63 to 1.02)	1.22 (0.85 to 1.76)	0.38 (0.22 to 0.66)	0.83 (0.67 to 1.02)
Decile 9	0.79 (0.62 to 1.00)	1.23 (0.86 to 1.77)	0.30 (0.16 to 0.55)	0.79 (0.64 to 0.97)
Decile 10, least wealth	0.46 (0.33 to 0.63)	1.63 (1.14 to 2.32)	0.25 (0.12 to 0.52)	1.33 (1.10 to 1.61)

Abbreviations. “ref”: reference category.

Table S3. Disability score ratios from the longitudinal submodel, for joint models with various association structures. Estimates presented are posterior means and associated 95% credible intervals.

	Current value parameterisation	Current value and slope parameterisation	Current value by age category interaction
Constant	0.02 (0.02 to 0.03)	0.02 (0.02 to 0.03)	0.02 (0.02 to 0.03)
Time (years)	1.03 (1.01 to 1.04)	1.02 (1.01 to 1.04)	1.02 (1.01 to 1.04)
Age category (ref: ≥50, <60y)			
≥60, <65y	0.92 (0.82 to 1.03)	0.92 (0.81 to 1.03)	0.92 (0.82 to 1.04)
≥65, <70y	1.19 (1.06 to 1.33)	1.18 (1.05 to 1.33)	1.18 (1.05 to 1.33)
≥70, <75y	1.72 (1.51 to 1.95)	1.70 (1.50 to 1.94)	1.71 (1.49 to 1.95)
≥75, <80y	3.04 (2.66 to 3.48)	2.99 (2.63 to 3.39)	3.03 (2.63 to 3.53)
≥80, <85y	5.70 (4.96 to 6.64)	5.62 (4.89 to 6.51)	5.69 (4.94 to 6.62)
≥85, <90y	9.75 (8.12 to 11.79)	9.51 (7.96 to 11.34)	9.57 (7.88 to 11.63)
Age category * time interaction			
≥60, <65y	1.05 (1.03 to 1.06)	1.05 (1.03 to 1.06)	1.05 (1.03 to 1.07)
≥65, <70y	1.10 (1.08 to 1.11)	1.10 (1.08 to 1.12)	1.10 (1.08 to 1.12)
≥70, <75y	1.18 (1.16 to 1.20)	1.18 (1.16 to 1.20)	1.18 (1.15 to 1.20)
≥75, <80y	1.22 (1.20 to 1.24)	1.23 (1.21 to 1.26)	1.22 (1.20 to 1.25)
≥80, <85y	1.27 (1.25 to 1.30)	1.29 (1.26 to 1.32)	1.28 (1.25 to 1.31)
≥85, <90y	1.27 (1.24 to 1.30)	1.28 (1.25 to 1.32)	1.27 (1.23 to 1.31)
Gender (ref: Male)			
Female	1.02 (0.95 to 1.09)	1.02 (0.95 to 1.09)	1.01 (0.94 to 1.08)
Race (ref: White or Caucasian)			
Black or African American	1.30 (1.17 to 1.44)	1.30 (1.17 to 1.45)	1.30 (1.17 to 1.45)
Other	1.15 (0.95 to 1.37)	1.15 (0.95 to 1.39)	1.15 (0.95 to 1.38)
Wealth category (ref: Decile 1, most wealth)			
Decile 2	1.10 (0.94 to 1.30)	1.10 (0.92 to 1.29)	1.11 (0.95 to 1.30)
Decile 3	1.27 (1.08 to 1.49)	1.27 (1.08 to 1.48)	1.27 (1.08 to 1.48)
Decile 4	1.74 (1.49 to 2.05)	1.74 (1.49 to 2.03)	1.75 (1.50 to 2.06)

Decile 5	1.86 (1.61 to 2.17)	1.87 (1.59 to 2.19)	1.87 (1.61 to 2.19)
Decile 6	2.23 (1.91 to 2.60)	2.25 (1.91 to 2.62)	2.25 (1.93 to 2.60)
Decile 7	3.06 (2.60 to 3.57)	3.07 (2.64 to 3.60)	3.05 (2.61 to 3.58)
Decile 8	3.71 (3.16 to 4.32)	3.70 (3.15 to 4.33)	3.72 (3.19 to 4.37)
Decile 9	5.28 (4.46 to 6.18)	5.31 (4.54 to 6.23)	5.31 (4.57 to 6.21)
Decile 10, least wealth	9.52 (8.12 to 11.25)	9.60 (8.22 to 11.24)	9.60 (8.15 to 11.34)
Disaster exposure			
Within previous 2 years	0.98 (0.93 to 1.03)	0.99 (0.92 to 1.04)	0.97 (0.90 to 1.02)

Abbreviations. “ref”: reference category.

Table S4. Hazard ratios from the survival submodel, for joint models with various association structures. Estimates presented are posterior means and associated 95% credible intervals.

	Current value parameterisation	Current value and slope parameterisation	Current value by age category interaction
Age category (ref: ≥50, <60y)			
≥60, <65y	2.40 (1.54 to 3.62)	2.54 (1.05 to 6.16)	2.58 (0.87 to 8.16)
≥65, <70y	3.58 (2.52 to 5.33)	3.78 (1.58 to 7.40)	3.31 (1.03 to 10.54)
≥70, <75y	3.80 (2.65 to 5.59)	4.11 (1.70 to 8.63)	3.73 (1.15 to 12.24)
≥75, <80y	5.81 (4.12 to 8.54)	6.42 (2.62 to 13.92)	5.75 (1.96 to 14.74)
≥80, <85y	7.88 (5.39 to 11.25)	7.76 (3.31 to 17.03)	7.62 (2.56 to 24.55)
≥85, <90y	9.88 (6.65 to 14.78)	10.08 (3.81 to 23.71)	9.94 (3.20 to 26.49)
Gender (ref: Male)			
Female	0.60 (0.56 to 0.65)	0.61 (0.53 to 0.68)	0.62 (0.53 to 0.72)
Race (ref: White or Caucasian)			
Black or African American	0.90 (0.80 to 0.99)	0.90 (0.72 to 1.11)	0.89 (0.67 to 1.10)
Other	0.74 (0.61 to 0.92)	0.75 (0.46 to 1.15)	0.69 (0.37 to 1.06)
Wealth trend across deciles			
Linear trend (0 = Decile 1; 9 = Decile 10)	1.13 (1.07 to 1.18)	1.15 (1.01 to 1.28)	1.12 (0.97 to 1.27)
Age category * wealth trend interaction			
≥60, <65y	0.93 (0.87 to 0.99)	0.92 (0.81 to 1.06)	0.92 (0.80 to 1.09)
≥65, <70y	0.91 (0.86 to 0.96)	0.91 (0.81 to 1.03)	0.93 (0.81 to 1.09)
≥70, <75y	0.93 (0.88 to 0.98)	0.91 (0.82 to 1.04)	0.93 (0.79 to 1.10)
≥75, <80y	0.91 (0.86 to 0.96)	0.89 (0.78 to 1.01)	0.91 (0.80 to 1.08)
≥80, <85y	0.89 (0.85 to 0.94)	0.89 (0.78 to 1.01)	0.91 (0.79 to 1.05)
≥85, <90y	0.88 (0.83 to 0.93)	0.87 (0.76 to 1.00)	0.90 (0.77 to 1.05)
Disaster exposure			
Within previous 21 days	0.96 (0.75 to 1.21)	0.94 (0.56 to 1.43)	0.94 (0.54 to 1.54)
Within previous 2 years, but not 21 days	1.03 (0.96 to 1.11)	1.02 (0.87 to 1.18)	1.02 (0.88 to 1.16)
Association parameters			

Current value	1.56 (1.45 to 1.65)	1.54 (1.41 to 1.66)	1.61 (1.15 to 2.32)
Current slope	na	1.62 (0.93 to 2.81)	na
Age category * current value interaction			
≥60, <65y	na	na	1.00 (0.63 to 1.47)
≥65, <70y	na	na	0.96 (0.64 to 1.41)
≥70, <75y	na	na	0.95 (0.57 to 1.45)
≥75, <80y	na	na	0.96 (0.58 to 1.54)
≥80, <85y	na	na	0.94 (0.60 to 1.35)
≥85, <90y	na	na	0.89 (0.56 to 1.40)

Abbreviations. “ref”: reference category. “na”: not applicable.

Table S5. Disability score ratios from the longitudinal submodel of a joint model fit to the subgroup of participants with a baseline disability score of 0, 1 or 2. Estimates presented are posterior means and associated 95% credible intervals.

	Disability score ratio (95% credible interval)
Constant	0.02 (0.02 to 0.02)
Time (years)	1.03 (1.02 to 1.04)
Age category (ref: ≥50, <60y)	
≥60, <65y	0.91 (0.82 to 1.01)
≥65, <70y	1.06 (0.95 to 1.17)
≥70, <75y	1.48 (1.32 to 1.65)
≥75, <80y	2.35 (2.09 to 2.65)
≥80, <85y	3.64 (3.18 to 4.16)
≥85, <90y	3.85 (3.23 to 4.62)
Age category * time interaction	
≥60, <65y	1.05 (1.04 to 1.07)
≥65, <70y	1.11 (1.10 to 1.13)
≥70, <75y	1.20 (1.18 to 1.22)
≥75, <80y	1.27 (1.24 to 1.29)
≥80, <85y	1.37 (1.34 to 1.40)
≥85, <90y	1.45 (1.41 to 1.49)
Gender (ref: Male)	
Female	0.99 (0.93 to 1.06)
Race (ref: White or Caucasian)	
Black or African American	1.24 (1.12 to 1.38)
Other	1.10 (0.92 to 1.31)
Wealth category (ref: Decile 1, most wealth)	
Decile 2	1.02 (0.88 to 1.18)
Decile 3	1.20 (1.04 to 1.38)
Decile 4	1.58 (1.37 to 1.82)
Decile 5	1.64 (1.41 to 1.89)
Decile 6	1.87 (1.61 to 2.16)
Decile 7	2.55 (2.22 to 2.94)
Decile 8	2.84 (2.44 to 3.28)
Decile 9	3.71 (3.21 to 4.28)
Decile 10, least wealth	5.26 (4.50 to 6.15)
Disaster exposure	
Within previous 2 years	1.04 (0.98 to 1.10)

Abbreviations. “ref”: reference category.

Table S6. Hazard ratios from the survival submodel of a joint model fit to the subgroup of participants with a baseline disability score of 0, 1 or 2. Estimates presented are posterior means and associated 95% credible intervals.

	Hazard ratio (95% credible interval)
Age category (ref: ≥50, <60y)	
≥60, <65y	2.14 (1.47 to 3.05)
≥65, <70y	3.15 (2.27 to 4.45)
≥70, <75y	3.29 (2.35 to 4.43)
≥75, <80y	5.45 (3.82 to 7.61)
≥80, <85y	6.92 (4.79 to 9.64)
≥85, <90y	8.21 (5.69 to 12.16)
Gender (ref: Male)	
Female	0.59 (0.56 to 0.63)
Race (ref: White or Caucasian)	
Black or African American	0.89 (0.80 to 0.99)
Other	0.75 (0.60 to 0.93)
Wealth trend across deciles	
Linear trend (0 = Decile 1; 9 = Decile 10)	1.12 (1.07 to 1.17)
Age category * wealth trend interaction	
≥60, <65y	0.94 (0.89 to 1.00)
≥65, <70y	0.93 (0.88 to 0.98)
≥70, <75y	0.95 (0.90 to 1.00)
≥75, <80y	0.91 (0.86 to 0.97)
≥80, <85y	0.90 (0.86 to 0.96)
≥85, <90y	0.91 (0.85 to 0.97)
Disaster exposure	
Within previous 21 days	0.92 (0.72 to 1.17)
Within previous 2 years, but not 21 days	1.03 (0.96 to 1.10)
Association parameters	
Current value	1.53 (1.46 to 1.60)

Abbreviations. “ref”: reference category. “na”: not applicable.

Appendix B. Supplementary materials for Chapter 4 paper

This appendix herein contains the supplementary materials for the following paper that was presented in Chapter 4:

Brilleman SL, Moreno-Betancur M, Polkinghorne KR, McDonald SP, Crowther MJ, Thomson J, Wolfe R. Longitudinal changes in body mass index and the competing outcomes of death and transplant in patients undergoing hemodialysis: a joint latent class mixed model approach. *Submitted for publication*.

Supplementary material for

"Longitudinal changes in body mass index and competing outcomes of death and transplant in patients undergoing haemodialysis"

Samuel L Brilleman, Margarita Moreno-Betancur, Kevan R. Polkinghorne,Stephen P. McDonald, Michael J. Crowther, Jim Thomson, Rory Wolfe

1. Details related to patient exclusions

Figure S1 provides a flowchart illustrating the derivation of the analysis sample. Of the 19,264 patients who initiated hemodialysis during the study period, we excluded the following: 536 patients who recovered kidney function, 879 patients with unknown height, race, or comorbidity status, 21 patients with extreme (<130 cm) baseline height, 1376 patients with an extreme (≥45 or <17.5 kg/m^2) BMI measurement during follow up. Since there were some overlaps for these exclusion categories (e.g. a patient might have had a missing height measurement *and* have recovered kidney function) these exclusions meant there were 16,585 patients remaining, of which 16,414 patients had at least one BMI measurement recorded prior to death, transplant or censoring and were therefore included in the main analyses.

2. Observed BMI trajectories

Observed longitudinal BMI trajectories for a random sample of 25 patients are shown in Figure S2.

3. Numbers of patients with BMI measurements in the 1st, 2nd, 3rd, 4th and 5th years

In the main manuscript we report that the percentage of patients with 1, 2, 3, 4, and 5 BMI measurements respectively (prior to death, transplant or censoring) was 18%, 21%, 17%, 14% and 30%. In Table 1 of the main manuscript we also report the specific number of patients with 1, 2, 3, 4, and 5 BMI measurements respectively. Here we provide some additional information about the frequency of the BMI measurements:

- Of the 12,449 patients who were still at risk of an event after 1 year of follow up, we found that 12,325 (99.0%) patients had a BMI measurement during their first year.

- Of the 9,196 patients who were still at risk of an event after 2 years of follow up, we found that 9,140 (99.4%) patients had a BMI measurement during their second year.

- Of the 6,588 patients who were still at risk of an event after 3 years of follow up, we found that 6,540 (99.3%) patients had a BMI measurement during their third year.

- Of the 4,511 patients who were still at risk of an event after 4 years of follow up, we found that 4,484 (99.4%) patients had a BMI measurement during their fourth year.

Of the 3,049 patients who were still at risk of an event after 5 years of follow up (i.e. those who were censored at the maximum follow up time), we found that 3,019 (99.0%) patients had a BMI measurement during their fifth year.

4. Computational details for the latent class joint model

4.1. Overview

The latent class joint model, defined in the main manuscript, was estimated using the 'Jointlcmm' command in the 'lcmm' R package [1–3]. Maximum likelihood estimates are obtained using a modified Marquardt algorithm (see the package documentation for details [2]). The package's default setting enforces strict convergence criteria, which we did not alter (tolerance of 1E-4 for each of: parameter stability, log-likelihood stability, and stability of the first derivatives). We chose to use the default square transformation of the baseline hazard parameters (argument 'logscale = FALSE') to ensure positivity of the baseline hazard throughout the estimation process.

Several possible extensions to the model were also explored, but often led to difficulties achieving convergence. Attempts were made to use a baseline hazard modelled with cubic M-splines (argument 'hazard = "splines"'), however, difficulties with convergence were encountered (even when using a small number of internal knots for the splines). Therefore, a simpler Weibull baseline hazard specification was used for the models in the main manuscript. We also encountered difficulties achieving convergence when the variance of the individual-level random effects was allowed to differ across latent classes.

4.2. Example code

In several supplementary files, we provide an example of the R code used to fit the latent class joint model. These files are named:

- "example_code_default.R"
- "example_code_gridsearch.R"
- "example_code_randominits.R"

The three files differ in their approach to starting / initial values for the parameters. The three approaches used for initial values are described in the next section.

4.3. Initial values

We found that we encountered difficulties with convergence unless reasonable initial values were used. In particular, our solution reported in the main manuscript was found using two strategies, described as follows.

Strategy A: example code shown in the supplementary file "example_code_default.R".

> A1. We first fit a model with the same structure as our intended latent class joint model, but with only one class (i.e. specifying the argument 'ng = 1' to the 'Jointlcmm' function in the lcmm R package). We denote this model: 'initmod'.
>
> A2. We then use the parameter estimates from the one class model ('initmod') to generate initial values for the M-class model we wish to fit (where M is the number of desired classes). This is achieved by specifying the arguments 'ng = M' and 'B = initmod' to the 'Jointlcmm' function. We use a maximum of 200 iterations. We denote this model: 'mod'.

A3. If the M-class model ('mod') converged, then we stop. Otherwise, if the M-class model ('mod') did not converge after 200 iterations then we do the following:

A3.1. We extract the final parameter estimates from the M-class joint model that did not converge.

A3.2. We then fit an M-class model for the longitudinal BMI data only (ignoring the death and transplant data) and we extract the final parameter estimates from that model.

A3.3. We substitute the estimates from Step A3.2 into the vector of parameter estimates from Step A3.1 and then refit the model using those parameter estimates as starting values for a new attempt at fitting the M-class joint model.

Strategy B: example code shown in the supplementary file "example_code_gridsearch.R". Note that Strategy B is the same as Strategy A, except that a grid search is used in the second step.

B1. Same as Step A1.

B2. We then use the parameter estimates from the one class model ('initmod') to randomly generate 3 sets of initial values for M-class model, we then run each of these 3 models for a maximum of 50 iterations. This is achieved by using the 'gridsearch' function in the lcmm package, with arguments 'rep = 3' and 'maxiter = 50'. We then take the model with the highest log-likelihood, and run that model until convergence or until the maximum number of iterations (200) is reached. Let us denote this model: 'mod'.

B3. Same as A3.

We then confirmed that Strategy A and Strategy B led to the same solution. Note that in most cases the model in Step A2/B2 did not converge, and the parameter estimates from an M-class model with just the longitudinal BMI data was required to produce starting values for the M-class joint model (i.e. Steps A3.1 through A3.3, or Steps B3.1 through B3.3, were required).

Additional Strategy: In response to a reviewer's suggestion, we then also ran the five-class joint model (i.e. our final model) with a variety of random initial values. Specifically, we used 27 randomly generated sets of initial values. The function to generate the initial values can be found in the supplementary file "example_code_randominits.R". From these 27 models, we found the following:

- 11 stopped abnormally (with no additional information provided by the lcmm package)
- 13 models did not converge after 200 iterations
- 3 models converged

Of the models that converged, one provided a seemingly nonsensical solution. The two other models converged to a common solution that was very similar, but not identical, to the solution found using Strategy A and Strategy B.

Compared with the final solution reported in the manuscript (found using Strategy A and Strategy B), the solution found with the random initial values had a slightly higher log-likelihood (-27324 vs -27434) and

slightly lower BIC value (55367 vs 55586). However, these differences are small relative to the decrease in BIC observed with adding additional latent classes; for example, the six-class solution reported in the manuscript had a BIC value of 54878. Moreover, the findings related to the predicted longitudinal trajectories and hazard functions (see Figure S3) were very similar to the results reported in the manuscript.

Since multimodality is such a critical issue in latent class models, one ideally wants to use many sets of randomly selected initial values. This can provide reassurance (but not certainty) that the final solution corresponds to a global maximum. Ideally, many sets of randomly selected initial values would also be used for the models estimated along the model building / selection / comparison process. However, we found that such an endeavour is somewhat hampered in latent class joint models by their computational complexity and - in particular – issues with convergence and long computation times (see the next section). Nonetheless, it is worth noting that the computation times would be shorter if fitting these models to a smaller dataset (for example less individuals and/or less longitudinal measurements), which would allow a greater number of models to be estimated in a given time.

4.4. Computation times

Fitting the latent class joint models was time consuming. For example, for the five-class joint model using Strategy A for specifying the initial values, it took approximately 30 hours to fit the final model using a single 2.5GHz core on the Monash University computing cluster. However, Steps A3.1 through A3.3 (see Strategy A in the previous section) of that process took only 3 hours, meaning that once reasonable starting values were used for the five class-specific longitudinal BMI trajectories the estimation time decreased dramatically.

5. Choice of individual-level random effects structure

We considered adding additional individual-level random effects to the model. Specifically, we considered including individual-level random effects for the coefficients of the cubic splines basis terms. However, this led to a model where the predicted BMI trajectory for each latent class was relatively stable/flat, with the latent classes primarily distinguished by different starting/average BMI values, or differences in the event rates, and not by differences in the *shapes* of the longitudinal BMI trajectories (see Figure S4). That is, in a model that included individual-level random effects for the cubic spline terms, variation in the shapes of the longitudinal BMI trajectories appeared to be attributed to between-individual (i.e. within-class) heterogeneity.

Accordingly, we chose to simplify the individual-level random effects structure in our model (that is, only include an individual-level random intercept). By doing this, we believe that differences in the shapes of the longitudinal profiles are exhibited primarily through between-class differences, and less absorbed by within-class (between-individual) variation. We believe that this approach more closely aligns with our study objectives, specifically to explore differences in the shapes of the class-specific longitudinal BMI trajectories, and how those are associated with differences in the class-specific rates of the competing events.

6. Results for the six-class joint model

Figure S5 shows the predicted BMI trajectories and cause-specific hazard functions, for each latent class, based on the six-class model. The covariate values used in the predictions are the same as used for the hazard functions in Figure 2 of the main manuscript. The difference here is that Figure S5 is for the six-class model, whereas Figure 2 in the main manuscript is for the five-class model.

As discussed in the main manuscript, we calculated Bayesian information criterion (BIC) values for models with a varying number of latent classes. The BIC values suggested that higher numbers of latent classes consistently resulted in a better fitting model. However, as the number of latent classes increased, the groups became less distinguished from one another (accompanied by a decrease in relative entropy, see Table 1 of the manuscript) and therefore less useful in terms of drawing meaningful conclusions from a clinical perspective. That is, based on a purely statistical criterion (the BIC), there is a suggestion that increasing numbers of classes are better, but from an interpretational perspective an ever-increasing number of latent classes was not useful. We therefore chose the five-class model as our final model.

It can be seen in Figure S5 that, for the six-class model, the main difference from the five-class model is that the "late BMI decline" class is split based on baseline BMI values and associated hazard rates for transplant. Since our primary interest is in the association between BMI and risk of death, we determined that increasing the number of latent classes to six was not warranted and we had similar conclusions from exploratory analysis with models having higher numbers of classes.

7. Cumulative incidence functions

The cause-specific hazard functions presented in Figure 2 of the main manuscript show the instantaneous rates (i.e. "hazard") of each event at time *t*, given that the individual is still at risk of the event. Alternatively, we can present cumulative incidence functions for each of the competing events; these are shown in Figures S6 (without 95% confidence limits) and S7 (with 95% confidence limits), for the same covariate profile as used for the hazard functions in Figure 2 of the main manuscript. The cumulative incidence functions show the cumulative risk (i.e. probability) of the event having occurred at any point up to time *t*.

Broadly speaking, the cause-specific hazard functions are useful for understanding the potential for etiological associations between the class-specific BMI trajectories and the occurrence of the competing events and thus more suited to the aims of our manuscript. On the other hand, the cumulative incidence functions are generally suited to understanding patient prognosis; for example, "what is the probability that a patient in latent class X will experience death within 5 years".

Importantly, the cumulative incidence function for one of the events, say death, depends on the hazard rates for both of the competing events. For example, whether a patient dies within 5 years depends partly on the rate of death, and partly on their rate of the competing event of transplant. Therefore, the association of a characteristic, say BMI, and the cumulative incidence function of an event, say death (without transplant), will depend on the associations between BMI and the hazard of both competing events, say death (without transplant) and transplant. Thus, in general, the association of a characteristic with the hazard of an event will be different to that with the cumulative incidence function of the same event, and in extreme cases could even be in opposite directions [4]. If the aims of our paper had been related to developing a model for patient prognosis, then we would have been interested in measures of

predictive performance for the fitted models. Moreover, issues such as non-proportional hazards would have been more relevant since they can have a significant impact on the prognostic performance of the fitted model.

For a more thorough discussion of the differences between cause-specific hazard functions and cumulative incidence functions, and how they each align with the intended aims of a study, we refer the reader to Koller et al. [5].

8. Goodness of fit of the final model (observed vs predicted longitudinal trajectories)

Figures S8 and S9 show observed and predicted, class-specific, longitudinal BMI trajectories. They are based on weighted means of the observed and predicted BMI values. The weighting refers to the estimated class membership probabilities for each individual, whilst the means are taken by splitting the distribution of observed measurement times into 15 quantiles (i.e. 15 "bins").

The plots show that the predicted mean longitudinal trajectories generally provide a good fit to the observed data. There is some discrepancy between the observed and marginal predicted BMI values beyond 2.5 years for the "rapid BMI decline" class (i.e. the black curve in Figure S8). However, this is probably due to the fact that this class has a relatively small number of patients overall, and has a high mortality rate early on in the follow up period. Therefore, the majority of the BMI measurements for this class are observed earlier in the follow up period. Because the bulk of the data is observed earlier in the follow up period the spline-based trajectory is seen to fit best in that region, whereas it has insufficient flexibility to capture the stabilising of the BMI curve after 2.5 years. Note however, that incorporating the subject-specific random intercept (i.e. the subject-specific predictions in Figure S9) resolves the discrepancy between the observed and predicted BMI values.

9. GRoLTS checklist

Table S2 shows a completed GRoLTS (Guidelines for Reporting on Latent Trajectory Studies) checklist for our study [6].

(To satisfy Item #15 of the GRoLTS checklist we have also included another file in our online Supplementary Materials entitled "table_full_model_estimates.txt". This plain text file includes an unformatted table of the entire list of parameter estimates from the final model.)

References

1. Proust-Lima C, Philipps V, Diakite A, *et al.* lcmm: Extended mixed models using latent classes and latent processes. R package version: 1.7.7. Published Online First: 2017.https://cran.r-project.org/package=lcmm

2. Proust-Lima C, Philipps V, Liquet B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software* 2017;**78**. doi:10.18637/jss.v078.i02

3. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Published Online First: 2017.https://www.R-project.org/

4. Andersen PK, Geskus RB, de Witte T, *et al.* Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology* 2012;**41**:861–70. doi:10.1093/ije/dyr213

5. Koller MT, Raatz H, Steyerberg EW, *et al.* Competing risks and the clinical community: irrelevance or ignorance? *Statistics in Medicine* 2012;**31**:1089–97. doi:10.1002/sim.4384

6. van de Schoot R, Sijbrandij M, Winter SD, *et al.* The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies. *Structural Equation Modeling: A Multidisciplinary Journal* 2017;**24**:451–67. doi:10.1080/10705511.2016.1247646

Table S1. Mean posterior probabilities of class membership, stratified by class membership (as determined by an individual's highest class-specific probability, and shown on the rows).

	Mean probabilities of class membership				
Class membership (% of patients)	Prob (Class A)	Prob (Class B)	Prob (Class C)	Prob (Class D)	Prob (Class E)
Class A (74.7%)	0.89	0.04	0.04	0.01	0.03
Class B (13.8%)	0.11	0.73	0.09	0.03	0.04
Class C (6.2%)	0.13	0.08	0.77	0.03	0.00
Class D (1.5%)	0.07	0.02	0.10	0.81	0.00
Class E (3.9%)	0.13	0.08	0.00	0.00	0.78

Table S2. GRoLTS checklist.

GRoLTS checklist item	Yes/No	Additional comments
1. Is the metric of time used in the statistical model reported?	Yes	Time metric is years.
2. Is information presented about the mean and variance of time within a wave?	NA	Exact time of the measurement (ANZDATA survey date - baseline date) is used in the analysis, so time-structured data is not relevant in this study.
3a. Is the missing data mechanism reported?	Yes	
3b. Is a description provided of what variables are related to attrition/missing data?	Yes	
3c. Is a description provided of how missing data in the analyses were dealt with?	Yes	Described in the Methods section in the manuscript, and supported by Figure S1 in the Supplementary Materials.
4. Is information about the distribution of the observed variables included?	Yes	Described in the model specification in the manuscript.
5. Is the software mentioned?	Yes	
6a. Are alternative specifications of within-class heterogeneity considered (e.g., LGCA vs. LGMM) and clearly documented? If not, was sufficient justification provided as to eliminate certain specifications from consideration?	Yes	We wish to allow for some between-individual variation within classes. Accordingly, we allow for this within-class heterogeneity through individual-level random effects, as described in the model specification.
6b. Are alternative specifications of the between-class differences in variance–covariance matrix structure considered and clearly documented? If not, was sufficient justification provided as to eliminate certain specifications from consideration?	Yes	Our supplementary materials describe our choice of structure for the individual-level random effects, as well as the predictions under a model with a more extensive individual-level random effects structure (including random effects for the spline terms).
7. Are alternative shape/functional forms of the trajectories described?	Yes	We believe cubic splines with 3 df provide sufficient flexibility to capture the underlying functional form of the longitudinal trajectories. In addition, the goodness of fit plots (observed vs predicted) suggest that this is the case.
8. If covariates have been used, can analyses still be replicated?	NA	The data for this study is not publically available.
9. Is information reported about the number of random start values and final iterations included?	Yes	In the Supplementary Materials.
10. Are the model comparison (and selection) tools described from a statistical perspective?	Yes	A subsection is contained in the Methods section of the manuscript.

11. Are the total number of fitted models reported, including a one-class solution?	Yes	Table 1 in the manuscript presents the different models that were considered. Note that the one-class solution is not appropriate for answering the research question in this study, since the one-class joint model corresponds to the assumption of no association between the longitudinal BMI trajectories and death/transplant event rates.
12. Are the number of cases per class reported for each model (absolute sample size, or proportion)?	Yes	Table 1 in the manuscript.
13. If classification of cases in a trajectory is the goal, is entropy reported?	Yes	Relative entropy is shown in Table 1. In addition, the mean posterior probabilities of class membership, stratified by class membership are presented in Table S1 in the Supplementary Materials.
14a. Is a plot included with the estimated mean trajectories of the final solution?	Yes	Figure 2 in the manuscript.
14b. Are plots included with the estimated mean trajectories for each model?	Yes	We provide plots of the six-class model, and an alternative model specification that includes additional individual-level random effects. It is infeasible to include plots of every model in our manuscript or supplementary.
14c. Is a plot included of the combination of estimated means of the final model and the observed individual trajectories split out for each latent class?	Yes	It is infeasible for us to plot all observed trajectories for the given sample size. But, we have provided plots of the mean predicted and mean observed BMI values across follow up; this answers a slightly different but nonetheless related question about goodness of fit.
15. Are characteristics of the final class solution numerically described (i.e., means, SD/SE, n, CI, etc.)?	Yes	Included in a separate .txt document in the supplementary materials.
16. Are the syntax files available (either in the appendix, supplementary materials, or from the authors)?	Yes	Example code is provided in the Supplementary Materials. The data is not publically available.

Figure S1. Flowchart showing numbers of patients excluded from the analysis.

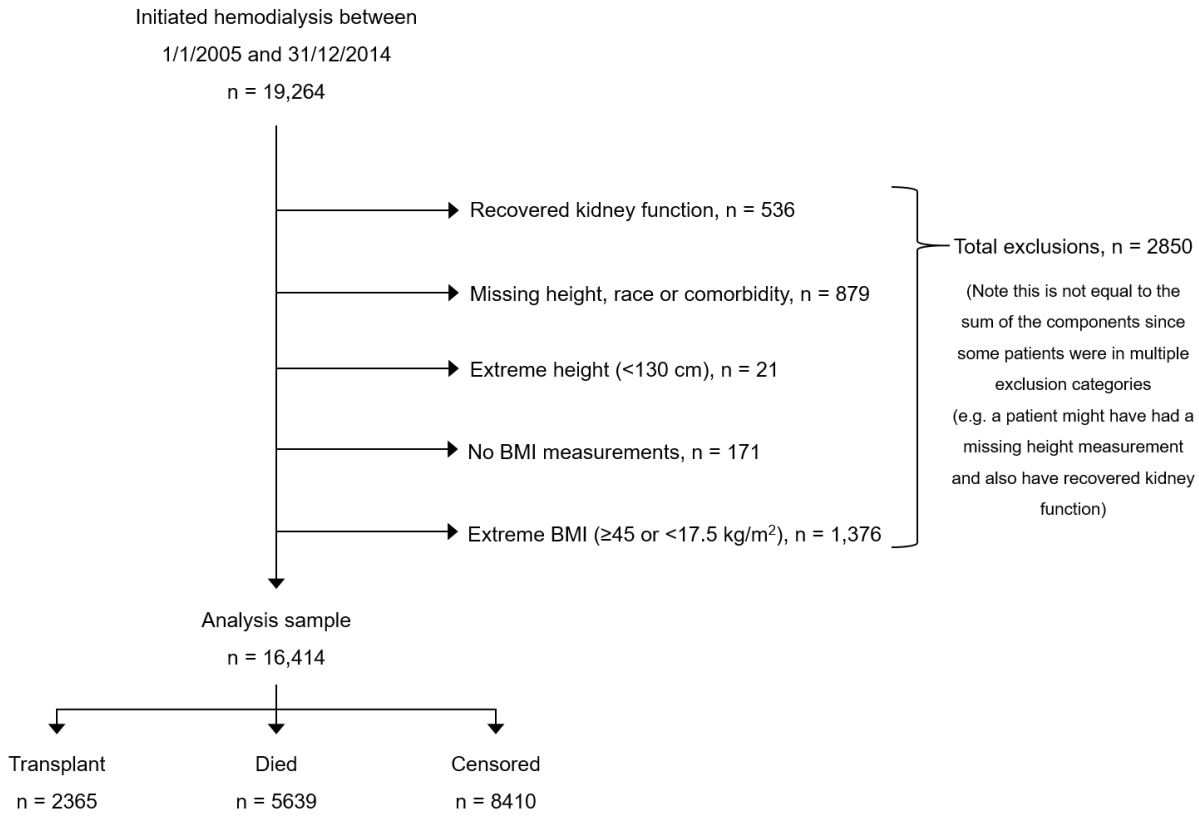


Figure S2. Observed BMI trajectories for a random sample of 25 patients.

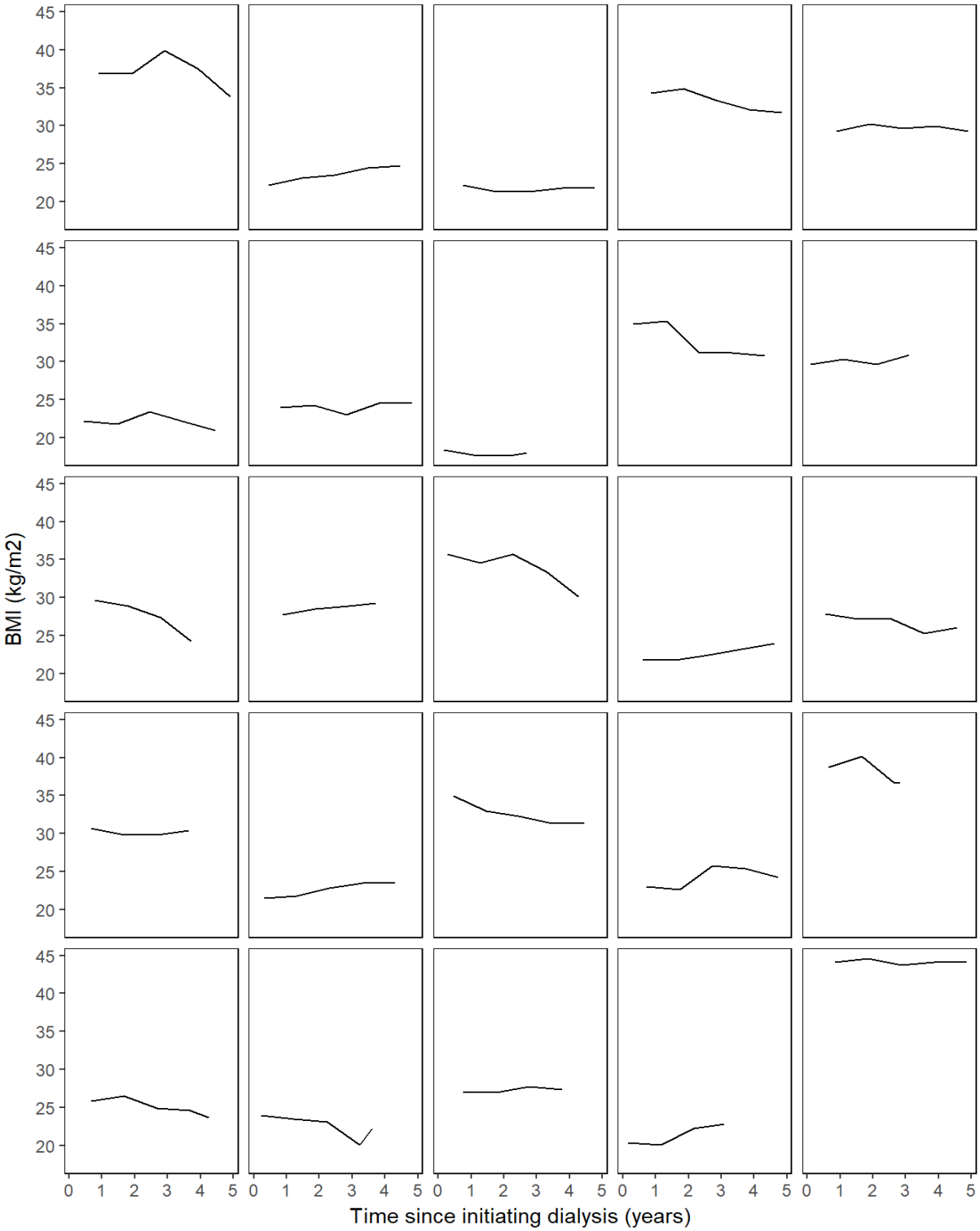


Figure S3. Predicted longitudinal BMI trajectories (left panel) and cause-specific hazard functions for death without transplant (middle panel) and transplant (right panel) for the alternative solution for the five-class model found using random initial values. The BMI predictions are on average (since no covariates were included in the BMI submodel), whilst the event outcome predictions are for a Caucasian male, aged ≤50 years, initiating RRT between 2005-09 with diabetic nephropathy, cerebrovascular disease and coronary artery disease.

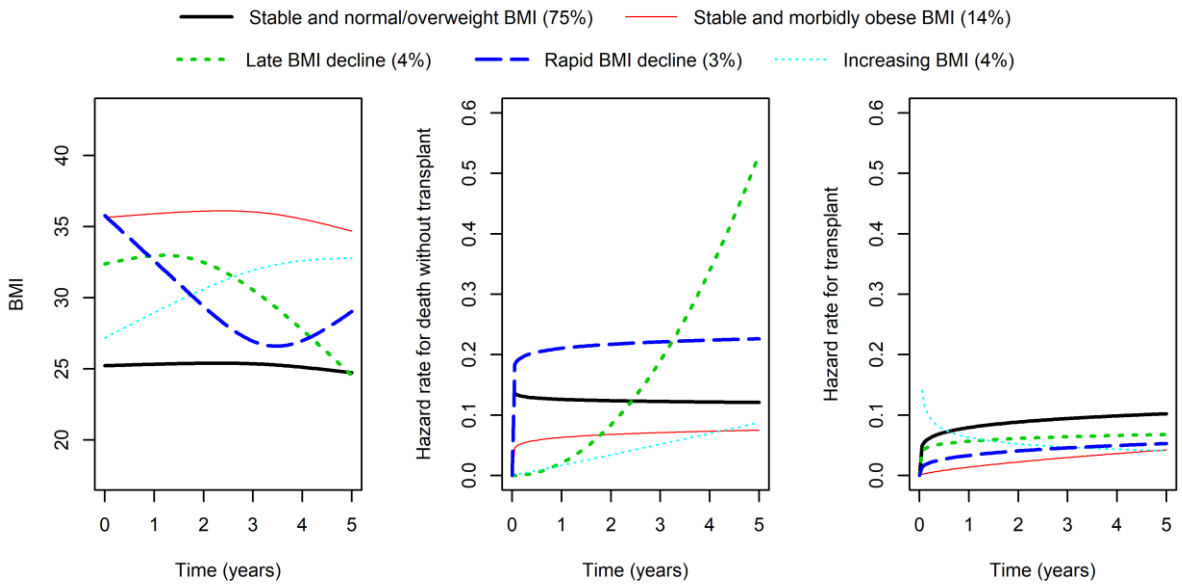


Figure S4. Predicted longitudinal BMI trajectories (left panel) and cause-specific hazard functions for death without transplant (middle panel) and transplant (right panel) for the five-class joint model after **including individual-level random effects for the cubic splines** (for the longitudinal BMI trajectories). The BMI predictions are on average (since no covariates were included in the BMI submodel), whilst the event outcome predictions are for a Caucasian male, aged ≤50 years, initiating RRT between 2005-09 with diabetic nephropathy, cerebrovascular disease and coronary artery disease.

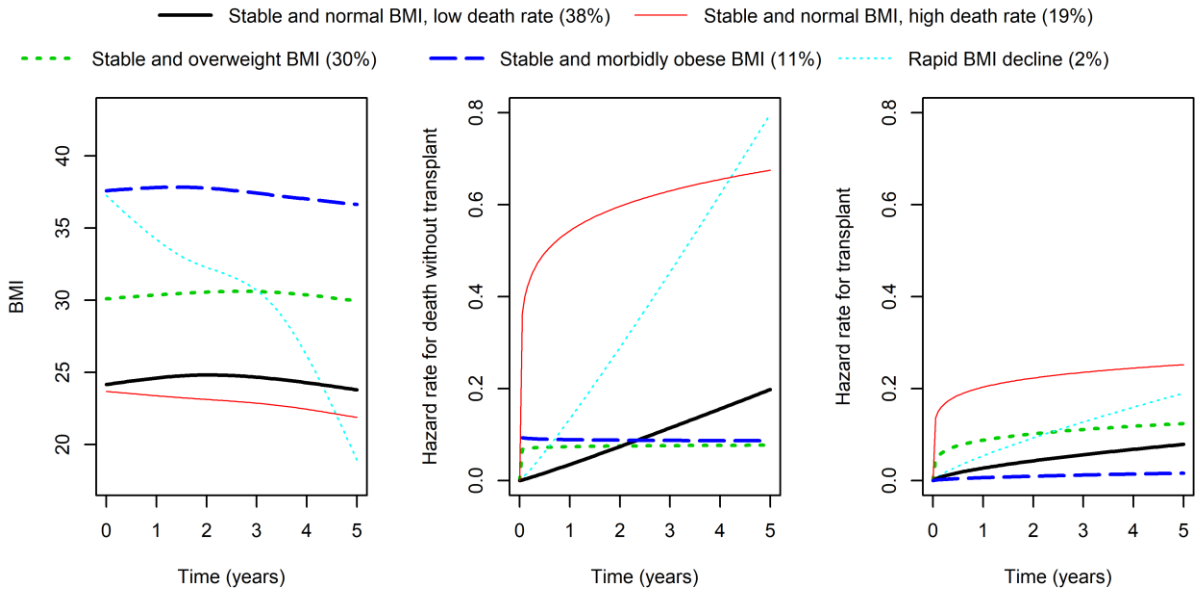


Figure S5. Predicted longitudinal BMI trajectories (left panel) and cause-specific hazard functions for death without transplant (middle panel) and transplant (right panel) from the **six-class** model. The BMI predictions are on average (since no covariates were included in the BMI submodel), whilst the event outcome predictions are for a Caucasian male, aged ≤50 years, initiating RRT between 2005-09 with diabetic nephropathy, cerebrovascular disease and coronary artery disease.

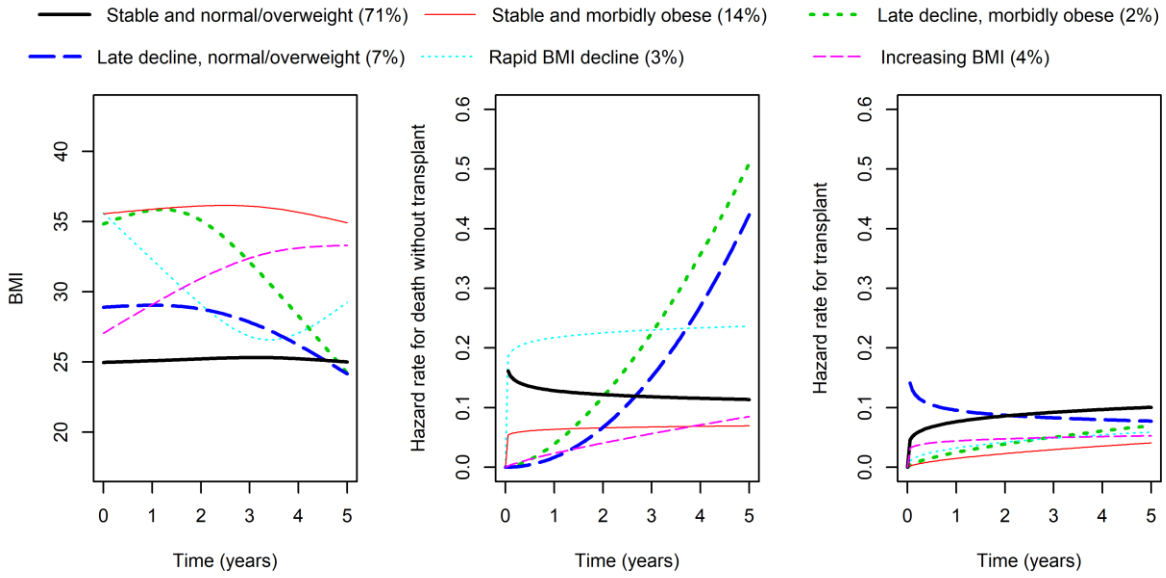


Figure S6. Predicted longitudinal BMI trajectories (left panel) and **cumulative incidence functions** for death without transplant (middle panel) and transplant (right panel) from the five-class model. The predictions are shown for each of the five possible latent classes. The BMI predictions are on average (since no covariates were included in the BMI submodel), whilst the event outcome predictions are for a Caucasian male, aged ≤50 years, initiating RRT between 2005-09 with diabetic nephropathy, cerebrovascular disease and coronary artery disease. These are the cumulative incidence functions for the same covariate profile as for the hazard functions shown in Figure 2 in the main manuscript.

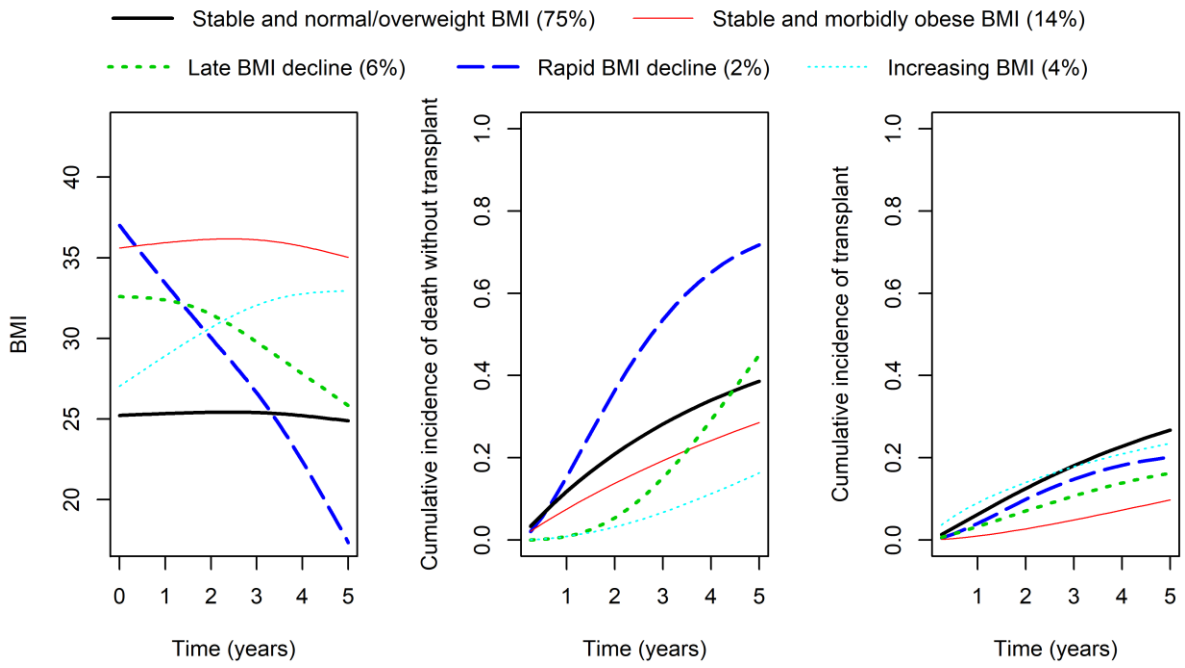


Figure S7. These are the same figures as described in Figure S6, but with 95% confidence limits included in the plots.

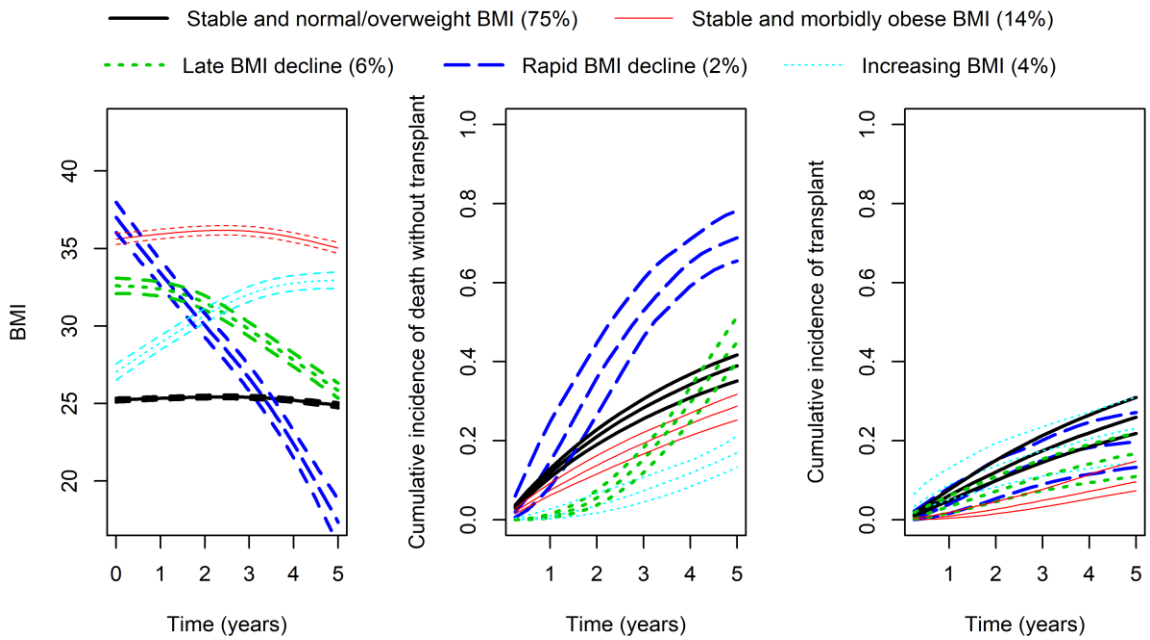


Figure S8. Observed and predicted BMI trajectories (**marginal** predictions). The plot shows class-specific BMI trajectories based on weighted means of the observed BMI data (with 95% confidence limits) and weighted means of the **marginal** predictions. The weighting is based on class membership probabilities, whilst the means are taken by splitting the distribution of observed measurement times into 15 quantiles (i.e. 15 “bins”).

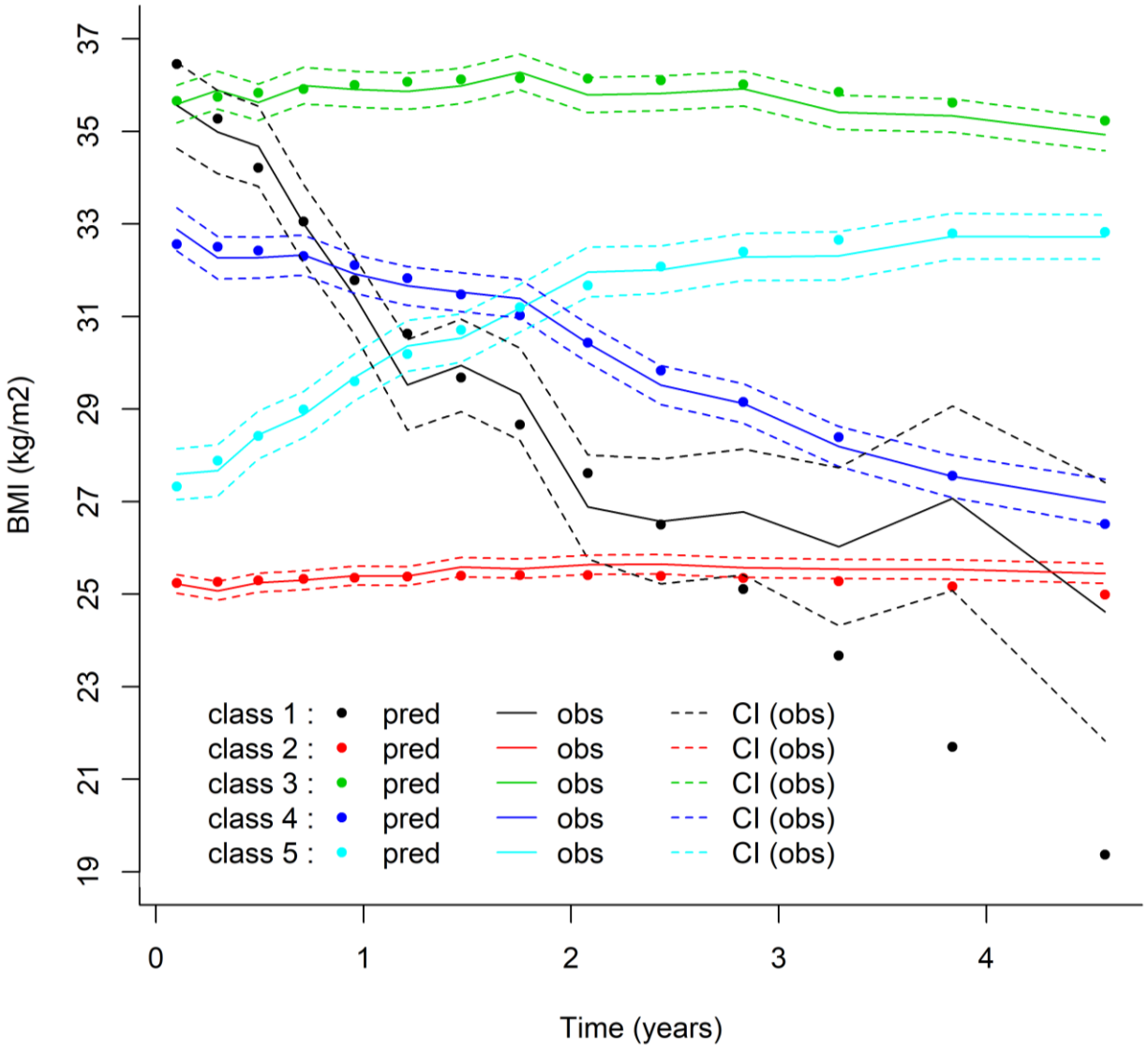
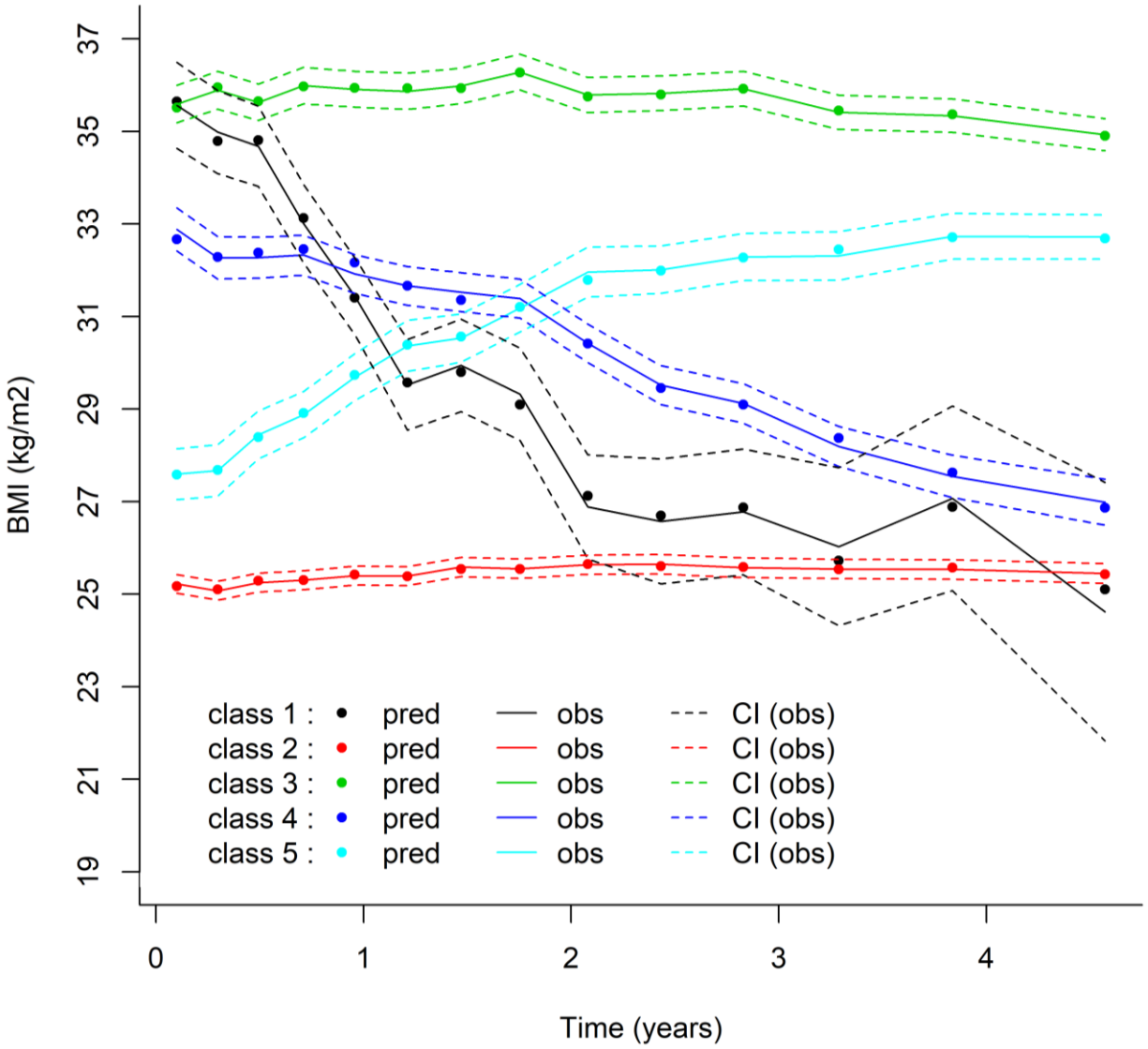


Figure S9. Observed and predicted BMI trajectories (**subject-specific** predictions). The plot shows class-specific BMI trajectories based on weighted means of the observed BMI data (with 95% confidence limits) and weighted means of the **subject-specific** predictions. The weighting is based on class membership probabilities, whilst the means are taken by splitting the distribution of observed measurement times into 15 quantiles (i.e. 15 "bins").



Appendix C. Supplementary materials for Chapter 6 paper

This appendix herein contains the supplementary materials for the following paper that was presented in Chapter 6:

Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J, Dunyak J, Al-Huniti N, Fox R, Hammerbacher J, Wolfe R. Joint longitudinal and time-to-event models for multilevel hierarchical data. *Submitted for publication*.

Supplementary material for: "Joint longitudinal and time-to-event models for multilevel hierarchical data"

Samuel L Brilleman, Michael J. Crowther, Margarita Moreno-Betancur, Jacqueline Buros Novik, James Dunyak, Nidal Al-Huniti, Robert Fox, Jeff Hammerbacher, Rory Wolfe

1. Further details on the model estimation

Our Bayesian specification requires prior distributions on all unknown parameters. We refer the reader to the documentation of the rstanarm R package for details on prior distributions, since our model was estimated using default priors implemented in the package. In brief, we used weakly informative normal distributions for each of the regression coefficients (fixed effects). The residual standard deviation (for the longitudinal outcome) was given a weakly informative half-Cauchy distribution. The B-spline coefficients for the log baseline hazard were given weakly informative Cauchy distributions. The weakly informative priors were only intended to reduce support given to values of the parameters that would seem implausible based on the scale of magnitude of the data. They were *not* intended to provide support to specific parameter values based on prior knowledge or expert opinion.

For estimation of the model parameters we ran four MCMC chains in parallel, each with 1000 sample iterations preceded by a warm up period of 1000 iterations (i.e. 2000 iterations in total, of which 50% were warm up). Although this number of iterations would seem small for a complex model estimated using a Gibbs sampler, the estimation in Stan is based on a Hamiltonian Monte Carlo (HMC) algorithm, not Gibbs sampling. The HMC results in much lower autocorrelation between subsequent MCMC draws compared with Gibbs sampling, and therefore is much more efficient in terms of the effective sample size per iteration. For example, for each of the models with an association structure based on the expected value, the effective sample size for the estimated association parameter was 4000.

A potential limitation of the proposed approach is the additional computational complexity. Additional clustering factors mean that there are an increasing number of cluster-specific parameters (i.e. random effects) to be estimated and therefore computation time increases. In our application with 430 patients having a total of 1209 lesions, there were 430 patient-specific parameters (intercept only) and 3627 lesion-specific parameters (intercept and two polynomial terms) to be estimated. Computation time for the models with an association structure based on the expected value ranged between 1.5 and 3 hours. The differences in computation time were partly related to the random nature of the different MCMC chains, and partly related to the type of summary function used in the association structure (i.e. the sum, average, maximum, or minimum of the level 2 clusters). The type of association structure is of course part of the model definition and therefore the choice of association structure will have an influence on the shape of the target posterior distribution, with some resulting posteriors easier for the MCMC sampler to explore (i.e. less extreme curvature in the posterior). When the association structure was based on both the expected value and the slope, the computation times were slightly longer; ranging between 2 and 5.5 hours. These times are based on 1000 warm up iterations, followed by 1000 sample iterations, on a standard quad-core desktop with a 3.30GHz processor and 8GB RAM.

2. Example code for fitting the model

The model in the paper can be easily estimated after downloading the rstanarm R package from the Comprehensive R Archive Network (CRAN). To download and install rstanarm, type the following into your R console:

```
install.packages("rstanarm")
```

And then an example of the code used to fit the model presented in Table 2 of the main manuscript would be:

```
library(rstanarm)
mod <- stan_jm(
  formulaLong = ldiam ~
    cat * poly(months, degree = 2) +
    (poly(months, degree = 2) | lesid_usubjid) +
    (1 | usubjid),
  dataLong = ipass$lesions,
  formulaEvent = Surv(progmnth, censor_p) ~ whostat,
  dataEvent = ipass$surv,
  seed = 9837355, time_var = "months", id_var = "usubjid",
  assoc = c("etavalue", "etaslope"), grp_assoc = "max")
```

Where ipass$lesions is a data frame containing the outcome and covariate data for the longitudinal submodel, and ipass$surv is a data frame containing the outcome and covariate data for the event submodel. Unfortunately, since the IPASS data used in the application in the main manuscript is not publically available, we cannot provide the reader with these data frames.

Appendix D. Vignettes for the simsurv R package

This appendix contains the vignettes for the **simsurv** R package (Brilleman, 2018b). There are two vignettes. Together, they provide more detailed information about the **simsurv** package, to accompany the material provided in Chapter 5 of the thesis.

The first vignette provides examples of usage of the simsurv package. In Chapter 5, the only example shown was simulating event times under a joint longitudinal and time-to-event model. However, in the vignette there are examples showing how to simulate event times from standard parametric survival distributions, two-component mixture distributions, and a survival model with non-proportional hazards.

The second vignette provides the technical background to the methods underpinning the **simsurv** package. This is similar to the details provided in Chapter 5, however, the vignette also includes explicit details on the parameterisations used for each of the distributions in **simsurv**.

How to use the simsurv package

Sam Brilleman

2018-03-09

Contents

Preamble

This vignette provides examples demonstrating usage of the **simsurv** package. For a technical vignette describing the methods underpinning the package, please see Technical background to the simsurv package.

Note that this package is modelled on the **survsim** Stata package. For comparability, the majority of the following examples are based on Crowther and Lambert (2012), which is the supporting paper for the **survsim** Stata package.

Usage examples

Example 1: Simulating under a standard parametric survival model

This first example shows how the **simsurv** package can be used to generate event times under a relatively standard Weibull proportional hazards model. This will be demonstrated as part of a simple simulation study.

The simulated event times will be generated under the following conditions:

- a monotonically increasing baseline hazard function, achieved by specifying a Weibull baseline hazard with a γ parameter of 1.5;
- the effect of a protective treatment obtained by specifying a binary covariate with log hazard ratio of -0.5;
- a maximum follow up time by censoring any individuals with a simulated survival time larger than five years.

The objective of the simulation study will be to assess the bias and coverage of the estimated treatment effect. This will be achieved by:

- generating 100 simulated datasets (ideally it should be more than 100 datasets, but we don't want the vignette to take forever to build!), each containing $N = 200$ individuals;
- fitting a Weibull proportional hazards model to each simulated dataset using the **flexsurv** package;
- calculating mean bias and mean coverage (of the estimated treatment effect) across the 100 simulated datasets.

The code for performing the simulation study and the results are shown below.

```
# Define a function for analysing one simulated dataset
sim_run <- function() {
  # Create a data frame with the subject IDs and treatment covariate
  cov <- data.frame(id = 1:200,
                    trt = rbinom(200, 1, 0.5))

  # Simulate the event times
  dat <- simsurv(lambdas = 0.1,
                 gammas = 1.5,
```

```
                betas = c(trt = -0.5),
                x = cov,
                maxt = 5)

  # Merge the simulated event times onto covariate data frame
  dat <- merge(cov, dat)

  # Fit a Weibull proportional hazards model
  mod <- flexsurv::flexsurvspline(Surv(eventtime, status) ~ trt, data = dat)

  # Obtain estimates, standard errors and 95% CI limits
  est <- mod$coefficients[["trt"]]
  ses <- sqrt(diag(mod$cov))[["trt"]]
  cil <- est + qnorm(.025) * ses
  ciu <- est + qnorm(.975) * ses

  # Return bias and coverage indicator for treatment effect
  c(bias = est - (-0.5),
    coverage = ((-0.5 > cil) && (-0.5 < ciu)))
}

# Set seed for simulations
set.seed(908070)

# Perform 100 replicates in simulation study
rowMeans(replicate(100, sim_run()))
```

```
##        bias    coverage
## 0.005858079 0.930000000
```

Here we see that there is very little bias in the estimates of the log hazard ratio for the treatment effect, and the 95% confidence intervals are near their intended level of coverage.

Example 2: Simulating under a flexible parametric survival model

Next, we will simulate event times under a slightly more complex parametric survival model that incorporates a flexible baseline hazard.

In this example we will use the publically accessible German breast cancer dataset. This dataset is included with the **simsurv** R package (see `help(simsurv::brcancer)` for a description of the dataset). Let us look at the first few rows of the dataset:

```
data("brcancer")
head(brcancer)
```

```
##   id hormon rectime censrec
## 1  1      0    1814       1
## 2  2      1    2018       1
## 3  3      1     712       1
## 4  4      1    1807       1
## 5  5      0     772       1
## 6  6      0     448       1
```

Now let us fit two parametric survival models to the breast cancer data:

- one Weibull survival model; and
- one flexible parametric survival model

The flexible parametric survival model will be based on the method of Royston and Parmar (2002); i.e. restricted cubic splines are used to approximate the log cumulative baseline hazard. This model can be estimated using the `flexsurvspline` function from the **flexsurv** package (Jackson (2016)).

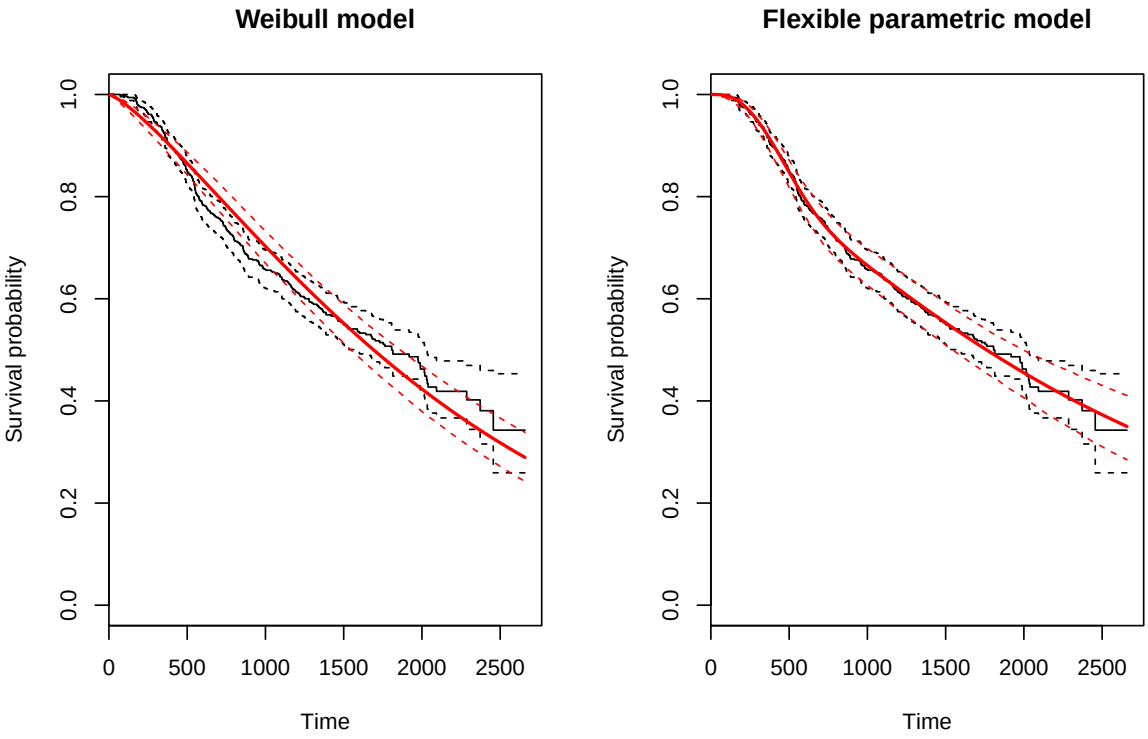
We will use three internal knots (i.e. four degrees of freedom) for the restricted cubic splines with the knot points placed at evenly spaced percentiles of the distribution of observed event time (obtained by specifying the argument `k = 3` in the code below). We can also estimate the Weibull proportional hazards model using the `flexsurvspline` function from the **flexsurv** package, by specifying no internal knots (i.e. specifying `k = 0`).

```
# Fit the Weibull survival model
mod_weib <- flexsurv::flexsurvspline(Surv(rectime, censrec) ~ hormon,
                                     data = brcancer, k = 0)

# Fit the flexible parametric survival model
mod_flex <- flexsurv::flexsurvspline(Surv(rectime, censrec) ~ hormon,
                                     data = brcancer, k = 3)
```

Now let us compare the fit of the two models by plotting each of the fitted survival functions on top of the Kaplan-Meier survival curve.

```
par(mfrow = c(1,2), cex = 0.85) # graphics parameters
plot(mod_weib,
     main = "Weibull model",
     ylab = "Survival probability",
     xlab = "Time")
plot(mod_flex,
     main = "Flexible parametric model",
     ylab = "Survival probability",
     xlab = "Time")
```



There is evidence in the plots that the flexible parametric model fits the data better than the standard Weibull model. Therefore, if we wanted to simulate event times from a data generating process similar to that of the breast cancer data, then using a Weibull distribution may not be adequate. Rather, it would be more appropriate to simulate event times under the flexible parametric model. We will demonstrate how the **simsurv** package can be used to do this. The estimated parameters from the flexible parametric model will be used as the "true" parameters for the simulated event times.

The event times can be generated under a user-specified log cumulative hazard function that is equivalent

to the Royston and Parmar specification used by the **flexsurv** package. First, the log cumulative hazard function for this model needs to be defined as a function in the R session. The **user-defined function passed to `simsurv` must always have the following three arguments:**

- `t`: scalar specifying the current time at which to evaluate the hazard
- `x`: a named list with the covariate data
- `betas`: a named list with the "true" parameters

Each of these arguments provide information that is used in evaluating the hazard $h_i(t)$, log hazard $\log h_i(t)$, cumulative hazard $H_i(t)$, or log cumulative hazard $\log H_i(t)$ (depending on which type of user-specified function is being provided). These three arguments (`t`, `x`, `betas`) can then be followed in the function signature by any additional arguments that may be necessary. For example, in the function definition below, the first three arguments are followed by an additional argument `knots`, which allows the calculation of the log cumulative hazard at time t to depend on the knot locations for the splines.

```
# Define a function returning the log cum hazard at time t
logcumhaz <- function(t, x, betas, knots) {

  # Obtain the basis terms for the spline-based log
  # cumulative hazard (evaluated at time t)
  basis <- flexsurv::basis(knots, log(t))

  # Evaluate the log cumulative hazard under the
  # Royston and Parmar specification
  res <-
    betas[["gamma0"]] * basis[[1]] +
    betas[["gamma1"]] * basis[[2]] +
    betas[["gamma2"]] * basis[[3]] +
    betas[["gamma3"]] * basis[[4]] +
    betas[["gamma4"]] * basis[[5]] +
    betas[["hormon"]] * x[["hormon"]]

  # Return the log cumulative hazard at time t
  res
}
```

Next, we will show how to use the `simsurv` function to simulate event times under the flexible parametric model. To demonstrate this, we will again generate the event times as part of a simulation study. The objective of the simulation study will be to assess the bias and coverage of the estimated log hazard ratio for hormone therapy. This will be achieved by:

- generating 100 simulated datasets (ideally it should be more than 100 datasets, but we don't want the vignette to take forever to build!), each containing $N = 200$ individuals. The simulated event times will be generated under our flexible parametric model (with the "true" parameter values taken from fitting a model to the German breast cancer data);
- fitting both a Weibull model and a flexible parameteric model to each simulated dataset;
- calculating the mean bias (across the 100 simulated datasets) in the log hazard ratio for hormone therapy under the Weibull model and the flexible parametric models.

```
# Fit the model to the brcancer dataset to obtain the "true"
# parameter values that will be used in our simulation study
true_mod <- flexsurv::flexsurvspline(Surv(rectime, censrec) ~ hormon,
                                     data = brcancer, k = 3)

# Define a function to generate one simulated dataset, fit
# our two models (Weibull and flexible) to the simulated data
# and then return the bias in the estimated effect of hormone
# therapy under each fitted model
sim_run <- function(true_mod) {
  # Create a data frame with the subject IDs and treatment covariate
  cov <- data.frame(id = 1:200, hormon = rbinom(200, 1, 0.5))
```

```r
  # Simulate the event times
  dat <- simsurv(betas = true_mod$coefficients, # "true" parameter values
                 x = cov,                       # covariate data for 200 individuals
                 knots = true_mod$knots,        # knot locations for splines
                 logcumhazard = logcumhaz,      # definition of log cum hazard
                 maxt = NULL,                   # no right-censoring
                 interval = c(1E-8,100000)) # interval for root finding

  # Merge the simulated event times onto covariate data frame
  dat <- merge(cov, dat)

  # Fit a Weibull proportional hazards model
  weib_mod <- flexsurv::flexsurvspline(Surv(eventtime, status) ~ hormon,
                                       data = dat, k = 0)

  # Fit a flexible parametric proportional hazards model
  flex_mod <- flexsurv::flexsurvspline(Surv(eventtime, status) ~ hormon,
                                       data = dat, k = 3)

  # Obtain estimates, standard errors and 95% CI limits for hormone effect
  true_loghr <- true_mod$coefficients[["hormon"]]
  weib_loghr <- weib_mod$coefficients[["hormon"]]
  flex_loghr <- flex_mod$coefficients[["hormon"]]

  # Return bias and coverage indicator for hormone effect
  c(weib_bias = weib_loghr - true_loghr,
    flex_bias = flex_loghr - true_loghr)
}

# Set a seed for the simulations
set.seed(543543)

# Perform the simulation study using 100 replicates
rowMeans(replicate(100, sim_run(true_mod = true_mod)))
```

```
##    weib_bias    flex_bias
## -0.013141112  0.007088905
```

Example 3: Simulating under a Weibull model with time-dependent effects

This short example shows how to simulate data under a standard Weibull survival model that incorporates a time-dependent effect (i.e. non-proportional hazards). For the time-dependent effect we will include a single binary covariate (e.g. a treatment indicator) with a protective effect (i.e. a negative log hazard ratio), but we will allow the effect of the covariate to diminish over time. The data generating model will be

$$h_i(t) = \gamma\lambda(t^{\gamma-1})\exp(\beta_0 X_i + \beta_1 X_i \times \log(t))$$

where X_i is the binary treatment indicator for individual i, λ and γ are the scale and shape parameters for the Weibull baseline hazard, β_0 is the log hazard ratio for treatment when $t = 1$ (i.e. when $\log(t) = 0$), and β_1 quantifies the amount by which the log hazard ratio for treatment changes for each one unit increase in $\log(t)$. Here we are assuming the time-dependent effect is induced by interacting the log hazard ratio with log time, but we could have used some other function of time (for example linear time, t, or time squared, t^2, if we had wanted to).

We will simulate data for $N = 5000$ individuals under this model, with a maximum follow up time of five years, and using the following "true" parameter values for the data generating model:

- $\beta_0 = -0.5$
- $\beta_1 = 0.15$
- $\lambda = 0.1$
- $\gamma = 1.5$

```
covs <- data.frame(id = 1:5000, trt = rbinom(5000, 1, 0.5))
simdat <- simsurv(dist = "weibull", lambdas = 0.1, gammas = 1.5, betas = c(trt = -0.5),
                  x = covs, tde = c(trt = 0.15), tdefunction = "log", maxt = 5)
simdat <- merge(simdat, covs)
head(simdat)
```

```
##   id eventtime status trt
## 1  1  1.547790      1   0
## 2  2  3.290966      1   1
## 3  3  3.002969      1   0
## 4  4  5.000000      0   0
## 5  5  4.227842      1   0
## 6  6  4.044081      1   0
```

Then let us fit a flexible parametric model with two internal knots (i.e. 3 degrees of freedom) for the baseline hazard, and a time-dependent hazard ratio for the treatment effect. For the time-dependent hazard ratio we will use an interaction with log time (the same as used in the data generating model); this can be easily achieved using the `stpm2` function from the **rstpm2** package (Clements and Liu (2017)) and specifying the `tvc` option. Note that the **rstpm2** package and **flexsurv** packages can both be used to fit the Royston and Parmar flexible parametric survival model, however, they differ slightly in their post-estimation functionality and other possible extensions. Here, we use the **rstpm2** package because it allows us to easily specify time-dependent effects and then plot the time-dependent hazard ratio after fitting the model (as shown in the code below).

The model with the time-dependent effect for treatment can be estimated using the following code

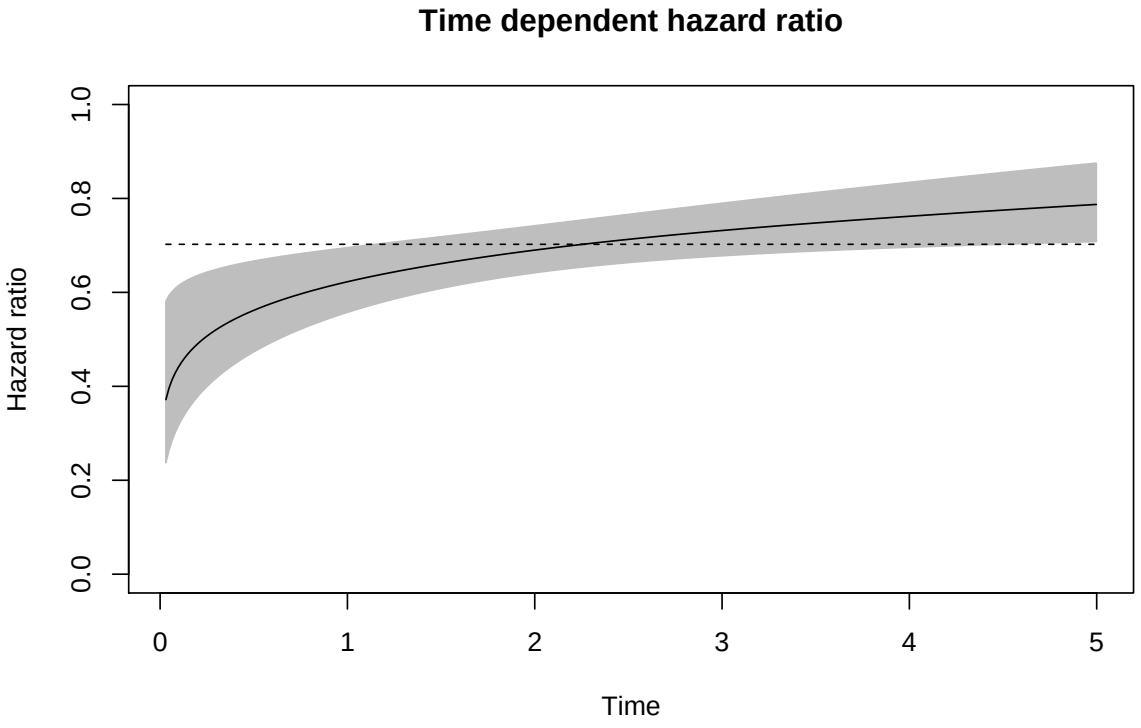
```
mod_tvc <- rstpm2::stpm2(Surv(eventtime, status) ~ trt,
                         data = simdat, tvc = list(trt = 1))
```

And for comparison we can fit the corresponding model, but without the time-dependent effect for treatment (i.e. assuming proportional hazards instead)

```
mod_ph <- rstpm2::stpm2(Surv(eventtime, status) ~ trt,
                        data = simdat)
```

Now, we can plot the time-dependent hazard ratio and the time-fixed hazard ratio on the same plot region using the following code

```
plot(mod_tvc, newdata = data.frame(trt = 0), type = "hr",
     var = "trt", ylim = c(0,1), ci = TRUE, rug = FALSE,
     main = "Time dependent hazard ratio",
     ylab = "Hazard ratio", xlab = "Time")
plot(mod_ph,  newdata = data.frame(trt = 0), type = "hr",
     var = "trt", ylim = c(0,1), add = TRUE, ci = FALSE, lty = 2)
```



From the plot we can see the diminishing effect of treatment under the model with the time-dependent hazard ratio; as time increases the hazard ratio approaches a value of 1. Moreover, note that the hazard ratio is approximately equal to a value of 0.6 (i.e. $\exp(-0.5)$) when $t = 1$, which is what we specified in the data generating model.

Example 4: Simulating under a joint model for longitudinal and survival data

This example shows how the **simsurv** package can be used to simulate event times under a shared parameter joint model for longitudinal and survival data.

We will simulate event times according to the following model formulation for the longitudinal submodel

$$Y_i(t) \sim N(\mu_i(t), \sigma_y^2)$$

$$\begin{aligned}
\mu_i(t) &= \beta_{0i} + \beta_{1i}t + \beta_2 x_{1i} + \beta_3 x_{2i} \\
\beta_{0i} &= \beta_{00} + b_{0i} \\
\beta_{1i} &= \beta_{10} + b_{1i} \\
(b_{0i}, b_{1i})^T &\sim N(0, \Sigma)
\end{aligned}$$

and the event submodel

$$h_i(t) = \delta(t^{\delta-1}) \exp(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \alpha\mu_i(t))$$

where x_{1i} is an indicator variable for a binary covariate, x_{2i} is a continuous covariate, b_{0i} and b_{1i} are individual-level parameters (i.e. random effects) for the intercept and slope for individual i, the β and γ terms are population-level parameters (i.e. fixed effects), and δ is the shape parameter for the Weibull baseline hazard.

This specification allows for an individual-specific linear trajectory for the longitudinal submodel, a Weibull baseline hazard in the event submodel, a current value association structure, and the effects of a binary and a continuous covariate in both the longitudinal and event submodels.

To simulate from this model using **simsurv**, we need to first explicitly define the hazard function. The code defining a function that returns the hazard for this joint model is

```
# First we define the hazard function to pass to simsurv
# (NB this is a Weibull proportional hazards regression submodel
# from a joint longitudinal and survival model with a "current
# value" association structure)
haz <- function(t, x, betas, ...) {
  betas[["delta"]] * (t ^ (betas[["delta"]] - 1)) * exp(
    betas[["gamma_0"]] +
    betas[["gamma_1"]] * x[["x1"]] +
    betas[["gamma_2"]] * x[["x2"]] +
    betas[["alpha"]] * (
      betas[["beta_0i"]] +
      betas[["beta_1i"]] * t +
      betas[["beta_2"]]  * x[["x1"]] +
      betas[["beta_3"]]  * x[["x2"]]
    )
  )
}
```

The next step is to define the true parameter values and covariate data for each individual. This is achieved by specifying two data frames: one for the parameter values, and one for the covariate data. Each row of the data frame will correspond to a different individual. The R code to achieve this is

```
# Then we construct data frames with the true parameter
# values and the covariate data for each individual
set.seed(5454) # set seed before simulating data
N <- 200       # number of individuals

# Population (fixed effect) parameters
betas <- data.frame(
  delta   = rep(2,    N),
  gamma_0 = rep(-11.9,N),
  gamma_1 = rep(0.6,  N),
  gamma_2 = rep(0.08, N),
  alpha   = rep(0.03, N),
  beta_0  = rep(90,   N),
  beta_1  = rep(2.5,  N),
  beta_2  = rep(-1.5, N),
  beta_3  = rep(1,    N)
)

# Individual-specific (random effect) parameters
b_corrmat <- matrix(c(1, 0.5, 0.5, 1), 2, 2)
b_sds     <- c(20, 3)
b_means   <- rep(0, 2)
b_z       <- MASS::mvrnorm(n = N, mu = b_means, Sigma = b_corrmat)
b         <- sapply(1:length(b_sds),
                    FUN = function(x) b_sds[x] * b_z[,x])
betas$beta_0i <- betas$beta_0 + b[,1]
betas$beta_1i <- betas$beta_1 + b[,2]

# Covariate data
covdat <- data.frame(
  x1 = stats::rbinom(N, 1, 0.45), # a binary covariate
```

```
  x2 = stats::rnorm(N, 44, 8.5)    # a continuous covariate
)
```

The final step is to then generate the simulated event times using a call to the `simsurv` function. The only arguments that need to be specified are the user-defined hazard function, the true parameter values, and the covariate data. In this example we will also specify a maximum follow up time of ten units (for example, ten years, after which individuals will be censored if they have not yet experienced the event).

The code to generate the simulated event times is

```
# Set seed for simulations
set.seed(546546)

# Then simulate the survival times based on the user-defined
# hazard function, covariates data, and true parameter values
times <- simsurv(hazard = haz, x = covdat, betas = betas, maxt = 10)
```

We can them examine the first few rows of the resulting data frame, to see the simulated event times and event indicator

```
head(times)
```

```
##   id eventtime status
## 1  1  7.676123      1
## 2  2 10.000000      0
## 3  3  1.142881      1
## 4  4  2.429828      1
## 5  5  6.901241      1
## 6  6  7.748474      1
```

```
## id eventtime status
##  1  4.813339      1
##  2  9.763900      1
##  3  5.913436      1
##  4  2.823562      1
##  5  2.315488      1
##  6 10.000000      0
```

Of course, we have only simulated the event times here; we haven't simulated any observed values for the longitudinal outcome. Moreover, although the **simsurv** package can be used for simulating joint longitudinal and time-to-event data, it did take a bit of work and several lines of code to achieve. Therefore, it is worth noting that the **simjm** package (https://github.com/sambrilleman/simjm), which acts as a wrapper for **simsurv**, is designed specifically for this purpose. It can make the process a lot easier, since it shields the user from much of the work described in this example. Instead, the user can simulate joint longitudinal and time-to-event data using one function call to `simjm::simjm` and a number of optional arguments are available to alter the exact specification of the shared parameter joint model.

References

Clements M, Liu X. (2017) rstpm2: Generalized Survival Models. R package version 1.4.1. https://CRAN.R-project.org/package=rstpm2

Crowther MJ, Lambert PC. Simulating complex survival data. *Stata J* 2012;**12**(4):674-687.

Jackson C. flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software* 2016;**70**(8):1-33. doi:10.18637/jss.v070.i08

Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* 2002;**21**(15):2175-2197. doi:10.1002/sim.1203

Technical background to the simsurv package

Sam Brilleman

2018-03-09

Contents

Preamble

This vignette herein describes the methodology used to simulate event times in the **simsurv** package. For a vignette related to usage of the package, including examples and code, please see How to use the simsurv package.

Introduction

The survival function for individual i is the probability that their true event time T_i^* is greater than the current time t. That is, the survival function can be defined as

$$S_i(t) = P(T_i^* > t)$$

Moreover, the corresponding probability of having failed at or before time t (i.e. having not survived up to time t) is the complement to the survival function. That is, the probability of failure is defined as

$$F_i(t) = P(T_i^* \leq t) = 1 - S_i(t)$$

If the survival time T_i^* is known to be drawn from some parametric distribution, then it also holds that the definition of the probability of failure, $F_i(t)$, is equivalent to the definition of the cumulative distribution function (CDF) for the distribution of event times. Moreover, probability distribution theory tells us that the CDF for a continuous random variable must follow a uniform distribution on the range 0 to 1 (Mood et al. (1973)). That is, $F_X(X) \sim U(0,1)$ where $F_X(.)$ denotes the CDF for the continuous random variable X. Similarly, the complement of the CDF for X must also follow a uniform distribution on the range 0 to 1, that is, $1 - F_X(X) \sim U(0,1)$.

These results therefore allow one to conclude that under a standard parametric distributional assumption for the event times T_i^* $(i = 1, \ldots, N)$, the survival probability for individual i at their true event time will be a uniform random variable on the range 0 to 1. That is,

$$S_i(T_i^*) = U_i \sim U(0,1)$$

It is then possible to extend these results to the setting of a proportional hazards model. Under a proportional hazards model the survival probability for individual i at their event time T_i^* can be written as

$$S_i(T_i^*) = \exp\left(-H_0(T_i^*)\exp(X_i^T\beta)\right) = U_i \sim U(0,1)$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the cumulative baseline hazard evaluated at time t, and X_i is a vector of covariates with associated population-level (i.e. fixed effect) parameters β.

Cumulative hazard inversion method (Bender et al. (2005))

Rearranging the equation for $S_i(t)$ and solving for t leads to the following general form for the inverted survival function

$$S_i^{-1}(u) = H_0^{-1}\left(-\log(u)\exp(-X_i^T\beta)\right)$$

where $S_i^{-1}(u)$ is the inverted survival function for individual i, $H_0^{-1}(u)$ corresponds to the inverted cumulative baseline hazard function, and X_i is a vector of covariates with associated population-level (i.e. fixed effect) parameters β.

Therefore, if the cumulative hazard function is invertible, we can easily simulate a new event time as

$$T_i^s = S_i^{-1}(U_i)$$

where T_i^s is the simulated event times for individual i, $S_i^{-1}(u)$ is the inverted survival function defined previously, and U_i is a random variable drawn from a $U(0,1)$ distribution. Note that if the cumulative baseline hazard is directly invertible then an analytic form will be available for $S_i^{-1}(u)$. That is, we can just plug in random draws of U_i and directly calcuate the simulated event times. Since independent draws of a $U(0,1)$ random variable are easily obtained using any standard statistical software, this method will be easy and fast for simulating event times.

This method was first proposed by Bender et al. (2005) and is commonly known as the *cumulative hazard inversion method*.

For the standard parameteric survival distributions included in the **simsurv** package (i.e. Weibull, exponential, Gompertz) an analytic form for $S_i^{-1}(u)$ does exist. Therefore, event times for these standard parametric distributions (assuming proportional hazards) are generated using the cumulative hazard inversion method. The parameterisations for each of these standard parametric distributions are shown next.

Exponential distribution

For the exponential distribution we have the following:

$$h_i(t) = \lambda\exp(X_i^T\beta)$$

$$H_i(t) = \lambda t\exp(X_i^T\beta)$$

$$S_i(t) = \exp\left(-\lambda t\exp(X_i^T\beta)\right)$$

$$S_i^{-1}(u) = \left(\frac{-\log(u)}{\lambda\exp(X_i^T\beta)}\right)$$

where $\lambda > 0$ is the rate parameter.

Weibull distribution

For the Weibull distribution we have the following:

$$h_i(t) = \gamma\lambda(t^{\gamma-1})\exp(X_i^T\beta)$$

$$H_i(t) = \lambda(t^{\gamma})\exp(X_i^T\beta)$$

$$S_i(t) = \exp\left(-\lambda(t^{\gamma})\exp(X_i^T\beta)\right)$$

$$S_i^{-1}(u) = \left(\frac{-\log(u)}{\lambda\exp(X_i^T\beta)}\right)^{1/\gamma}$$

where $\lambda > 0$ and $\gamma > 0$ are the scale and shape parameters, respectively.

Gompertz distribution

For the Gompertz distribution we have the following:

$$h_i(t) = \lambda\exp(\gamma t)\exp(X_i^T\beta)$$

$$H_i(t) = \frac{\lambda(\exp(\gamma t)-1)}{\gamma}\exp(X_i^T\beta)$$

$$S_i(t) = \exp\left(\frac{-\lambda(\exp(\gamma t)-1)}{\gamma}\exp(X_i^T\beta)\right)$$

$$S_i^{-1}(u) = \frac{1}{\gamma}\log\left[\left(\frac{-\gamma\log(u)}{\lambda\exp(X_i^T\beta)}\right)+1\right]$$

where $\lambda > 0$ and $\gamma > 0$ are the shape and scale parameters, respectively.

Numerical root finding

If the cumulative baseline hazard function is not invertible, then numerical root finding can be used to solve to t. This method has been discussed by both Bender et al. (2005) and Crowther and Lambert (2013). In **simsurv** this is required for the two-component mixture distributions (assuming proportional hazards). An analytical form is available for the survival function of each of these distributions, but numerical root finding must be used to invert the survival function. In practice, the **simsurv** package uses the `stats::uniroot` function based on the method of Brent (1973). This means iteratively finding a solution to the equation $S_i(t) - U_i = 0$.

The two-component mixture distributions in **simsurv** are parameterised in the same way as the **survsim** Stata package (Crowther and Lambert (2002)). That is, they are additive on the survival scale, with a parameter defining the mixing proportions, i.e.

$$S_0(t) = \pi S_{01}(t) + (1-\pi)S_{02}(t)$$

where $S_0(t)$ is the baseline survival function, $S_{01}(t)$ and $S_{02}(t)$ are baseline survival functions for the two component distributions, and $0 \leq \pi \leq 1$ is the mixing parameter. The specific parameterisations for the hazard and survival functions of each of the two-component mixture distributions in **simsurv** are shown next.

Exponential mixture distribution

For the two-component exponential mixture distribution we have the following:

$$h_i(t) = \left[\frac{\pi\lambda_1 \exp(-\lambda_1 t) + (1-\pi)\lambda_2 \exp(-\lambda_2 t)}{\pi \exp(-\lambda_1 t) + (1-\pi)\exp(-\lambda_2 t)}\right] \exp(X_i^T \beta)$$

$$H_i(t) = -\log\left[\pi \exp(-\lambda_1 t) + (1-\pi)\exp(-\lambda_2 t)\right] \exp(X_i^T \beta)$$

$$S_i(t) = \left[\pi \exp(-\lambda_1 t) + (1-\pi)\exp(-\lambda_2 t)\right]^{\exp(X_i^T \beta)}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the rate parameters for the component exponential distributions.

Weibull mixture distribution

For the two-component Weibull mixture distribution we have the following:

$$h_i(t) = \left[\frac{\pi\gamma_1\lambda_1(t^{\gamma_1-1})\exp(-\lambda_1(t^{\gamma_1})) + (1-\pi)\gamma_2\lambda_2(t^{\gamma_2-1})\exp(-\lambda_2(t^{\gamma_2}))}{\pi \exp(-\lambda_1(t^{\gamma_1})) + (1-\pi)\exp(-\lambda_2(t^{\gamma_2}))}\right] \exp(X_i^T \beta)$$

$$H_i(t) = -\log\left[\pi \exp(-\lambda_1(t^{\gamma_1})) + (1-\pi)\exp(-\lambda_2(t^{\gamma_2}))\right] \exp(X_i^T \beta)$$

$$S_i(t) = \left[\pi \exp(-\lambda_1(t^{\gamma_1})) + (1-\pi)\exp(-\lambda_2(t^{\gamma_2}))\right]^{\exp(X_i^T \beta)}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the scale parameters, and $\gamma_1 > 0$ and $\gamma_2 > 0$ are the shape parameters, for the component Weibull distributions.

Gompertz mixture distribution

$$h_i(t) = \left[\frac{\pi\lambda_1 \exp(\gamma_1 t)\exp\left(\frac{-\lambda_1(\exp(\gamma_1 t)-1)}{\gamma_1}\right) + (1-\pi)\lambda_2 \exp(\gamma_2 t)\exp\left(\frac{-\lambda_2(\exp(\gamma_2 t)-1)}{\gamma_2}\right)}{\pi \exp\left(\frac{-\lambda_1(\exp(\gamma_1 t)-1)}{\gamma_1}\right) + (1-\pi)\exp\left(\frac{-\lambda_2(\exp(\gamma_2 t)-1)}{\gamma_2}\right)}\right] \exp(X_i^T \beta)$$

$$H_i(t) = -\log\left[\pi \exp\left(\frac{-\lambda_1(\exp(\gamma_1 t)-1)}{\gamma_1}\right) + (1-\pi)\exp\left(\frac{-\lambda_2(\exp(\gamma_2 t)-1)}{\gamma_2}\right)\right] \exp(X_i^T \beta)$$

$$S_i(t) = \left[\pi \exp\left(\frac{-\lambda_1(\exp(\gamma_1 t)-1)}{\gamma_1}\right) + (1-\pi)\exp\left(\frac{-\lambda_2(\exp(\gamma_2 t)-1)}{\gamma_2}\right)\right]^{\exp(X_i^T \beta)}$$

Extending to time-dependent effects or user-defined hazard functions

If one can obtain an algebraic closed-form solution for the inverse cumulative baseline hazard, $H_0^{-1}(u)$, then a major benefit of the cumulative hazard inversion method is that it is simple and computationally efficient. Moreover, it can be used to generate survival times for a variety of standard parametric baseline hazards, for example, the exponential, Weibull or Gompertz distributions. Even if the cumulative baseline hazard cannot be inverted analytically then one can still use numerical root finding, as described in the previous section, to numerically invert the survival function and solve for t.

However, using numerical root finding still requires an analytical form for the survival function. For some complex data generating processes it may not be possible to obtain a closed-form solution to the cumulative baseline hazard $H_0(t)$ and therefore the form of $S_i(t)$ will also be intractable. Crowther and Lambert (2013) therefore proposed an extension to overcome these issues. Their extension incorporates

both numerical root finding and numerical quadrature. The numerical quadrature is used to evaluate the cumulative hazard in situations where it cannot be evaluated analytically.

An example of a situation where their method is required is the introduction of time-dependent effects on the hazard scale (i.e. non-proportional hazards). The introduction of time-dependent effects often leads to an intractable integral when evaluating the cumulative hazard. The method therefore involves iterating between numerical quadrature and numerical root finding until an appropriate solution for t is obtained. This is the method used by the **simsurv** package when the user supplies their own hazard or log hazard function for generating the event times, or when they are simulating with time-dependent effects (i.e. non-proportional hazards). The numerical integration is based on Gauss-Kronrod quadrature with a default of 15 nodes (although the user can choose between 7, 11, or 15 nodes). For further details on the method we refer the reader to the the paper by Crowther and Lambert (2013).

References

Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005;**24**(11):1713-1723. doi:10.1002/sim.2059

Brent R. *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Crowther MJ, Lambert PC. Simulating complex survival data. *Stata J* 2012;**12**(4):674-687.

Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013;**32**(23):4118-4134. doi:10.1002/sim.5823

Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics*. McGraw-Hill: New York, 1974.