THE UNIVERSITY *of* EDINBURGH

A scalable formulation of joint modelling for longitudinal and time to event data and its application on large electronic health record data of diabetes complications

Ioanna Thoma

THE UNIVERSITY *of* EDINBURGH

Thesis submitted for the degree of Doctor of Philosophy

College of Medicine and Veterinary Medicine

Edinburgh Medical School

University of Edinburgh

February 2023

Abstract

Introduction: Clinical decision-making in the management of diabetes and other chronic diseases depends upon individualised risk predictions of progression of the disease or complications of disease. With sequential measurements of biomarkers, it should be possible to make dynamic predictions that are updated as new data arrive. Since the 1990s, methods have been developed to jointly model longitudinal measurements of biomarkers and time-to-event data, aiming to facilitate predictions in various fields. These methods offer a comprehensive approach to analyse both the longitudinal changes in biomarkers, and the occurrence of events, allowing for a more integrated understanding of the underlying processes and improved predictive capabilities. The aim of this thesis is to investigate whether established methods for joint modelling are able to scale to large-scale electronic health record datasets with multiple biomarkers measured asynchronously, and evaluates the performance of a novel approach that overcomes the limitations of existing methods.

Methods: The epidemiological study design utilised in this research is a retrospective observational study. The data used for these analyses were obtained from a registry encompassing all individuals with type 1 diabetes in Scotland, which is delivered by the Scottish Care Information - Diabetes Collaboration platform. The two outcomes studied were time to cardiovascular disease (CVD) and time to end-stage renal disease (ESRD) from T1D diagnosis. The longitudinal biomarkers examined in the study were glycosylated haemoglobin (HbA_{1c}) and estimated glomerular filtration rate (eGFR). These biomarkers and endpoints were selected based on their prevalence in the T1D population and the established association between these biomarkers and the outcomes.

As a state-of-the-art method for joint modelling, Brilleman's `stan_jm()` function was evaluated. This is an implementation of a shared parameter joint model for longitudinal and time-to-event data in Stan contributed to the `rstanarm` package. This was compared with a novel approach based on sequential Bayesian updating of a continuous-time state-space model for the

biomarkers, with predictions generated by a Kalman filter algorithm using the `ctsem` package fed into a Poisson time-splitting regression model for the events. In contrast to the standard joint modelling approach that can only fit a linear mixed model to the biomarkers, the `ctsem` package is able to fit a broader family of models that include terms for autoregressive drift and diffusion. As a baseline for comparison, a last-observation-carried-forward model was evaluated to predict time-to-event.

Results: The analyses were conducted using renal replacement therapy outcome data regarding 29764 individuals and cardiovascular disease outcome data on 29479 individuals in Scotland (as per the 2019 national registry extract). The CVD dataset was reduced to 24779 individuals with both HbA_{1c} and eGFR data measured on the same date; a limitation of the modelling function itself. The datasets include 799 events of renal replacement therapy (RRT) or death due to renal failure (6.71 years average follow-up) and 2274 CVD events (7.54 years average follow-up) respectively. The standard approach to joint modelling using quadrature to integrate over the trajectories of the latent biomarker states, implemented in `rstanarm`, was found to be too slow to use even with moderate-sized datasets, e.g. 17.5 hours for a subset of 2633 subjects, 35.9 hours for 5265 subjects, and more than 68 hours for 10532 subjects. The sequential Bayesian updating approach was much faster, as it was able to analyse a dataset of 29121 individuals over 225598.3 person-years in 19 hours. Comparison of the fit of different longitudinal biomarker submodels showed that the fit of models that also included a drift and diffusion term was much better (AIC 51139 deviance units lower) than models that included only a linear mixed model slope term. Despite this, the improvement in predictive performance was slight for CVD (C-statistic 0.680 to 0.696 for 2112 individuals) and only moderate for end-stage renal disease (C-statistic 0.88 to 0.91 for 2000 individuals) by adding terms for diffusion and drift. The predictive performance of joint modelling in these datasets was only slightly better than using last-observation-carried-forward in the Poisson regression model (C-statistic 0.819 over 8625 person-years).

Conclusions: I have demonstrated that unlike the standard approach to joint modelling,

implemented in `rstanarm`, the time-splitting joint modelling approach based on sequential Bayesian updating can scale to a large dataset and allows biomarker trajectories to be modelled with a wider family of models that have better fit than simple linear mixed models. However, in this application, where the only biomarkers were HbA_{1c} and eGFR, and the outcomes were time-to-CVD and end-stage renal disease, the increment in the predictive performance of joint modelling compared with last-observation-carried forward was slight. For other outcomes, where the ability to predict time-to-event depends upon modelling latent biomarker trajectories rather than just using the last-observation-carried-forward, the advantages of joint modelling may be greater.

This thesis proceeds as follows. The first two chapters serve as an introduction to the joint modelling of longitudinal and time-to-event data and its relation to other methods for clinical risk prediction. Briefly, this part explores the rationale for utilising such an approach to manage chronic diseases, such as T1D, better. The methodological chapters of this thesis describe the mathematical formulation of a multivariate shared-parameter joint model and introduce its application and performance on a subset of individuals with T1D and data pertaining to CVD and ESRD outcomes.

Additionally, the mathematical formulation of an alternative time-splitting approach is demonstrated and compared to a conventional method for estimating longitudinal trajectories of clinical biomarkers used in risk prediction. Also, the key features of the pipeline required to implement this approach are outlined. The final chapters of the thesis present an applied example that demonstrates the estimation and evaluation of the alternative modelling approach and explores the types of inferences that can be obtained for a subset of individuals with T1D that might progress to ESRD. Finally, this thesis highlights the strengths and weaknesses of applying and scaling up more complex modelling approaches to facilitate dynamic risk prediction for precision medicine.

Lay summary

In medicine, it is very often important to be able to predict who is going to get a disease in the future, or more particularly, who has a high risk of developing complications shortly after diagnosis. In order to do this, we can use data from people without the condition who go on to develop the condition and data from people who stay free of the condition. If we have measurements made on such individuals at a single point in time, or even better, at many points in time, this can help us to improve our ability to predict who is at highest risk and tailor interventions to those who need them most. Moreover, these data points are likely to be observed at different points in time for different people, and there also are differing numbers of repeated measurements, which complicates modelling.

For the task of clinical risk prediction, we use measurements called biomarkers, that we might measure on a regular check-up, such as blood pressure, how much cholesterol is in the blood, or how much sugar is in the urine, for example. When utilising observations from individuals for prediction purposes, the first step requires the development of a risk prediction model. That involves bringing all the data together in intelligent (statistical) ways to maximise our ability to predict. The statistical methods for bringing a single time point of measurement per individual together into a prediction model are well established. What could be better established are the methods for binding multiple measures together into a statistical model. Yet this is of increasing importance for scientists to be able to do, especially now that with the advent of electronic healthcare records, we often have many points of data captured routinely available just from the regular clinical follow-up of patients.

The statistical methods for analysing repeated measurements of biomarkers from individuals in order to predict future risk of an event is an active area of research and comprises the topic of my thesis. It concerns taking an established method that has been applied to exploit multiple biomarker measurements for the prediction of time-to-event, called joint modelling and trying to apply it to a very large dataset of biomarker data from people with type one

diabetes so as to predict each one's risk of cardiovascular disease and kidney disease, as a consequence of their diabetes.

In this thesis, I first attempted to use a joint modelling approach to predict the risk of developing cardiovascular disease. However, I found that it takes far too long to construct such models on a computer. Days and days of intense calculations on a powerful computer were required even for a cut-down dataset. Because of that, I went on with my approach to apply more novel Bayesian methods that split the follow-up time in a new way and annotate the follow-up time with the data from the patient and then conduct a statistical computation so as to maximise prediction. This approach is enabled by recently emerged software allowing us to exploit advances in Bayesian computations.

I decided in this part of the thesis to attempt to apply this new method to the problem of predicting who has a high risk of kidney failure among those with type one diabetes residing in Scotland. I applied this method, compared it to the standard joint model, and found that the quality of the prediction obtained was similar. However, the increment in time-to-event prediction was slight; or at least, it was not enough to help in clinical decision-making.

However, in doing this, I did refine and hone the techniques for it, and I was able to produce new code that others will be able to use for different problems that might be inherently easier to predict than cardiovascular disease or kidney disease in diabetes. This required me to think through the maths involved and underlying these new statistical methodologies and organise the data in a way that allows us to use them.

In summary, the present thesis discusses the benefits of modelling longitudinal measurements while acknowledging the complexities involved in handling such data. The work that I have produced in my thesis will be helpful to others, saving them time by avoiding attempting to laboriously apply joint modelling where it simply will not perform satisfactorily and pointing them to newer methods that are much more likely to be useful for their risk prediction endeavours.

Acknowledgments

A number of people have been instrumental in the successful completion of this thesis. First and foremost, I am grateful to my primary and secondary supervisors, Professor Helen Colhoun and Dr Athina Spiliopoulou, who allowed me to develop my knowledge and never let me stray from my path. Furthermore, a lot of gratitude is due to Professor Paul McKeigue, a passionate mentor with a great understanding of his students and admirable work ethic, who supported me throughout my PhD ride.

Research is a decisively collaborative endeavour, in which intellectual exchange is front and centre. I would like to acknowledge each and every single member of Diabetes Medical Informatics & Epidemiology Group, all of whom have assisted me shape and form, in some way, namely Stuart McGurnaghan, Luke Blackbourn, Andrii Iakovliev, Thomas Caparrota, Joe Mellor, Sara Hatam, Svenja Moser, Katharina Diernberger. I would like to extend my gratitude to a friend and key collaborator on this project, Dr Marco Colombo, for his generous support and invaluable advice throughout this journey.

Furthermore, I would like to thank the Precision Medicine Doctoral Training Programme team and Medical Research Council for funding me and making this research project possible. Furthermore, I would like to acknowledge my thesis committee members who participated in my annual reviews, Professor Caroline Hayward and Professor Lynne Regan for their valuable feedback and careful evaluation.

Of course, I would like to thank my family and explicitly my dear sister, Pepi, and all of my friends, both inside and outside the University, for reminding me what matters most at the end of the day. Thank you all for the kind disposition I have experienced during the past few years.

Lastly, I would like to spend my greatest thanks on my partner, Orfeas Stefanos Thyfronitis Litos, who has simply never failed me. Deeply grateful for your all-round support and

affection.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Ioanna Thoma

27/02/2023

Abbreviations

Term	Abbreviation
Akaike Information Criterion	AIC
Albumin to Creatinine Ratio	ACR
Area Under the ROC Curve	AUC
Bayesian Applied Regression Modeling via Stan	rstanarm
Cardiovascular Disease	CVD
Concordance Statistic	C-statistic
Continuous Time Structural Equation Modelling	ctsem
Diabetic Kidney Disease	DKD
Electronic Healthcare Records	EHR
End-Stage Renal Disease	ESRD
Glycated Haemoglobin	HbA_{1c}
Hazard Ratio	HR
Last Observation Carried Forward	LOCF
Receiver Operating Characteristic	ROC
Renal Replacement Therapy	RRT
Type 1 Diabetes	T1D
Type 2 Diabetes	T2D
estimated Glomerular Filtration Rate	eGFR

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This doctoral thesis scrutinises the potential of an R implementation of joint modelling for longitudinal data of biomarkers and clinical outcomes. Additionally, the thesis assesses whether joint modelling approaches can be used at scale to harness the longitudinal information in biomarker trajectories and improve time-to-event prediction.

The motivations for undertaking a joint modelling approach to estimate time-to-event determined by the longitudinal nature of predictive biomarker data include the following two reasons:

- We are interested in how underlying changes in a biomarker (outcome A) influence the occurrence of an event (outcome B).
- Joint models are naturally suited to the task of dynamic risk prediction. We wish to build models where predictions of event risk can be updated as new biomarker measurements become available.

Risk prediction models with time-varying biomarker data are usually developed based on the last available biomarker observation. To make use of today's computational potential and ameliorate public health, researchers develop and optimise machine learning and statistical

methods designed to exploit hidden information in large collections of medical data and to learn meaningful patterns from them. The joint modelling approach has been proposed as a means to incorporate longitudinal data into the survival function estimation since computational advancements at 1990s. However, researchers are currently limited in their ability to fit these models routinely. Despite methodological improvements, a coherent and flexible modelling framework that encapsulates an scalable effective multivariate joint model is lacking.

The remainder of Introduction elaborates on various concepts mentioned in the following chapters, such as what diabetes is for the non-clinical reader, the concept of precision medicine, the context of survival analysis and of joint modelling used for prediction of clinical events, as well as today's role of biomarker data in risk prediction.

1.1 Overarching framework

A common problem in epidemiology and clinical medicine is the construction of prediction models for future clinical events. These prediction models are typically used to differentiate individuals at high risk of future events, so preventive therapies can be introduced. This idea of tailoring intervention to those who are at the highest risk is one of the key concepts of precision medicine, also known as tailored medicine.

Clinical prediction models have traditionally used characteristics measured at a single time to predict future event status over some variable follow-up period. These models are generally known as survival models and typically comprise Cox and Poisson regression.

The advent of electronic health care records (EHRs) means that there are often large datasets of clinical measurements in multiple time points for individuals who develop the disease in the upcoming period. These measurements are potentially predictive characteristics related to the outcome. In this context, these likely predictive characteristics are referred to as prognostic biomarkers of a condition that might be of interest.

However, traditional methods for developing risk models may not be well adapted to handling irregularly collected clinical covariates (Goldstein et al. 2017). The advantage of using multiple longitudinal measures in risk prediction models as a means to incorporate more clinical information is established. Nevertheless, numerous attempts at modelling future events in the absence of new data involve extrapolations based on single time points, such as carrying the last observation forward. Inference based on single time points can be less accurate as opposed to building longitudinal models that capture the underlying biological behaviours (Molnar, Hutton, and Fergusson 2008). Moreover, longitudinal models can also be used to examine the biological mechanisms behind the pathogenesis of diseases. Albeit interesting on its own right, this is not a direction that has been explored herein.

Since the 1990s, there have been several remarkable attempts that exploit more fully the longitudinal and serial nature of predictive biomarkers, introducing a class of models known as joint modelling of longitudinal and time to event data (Wulfsohn and Tsiatis 1997; Tsiatis and Davidian 2004; Brilleman et al. 2018). In modern electronic healthcare records contain a great variety of longitudinal biomarker data that can be harnessed to predict and possibly prevent future adverse events. This idea is central to the concept of precision medicine, which attempts to tailor interventions to those individuals who need them most. Regrettably, the collected data lack a standard format that can be conveniently used by statistical programs and relevant prediction algorithms, making their use non straightforward and their processing heavy. As a result, there have been only a few successful attempts in applying joint modelling in large datasets of prognostic biomarkers for predicting clinical adverse events (Asar et al. 2015; Long and Mills 2018).

1.2 Epidemiological and clinical contexts

Diabetes mellitus (diabetes (Greek): diabaínō, pass through/siphon; mellitus (Latin): honey sweet) is a heterogeneous, multifactorial metabolic disease, characterised by chronic hyper-

glycemia (Lammert et al. 2014) and has two *main* types: type 1 and type 2. Type 1 diabetes (T1D), also known as autoimmune diabetes, is characterised by insulin deficiency due to pancreatic β-cell loss (Katsarou et al. 2017). Type 2 diabetes (T2D) is associated with lifestyle, primarily diet and lack of exercise, and occurs when insulin production does not suffice for exceptionally high needs, as a consequence of insulin resistance. The onset of T1D is usually earlier in life compared with T2D, with 50% of individuals being diagnosed in childhood.

Diabetes is one of the most prevalent diseases of the 21^{st} century and a major determinant of additional complications. Long-term complications of T1D, including nephropathy, retinopathy, neuropathy and vascular disease can be life-threatening and may greatly compromise the quality of life. Diabetic kidney disease (DKD) remains a leading cause of early mortality in people with T1D, while the risk of developing CVD continues to increase threefold relative to the general population (Schofield, Ho, and Soran 2019).

However, susceptibility to diabetes complications varies significantly between individuals. Although the pathogenesis of diabetes complications is complex, a number of factors that increase the risk for development have been identified. However, existing risk factors of complications only explain to some extent the inter-individual variation in risk and patterns of complications (Deshpande, Harris-Hayes, and Schootman 2008). Some patterns reflect variability in environmental factors, genetics or both. It is well known that the chronic complications of diabetes are all strongly associated with hyperglycaemia. Furthermore, the degree of hyperglycaemia may change over time, depending on the extent of the underlying disease process (Association and others 2006).

A cure for T1D is not available, and patients depend on lifelong insulin injections. Novel approaches to insulin treatment, such as insulin pumps, continuous glucose monitoring and hybrid closed-loop systems, are in development. If diabetes is well controlled, the risk of complications is reduced. However, despite this fact, the majority of T1D patients still

experience significant microvascular and macrovascular complications.

Being able to differentiate between more and less susceptible individuals is critical for clinical practice and the furtherance of understanding diabetes. Data linkage with electronic healthcare records has enabled the study of inter- and intra-individual variation with much finer granularity. Dynamic disease risk predictions at an individual level can be achieved by exploiting personalised clinical profiles, particularly through the increasing amount of available longitudinal biomarker data and computational developments in the realm of survival modelling.

Therefore, the statistical survival analysis described in the following sections has been approached from the angle of modelling longitudinal and time to event data *simultaneously* to assess whether we can obtain more accurate estimates of the risk of progression to a complication. The advent of statistical software that handles time-updated data, gives rise to more precise methods to estimate the underlying hazard of event.

1.2.1 Precision medicine aspects

The primary goal of precision medicine is to build a solid foundation for identifying differential risk amongst groups so that they can be treated accordingly. Tools that stratify patients conditional on risk can enable clinicians to perform better evaluations regarding the rate of progression to a disease, which in turn offers more appropriate treatment allocation.

The precision medicine initiative was first seen in cancer therapy, particularly for managing certain cancer types, such as breast and ovarian cancers. It started with the realisation that the harm caused by complex anti-cancer drugs could be reduced if the drugs were designed to target only a particular protein in the developing tumour, should the protein be expressed. However, the genomic heterogeneity in tumour development limited this application. In some cases, tumours had started to mutate and develop primary resistance against the targeted molecules, discouraging researchers and funding bodies (Tannock and Hickman 2016).

Moreover, the area of pharmacogenomics, which aims to determine how genetic variations might influence individual responses to medications, can significantly benefit precision medicine. Genetic testing for guiding treatment allocation is becoming increasingly available across diverse areas of medical care. These tests could assign more effective drugs to patients earlier in their treatment. However, this individualistic approach, i.e., addressing drugs to each patient separately, involves substantial cost, which poses further limitations to research programs (Krzyszczyk et al. 2018).

Additional initiatives in precision medicine include growing replacement tissue, molecular profiling of microbes and personalised diets. These approaches highlight the broader scope of precision medicine beyond cancer research. Furthermore, more recently, precision medicine has been relevant in addressing interventions for COVID-19. As we learn more about COVID-19, we can consider more targeted ways of preventing and treating infections (Zhou et al. 2021) and protecting the most sensitive groups.

A lesson already learned from the response to the pandemic is that systematic resilience, primarily at an individual level and secondary at healthcare infrastructure might sabotage personalised interventions. The overall response to COVID-19 shows that the incentive has primarily been to treat the public as if everyone is at the same risk rather than switching to a more tailored and efficient intervention once data started to accumulate.

Arguably, the precision medicine experimentation, especially in cancer research and drug development, has contributed to improving treatment effects, resulting in considerable patient benefits, albeit its implementation is still slow and complex. These unfavourable facts should not undermine the potential positive impact of precision medicine in enhancing treatment outcomes and patient well-being. This evidence highlights the necessity to formalise the mathematical basis for personalised medicine applications. Harnessing the growing availability of data is pivotal to this goal. Having approaches ready to use when data and statistical methods start to scale to high dimensional data would enable swift adoption and considerably

ameliorate risk prediction of clinical outcomes.

In conclusion, personalised medicine, albeit a promising initiative, is years away from reaching its full potential. Although it holds a great promise, there are still challenges and complexities that are not quick to resolve. Personalised medicine might become more feasible and and its use is normalised, once there has been a revolution in the way we assess the quality of large volumes of information, and in how we collect and define measures used in modelling techniques that allow real-time interventions.

1.2.2 Types of biomarker data

The term biomarker has been employed for many years in biomedical research, referring to any observation that could be used as an *indication* of an underlying physiological state.

The growing availability of sequential measurements for patients and further advances in computational statistics have provided powerful methods for quantifying the underlying, unknown biological processes conditional on a number of risk factors. However, the quality of predictions depends on the quality of the biomarker data and the underlying relationship between biomarkers and outcomes.

Increased glucose levels is a marker for the development of diabetes, for example, and a rise in prostate-specific antigen (PSA) indicates risk of developing prostate cancer (Shortliffe et al. 2014). Furthermore, changes in microRNA levels in the blood and other body fluids (miRNAs) have been linked to a diverse set of diseases (Condrat et al. 2020), including type 1 and type 2 diabetes, pre-diabetes, insulin resistance, obesity and metabolic diseases. Modern methods for analysing molecules have resulted in a vast expansion of the list of known biomarkers, offering greater visibility to the inner workings of various physiological mechanisms.

Biomarkers are categorised according to their types and characteristics (Food, Administration, and others 2020). The *Biomarkers, EndpointS and other Tools* (BEST) *Glossary* defines seven biomarker categories: susceptibility/risk, diagnostic, monitoring, prognostic, predictive,

pharmacodynamic/response, and safety. This thesis focuses on the use of prognostic and predictive biomarkers to assess whether they contribute to determining the correct time to event in survival analysis.

1.3 Quality of prediction

Longitudinal studies are commonplace in medical research, where the interest often lies in the interrelationships between variables. Predictive modelling of time to event aims at developing tools that can be used for individual prediction of the probability of an event's occurrence (or recurrence). The accuracy of such guess depends on what is known about the subjects and their environments. The growing availability of data from electronic healthcare records consistently contributes to decipher how those risk factors and observations affect time to event and risk propensity.

Predictive models should ideally use all relevant and available data. However, various model assumptions may not be realistic long-term, causing the model to deviate from original assumptions and projections. The data quality and the biomarker's prognostic power highly determine the model's predictive ability. However, there is a limit on how many predictor variables can be included in the modelling phase (Kuijk et al. 2019). A clinical model with too many predictors is more likely to be overfitted on the data used, i.e., the model performs adequately on the train data, but performs poorly on new datasets. Therefore, using appropriate methods to avoid issues like overfitting and multicollinearity is paramount (Frost and others 2017).

From a public health perspective, prediction models may contribute to targeting preventive interventions towards predisposed individuals. Furthermore, they may stimulate the development of tailored therapeutic approaches when the reversal of complications is still feasible. For the outcomes studied, targeted interventions may be beneficial in slowing down the progression of diabetic kidney disease (DKD) or preventing the recurrence of CVD events

in the future.

In clinical practice, predictive models facilitate informed diagnoses and decision-making about treatment. Arguably, greater exploitation of biomarker data that are routinely collected over the course of the disease has revolutionised diagnosis and timely intervention, but it has so far made limited contributions to improving the prediction of major clinical outcomes.

1.4 Using a joint modelling approach for risk prediction

Survival analysis specifies the time to occurrence of a clinical outcome of interest, such as death or the development of a condition, while also accounting for serial measurements of one or more biomarkers pertinent to the outcome, or are surrogate variables of some potential interest.

Routinely measuring clinical and biomarker data in individuals who receive medical care for chronic disease is standard practice. Such longitudinal profiles may provide valuable insights into underlying susceptibility, according to which informed decisions can be taken, such as interventions that may delay the onset of complications.

However, coupling longitudinal and survival observations is a demanding and complex task, especially when modelling of multiple biomarkers is involved. These challenges are mostly met when biomarker measurements are sporadic and different measures are recorded asynchronously.

In addition, it is important to account for the study design when designing a modelling approach. Information sourced from EHRs deviate from longitudinal and time to event data, typically met in randomised clinical trials (RCTs) and traditional cohort studies in frequency and quality. RCTs are more likely to have specific time points of follow-up for all subjects, while the data used in the given analyses were measured at irregular time points and were lacking a standard format, as opposed to less noisy data stemming from more conventional

observational studies. This is an argument that supports employing joint modelling that better handles scattered observations.

1.4.1 Modelling prerequisites

The most widely used method for modelling longitudinal data is the linear mixed-effects model (LMM). LMMs are well-suited for repeated observations that exhibit correlation. However, there might be limited understanding of the inter-relationships amongst the various components in the model. As discussed later in the thesis, the LMM is a special case of a broader family of continuous-time structural equation modelling (Hoyle 2012), that can be extended to provide a better fit to the longitudinal data.

Such models are applicable in settings where individuals are followed up for a long period of time to monitor disease progression or development. This progression is typically assessed via repeated measurements of biomarkers pertinent to the medical condition. For the management of chronic diseases, clinicians need to be able to make dynamic risk predictions that are updated *as new biomarker observations arrive*. This is a key requirement for precision medicine to realise its potential.

Moreover, time to event data require special treatment because the outcome of interest is composite; whether or not an event occurred and also *when* that event occurred. By the end of an observational study, some individuals will have yet to experience the event of interest. Therefore, their true time to event is still in the process of being determined. In survival analysis or time to event analysis, outcomes that are still uncertain are treated as censored observations. Typically, censoring is observed when the event of interest has not occurred for certain individuals within the study period, or they are lost to follow-up. Censoring affects the estimation of event probabilities for the entire population, hence it must be appropriately accounted for to avoid biased calculations.

The central assumption in survival analysis is that of non-informative censoring: individuals

who are censored for some reason are assumed to have the same probability of experiencing the event as individuals who remain in the study.

However, the occurrence of time to event may induce informative censoring, as discussed by Wu and Carroll (1988), Schluchter (1992), and other authors. For example, when they entered the study, individuals with more severe renal disease may experience the event much earlier than healthier subjects. They might also have more biomarker data during recruitment for the same reason. Therefore, sharper declines of the biomarkers and other relevant patterns, such as fluctuations, must be as explicitly modelled as possible to avoid biased estimations of quantities of interest.

In more detail, in routine health data, various risk factors and biomarkers are typically measured at irregular times, with the frequency of the measurements depending on multiple reasons. The way that information is missing is on its own informative. For example, sicker patients may be assessed more often. One of the inherent problems in EHRs is *non-ignorable missingness* (Sperrin, Petherick, and Badrick 2017).

Missingness, irregularly spaced observations (measurements taken on different dates for different biomarkers), informative presence (what the presence of a particular observation says about health status), delayed entry, selection bias, complex correlation structures, mixtures of time-varying and time-invariant covariates are all instances of elaborate data generation processes. These factors should either be addressed during the study design or appropriately accounted for using appropriate statistical modelling techniques.

As such, there is a partial likelihood which cannot be maximised to predict time to event, as we would need all covariate data at all failure times to do so, and in practice, data are only available intermittently over follow-up for each subject. To handle missing data, usually a type of imputation is employed, e.g., the last-observation-carried-forward (LOCF) technique (Carpenter and Kenward 2007; Prentice 1982). However, naive extrapolations could result in biased estimation of model parameters (Lachin 2016). Therefore, a more sophisticated

framework in which the features of censoring and missing data may be incorporated is required and presented next.

As per Tsiatis and Davidian (2004), models for the possibly error-prone longitudinal process and the hazard of the possibly censored time to event can be defined to depend jointly on shared, underlying random effects.

A familiar example is that of HIV clinical trials (Tsiatis, Degruttola, and Wulfsohn 1995), where covariates, including treatment assignment, demographic information, and physiological characteristics are recorded at baseline, and measures of immunologic and virologic status, such as CD4 count and viral RNA copy number are taken at subsequent clinic visits. Time to progression to AIDS or death is also recorded for each participant, although some subjects may withdraw early from the study or fail to experience the event by the time of study closure.

An implication has been that the joint modelling construction allows the biomarker trend to vary with time and induces a within-subject autocorrelation structure that may be thought of as arising from evolving biological fluctuations in the process.

Two-staged joint models were proposed by Wulfsohn and Tsiatis (1997), who utilised numerical integration (via low-dimensional Gauss-Hermite quadrature) as part of an expectation-maximisation algorithm to circumvent the dependence on the unobserved random effects on the parameter estimates by treating them as missing data.

Using a random slope and random intercept usually is the simplest form to specify the true longitudinal process. In the study conducted by Wulfsohn and Tsiatis (1997), a linear mixed-effects model performed well enough, hence it became the de facto approach in the early years of joint models. However, when the number of longitudinal components, the complexity of the random effects structure, or both, grows, this approach becomes less tractable due to the inherent computational burden, which has led to alternative fitting procedures being utilised: for instance, Monte Carlo techniques (Henderson et al., 2000; Lin et al., 2002; Hickey et al., 2018a) and Laplace approximations (Rizopoulos et al., 2009). However, as Hickey et

al. (2018a) draw attention to, the increasing volumes of data collected by clinical trials with many longitudinal responses and increasingly complex electronic healthcare records would likely require approximate methods for the numerical integration in the estimation step.

Furthermore, outcome data are typically skewed rather than normally distributed. They may comprise several early events and relatively few late events, and vice versa, or may form a U-shaped curve depending on the health condition. Hence, a time to event model must account for three potential outcomes (survival, failure, censoring) instead of two (survival, failure). Censored survival times can lead to an underestimation of the true, albeit unknown, time to event.

These features of longitudinal and survival data have made the suite of modelling techniques, known as joint models for longitudinal and time to event data emerge, where the event time distribution and the longitudinal data are taken to depend on a common set of latent random effects. In the literature, precise statement of the underlying assumptions typically made for these models has been rare.

What is the relationship between the features of the longitudinal biomarker and the time to progression?

Conventionally, the hazard depends linearly on the history of the longitudinal biomarker up to the current time point thus, it depends linearly on the current value, but other forms of relationship/specifications are possible with joint modelling.

Formalising these objectives is straightforward in terms of the 'idealised' data, but addressing them in practice is complicated by the nature of the data actually collected, e.g., the observed values are not the true values, but instead, we might observe errors and randomness.

Today's opportunity to exploit a stream of individual-level data stemming from various continuous monitoring methods, such as wearable electronic devices and EHRs can be an excellent application for capturing the underlying dynamics of biological mechanisms and

better estimate risk. Joint modelling of longitudinal and time to event data has been fundamental to the application of precision medicine for several decades now (Wulfsohn and Tsiatis 1997; Henderson, Diggle, and Dobson 2000; Tsiatis and Davidian 2004; Rizopoulos 2012; Gould et al. 2015). Precision medicine must exploit time-updated and correlated biomarker data to maximise time to event prediction. However, the predictive performance of this statistical method has been shown to heavily depend on the application and the type of prediction.

In joint modelling of longitudinal and time to event data, an observation up to each time point is handled explicitly. Furthermore, the timings of the observations are taken into account rather than just the observations themselves. Hence, the joint model offers a way to simultaneously characterise the relationship between the longitudinal process and time to event for each subject. However, for more than one longitudinal process (multivariate joint models) and large datasets, the increased dimensionality and complexity of random effects translate into increased computing time, hampering the implementation of many classical approaches. As a solution, Bayesian methods can be used to address these challenges and help with model fitting.

Researchers have been attempting to develop techniques to calculate the probability of transitioning to a disease based on observations of patient-specific parameters since the 1990s. Various Bayesian programs have been developed since then, many of which are accurate in selecting competing explanations of disease states. However, although probabilistic models are of great promise in explaining data structures in medical records, they are restrictive when dealing with more complex datasets, and existing Bayesian implementations of continuous-time models suffer from rather high computing times.

1.5 Thesis overview

1.5.1 Thesis objectives

A fundamental prerequisite for precision medicine is to develop performant models of highly predictive power for the risk of adverse outcomes and the effectiveness of interventions.

The thesis objectives concern the development and evaluation of the performance of modelling techniques for longitudinal data that go beyond the conventional LOCF approach, utilising a large modern dataset of longitudinal biomarkers and event data.

The thesis explores the limitations of a developed methods for joint modelling, and evaluates a scalable construction that exploits in a timely manner large electronic healthcare records of multiple biomarker and event data measured intermittently.

My analysis's primary objective has been to understand better the methods and approaches that can maximally exploit the information hidden in the available data. Furthermore, conducting this research helps understand deeper the variability and uncertainty in risk prediction and its association with translational studies such as precision medicine's applications.

The models described herein were assessed in terms of increment of risk prediction of developing complications of diabetes, developed on a sizeable nationwide dataset of electronic linkage records obtained in Scotland.

To that regard, model implementations that are more likely to identify longitudinal patterns in complex biomedical data have been explored. Furthermore, the evaluation of the predictive performance of the constructed models has been done in two ways:

- by generating out-of-sample predictions, i.e., estimating probabilities of events of a population different from the one used to train the model (a.k.a. previously unseen individuals) and
- by obtaining forward predictions for individuals included in the training sample up to a

landmark time.

The latter is implemented by censoring part of the analysis dataset after a landmark time point. Typically, complications do not occur close to the date of diagnosis but are expected after a particular age and diabetes duration (ten years or more). Therefore forward prediction is better suited to an observational setting since previously unseen individuals could comprise a population with different covariate profiles (measured or unmeasured), and it might include people who have been recently added to the registry with newly diagnosed diabetes.

1.5.2 Structure of this thesis

This thesis is structured as follows:

Chapter 1 introduces the rationale for precision medicine in T1D. High variation at risk of developing complications among people with T1D and today's great opportunity to obtain insights from massive data collections are discussed here. Major challenges and barriers to the implementation of a suitable framework that integrates observational data and allows individualised predictions are also described. A literature review that closely relates to the researched topics has been conducted and comprises part of the introduction, providing relevant sources of evidence to support methodology and findings.

Chapter 2 gives an epidemiological background first related to the biological risk of developing T1D complications, and continues with a methodological background, discussing approaches to time to event analysis, joint model formulations, and the Bayesian framework used for making inferences and calculating event probabilities. Finally, the chapter ends with an overview of current implementations and endeavours to manipulate this type of data, highlighting the very present trade-off between statistical and computational efficiency in large-scale applications such as this one.

Chapter 3 describes the data sources used for conducting the analyses.

Chapter 4 describes the state-of-the-art joint modelling functionality as implemented in `rstanarm`. It introduces the application of joint models in quantifying the risk of developing cardiovascular disease (CVD) among people with T1D. Assumptions, model formulation, the range of possible association structures, model performance and diagnostics are discussed in order to test different fitting settings. The analysis concludes that the joint model fitted with the modelling function `stanjm()` is quite restrictive and computationally expensive, albeit its elegant mathematical formulation.

In light of the aforementioned outcome, I followed an alternative approach based on Bayesian sequential updating, described in chapter 5, where I broke down the total follow-up time into shorter consecutive intervals of time. I call this formulation the time-splitting joint model. The rationale for this splitting is that since the rate of risk is time-varying, the probability of an event may be estimated more efficiently as a stochastic process. Structural equation modelling, as implemented in the `ctsem` package, allows for a broader family of hierarchical models that include autoregressive drift (gradual smooth changes) and diffusion (sudden perturbations) to be used to fit the biomarker data. A joint model comprising a Poisson regression model dependent upon time-updated covariates that have been specified more dynamically is likely to infer time to event more accurately than conventional LOCF models.

The functionality of the `ctsem` package and its application to the diabetes dataset is demonstrated in chapter 6, with the objective of modelling the composite outcome of time to renal replacement therapy (RRT) and death ascribed to renal failure among people with T1D. A two-stage approach based on Kalman filter updates and a time-split Poisson counting process is evaluated as a scalable alternative to `rstanarm`'s joint model implementation.

In this exemplar, the observed events are modelled as a counting process over many short person-time intervals and the biomarker values at the start of each interval are imputed by forward updating, *using only observations up to the start of each interval*, from a class of models known as hierarchical continuous-time dynamic models. As a baseline for comparison,

a last-observation-carried forward model was evaluated for the prediction of time to event.

Chapters 7, 8 and 9 describe the analysis pipeline that produced the thesis results, present various findings for the ESRD analyses and elaborate on the evaluation of the predictive performance and calibration of the different models tested.

Chapter 10 contains all final remarks and concluding summary, in addition to contributions towards the fields of biostatistics and precision medicine. Moreover, insights regarding future steps are briefly considered.

Moving forward, chapter 2 elaborates further on the background, on which I review the statistical theory and I briefly mention some details about the aetiology of diabetes complications and how these matter. Since the focus of this thesis has been on statistical approaches for improving risk prediction, providing an extensive review on the epidemiology of employing biomarkers in clinical research and diabetes complications falls out of the scope of a methodological thesis. Therefore, the following chapter primarily focuses on the background needed for the statistical modelling, reviewing prerequisite material for the implementation of joint models and the theory underlying various approaches to survival analysis. Prior to that, a brief analysis of past attempts to fit survival models for risk of CVD and ESRD is given to acquire a more integrated view on the topic.

Chapter 2

Background

This chapter provides additional contextual background to the importance of these clinical outcomes in people with diabetes. Henceforth, I examine prior research endeavours pertaining to developing predictive models for complications, specifically CVD and ESRD in T1D. I further present various modelling techniques employed in survival analysis, focusing on Cox's proportional hazards and Poisson regression models, along with computational methodologies emerged. Finally, I provide an overview of the approaches typically utilised to evaluate the predictive performance of such models.

2.1 Susceptibility to complications of T1D

Diabetes mellitus is a group of metabolic diseases. Several pathogenic processes are involved in the development of diabetes. A mixture of genetic predisposition and environmental factors contributes to the onset of diabetes and the risk of developing complications. Most types of diabetes are associated with defects in insulin secretion and share common long-term complications due to chronic hyperglycemia.

Elevated blood glucose causes particular damage to the cardiovascular and nervous systems (Lammert et al. 2014). The level of glycaemic control as measured by the degree of nonenzy-

matic glycation of haemoglobin (HbA_{1c}) is primarily determined by patient behaviour, insulin therapy and non-insulin anti-diabetic agents. T1D is additionally determined by the level of residual β-cell function, which in turn is partly under genetic control (Yu et al. 2019; Gubitosi-Klug et al. 2021). Poorer glycaemic control is further associated with developing diabetic kidney disease (DKD), amongst other vascular complications. However, even for a given level of glycaemic control and diabetes duration, the variability in risk of developing complications indicate that other factors must be relevant (Harjutsalo and Groop 2014). For example, familial clustering and heritability in DKD have highlighted an underlying genetic susceptibility (Gu 2019).

Major risk factors for developing CVD and DKD are age, HbA_{1c}, waist-to-hip ratio, blood pressure and non-HDL cholesterol. Furthermore, cardiovascular risk markers are predictive of the development of diabetic peripheral neuropathy (DPN), particularly of large-nerve fibre dysfunction, which may account for the high mortality rate in patients with an abnormal vibration perception threshold (VPT) (Elliott et al. 2009). Macroalbuminuria, peripheral and autonomic neuropathy are found to be the most relevant factors for mortality in individuals with T1D (Soedamah-Muthu et al. 2008).

Previous studies of the relationship of C-peptide to microvascular complications, such as the Maser et al. (1992), have been too small and underpowered to confirm the association (Williams et al. 2019). Recent results obtained from a large representative cohort of individuals with T1D in Scotland (Jeyam, Colhoun, et al. 2021) suggest that even minimal residual C-peptide secretion could have clinical benefit in T1D. Interestingly, this is in contrast to the results obtained by a follow-up study of the Diabetes Control and Complications Trial (DCCT) intensively treated cohort, where an effect on hypoglycemia (low blood glucose) was seen only at C-peptide levels $\geq$ 130 pmol/L (Nathan and Group 2014). Such findings emphasise the importance of early determination of factors that increase the risk of complications.

Part of the interindividual variation in developing complications reflects fluctuation in some

of the fundamental biological pathways involved, such as advanced glycation end-products (AGE) formation. In this case, the glucose in the blood attaches itself to proteins without the aid of an enzyme. As the glucose builds up, resulting in hyperglycaemia, it attaches itself to the amino-acids and proteins which in turn get modified. AGEs can cause damage throughout the body, including the blood vessels. They can cause damage to vessels' walls, and when this occurs in large and medium sized arteries, it results in accelerating arteriosclerosis, or hardening of the vessels. Small damaged vessels may lead in turn to retinal, renal and nerve damage.

Other mechanisms linking T1D with accelerated atherosclerosis, cardiomyopathy, and increased post-myocardial infarction mortality rates include prolonged increases in reactive oxygen species (ROS) production in diabetic cardiovascular cells. Insulin resistance causes excessive cardiomyocyte ROS production by increasing fatty acid flux and oxidation, which causes cardiomyopathy (Shah and Brownlee 2016). Although the clinical correlations which link diabetes with its complications are increasingly understood, more studies of new biomarkers to measure end-products and ROS production are needed.

There is also likely variation between individuals in the extent to which high glucose levels activate these various pathways. HbA_{1c} does not reflect most glycative and oxidative chemical pathways that cause complications (Beisswenger 2012). Furthermore, there is likely variability in the susceptibility of organs to damage. Current biomarkers, retinal examinations and albuminuria cannot detect early tissue damage. Such glucose-induced tissue damage is hypothesised to be mediated through a range of pathways, including increased flux of glucose through the hexosamine biosynthetic pathway (HBP) (Buse 2006). Previous studies have proposed that glycaemia-related damage in T1D is at least partly attributable to increased flux through the HBP (Ferranti et al., n.d.).

Other pathways include enzymatic glucose modification of proteins, so-called N-linked glycosylation. Altered N-glycosylation profiles are emerging as a novel risk factor contributing to

complications development. Evidence indicates that diabetes patients can be distinguished based on N-glycome composition (Rudman, Gornik, and Lauc 2019). Such modifications can be described in terms of the amount of N-acetylglucosamine (GlcNAc), fucosylation, galactosylation, and sialylation (N-acetylneuraminic [sialic] acid), as well as branching. Serum N-glycome has been previously shown to regulate EGF receptor and TGF-β signaling pathways that are generally considered important in mediating DKD (Bermingham et al. 2018). Also, Bermingham et al. (2018) studied whether the altered N-glycan profile in T1D is part of the mechanism of glucose-induced kidney damage. They found substantial alterations in the relative abundance of released total and IgG N-glycans in serum, along with associations between N-glycans and both higher albumin-to-creatinine ratio (ACR) and steeper estimated glomerular filtration rate (eGFR) slope in patients with T1D. Although elevated HbA_{1c} is associated with an altered N-glycan profile in individuals with T1D, none of the N-glycan peaks has been found to be prognostic of future renal function decline independently of HbA_{1c} (Colombo et al. 2021).

2.2 Previous literature on prediction of CVD and ESRD in diabetes

This section provides a concise overview of prior endeavours in modelling the risk of developing CVD and progression to ESRD using eGFR as a predictive biomarker. It further elucidates how the reviewed articles motivate the research conducted.

2.2.1 Epidemiology of eGFR and renal disease

Current clinical guidelines in the UK use the estimated glomerular filtration rate as an overall measure of kidney function and recommend that a patient who is losing kidney function at a relative rate of at least 5% per year should be referred for specialised treatment (Health and Excellence 2021). Therefore, progression to ESRD has been characterised by the rate of

change in a person's kidney function as measured by eGFR, an adjusted version of the serum creatinine level found in a blood sample.

Multiple studies have confirmed and consolidated that cystatin C is more accurate than creatinine for measuring kidney function, although it is not routinely available (Randers and Erlandsen 1999; Lopez 2015). Shlipak et al. (2009) showed that in elderly persons, cystatin C predicted substantially more significant kidney function declines than creatinine. Therefore, it has been argued that cystatin C is a valid marker of glomerular filtration rate (GFR) and might be more accurate than using creatinine in some age groups. However, it is not routinely collected and this places eGFR in a better place for usage in longitudinal studies.

Furthermore, β-trace protein, also known as Lipocalin type prostaglandin D synthase, has recently emerged as a novel marker of GFR, representing a more sensitive marker for mild kidney dysfunction than serum creatinine (White, Ghazan-Shahi, and Adams 2015). In this regard, β-trace protein has been proposed as an alternative marker to cystatin C for estimating renal function (Orenes-Pinero et al. 2013). Beyond its role in estimating renal function, β-trace protein has also been advocated as a novel biomarker for cardiovascular risk prediction, which may be relevant to modelling diabetes-related complications. Notwithstanding the importance of these findings, longitudinal data on β-trace protein and cystatin C are typically hard to collect. When available, these usually are helpful adjuncts to serum creatinine in quantifying the risk of progression to DKD.

Circulating kidney injury molecule (KIM)-1 has also been examined as an early biomarker of renal decline in diabetic kidney disease. Gohda et al. (2020) found that serum KIM-1 was associated with a lower eGFR ($<$60 mL/min/1.73 m^2) after adjustment for covariates in patients with diabetes.

More recently, Colombo et al. (2019) attempted to identify a panel of biomarkers for improving the prediction of renal disease progression in T1D. Colombo et al. (2019) employed a penalised Bayesian approach to analyse 297 circulating biomarkers on 859 individuals from the Scottish

Diabetes Research Network Type 1 Bioresource (SDRNT1BIO) and 315 individuals from the Finnish Diabetic Nephropathy (FinnDiane) study, all with entry eGFR between 30 and 75 mL/min/1.73 m^2. They reported that only a sparse panel of CD27 and KIM-1 contains most of the predictive information for eGFR progression, with a modest increment in prediction of renal disease beyond including age, sex, diabetes duration, baseline eGFR and length of follow-up as covariates in the models. The presented evidence supports the argument that eGFR is the most reliable predictive biomarker linked to future changes in renal function.

2.2.2 Time to ESRD based on longitudinal eGFR: understanding the impact of covariates on survival probabilities

It is common practice in survival analysis to use a Cox's proportional hazards model, also known as the Cox regression model (Cox 1972; Bellera et al. 2010; Abd ElHafeez et al. 2021). The Cox proportional hazards model allows us to examine the relationship between covariates (independent variables) and the hazard function, which represents the probability of experiencing an event at a particular time given survival up to that moment. The model assumes that the hazard function is a product of two components: a baseline hazard function that depends only on time and a set of covariate variables, where their effects are independent of time.

The key assumption of the Cox regression model is the proportionality assumption, which states that the hazard ratio between any two individuals remains constant over time. This implies that the effect of covariates on the hazard function is constant over time, and the hazard curves for different individuals only differ by a constant factor.

The Cox regression model estimates the hazard ratio, which quantifies the relative change in hazard for a one-unit change in a covariate, adjusting for the remaining covariates in the model. It does not require the specification of the underlying distribution of survival times, making it a non-parametric or semi-parametric model.

For the modelling of clinical outcomes, various statistical methods have been employed and extended over time. Based on the components drafted in the preliminary skeleton of this thesis, a comprehensive search of the literature was performed using the PubMed and Google Scholar search engines. Multiple keywords related to the topic were used. To the best of my knowledge, the most up-to-date scoping review with respect to *predictive modelling of clinical progression to chronic kidney disease (CKD)* has been conducted by Lim et al. (2022), which is very useful material for reviewing the existing evidence regarding this area.

The data sources of this review are Medline, EMBASE, CINAHL and Scopus from years 2011 to 17th February 2022. They have identified 516 studies, 33 of which were included in the review in full. All selected articles built statistical or computational model(s) to predict the risk of developing CKD. The patient data acquired by these studies were sourced from various healthcare systems, adding diversity which enhances the quality of the conducted research. The reviewers concluded a lack of reporting consistency regarding the details of developing risk prediction models for CKD. In the context of my specific interest, it was also observed that Cox regression modelling was the prevailing method employed, according to the systematic review.

Despite the relatively low number of studies considered, the existing literature implies that Cox regression models are most commonly used in survival analysis to investigate and infer the relationship between covariate data and survival outcomes. The work that is described in the thesis explains how the traditional Cox approach can be expanded in order to improve the yielded prediction of time to event.

Foremost, the study conducted by Diggle, Sousa, and Asar (2015) with respect to modelling longitudinal eGFR to determine progression to kidney disease has been the closest to my analysis. They utilised a random intercept and a continuous time, non-stationary stochastic process to obtain a predictive distribution of the the underlying rate of change of kidney function given via eGFR trajectories. More details about the particular type of modelling

follow in chapter 6.

2.2.3 CVD risk models developed for T1D

Cardiovascular disease (CVD) is one of the major causes of death worldwide (Ezzati et al. 2005). Individuals with T1D face an elevated susceptibility to developing CVD. In 1961, the coining of the term 'risk factor' resulted from the identification of an initial set of traditional risk factors for coronary heart disease in the Framingham Heart Study (Kannel et al. 1961). The important risk factors identified by the Framingham study were age (males ≥ 45 years or females ≥ 55 years), male sex, hypertension, dyslipidemia, smoking, and diabetes mellitus (Khambhati et al. 2018).

Soedamah-Muthu et al. (2008) identified important risk factors associated with high CVD mortality, including age, waist-to-hip ratio, pulse pressure, and non-HDL cholesterol, among others. Analysis of longitudinal biomarker data in the context of CVD risk prediction is of utmost importance. By considering biomarker measurements over time, researchers and clinicians can better understand the dynamic relation between biomarkers and outcomes over time, i.e., temporal patterns and trajectories of biomarkers and their predictive value in identifying individuals at risk of CVD. Overall, the analysis of longitudinal biomarker data helps uncover the relationship between biomarker changes and CVD occurrence, thereby improving the accuracy and effectiveness of risk assessment and management strategies.

Stevens et al. (2021) carried out a systematic methodological review of the modelling of cardiovascular disease (CVD) risk using longitudinal biomarker data and risk factor trajectories. For this purpose, they screened MEDLINE-Ovid from inception until 3 June 2020. Key search terms focused on data type, modelling type and disease area, including search terms such as longitudinal, trajectory and cardiovascular, respectively. Studies were selected to meet the following inclusion criteria: longitudinal individual data in adult patients with more than three time points and a CVD or mortality outcome.

The researchers identified 2601 studies, 80 of which were included in the review in full. Four statistical approaches were identified for modelling the longitudinal data: 4% compared time points with simple statistical tests, 50% used single-stage approaches, such as single-time analysis or including summary measures in the survival models, 36% used two-stage approaches, including an estimated longitudinal parameter in survival models, and 10% used joint modelling, which modelled the longitudinal and survival data together. The 80 studies included in the article are given in the following link by model and study characteristics.

The particular systematic review has primarily highlighted that single-stage models are heavily employed in CVD risk prediction studies for modelling longitudinal data, risking the induction of bias in retrospective prediction of time to event, as the likelihood of event is conditioned upon the full longitudinal trajectory, which might comprise data points during the prediction time. Furthermore, they have affirmed the benefit of employing two-stage and joint approaches, which utilise the available data resources more cautiously, and they are of increasing popularity as more exemplar cases become available.

The study of Vistisen et al. (2016) has contributed significantly at identifying a number of risk factors for CVD risk prediction, including age, sex, diabetes duration, systolic blood pressure, low-density lipoprotein cholesterol, HbA_{1c}, albuminuria, glomerular filtration rate, smoking, and exercise and using them as covariates in a Poisson model for a 5-year CVD event. However, the model was poorly calibrated when validated using an external population with T1D, which indicates overfitting issues and the need for developing more robust approaches.

Furthermore, markers predicting CVD risk in T1D, or included in the reviewed risk models for CVD as covariates have been reported by Livingstone et al. (2012) and Rawshani et al. (2017).

Several additional studies related to the current state of research on CVD risk prediction have been identified and reviewed in the following section. The studies were selected based on their relevance to the topic.

The objectives of the study conducted by Jia et al. (2019) has been to identify key risk factors that affect the prediction of CVD risk and to develop a 10-year CVD risk prediction model using the identified risk factors. The study found that heart rate was determined as a novel risk factor, and a CVD risk model incorporating factors such as age, sex, body mass index (BMI), hypertension, systolic blood pressure (SBP), tobacco use per day, pulse rate, and diabetes was constructed. Therefore, the inclusion of heart rate as a predictor extended the predictive ability of existing risk factors.

Yang et al. (2022) developed a 10-year CVD risk prediction model applicable to diverse regions of China by excluding blood lipids as predictors. The study involved a large cohort of individuals and found that after recalibration, the model's calibration performances improved. The results were comparable for both women and men. The development of flexible recalibration models could enable more widespread use of CVD risk prediction in different regions using healthcare records of the Chinese population.

Furthermore, the objective of Sung et al. (2019) was to compare the performance of a Cox hazard regression model with a Recurrent Neural Network (RNN) - Long Short-Term Memory (LSTM) model (Sherstinsky 2020) based on survival analysis, to demonstrate the increment deep learning might bring in prediction. The study found that the deep learning model outperformed the Cox regression model in terms of time-dependent area under the curve for both females and males at 2 years. Layer-wise Relevance Propagation (LRP) (Montavon et al. 2019) analysis revealed that age, SBP, and diastolic blood pressure (DBP) were the variables with the greatest effect on the outcome of CVD.

McGurnaghan et al. (2021) recently developed a Poisson regression model for individualised predictions for developing CVD over a 10-year follow-up period. Their 10-year CVD risk prediction tool could facilitate discussions regarding appropriate statin prescribing. For model derivation, the Scottish cohort with T1D has been used, which has been previously described by Livingstone et al. (2012). The person-time of the study cohort was split into one-year

intervals, and then the model was externally validated using the Swedish National Diabetes Register. Covariates were included as baseline measurements, apart from age which was time-updated.

Barbieri et al. (2022) compared deep learning extensions of survival analysis models with Cox proportional hazards models for predicting CVD risk using national health administrative datasets. The study found that the deep learning models outperformed the Cox proportional hazards models in terms of explained variance, calibration, and discrimination. Hazard ratios estimated by the deep learning models aligned with known CVD risk factors, such as tobacco use, hypertension, chest pain, and diabetes.

Cooper, Wells, and Mehta (2022) investigated whether accounting for the competing risk of non-CVD death improves the performance of CVD risk-prediction equations in older adults. The study analysed a large cohort of older individuals and found that standard Cox models overestimated the 5-year CVD risk, while Fine-Gray models (Austin, Steyerberg, and Putter 2021), which consider competing risks, were overall better calibrated. The study suggests that new CVD risk equations that take competing risks into account should be considered for people aged over 65 years.

According to the broader literature, most discoveries made nowadays are implemented by using more elaborate methods since conventional methods have been fully exploited. The scope of the thesis primarily revolves around assessing whether specific methodological advancements result in a considerable increment in risk prediction of two major complications contributing to the increasing mortality rates in individuals with T1D. Before proceeding with the introduction and the assessment of the specific methods that expand the traditional Cox regression, I briefly introduce the main advancements that have led to the rise of joint modelling, as a statistical technique.

2.3 Modelling time to event: survival analysis

Prevalent mathematical concepts used in survival analysis are presented below:

- Survival Function, $S(t)$: the probability that an individual will survive beyond time t $\Pr(T > t)$. When we know the outcomes for all subjects, we can estimate the survival function empirically. At any time x, the survival function is given by the proportion of subjects that have not experienced the event yet. In theory, the survival function is smooth, but in practice, the estimation of the survival function based on observed data is discretised using, for example, the Kaplan-Meier survival curve estimation (Kaplan and Meier 1958; Goel, Khanna, and Kishore 2010).

- Probability Density Function, $f(t)$: the relative likelihood that an individual will die at a time t, and it is the derivative of the cumulative probability function.

- Cumulative Probability Function, $F(t)$: the probability that an individual will have a survival time less than or equal to t $\Pr(T \leq t)$. The survival function and the cumulative probability function sum to 1, $S(t) = 1 - F(t)$.

- Hazard Function, $h(t)$: the instantaneous hazard rate of experiencing an event at time t conditional on having survived to that time. The hazard function is also given by $h(t) = \frac{f(t)}{S(t)}$, i.e., the instantaneous hazard equals the unconditional density function of experiencing the event at time t, scaled by the event-free fraction at time t.

- Cumulative Hazard Function, $H(t)$ (or the cumulative hazard rate): the integral of the hazard function from time 0 to time t, which equals to the area under the curve $h(t)$ between time 0 and time t. The cumulative hazard is the total accumulated risk of experiencing the event of interest that has been gained by progressing to time t.

An additional connection between $H(t)$ and $S(t)$ can be written in the following two ways:

- $H(t) = -log[S(t)]$: The cumulative hazard function equals the negative log of the survival function.

- $S(t) = e^{-H(t)}$: The survival function is the exponentiated negative cumulative hazard function. An increase in $h(t)$, the instantaneous hazard will lead to an increase in $H(t)$, the cumulative hazard, which translates into a decrease in $S(t)$, the survival function.

The most common method of depicting survival data is the Kaplan-Meier curve (Kaplan and Meier 1958; Rich et al. 2010; Goel, Khanna, and Kishore 2010), which is a graphical representation of survival probabilities over time; the estimated survival function is based on event status and follow-up length. The Kaplan-Meier survival curve is based on observed survival times and censoring information, and graphically it appears as a step function, with a step down at each event occurrence. The Kaplan-Meier estimator breaks up the estimation of $S(t)$ into a series of steps (intervals) based on known event times. Observations contribute to the estimation of $S(t)$ until the event occurs or until a subject is censored. For each step, the probability of surviving until the end of that interval is calculated, given that subjects are at risk at the beginning of the interval. The estimated $S(t)$ for every value of t equals the product of surviving each interval up to and including time t.

The main assumptions of this non-parametric method, in addition to a non-informative censoring mechanism is that censoring occurs after failing and that there is no cohort effect on survival, so subjects have the same survival probability regardless of when they entered the study. The estimated $S(t)$ from the Kaplan-Meier method can be plotted as a stepwise function with time on the X-axis, as shown in figure 2.1.

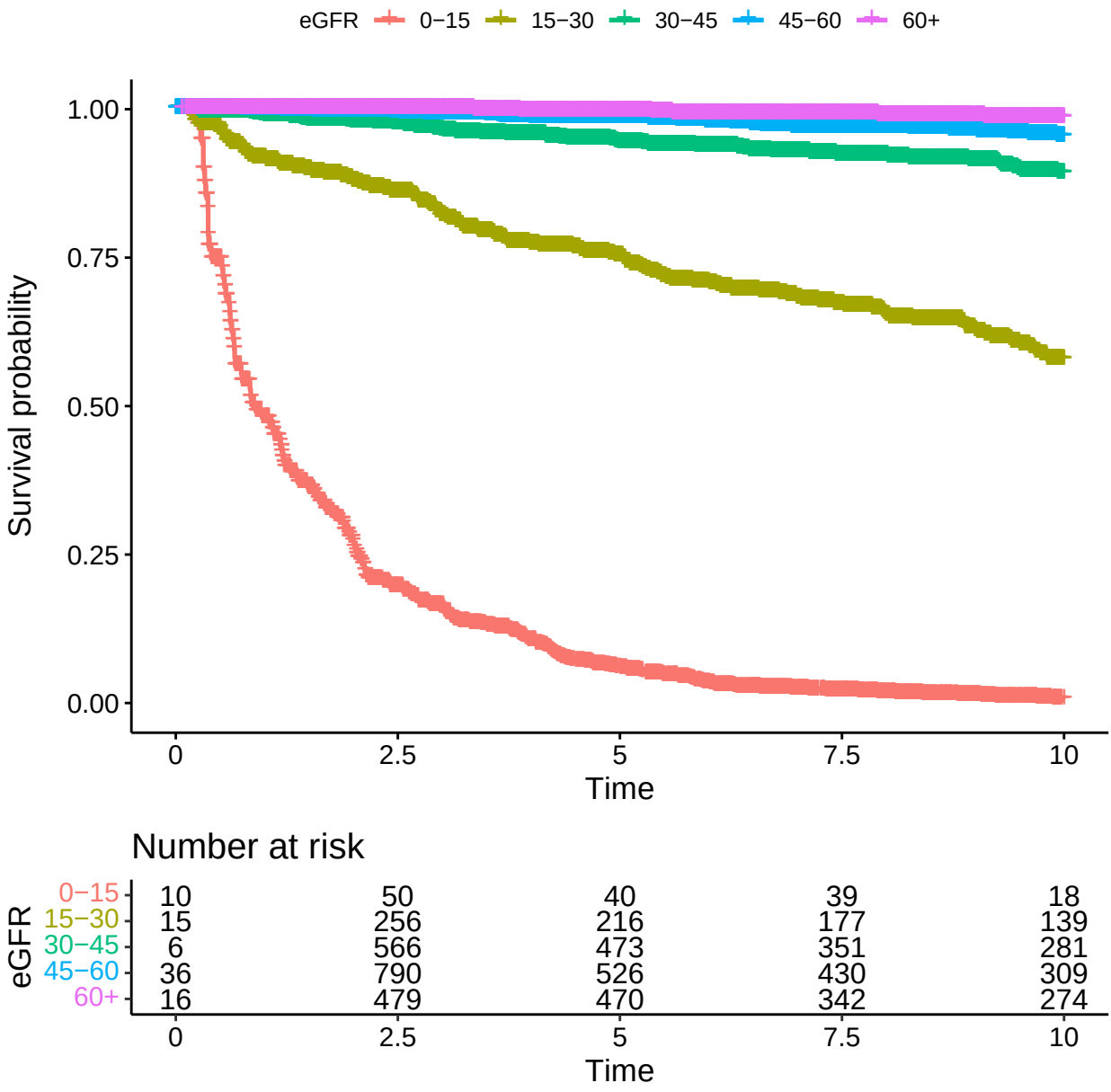


Figure 2.1: Non-parametric time-to-event estimates for 2633 individuals with T1D, based on observed eGFR values spanning a time period of 10 years split into 5 classes.

Let T be a non-negative random variable representing the waiting time until the occurrence of an event. Then, we define the hazard rate of event occurrence as the conditional probability that the event will occur in the interval $[t, t + dt)$, where dt is the length of this time window, given that an event has not happened until time t. The event rate λ refers to an individual subject, giving the conditional probability of an event per unit of time. Mathematically, $\lambda = \lim_{\delta t \to 0} \frac{P(t \leq T < t+\delta t | T \geq t)}{\delta t}$. The conditional probability divided by dt shows the rate of event occurrence per unit of time. It is also called instantaneous event rate to emphasise the fact that it refers to a specific moment in time.

An exponential distribution of time to event is equivalent to a Poisson process for the arrival of events, with a constant arrival rate λ. Hence, the Poisson arrival process and exponential time to event are equivalent terms. Splitting time into short intervals allows the survival curve to be modelled as a piecewise exponential curve. Over any short interval, the hazard rate is approximately constant.

The most widely used method to analyse time to event data is the Cox proportional hazards model, introduced in 1972 (Cox 1972), which is a statistical regression model that describes the relationship between an event occurring, as expressed by the baseline hazard function denoted by $h_0(t)$, and a set of covariates X_i's and corresponding coefficients β's, as given by the following formula:

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p).$$

This formulation allows for estimating hazard ratios, which represent the relative risk associated with different covariates.

The Cox model is often described as a semi-parametric approach. In semi-parametric models, certain aspects of the model are specified parametrically, i.e., specific functional forms or distributions are assumed for certain model parameters, while other components of the model are left non-parametric, allowing for more flexibility in specifying complex relationships.

The Cox model does not make assumptions about the underlying distribution of survival times when estimating the hazard ratios. In particular, the Cox model assumes that the hazard ratio remains constant over time, indicating that the 'proportional hazards' assumption holds (Kumar and Klefsjö 1994). This assumption implies that the hazard functions for different individuals may vary, but the ratio of their hazards remains constant over time.

On the other hand, the Poisson model is typically used for analysing count data, where the outcome of interest is the number of events occurring within a *specified time interval*. In the context of survival analysis, the Poisson model is employed to estimate event rates and assess the association between covariates and event occurrence. The Poisson model assumes that the hazard ratio or event rate is constant over time, which may not hold true in some applications. The Poisson rate parameter λ is the event rate: defined as the expected number of occurrences per unit of time as a function of the covariates.

The formula gives a Poisson regression model

$$log(\lambda) = \alpha + b_1 x_1 + b_2 x_2 + ... + b_n x_n, \tag{2.1}$$

where the natural log transformation of the event rate (left-hand side) is a linear combination of the explanatory variables (right-hand side). In particular,

- x_i is the ith explanatory variable $(i = 1, ..., n)$
- λ is the expected value of the mean, i.e., the expected event rate dependence on the covariates in the study for an individual with a particular set of values for $x_1, x_2, ..., x_n$
- α is the estimated constant term, i.e., the intercept, that provides an estimate of the log event rate when all x_i's of the equation are equal to 0 (the log of baseline hazard)
- b_i are the estimated Poisson regression coefficients (log hazard ratios).

The equivalence between piecewise-exponential models and Poisson models:

The Poisson parameter λ and the hazard ratios (HRs) are closely related in the context of modelling time-to-event. When comparing two groups using HRs, we are essentially comparing their hazard functions over time. The HR reflects how the instantaneous risk of an event differs between the two groups at a given time point. The event rate λ is a component of the hazard function $h(t)$, and the HR compares the hazard functions between groups. Therefore, the event rate λ contributes to the HRs when they are compared between different groups.

The estimated HRs from a piecewise-exponential model can be approximated using Poisson models fitted on finely split time intervals (one interval for each event). A Poisson regression model can replicate a piecewise-exponential model, in which case, the aforementioned issue of non-constant hazard is overcome by the time splitting (Dickman et al. 2004; Rodriguez 2007; Li et al. 2016). A piecewise-exponential model is a type of survival model that estimates the HR, which represents the instantaneous risk of an event occurring at a specific time, given that the event has not yet happened. Such models, like Cox regression models, are particularly applicable when the hazard rate is not constant over time and may vary at specific time intervals or time points, as in the case of developing risk of complications in T1D. The piecewise-exponential model can be effectively approximated by a Poisson regression model that divides the follow-up time into consecutive time intervals, where within each interval, the event rate is constant. This enables the estimation of different hazards in each time window, capturing potential changes in risk over follow-up time.

Considering an alternative viewpoint, the hazard ratio (HR) and event rate ratio, or incident rate ratio (IRR), are both measures of effect size; the value that quantifies the strength of the relationship between two variables (Sullivan and Feinn 2012). In the Cox proportional hazards model, the HR is used to quantify the effect of a covariate on the hazard function. It represents the ratio of the hazards between two groups of a covariate. For example, an HR of 1 indicates no difference in the hazards between the groups, and an HR greater than 1 indicates an increased hazard in the reference group compared to the comparison group.

On the other hand, the event rate ratio is commonly used in Poisson regression models to quantify the effect of a covariate on the incidence rate. It represents the ratio of the incidence rates between two groups or levels of a covariate (Wilson 2021). An IRR of 1 indicates no difference in the event rates between the groups, and an IRR greater than 1 indicates an increased event rate in the reference group compared to the comparison group.

Consequently, the HR measures the relative risk of an event occurring at any given time, while the IRR measures the relative rate of events occurring over a specific time period. Therefore, the Cox regression model can be seen as a a special case of Poisson regression in which the intercept has been eliminated by setting the baseline hazard rate in each interval to its maximum likelihood value.

It is important to note that although the parameter estimates and standard errors are effectively interchangeable between Cox and Poisson regression models, the two models do not make the same assumption a priori; the Poisson models assume that the HRs are constant within intervals containing the event times, while the Cox models do not make this assumption.

The Poisson model assumes that the event rate or event occurrence follows a Poisson distribution. For a given individual with covariate values x, the event rate $\lambda(t|x)$ at time t is modelled as $\lambda(t|x) = \lambda_0(t) \times exp(\beta x)$, where $\lambda_0(t)$ is the baseline event rate and β is the vector of regression coefficients.

The association between the Cox proportional hazards and the Poisson model is the exponential term $exp(\beta x)$, which captures the effect of covariates on the hazard or event rate. This term represents the multiplicative effect of the covariates on the hazard or event rate. For a Cox proportional hazard model, the inclusion of a time-varying covariate would take the form of $h(t) = h_0(t)\exp(\beta_1 x_1(t))$.

In conclusion, the Cox model can be effectively approximated by the Poisson model, considering it as a special case of a parametric Poisson regression model. In this approximation, the baseline hazard $h_0(t)$ is represented by the intercept term associated with time t, and the

coefficients correspond the log HRs associated with each covariate.

Two papers published separately in the early 1980s, Holford (1980) and Laird and Olivier (1981), showed that the likelihood of the Cox model was technically equivalent to that of a Poisson regression model in which individuals are censored at the first event, with an intercept term for each person-time interval.

They independently formulated that a Poisson arrival process with a constant hazard/event rate implies an exponential survival function (cumulative probability of time to failure) and vice versa. Let t_{ij} denote the time elapsed by the i-th individual in the j-th interval, that is, between $[t_{j-1}, t_j]$. The piecewise exponential model may then be fitted to the data by treating the events as if they were independent Poisson observations with means $\mu_{ij} = t_{ij}\lambda_{ij}$, where λ, the Poisson parameter, is explicitly linked to each different time interval. The observations in each person-time interval are obviously not independent draws from a Poisson distribution. For all person-time intervals, the event indicators are all zero, except for the last interval, which can take the value 0 or 1. German Rodriguez (2007) has elaborated on this theory and constitutes a foundation of this work.

To fit the Cox model, the dataset is split at each time point where at least one individual has an event to construct a dataset in which there is one observation for each of the resulting person-time intervals. The Cox regression model can be extended to model time-updated covariates, but it cannot estimate the baseline hazard rate. In some situations, such as a drug clinical trial, we only need to estimate rate ratios and do not need to estimate the baseline hazard rate. Nevertheless, in other situations, such as clinical risk prediction, we need the baseline hazard rate in order to make realistic estimations.

Heterogeneity of susceptibility due to risk factors that cannot be measured, i.e., is unknown or not represented in available data, is commonly referred to as frailty; a quantity that varies among individuals (Vaupel, Manton, and Stallard 1979). In many diseases, frailty variation is a possible explanation for an early peak in incidence (Zarulli 2016). Not accounting for

unobserved frailty variation in statistical analyses may lead to misleading comparisons of hazard and incidence rates (Aalen et al. 2015). The statistical interest in frailty stems partly from the fact that it can lead to waning effects of the predictor under investigation over time. Observational studies commonly show that the effects of various risk factors may wane within more extended follow-up periods. For instance, the risk ratio attributable to smoking and blood pressure decreases as the population transitions from younger to older since individuals at high risk have been removed from the study. The frailest group experiences the event first because they are at higher risk. Hence, unmeasured susceptibility decreases faster in individuals with high risk, who are smokers and have high pressure (Moolgavkar 2015). Although likely, frailty frequently influences susceptibility and severity of progression to complications in individuals with any health conditions, methods for allowing for it in survival analysis are still not fully developed.

Although Cox regression is still widely used, statistician Bendix Carstensen has argued that it is now rarely the most appropriate method for survival analysis ("Who Needs the Cox Model Anyway?" 2018). With powerful computers and efficient algorithms for fitting Poisson regression models, there is no longer any computational advantage over Poisson regression. In a Poisson regression, the baseline hazard can be specified as constant, modelled as a linear relationship to the timescale, or modelled as a spline function allowing the model to adapt as required to fit the data.

2.3.1 More flexible modelling of the survival function

The effects of the covariates on the hazard and survival probabilities can also be approximated by a functional form such as a spline (Whittemore and Keller 1986; Gray 1992) that allows for non-monotonic relationships to be captured. Splines allow for time-dependent effects, non-linear effects and interactions between covariates (Fauvernier et al. 2019). Thus, splines approximate the true underlying relationship between the covariates and the survival function. Splines are a compromise between fitting a straight line (effectively one parameter for every

failure time point) and fitting a smoothed function that allows one to fit the data more closely and capture non-linear, complex patterns.

One possible way to implement such penalised models is to use the described approximation of the survival likelihood by a Poisson likelihood by splitting the data into person-time intervals. Dividing the time of follow-up into smaller intervals (having one longitudinal observation per interval), and then fitting piecewise-polynomial functions within each person-time interval generates a smooth curve that approximates the true effect of the predictor-biomarker and might estimate better the overall relationship between the longitudinal and time to event processes.

The fit of splines heavily depends on the number of knots (degrees of freedom) specified. The R package `mgcv` (Mixed GAM Computation Vehicle with Automatic Smoothness Estimation) allows the number of knots to be learned effectively from the data, and can be used to fit such penalised models. Restricted cubic splines may be best for dynamic forward projection, which restricts the function to be linear at the ends where data are sparse.

Furthermore, Royston-Parmar parametric models, also known as Royston-Parmar flexible parametric survival models are a class of statistical models used in survival analysis (Ng et al. 2018). These models were developed by Patrick Royston and Mahesh Parmar as an extension of the traditional Cox proportional hazards model.

The Royston-Parmar models allow for more flexible modelling of the baseline hazard function, providing a wider range of shapes and associations. They model the baseline log cumulative hazard on the proportional hazard scale. They are specified by a series of spline functions that approximate the baseline hazard, allowing for more dynamic modelling of the hazard over time and overcoming the restrictive assumption of constant hazard ratios over time.

This flexibility allows for capturing time-varying effects and non-proportional hazards, which can be important in certain survival scenarios. These models have been shown to be particularly useful when the baseline hazard exhibits complex shapes or when there are non-linear

relationships between covariates and the hazard. However, such modelling technique is not utilised herein, primarily due to the risk of overfitting the baseline hazard function and the difficulties of interpreting a model with multiple time-dependent effects, because the hazard ratios are dependent on more than one covariates (Baade et al. 2015).

A survival dataset reformatted as person-time intervals enables us to incorporate exposures with values that change over time. Time-varying covariates can be included in survival models by changing the unit of the analysis from the individual to the period of time when the exposure is constant (Zhang et al. 2018).

To allow for unequal lengths of person-time intervals, the Poisson regression model needs to include the log of the length of each interval, as an *offset* term in the formula. An offset term is a term in the model defined to have a coefficient of 1. With many short interval lengths, the continuous-time model can be approximated arbitrarily close by a Poisson time-splitting model with sufficiently short interval lengths.

In an accelerated failure time model (Wei 1992) in which the hazard rate increases linearly with follow-up time can be represented by including a covariate term for the time of follow-up. For instance, the human lifespan, where mortality is high in the first years of life, and low in later childhood and young adulthood, before rising with increasing steepness in later life, could be modelled with an appropriate spline function for the hazard rate.

By splitting time into short intervals (the ends must join up), we consider the constant hazard rate within each interval and specify a survival function that is piecewise-exponential.

Therefore, the time-splitting approach is easy to implement and allows one to fit different hazard functions and implicitly estimate different hazard ratios for each person-time interval. Hence, by approximating any underlying survival function by a piecewise-exponential distribution, there is no need for parametric models of other survival functions, such as the Weibull distribution, where the hazard rate is a linear function of time. This can be addressed by splitting time into short intervals and including follow-up time as a time-updated covariate.

Effectively, the time-splitting approach makes any other parametric survival model, which makes assumptions about the shape of the hazard in advance, obsolete. This is one of several advantages of the Poisson time-splitting approach.

Thus, the time-splitting approach makes it possible to incorporate time-updated covariates in the model, fixing the covariate values at the start of each interval. With these types of regression, time intervals in which no events occur do not contribute to the likelihood and are dropped from the analysis.

One area for improvement with the time-splitting approach is that the size of the dataset can become impractically large if we say, split millions of people into a hundred person-time intervals each. This can be overcome by thinning observations (person-time intervals) where no event occurs (Gratton et al. 2015). With a Poisson regression, this is of no importance.

Last but not least, parametric approaches rely on the maximum likelihood to estimate parameters. Akaike Information Criterion (AIC) is commonly used to compare models run with different parametric forms, with the lowest AIC being indicative of the best fit to the data.

2.4 Joint modelling of longitudinal biomarker data and time to event

Joint modelling of time to event and longitudinal data was originally developed during the 1990s to make updated predictions from measurements of biomarkers of immune cells about progression to AIDS in HIV-positive individuals (Tsiatis, Degruttola, and Wulfsohn 1995). Other applications are to:

- investigate the relationship between a biomarker and the risk of occurrence of the clinical outcome,
- improve estimation in either or both outcomes as compared to separate analyses of the

two outcomes.

Wearable devices and continuous monitoring is an exemplar application of precision medicine and joint modelling is ideally suited for the task of dynamic prediction.

A shared-parameter joint model consists of two submodels: a longitudinal submodel, such as a linear mixed-effects model (A) and a time to event submodel (B), such as a proportional hazards model, which is linked using an association structure. Time-updated biomarker observations provide information about the last time a subject was observed and remained event-free. Time is interpreted as time elapsed since entry: the unique time an individual enters the study is time zero, and after that, the time elapsed since the last observation. The formulation of a joint model is shown below:

$$\text{(A): } X_{im} = f(t_{im}, \boldsymbol{b}_i, \boldsymbol{\beta}_L) + \epsilon_{im}$$

In this formula, X_{im} is the m-th longitudinal biomarker for the i-th individual at time t_{ij}, $f(\cdot)$ represents the functional form relating the biomarker to time, random effects $\boldsymbol{b}_i$ capture the individual-specific deviations from the population average, $\boldsymbol{\beta}_L$ represents the fixed effects coefficients, and ϵ_{im} is the residual error term.

$$\text{(B): } h_i(t) = h_0(t) \exp(\boldsymbol{\beta}'_S \boldsymbol{x}_i + \alpha b_i)$$

where $h_i(t)$ is the hazard function for the i-th individual at time t, $h_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}$ represents the fixed effects coefficients for the covariates $\boldsymbol{x}_i$, and α captures the association between the random effect b_i that link the longitudinal submodel and the hazard.

Biomarker trajectories are modelled as random effects, given the observed values.

> If an effect is assumed to be a realised value of a random variable, it is called a random effect. (LaMotte, 1983)

Inherently, random effects capture the random or unobserved factors that introduce variability among the groups or levels.

Before I discuss the inner mechanisms of joint modelling in the following chapter, where I showcase its application on a diabetes dataset, I give a high-level explanation of how the joint modelling approach has progressed as a technique over the past few years.

Joint modelling can be described as a parallel estimation of two or more statistical models, which traditionally would have been fitted separately. I have studied the shared-parameter joint model implementation for longitudinal and time to event data fitted via the R package `rstanarm`. The application of joint modelling via `rstanarm` is described in chapter 4.

A shared-parameter joint model consists of a linear mixed-effects model called the longitudinal submodel and a time to event submodel. The longitudinal submodel analyses the patterns of the component that has been measured repeatedly over time. The time to event submodel specifies the time until an event of interest occurs, such as death or disease progression. In particular, the longitudinal submodel estimates the fixed effects for covariates and random effects for individuals. The survival submodel specifies the hazard function and translates the effects of the covariates into an event likelihood. Fixed effects are constant across individuals, while random effects vary. For instance, if each subject receives both the drug and placebo on different occasions in a clinical trial, then the fixed effects term would estimate the effect of the drug, while the random effects would capture how the response differs among subjects, i.e., random effects are used to estimate the variability among individuals.

The association structure defines the relationship between the longitudinal processes and the time to event. The joint model explicitly builds the joint distribution of the longitudinal and survival outcomes to yield survival probabilities. Because the biomarkers are measured only at intervals (intermediately) and usually with error, the biomarker trajectories are not observed directly. Joint modelling accounts for the error measurements inherent in the longitudinal component, whereas more straightforward modelling techniques in which the last observation

is carried forward as a time-varying covariate in a Cox regression or Poisson regression do not allow for this.

A primary limitation of such a naive approach arises when biomarkers are not updated that often, and when they do, it usually happens because the subject has transitioned to the end stage of a disease, which changes the baseline assumptions on which the model has been built on. In particular, diseases such as diabetic kidney disease can remain asymptomatic for many years.

The premise for joint modelling is to specify appropriate submodels for the longitudinal component(s), and the time to event, coupled by a dependence of the survival function on the biomarkers: the so-called *shared-parameter model*. The model is fitted by maximising the likelihood of the model parameters given the longitudinal measurements and time to event observations. Hickey et al. (2016) has discussed recent developments in this area. It provides appropriate context for scrutinising the state-of-the-art approaches, as such, the `rstanarm` implementation.

The likelihood of the standard joint model is derived under the assumption that the longitudinal and time to event processes are independent, given the random effects. I.e., unobserved biomarker values, (also called latent parameters) comprise part of the information that is included in the association structure, that links longitudinal and time to event data. In addition, the longitudinal observations themselves are considered independent for each individual, given these random effects. Given these assumptions, the random effects must be integrated out in order to derive the joint likelihood. However, obtaining a closed-form solution to the integral, that sums over the random effects, is not trivial, thus numerical methods are employed to approximate the integral over the random effects.

To that end, the standard approach to joint modelling requires the likelihood to be computed by integration over these unobserved trajectories using computationally intensive numerical techniques, such as Gauss–Hermite quadrature (Pinheiro and Bates 1995).

2.4.1 Two-stage approaches and immortal-time bias

Joint modelling has emerged due to the computational developments of the last years: it has started to impact the modelling of various clinical outcomes, including these studies Baart et al. (2021); Brown (2003); Ibrahim, Chu, and Chen (2010).

The two-stage approach, which entails fitting the longitudinal and survival outcomes separately is the most common solution to overcome the computational requirements of maximising the joint likelihood. However, substantial research on this topic (Tsiatis and Davidian 2004; Ye, Lin, and Taylor 2008; Rizopoulos 2012; Mauff et al. 2017) indicates that this approach results in biased estimates.

Recently, Mauff et al. (2020) have proposed an adaptation of the standard two-stage approach, which eliminates bias and substantially reduces computational time. They used a correction factor based on importance sampling theory (Press et al. 2007). This correction factor allows for reweighing each draw of the Markov Chain Monte Carlo (MCMC) sample obtained from the Bayesian estimation of the two-stage approach, such that the resulting estimates more closely approximate those yielded by the shared-parameter multivariate joint model. The weights are given by the target distribution, i.e., the full posterior distribution of the multivariate joint model, divided by the product of the posterior distributions for each of the two stages, evaluated for each iteration of the MCMC algorithm. Before the use of the correction factor, the two-stage approach is itself modified: where before, in the second stage, only the parameters of the survival submodel were updated, now the random effects are updated as well. The iterative reweighing sampling increases linearly and is much faster than the quadratic computations. These adaptations combined achieve unbiased estimates in a fraction of the time required to compute the full multivariate model.

Two-stage models, in which the longitudinal submodel is fitted first, and the results are plugged into the event submodel, have been shown to give biased results (Faucett and Thomas 1996; Tsiatis and Davidian 2004; Ye, Lin, and Taylor 2008; Rizopoulos 2012). This may be because

this procedure uses information from the future (Anderson et al. 1983). Immortal-time bias (Ho et al. 2013; Yadav and Lewis 2021) can be introduced in the likelihood computation when the follow-up time for certain subjects is incorrectly computed, leading to biased estimates. Using information from the future retrospectively to specify the time to event notoriously gives rise to immortal-time bias (Suissa 2007; Lévesque et al. 2010).

In the survival analysis setting, landmark analysis refers to the practice of designating a time point occurring during the follow-up period, known as the *landmark time* and analysing only those subjects who have survived until such landmark time (Anderson et al. 1983). Dafni (2011) has extensively described the landmark time approach. The inclusion of a landmark point ensures that the model uses only observations up to the landmark time to predict events after that time.

2.4.2 Options for coupling longitudinal and survival processes

A variety of different aspects of the longitudinal process may be related to the risk of event. Some attention has been given to the selection of the most appropriate functional form to link the two processes. Among the most prominent association structures are

- the current value of the biomarker variable,
- the current slope of the longitudinal profile, and
- the area under such profile.

The current value refers to models inserting the current state of the longitudinal variable into the time to event submodel and is used when the current overall value of the longitudinal trajectory is believed to affect the risk of event. The current slope which is the first derivative of the population trajectory, could also be inserted into the time to event submodel in conjunction with the current value or alone. It could also be the case that both the true underlying value and the area under the curve may affect time to event. This is usually used

to model the effect of the rate of change of the longitudinal variable on the risk of event. Moreover, assuming a past value of the biomarker to be related to the survival at time t may be biologically sensible as well. Thus the link structure can be extended with a lag effect (Andrinopoulou and Rizopoulos 2016).

For each biomarker that has been used, we have explicitly assumed that the underlying statuses are related to the biomarkers' current (actual) value, to work with the simplest association structure and the one most typically met in the relevant literature. However, this particular functional form may not be adequate in describing the association structure between the outcomes in all settings. To this end, alternative association structures have been proposed (Lin, Taylor, and Ye 2008; Brown, Ibrahim, and DeGruttola 2005; Dimitris Rizopoulos et al. 2014; Campbell et al. 2021). It could be the case that there is a lag effect or a cumulative effect of the biomarker on survival. For example, it could be that the total HbA_{1c} of your lifetime predicts your risk of developing a diabetes complication as opposed to what was observed in the last few weeks.

The (log) hazard of the event at time t can be associated with the value of the linear predictor of the longitudinal model evaluated at time t in various ways. The most common association structure met in the joint modelling literature to date is the linear association of the log hazard of an event at time t and the linear predictor of the longitudinal submodel at the same time, and this is the option employed in my analysis.

The dependence between the longitudinal and event submodels is captured through the association structure, which can be specified in a number of ways. The simplest parametric association structure between the measurement and event processes is

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

and 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. In a situation where the longitudinal submodel is based on an identity link function and normal error distribution (LMM), the current value association structure can be viewed as a method for including the true underlying value of the biomarker as a time-varying covariate in the event submodel.

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

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

The functional form depends on parameters a for each biomarker included in the model and the number of sequential data each biomarker has. There is, in fact, a whole range of possible association structures, many of which have been discussed in the literature (Crowther, Lambert, and Abrams 2013; Hickey et al. 2016; Campbell et al. 2021) and in section 4.2.1.2. As of the present moment, literature suggests that the most intricate aspect might be the individual-level variance, wherein each patient is allowed a distinct residual error term in the linear mixed model.

2.5 Bayesian methods for statistical computation

Classical model-fitting methods maximise the likelihood of the parameters given the data. The issue arises when the so-called *nuisance variables* have a marginal likelihood that is not trivial to be integrated over the parametric space, also known as individual-level parameters (Berger, Liseo, and Wolpert 1999). In standard Poisson regression models, there are no nuisance variables. Note that treating the deaths as Poisson conditional on exposure time leads to exactly the same estimates (and standard errors) as treating the exposure times as

censored observations from an exponential distribution. This result will be exploited below to link survival models to generalised linear models with Poisson error structure ("Lecture Notes on Generalized Linear Models," n.d.). In generalised linear mixed models, it is straightforward to maximise the likelihood of the model parameters because the nuisance variables (random effects) can be integrated out analytically. However, classical model-fitting methods fail when there are many nuisance variables, and there is no analytical solution. Therefore, Bayesian computational methods facilitate inference in complex models that might have been challenging to analyse using traditional methods. With a model specified in a Bayesian form, with a prior distribution, representing preexisting knowledge, and the likelihood of the data, it is possible to compute the posterior distribution of all the parameters, including the nuisance variables, by Markov chain Monte Carlo algorithms.

2.5.1 Sampling from posterior distribution

MCMC algorithms are a class of methods used for sampling from complex probability distributions. A succinct survey of Markov chain results has been made by Roberts and Rosenthal (2004). Each algorithm has its own strengths, limitations, and applicability to different types of problems. The choice of which MCMC algorithm to use depends on the specific scenario, the structure of the target distribution, and computational considerations. The Gibbs sampler has been the most widely used MCMC algorithm in Bayesian statistics until 2015. The key idea behind Gibbs sampling is that each variable is sampled conditionally on the current values of the other variables. Thus, the joint distribution is effectively constructed by iteratively updating each variable; one variable at a time. Gibbs sampling has also been used to estimate increasingly complex joint models as per Goudie and Mukherjee (2016). For example, those with more than one longitudinal biomarkers (Rizopoulos and Ghosh 2011), two-part longitudinal submodels (Hatfield, Boye, and Carlin 2011), or complex association structures (Mauff et al. 2017).

General-purpose computer programs for Gibbs sampling include BUGS (Lunn et al. 2000)

and JAGS (Depaoli, Clifton, and Cobb 2016). Specific programs for joint modelling that have used Gibbs sampling include JM (Rizopoulos 2010) package implemented in R, and the PROC NLMIXED (Wolfinger 1999) library in SAS. The first implements frequentist joint models that feed the fixed and random effects of the longitudinal submodel into the time to event submodel. The latter allows the fitting of a range of non-linear mixed models. Moreover, the package JMBayes (D Rizopoulos 2014) has attracted growing attention for fitting joint models under a Bayesian framework, with a focus on correcting for non-random dropout. A more comprehensive review of software implementations and computational approaches has been done by Furgal, Sen, and Taylor (2019).

Although there has been an undeniable surge in computational statistics that improve existing modelling methods, there has also been a steep increase in the complexity of biomedical datasets to be analysed. Stan is an innovative, broad purpose probabilistic programming language, with which a user can input a model to receive the corresponding probability distribution. I have employed the Bayesian software package Stan (Carpenter et al. 2017) via `rstanarm` which has introduced a variant of the Hamiltonian Monte Carlo (HMC) (Hoffman, Gelman, and others 2014; Gelman, Lee, and Guo 2015) to estimate posterior distributions.

The Hamiltonian Monte Carlo (HMC) algorithm is typically used to sample the posterior distribution given the data and a model. A salient advantage of HMC is that it updates all parameters simultaneously, whereas algorithms implemented in BUGS (1996) and JAGS (2007) can sample only one parameter at a time. That has been a major bottleneck in the computation of joint models. HMC, albeit being a Markov chain Monte Carlo (MCMC) algorithm, avoids the random walk behaviour and sensitivity to correlated parameters that hound many MCMC programs, by taking a series of steps informed by *first-order gradient information*.

More formally, sampling aims to draw from a density $p(\theta)$ for parameters θ. This is typically a Bayesian posterior $p(\theta|y)$ given data y, particularly a Bayesian posterior coded as a Stan

program. The Hamiltonian Monte Carlo algorithm starts at a specified initial set of parameters θ. In Stan, this value is either user-specified or generated randomly. Then, for a given number of iterations, a new momentum vector is sampled, and the current value of the parameter θ is updated using the leapfrog integrator with discretisation time ϵ and number of steps L, according to the Hamiltonian dynamics. Then a Metropolis acceptance step is applied, and a decision is made on whether to update to the new state (θ^*, ρ^*) or keep the existing state.

Therefore, HMC focuses only on explicit regions of the parameter space, using its adaptive variant, the No-U-Turn Sampler (NUTS). Efficient NUTS features allow the algorithm to converge to the target distributions much more quickly than random walk Metropolis or Gibbs sampling. The method is built upon a rich theoretical foundation that makes it uniquely suited to high-dimensional settings and non-trivial probability distributions (Betancourt 2017) of latent longitudinal and time to event processes.

NUTS is a recursive algorithm which adapts the path lengths in Hamiltonian Monte Carlo. It stops when it hits a U-turn in the trajectory and when there is divergence. In practice, there is an upper limit of how many trees it doubles (the default is ten usually). When it reaches the maximum tree depth and still has not got a U-turn, the user gets a warning. Usually, this is an indication that NUTS is taking too small steps, resulting in poor exploration of space.

In short, the main benefit of Stan is that it uses gradient information to find its way around the posterior. It uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior, as per Betancourt and Girolami (2013); Neal (2012). An approximate Hamiltonian dynamics simulation based on numerical integration is employed, which is then corrected by performing a Metropolis acceptance step.

Therefore, the objective is to eliminate the nuisance variables. The problem of computing the posterior distribution of the data given the parameters is dealt with by using Bayesian techniques that specify a posterior distribution by integrating the random effects out. Therefore, the approximation of the posterior happens by sampling from a distribution that resembles

the distribution of the parameters.

The frequentist approximation is typically equal to maximising the likelihood of the fixed-effects parameters. An alternative, more efficient method is to employ a Bayesian approach to sample the posterior distribution of the regression coefficients.

2.5.2 Descendants of Gibbs sampling

Markov Chain Monte Carlo algorithms are a broad method for sampling from a probability distribution, and can be used to obtain draws from the joint posterior distribution of model parameters given the likelihood and the prior.

The software most widely used for Bayesian analysis before Stan (which proposed alternatives to Gibbs sampling) were BUGS (Lunn et al. 2000) and JAGS (Depaoli, Clifton, and Cobb 2016) and this contributed to the widespread preference for Gibbs sampling, which was the main estimation algorithm in those software packages.

The issue with Gibbs sampling is that because only one parameter can be updated at a time, mixing of the sampler can be very slow if there is high posterior correlation between the parameters, which usually is true. Since 2015 these earlier methods have been supplanted by a more efficient class of algorithms for sampling a posterior distribution that can update all parameters simultaneously. These algorithms are implemented in the program Stan (Carpenter et al. 2017), making it a very powerful tool. The basic principle is to compute the gradient of the log posterior distribution with respect to all parameters, and to use the gradient to propose an update of the parameters. Physicists developed the basic algorithm for using gradients to propose updates as Hamiltonian Monte Carlo (HMC) (Duane et al. 1987).

Further details of the algorithm are beyond the scope of this thesis. The useful takeaway of such experiment is that when the sampling algorithm used by Stan does not work, this is usually obvious and easily detected by built-in diagnostics. This is another advantage over earlier methods.

Moreover, Stan can also implement an algorithm for drawing approximate samples from the posterior, called Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al. 2015). Variational inference replaces the true posterior distribution with an approximate distribution, with two options:

- Variational inference with independent normal distributions (mean-field approximation)
- Variation inference with a multivariate normal distribution (full rank approximation)

However, these algorithms do not always work reliably.

2.5.3 Diagnostics

The ability to obtain valid samples from the posterior relies on two factors:

- that the algorithm converges from its starting point to the target posterior distribution,
- that the correlation between subsequent draws in the MCMC chain is negligible, so that the complete set of MCMC draws can be considered a random sample from the target posterior.

The amount by which autocorrelation within the chains increases uncertainty in estimates can be measured by a metric called effective sample size (ESS). Then, given independent samples, the Central Limit theorem (CLT) premise bounds the uncertainty in estimates, based on the number of samples N. Given dependent samples, the number of independent samples is replaced with the effective sample size N_{eff}, which is the number of independent samples with the same estimation poIr as the N autocorrelated samples. In practice, the probability function in question cannot be tractably integrated, and thus the autocorrelation cannot be calculated, nor the effective sample size. Instead, these quantities must be estimated from the samples themselves.

To sum up, this section has discussed one practical issue concerning advanced statistical modelling: computing time, i.e., the feasibility of actually applying a particular piece of

software to a large dataset, which in turn incites using scalable Bayesian methods.

2.6 Methods of evaluation of predictive performance: a short recap

This section discusses how predictive performance is typically evaluated in survival analysis. This is typically related to the spread in predictions: how well we can separate low-risk from high-risk subjects. The most common quantitative metric used for model evaluation in survival analysis is the concordance (C)-statistic, the so-called Harell's C-statistic. In time to event analysis, C-statistics are used to quantify the ability of the yielded risk score in discriminating among subjects with different event times. The C-statistic provides a universal assessment of a fitted survival model for the continuous event time, i.e., providing a probability of correctly classifying a case-control pair as such, rather than assessing the prediction of t-year survival for a fixed time (Uno et al. 2011).

The C-statistic and the area under the ROC curve (AUROC), or AUC in short, are identical, for binary outcomes. This is a measure of discriminative ability, where the curve represents the proportion of the correct predictions (true positive rate) over the false positive rate for consecutive cutoffs for the predicted probability of the outcome of interest. If the prediction is higher than a cutoff, the subject is classified as positive, otherwise, as negative. The area under the ROC curve can be interpreted as the probability that an individual who experiences the outcome has been given correctly a higher probability of the outcome, than a randomly chosen subject without the outcome (Hanley and McNeil 1982). As a consequence, the AUC is upper bound by 1.

In this regard, other metrics have been proposed, including the Lorenz curve (Lorenz 1905), which shows the true positive subjects versus total classified as positive, the Gini index, usually given as a summary statistic of the Lorenz curve and shows the total number of outcomes

missed by the cumulative proportion of negative classifications, the discrimination slope, which is equal to the mean difference between outcomes predicted, and Harrell's C-statistic (the most widely used C-statistic), which considers pairs of subjects at risk at any time during follow-up (Harrell Jr et al. 1984). I have primarily focused on measures that are in wide use in medical research nowadays.

Moreover, R^2 is the percentage of variance explained by the model and it is commonly used to quantify the predictability of the outcome. That is, R^2 is the fraction by which the variance of the errors is less than the variance of the dependent variable. In a simple regression model, it is the square of the correlation between the dependent and independent variables. In a multiple regression model, R^2 is determined by pairwise correlations among all the variables (including correlations of the independent variables with each other, as well as with the dependent variable). However, prognostic models usually only have R^2 around 20–30%. This indicates that substantial uncertainty remains at the individual level: we can only provide probabilities, and we are far away from providing certainty on the individual outcome (Altman and Royston 2000; Vachon et al. 2007).

As an example, in a clinical trial study for a new drug development, it might be observed that there are highly variable effects on individual patients, in comparison to standard treatments, and yet have statistically significant benefits in an experimental study of thousands of subjects. That is to say, the amount of variance explained when predicting individual outcomes could be small, and yet the estimates of the coefficients that measure the drug's effects could be significantly different from zero (as measured by low p-values) in a large sample.

Furthermore, it is worth mentioning that the C-statistic is considered independent of the event rate, i.e., the incidence of the outcome, in contrast to R^2 or the Brier score, for example, which makes it quite insensitive to increments in predictive performance when including accurate predictors, such as biomarkers.

The net reclassification improvement (NRI) has been proposed for assessing the incremental

value of adding a biomarker to competing risk prediction models (Pencina, D'Agostino Sr, and Steyerberg 2011; Uno et al. 2013). For time to event and binary data, the NRI for events is the difference in sensitivity (true positive rate), and the NRI for non-events is the change in specificity (true negative rate). Note that an increase in AUC does not necessarily correspond to an increase in both sensitivity and specificity (Van Calster et al. 2014). However, the C-statistic, or AUC, is the most commonly reported measure, as it is insensitive to miscalibration.

2.6.1 Calibration: a key property of predictive modelling

By the term calibration in statistical analysis we mean the degree of the agreement between observed and predicted values. Only when a model is calibrated well enough, we can interpret its outputs as informative, standalone probabilities. Such a task is critical in the sense that we want to thoroughly understand our model's predictions in order to improve its precision, so that is sensitive to correctly predicting true events.

Graphical inspection is a common way to assess calibration. Reliability curves can help us understand whether there is e.g., a general trend in predictions being too extreme, an indication of overfitting, etc. Furthermore, we can plot results for subjects grouped by similar probabilities, creating bins from 0 to 1. This allows us to assess calibration by comparing the mean observed proportions per group to the mean predicted outcome. For example, we can plot the observed outcome in groups defined by quintile or by decile of predictions, as explained later on in 8.2.1. Then, the plot results in a graphical illustration of the Hosmer-Lemeshow goodness-of-fit test. Note that the choice of groups is important for the visual impression of calibration: if small groups are plotted, the variability is usually more considerable.

2.6.2 Evaluating calibration of a Poisson regression model on test data

As described earlier, in survival analysis, it is common to model the occurrence of event as a binary variable over many person-time intervals. The event cannot occur in interval i, if it occurred in any interval prior to i, as it would have been censored at that earlier time. Thus, the probability of an event occurring in interval i is not independent of whether an event occurred in interval $i-1$. In this context, calibration is the statistical consistency between a dataset's predicted number of events and the observed number of events. As theory guarantees, fitting a Poisson regression, or any other model with the likelihood in the exponential family guarantees that the observed and expected numbers of events equate exactly.

Therefore, the expected number of events must agree with the observed number of events when fitting a Poisson regression model to estimate the event rate on the training set, i.e., subjects with known outcomes. This observation showcases the perfect calibration we expect the *trained* model to have and can be established in two ways: 1) by taking the sum of the fitted values of the generated Poisson model, 2) by adding the output values of the build-in `predict()` function in R. Both approaches should return the number of observed events in the training data for the time period the model is fitted on.

The function `predict()` with type = *response* and input a generalised linear model with likelihood from a Poisson distribution returns the hazard λ scaled by the interval lengths t (exposure time), as shown by the following relationship

$$\exp(\log(\lambda) + \log(t)) = \exp(\log(\lambda)) \times \exp(\log(t)) = \lambda \times t.$$

For a more systematic review of performance measures, I refer the reader to the work done by Habbema and Hilden (1981). In summary, many performance measures are related to each

other, and there is a debate about the appropriate metrics for each analysis, with a thorny problem being the concept of calibration, which should receive more attention, especially when externally validating prediction models (Collins et al. 2014).

Chapter 3

Data sources used in this thesis

The following chapter addresses a number of questions, such as what is the source of the biomarker and outcome data, who is the population considered and how did they get to this database.

The main source of diabetes data comes from NHS Scotland's national patient record for diabetes care, called Scottish Care Information - Diabetes Collaboration platform (SCI-Diabetes). All newly diagnosed patients given a diabetes code in primary care, including services provided by general practitioners (GPs), clinics and hospitals have a record created in SCI-Diabetes. Therefore, people are incepted into the database when they first receive a diagnosis of diabetes by any point of contact with the health care system.

Furthermore, all patients are assigned a Community Health Index (CHI) number, which is used as the key identifier on all healthcare record systems across Scotland. This allows linkage of the primary SCI-Diabetes databases to other key sources of data for research purposes.

It is estimated that the coverage of the diabetes population residing in Scotland by SCI-Diabetes is more than 99% nowadays. Such large-scale datasets provide great advantages to researchers.

This linkage has given rise to the Diabetes Research Platform, a population-based cohort of people with diabetes in Scotland, contains data on 528 721 individuals either alive or not observable anymore, primarily between January 1984 and April 2020. The diabetes electronic healthcare record in Scotland was used in some parts of the country since the mid-1990s but did not reach >95% coverage of the population of Scotland until 2006. For my analyses, I have used a dataset which includes 472 648 individuals with T1D entered any time between 2006 and 2020, from whom 32317 have been assigned as having T1D between January 2008 and January 2018, and in whom there were approximately 4 million person-years of follow-up (McGurnaghan et al. 2022).

For the same cohort of individuals, I have used linked datasets including the Scottish Morbidity Records (Harley and Jones 1996) that cover inpatient (SMR01) and outpatient (SMR00) attendances, and deaths data from National Records of Scotland (NRS) (Team 2013).

I have further used the Scottish Renal Registry (Simpson 1993) as outcome data source that includes individuals with established renal failure who receive renal replacement therapy.

Participants' observation period stops in case they leave the country or die.

Examples of recent studies that have utilised the Diabetes Research Platform and electronic health care records provided by SCI-Diabetes include McKeigue et al. (2022); Höhn et al. (2022); Captieux et al. (2021); Prigge et al. (2022), McGurnaghan et al. (2021), O'Reilly et al. (2021). The objectives observed in these studies have varied, ranging from the relation of COVID-19 to T1D, to widening socioeconomic disparities in ketoacidosis incidents.

Furthermore, McGurnaghan et al. (2022) provides a detailed description of the cohort according to the type of diabetes, including the median (IQR) frequency of each measure from 2006 to 2020 and the percentage of missingness.

More specific cohort characteristics are given in the respective chapters that describe each analysis.

The two outcomes studied were incident cardiovascular disease (CVD) and time to end-stage renal disease (ESRD). The longitudinal biomarkers included in the models were glycosylated haemoglobin (HbA_{1c}), which reflects the average blood glucose over the past three months, and estimated glomerular filtration rate (eGFR), which is an estimated index of the overall kidney function. For the national dataset, all biochemistry results were from SCI-Diabetes; eGFR was derived from creatinine results from SCI-Diabetes, and it was then calculated using the CKD-Epi equation, given by the formula 6.1.

There is a slight excess of males in the cohort (55%). The average age during the follow-up period is 47 years for T1D. The average duration of diabetes during the follow-up period is 18 years. On average, people have at least one reading per year for each year of follow-up. Thus, the database is a very rich source of longitudinal trajectories of these characteristics in diabetes.

3.1 Sources of event and biomarker data

3.1.1 CVD analysis

Entry criteria, exclusion criteria: CVD events are captured in the database through hospital admissions, with a discharge code mentioning CVD. The coding system used is the International Classification of Disease version 10 codes (ICD-10). In particular, CVD outcome data were acquired through linkage to the Scottish Morbidity Records, the National Health Service admissions dataset and the death registrations with mention of CVD as a cause of death, held by the General Register Office for Scotland.

CVD was defined as any hospital admission or death due to the following clinical concepts: myocardial infarction, stroke, unstable angina, transient ischaemic attack or peripheral vascular disease, any coronary, carotid or peripheral artery revascularisations, major amputation procedures, or acute coronary heart disease. See the Appendix for the International Classifi-

cation of Disease version 10 codes (ICD-10) and Office of Population Censuses and Surveys Classification of Interventions and Procedures (OPCS-4) codes within this definition. It is important to note that we anticipate a heterogeneous association between biomarkers used for prediction and the endpoints defining CVD. While some biomarkers may show consistent associations across different CVD endpoints, others may exhibit differential associations or contradictory patterns. This could be due to the complexity of CVD, which encompasses a range of conditions and biological pathways.

Overall, 27527 individuals with T1D were included in the final cohort for analysis following the exclusion of 1952 individuals with a previous history of CVD. There were 2790 CVD events in the years 2008-2018.

For the CVD analysis, I have used all participants' test results for HbA_{1c} levels (a test regularly done once or twice a year to see how well patients controlled their blood sugar over the previous three months). Baseline measurements of HbA_{1c} were taken nearest to and prior to the start of the study and at most 24 months before individuals' entry. After entry, HbA_{1c} was updated asynchronously for each subject according to their individualised profiles. All other covariates, including sex, diabetes duration and current age, were also derived from SCI-Diabetes.

3.1.2 ESRD analysis

ESRD is a condition characterised by a significant and permanent loss of kidney function. ESRD is typically diagnosed when an individual's glomerular filtration rate (GFR), a measure of kidney function, falls below 15 mL/min/1.73 m^2. At this stage, the kidney's ability to maintain proper function is severely compromised.

Renal replacement therapy (RRT) refers to the medical interventions used to replace the lost kidney function in individuals with ESRD. It is a life-sustaining treatment that aims to manage further complications and maintain the overall status of patients with ESRD.

In the context of ESRD, RRT has been used as an indicator (or proxy) for the occurrence of the outcome of interest. RRT serves as a marker for the presence of ESRD and is used to define the population of interest in many studies and clinical research related to this condition. In other words, when an individual with ESRD receives RRT, it is seen as an indication that they have reached the stage where their kidney function has significantly deteriorated and requires ongoing treatment to sustain life. By using RRT as an indicator, researchers and clinicians can identify and study individuals who have progressed to the advanced stage of kidney disease requiring intervention.

For the definition of RRT, I have used a multi-step process which looks for one of:

- hospital records of dialysis-related diagnosis/operation, if the person is on drugs, they would usually be on dialysis
- hospital renal transplant diagnosis/operation records
- records of the Scottish Renal Registry

and then takes the earliest of these records as the start date of RRT.

For the ESRD/RRT analysis (terms are used interchangeably as of this point), I have defined the composite outcome comprising (a) initiation of renal replacement therapy, (b) deaths with a mention of renal failure, after chronic renal disease, in the death certificate. In total, there were 799 RRT events, of which 473 were fatal: those 473 individuals are censored after initiation of RRT. No one died before starting RRT among those belonging to the dataset.

The incidence rate of the composite definition of event for renal failure is 799 events over 225598.3 person-years, 4 for every 1000 people. Additionally, I censor the population after the first event in both analyses.

The progression to renal failure was based on estimated glomerular filtration rate (eGFR) measurements of people with T1D, derived from SCI-Diabetes. There is a high capture of the eGFR rate in the database, with at least one measure each year in those with T1D.

It is crucial to implement specific serum creatinine cleaning routines before eGFR is calculated using equations, such as the Modification of Diet in Renal Disease (MDRD) or Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equations. This step is necessary to ensure accurate and reliable eGFR calculations.

Creatinine is a waste product produced during normal wear and tear in the muscles and is usually cleared by the kidneys. Further details on longitudinal eGFR measurements are given in chapter 5.

In the next chapter, I detail the application of joint modelling using `rstanarm` to predict CVD in the T1D cohort. First, I provide some details on how the data were set up for modelling and on the number of events available over the person-years of follow-up. Then, I provide the theoretical background, particularly how `rstanarm` is implemented and the underlying mathematical theory of the models built. I then describe the results of applying the joint models for CVD prediction.

Chapter 4

Application and results of using `rstanarm` to predict CVD risk in T1D

4.1 Data characteristics pertinent to the CVD risk model

The national cohort is a 2019 extract from the Scottish Care Information - Diabetes (SCI-Diabetes) dataset (Livingstone et al. 2012). The platform has been described in chapter 3 and is based on electronic healthcare records of people with diabetes in Scotland.

Briefly, since 2004, SCI-Diabetes has collated national demographic and clinical data for over 99% of people with an assigned diagnosis of diabetes.

Assignment of diabetes and person-time in the database: Every time an individual has a clinical encounter in a diabetes clinic, this is captured in SCI Diabetes, as well as every time they go to the hospital, they are captured in SMR01. If this person dies, it is captured by the Death registry data, and they stop being part of the platform. Additionally, a person stops being observable if they leave the country.

The study cohort comprises adults 18 years and above, who were diagnosed with type 1 diabetes before their 50th year. Those with a previous CVD event were omitted. The study started on 1 January 2008 and ended on 1 January 2018.

The baseline characteristics of the T1D population with CVD outcome data are given in table 4.1.

	Patient cohort			
Variable	Population	In study	CVD	No CVD
Sex				
Male	16615	14821	1316	13505
Female	12981	11662	958	10704
Age (years)				
0-20	18.35 (0.59)	18.25 (0.53)	18.23 (0.37)	18.25 (0.53)
20-50	35.06 (8.53)	38.11 (15.75)	40.35 (7.17)	34.72 (8.52)
50+	59.67 (7.86)	60.59 (8.47)	62.95 (9.09)	59.97 (8.19)
Diabetes duration (years)				
1-5	1.57 (1.75)	1.23 (1.67)	1.56 (1.71)	1.22 (1.66)
6+	19.4 (10.42)	19.26 (10.67)	25.13 (12.42)	18.56 (10.22)
Mean HbA_{1c} (mmol/mol)	Unknown	71.44 (18.46)	77.74 (20.21)	70.86 (18.16)
Mean eGFR (mL/min/1.73 m^2)	Unknown	95.64 (23.3)	78.87 (24.48)	97.22 (22.56)
Follow-up (years)	9.02 (2.05)	8.07 (2.91)	5.34 (2.73)	8.32 (2.8)

Table 4.1: Demographics of individuals with CVD outcome data. Those with history of CVD are not included in the study.

A cohort of 26483 subjects with T1D and without prior CVD was defined. Each individual's entry date was defined as the latest of the study start date, date of diagnosis, date of turning 18 years old or date of first coming under observation with T1D in the registry. The end date was defined as the earliest of the study ending date, date of death, date of incident CVD or ceasing to be under observation in the diabetes registry.

There were 2274 CVD events during 199,552 person-years of follow-up. CVD outcome data were established from hospital admissions and death registrations. In particular, CVD outcome data were acquired through linkage to the Scottish Morbidity Records, the National Health Service admissions dataset and the death registrations held by the General Register Office for Scotland. CVD was defined as any hospital admission or death due to myocardial infarction, stroke, unstable angina, transient ischaemic attack or peripheral vascular disease; or any coronary, carotid or peripheral artery revascularisations; or major amputation procedures; or any death due to these conditions; or acute coronary heart disease.

The factors used as predictors for CVD risk include sex, age and diabetes duration at baseline the latter taken as time from T1D diagnosis, routine measurements of HbA_{1c} and eGFR, along with the time of measurement taken with respect to time of entry to the study. Based on reviewing the relevant literature, the choice of covariates for the models of interest aligns with previous modelling endeavours aiming to predict time to CVD. Moreover, this selection emphasises the need to maintain computational efficiency in model fitting. Instead of incorporating all the available information that exists in the platform, the focus is on including a subset of covariates that strike a balance between predictive accuracy and computational burden.

4.2 Joint modelling with a new Bayesian program: `stan_jm()`

At the time of beginning this thesis, a state-of-the-art implementation of the standard approach to joint modelling (based on integrating over biomarker trajectories) had been released as the R function `stan_jm()`, bundled with the R package `rstanarm`, which provides an implementation in Stan of algorithms for fitting mixed-effects regression models (Brilleman 2022). The `stan_jm()` function can be used to fit a joint model (also known as a shared parameter model) for longitudinal and time to event data under a Bayesian framework. The underlying estimation is carried out using the Bayesian C++ package Stan. My first step was to investigate whether this program could overcome the barrier of fitting joint models for time to CVD, in terms of inference reliability, and computational power.

The `rstanarm` package provides users with the capability to fit models using standard R syntax, eliminating the need for manually coding directly in Stan. Its functionality allows for modelling multiple longitudinal outcomes of multiple biomarkers, although being bound to observations measured at the same time points (at individual level), as explained later in this chapter.

4.2.1 Shared-parameter joint model formulation

The longitudinal submodel is a linear mixed-effects model (LMM) aiming to describe the shapes of subject-specific longitudinal trajectories. With continuous biomarker data, I have fitted univariate and multivariate LMMs to model synchronous measurements of HbA_{1c} and eGFR and link these longitudinal profiles to the underlying risk of developing CVD in a large T1D population residing in Scotland.

Typically, a linear mixed-effects model consists of two components: an overall slope that represents the slope of the average individual, and individual-specific random deviations from

the overall slope. LMMs generally have random intercepts, random slopes, or both. The overall slope refers to the average change in the outcome variable associated with a unit change in the predictor variables for the population as a whole, accounting for individual variability, i.e., it takes into account both the fixed effects and the random effects (individual-specific effects) in the model. The random effects are typically assumed to follow a multivariate Gaussian distribution. This assumption allows for capturing the individual-specific variations around the fixed effects. It generally allows for flexibility in specifying the correlation structure among the random effects.

The survival submodel utilises the time to event data, typically using a parametric model. Joint estimation of these submodels is conditional on the assumption that they are correlated via individual-specific parameters, i.e., individual-level random effects (Brilleman et al. 2018). The joint model combines the longitudinal submodel and the survival submodel by incorporating the random effects from the longitudinal submodel into the survival submodel.

The LMM specifies a correlation structure among observations belonging to a specific individual, as opposed to observations from others, because of the inclusion of both fixed and subject-specific (random) effects. Since the longitudinal submodel explicitly expresses the co-dependency among covariates, it can better capture the trajectory of each predictor by considering not only past measurements of that predictor itself but also past measurements of correlated risk factors, and these can be at different times. From a high-level perspective, the joint model allows for modelling the correlation among multiple longitudinal processes, such as multiple repeated measurements or multiple longitudinal biomarkers by specifying a correlation structure that accounts for the dependencies between these processes. This correlation structure can be modelled through the covariance matrix of the random effects.

The longitudinal component of a joint model gives one a rigorous way of modelling the entire trajectory of a longitudinal biomarker-covariate through time. Time-updated measurements of covariates can be included in conventional survival models, but this approach is ad-hoc, as

one needs to specify how the time updates are dealt with.

A *demonstration* of the `rstanarm` joint model formulation follows, as it has been a central component of my approach to joint modelling. Notation and descriptions are sourced from the package documentation and the relevant vignette (Brilleman 2022).

4.2.1.1 Joint model components

Let $x_{ijm}(t) = x_{im}(t_{ij})$ correspond to the observed value of the $m^{th} \quad (m = 1, ..., M)$ biomarker for individual $i \quad (i = 1, ..., N)$ at time point t_{ij}, $j = 1, ..., n_{im}$. A (multivariate)[1] generalised linear mixed model is specified that assumes $x_{ijm}(t)$ follows a distribution in the exponential family with mean $\mu_{ijm}(t)$.

The linear predictor is defined as

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

where $\mathbf{u}_{ijm}^T(t)$ and $\mathbf{z}_{ijm}^T(t)$ are covariates, including age at baseline, sex, diabetes duration at baseline, baseline HbA_{1c}, time-updated HbA_{1c} and eGFR, which likely include some function of time with population (fixed) $\boldsymbol{\beta}_m$ and individual-specific parameters $\boldsymbol{b}_{im}$, respectively, and g_m is a known link function.

The vector $\boldsymbol{\beta} = \boldsymbol{\beta}_{m;m=1,...,M}$ denotes the population-level parameters across the M longitudinal submodels.

We further assume

[1]We specify a longitudinal submodel for each biomarker we want to model. The distribution and link function may differ among the M longitudinal submodels. It is assumed that the dependence, i.e., the correlation across the various longitudinal biomarkers is expressed through a shared multivariate normal distribution for the individual-specific parameters.

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

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

Furthermore, let $T_i = \min(T_i^*, C_i)$ be an event time, where T_i^* denotes the true event time for subject i and C_i denotes the censoring time. The event time may be unobserved.

Thus, the event indicator $d_i = I(T_i^* \leq C_i)$ takes the value 1, if an event is observed and 0, if the subject is censored. The hazard of the CVD event is specified using a parametric proportional hazards regression model of the form:

$$h_i(t) = h_0(t; \boldsymbol{\omega}) \exp(\mathbf{w}_i^T(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \sum_{q=1}^{Q_m} f_{mq}(\boldsymbol{\beta}, \boldsymbol{b}_i, \alpha_{mq}; t))$$

where $h_i(t)$ is the hazard of the event for subject i at time t, $h_0(t; \omega)$ is the baseline hazard at time t given parameters $\boldsymbol{\omega}$, $\mathbf{w}_i^T(t)$ denotes the individual-specific covariates with a vector of regression coefficients $\boldsymbol{\gamma}$, i.e., log hazard ratios, $f_{mq}(.)$ are a set of known functions for $m = 1, ..., M$ and $q = 1, ..., Q_m$[2], and the α_{mq} are regression coefficients (log hazard ratios). The correlation between the longitudinal and survival processes is captured via shared random effects.

The survival probability of individual i still being event-free at time t

$$S_i(t) = \text{Prob}(T_i^* \geq t) = \exp(-H_i(t))$$

where $H_i(t) = \int_{s=0}^{t} h_i(s)ds$ is the cumulative hazard for individual i.

We assume the baseline hazard $h_0(t; \omega)$ is modelled parametrically. In the `stan_jm()` modelling

[2]The association structure may differ among biomarkers.

function, the baseline hazard might be specified as either: an approximation using B-splines on the log hazard scale (the default); a Weibull distribution, or an approximation using a piecewise constant function on the log hazard scale (sometimes referred to as piecewise exponential).

Currently, limited post-estimation functionality is available for models estimated with a piecewise constant baseline hazard, so this is the least preferable choice. Parametric baseline hazard used for the event submodel: cubic b-splines approximation estimated for the log baseline hazard (default option among Weibull distribution, piecewise constant baseline hazard). Furthermore, by default, the degrees of freedom are set to six.

4.2.1.2 Association structure and joint likelihood

The longitudinal and event submodels are linked via a set of functions that may each be conditional on the population-level parameters from the longitudinal submodel $\boldsymbol{\beta}$, the individual-specific parameters $\boldsymbol{b}_i$ and the population-level parameters α_{mq} for $m = 1, ..., M$ and $q = 1, ..., Q_m$. The α_{mq} are referred to as the association parameters since they quantify the strength of the association between the longitudinal and event processes.

The `stan_jm()` modelling function allows for the following association structures:

- Current value (of the linear predictor or expected value)

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

- Current slope (of the linear predictor or expected value)

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

- Area under the curve (of the linear predictor or expected value)

$$f_mq(\boldsymbol{\beta}, \boldsymbol{b}_{im}, \alpha_{mq}; t) = \alpha_{mq} \int_0^t \eta_{im}(u)du$$

- Interactions between different biomarkers (for some m = m' or m $\neq$ m')

$$f_mq(\boldsymbol{\beta}, \boldsymbol{b}_{im}, \alpha_{mq}; t) = \alpha_{mq}\eta_{im}(t)\eta_{im'}(t)$$

- Interactions between the biomarker (or its slope) and observed data (for some covariate value $c_i(t)$)

$$f_mq(\boldsymbol{\beta}, \boldsymbol{b}_{im}, \alpha_{mq}; t) = \alpha_{mq}c_i(t)\eta_{im}(t)$$

- Lagged values for any of the above. That is, replacing t with $t - u$ where u is some lag time, such that the hazard of the event at time t is assumed to be associated with some function of the longitudinal submodel parameters at time $t - u$.

More than one association structure can be specified; however, not all possible combinations are allowed.

The joint likelihood of the shared-parameter model is evaluated by computing the area under the curve of the estimated hazard rate through time. When the hazard rate is evaluated at a set of time points chosen to give an approximation to the area under the curve is known as quadrature estimation. Gauss-Kronrod quadrature with Q nodes is used to approximate the necessary integrals and ultimately evaluate the cumulative hazard and the survival probability (Bianconcini 2014). The accuracy of the numerical approximation can be controlled using the number of quadrature nodes, specified through the `qnodes` argument in `stan_jm()`. Using a

higher number of quadrature nodes will result in a more accurate approximation. However, this estimation becomes very challenging from a computational point of view, due to the large vector of random effects involved in the numerical integration of the density of the survival outcome. When the linear predictor is time-fixed, there is a closed-form expression for both the hazard rate and survival probability in almost all cases (the single exception is when B-splines are used to model the log baseline hazard). When there is a closed-form expression for both the hazard rate and survival probability, there is also a closed-form expression for the (log) likelihood function. When the linear predictor is time-varying, there is not a closed-form expression for the survival probability, and this, in turn, creates the requirement for quadrature approximation. As the number of biomarkers increases, the number of points required for quadrature scales exponentially, thus becomes computationally intractable.

4.3 Findings: computational tractability

The endeavour to establish a concrete pipeline of how to fit a shared-parameter joint model on longitudinal and survival outcomes in the T1D population dataset using `rstanarm` (version 2.19.3) proved futile. Despite some successful fitting on a reduced-scale dataset that utilised a single biomarker outcome[3], leveraging large-scale data from such a rich database with this implementation ultimately was to no avail.

The subsequent section encapsulates a substantial amount of trial and error, which, albeit challenging, offered valuable insights and enriched my understanding of feature engineering and what would comprise a streamlined analysis. These findings could serve as an exemplary case of this type of modelling, that could guide future attempts in similar modelling attempts. While it has been evident that extensive debugging and settings refinement were necessary, it is important to acknowledge that the current framework represents the backbone of a streamlined analysis, and holds the potential to provide valuable guidance for modelling and

[3]Intermediate results are shown in embedded notebooks in Chapter 4 and the Appendix.

scaling up endeavours in the future.

I aimed to compare the fit and performance of the more sophisticated implementation of the joint model to a simpler Poisson model for the CVD hazard, where the same biomarker observation is used until the next one becomes available. A joint model, on the other hand, explicitly specifies a longitudinal trajectory for the biomarker, to prudently specify the rate of event.

To initiate the analysis, this joint model implementation requires the different longitudinal biomarkers to be measured on the same date, which regrettably results in significant data loss. Furthermore, this implementation is inevitably computational cumbersome due to the quadrature calculations. Running times ranged from days to weeks, even for the simple parameterisations (current value) and for a reduced number of individuals as input. Parallelisation of the code helped to reduce the running time. However, as these models are also memory intensive, only a few parallel processes could be run at any one time using a powerful server with 2Tb of memory.

To better understand how `rstanarm` functionality works, in particular the `stan_jm()` and `posterior_survfit()` functions and to diagnose convergence and other sampling issues, I analysed various subsets of the data (10%, 20%, 40%, 50%) to acquire a deeper understanding of the procedure, and also limit running time and memory usage. On every subset, the ratio of events was maintained to represent the entire population[4].

The format of the data is one row per observation for each individual. Since biomarker data are updated asynchronously, this translates into intervals of varying length for each individual.

Every new observation of an individual teaches the survival submodel that the subject is still event-free at that time point. Furthermore, working with various subsets of the dataset, in addition to the entire dataset has allowed me to assess what sample sizes would be sufficient for specifying the model parameters.

[4]https://github.com/IoannaThoma/PhD/tree/main/code.

Working with a smaller dataset on which model fitting would complete quickly, simplified debugging and model diagnostics. I experimented with gradually increasing sample sizes; however, running times of posterior sampling remained a stumbling block, as it did not increase linearly with the number of individuals. The `rstanarm` methodology is designed to calculate the random effects of both the training and testing datasets, in order to produce the latent biomarker values at various time points needed for the specification of the survival function of each subject. For example, 17.5 hours were needed for fitting a model on a subset of 2633 subjects, 35.9 hours for 5265 subjects, and more than 68 hours were needed for 10532 subjects. Various edge cases needed special treatment, such as subjects who had limited data, despite being followed up full-time.

Furthermore, the sampling of the posterior distribution has been very CPU and memory intensive, taking up weeks to converge even when parallelised.

This heavily depends on model parameterisation and the number of longitudinal data. I provide some more details regarding my computational experience in the following sections.

The modelling function `stan_jm()` (version 2.19.3) ran four randomly initialised Markov chains, each for 2000 iterations (including a warm-up period of 1000 iterations that are discarded), in its default settings. Several important checks (not mixing chains, trace plots, etc.) are necessary to ensure that there are no problems with the MCMC procedure used to get the samples.

An exemplar case of fitting a joint model with `stan_jm()` is presented below. The subsequent embedded notebook pages showcase the process of fitting the model on a scaled-down dataset. It demonstrates several model fitting procedures, allowing for a better understanding of the methods and libraries employed.

Fitted model outputs and diagnostics

Ioanna Thoma

2023-07-03

1 Methods and results

The models shown below use the same 10% of the data.

1.1 Continuous-time approach to joint modelling, using the `rstanarm` package

We use `stan_jm()` to fit a univariate joint model to the longitudinal biomarker HbA_{1c} and time to CVD event. A linear mixed model is specified for the biomarker with an individual-specific intercept and slope. The event model includes age, sex, duration of diabetes, and a 3-year average HbA_{1c} prior to entry as baseline covariates. The log hazard of CVD at time t is modelled as dependent on the current value of the biomarker.

Table 1: Runtime in minutes of fitting a stan_jm() model with 2000 iterations (default option)

	warmup	sample	total
chain:1	150.5	85.9	236.4
chain:2	155.5	85.7	241.2
chain:3	156.3	86.1	242.4
chain:4	121.6	89.7	211.3

```
kable(s.base10jm2000[grep("Event|Assoc", rownames(s.base10jm2000)),
                     c(1, 3, 9)],
      digits = c(2, 2, 0), "simple",
      caption = "Continuous-time joint model fitted with stan_jm()")
```

Table 2: Continuous-time joint model fitted with stan_jm()

	mean	sd	n_eff
Event|(Intercept)	-9.99	0.39	4634
Event|genderFemale	-0.23	0.13	6956
Event|entry.age	0.06	0.01	4390
Event|diabetes.duration	0.03	0.01	4474
Event|entry.hba1c.med	0.02	0.00	5317
Event|b-splines-coef1	-1.84	0.42	7300
Event|b-splines-coef2	-0.78	0.46	6297
Event|b-splines-coef3	-1.42	0.43	5510
Event|b-splines-coef4	0.16	0.36	5648
Event|b-splines-coef5	-0.74	0.38	6109
Event|b-splines-coef6	0.13	0.33	6299

	mean	sd	n_eff
Assoc|Long1|etavalue	0.03	0.01	7239

1.2 Continuous-time approach to joint modelling, using the CmdStanR package

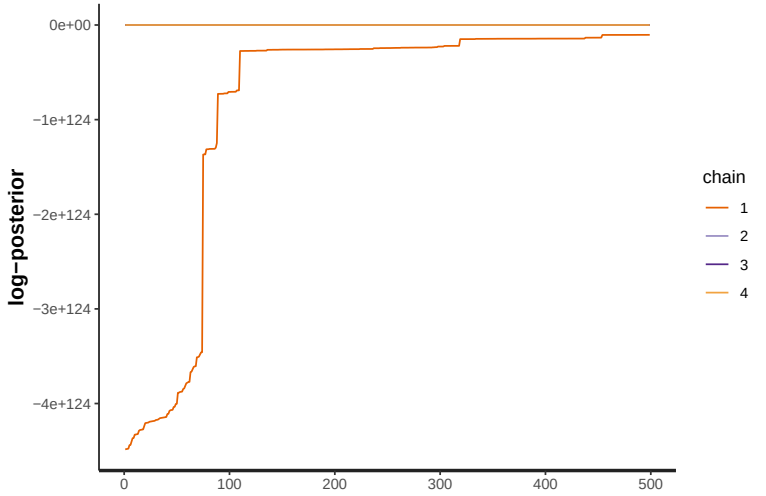
We use CmdStanR to see if we can obtain any speed up. The cmdstan_model() function creates a new CmdStanModel object from a file containing the Stan program. Under the hood, CmdStan is called to translate a Stan program to C++ and create a compiled executable.

```
newcmdstan.no.par.chain <- FALSE
if(newcmdstan.no.par.chain) {
load("~/jm.standata10.Rdata")

inits.jm <- get_inits(base10jm2000$stanfit) # list with one component per chain

## within-chain parallelisation disabled
stan_file <- file.path("/tmp/ioanna/mini-jm-stan-code-extracted.stan")

mod.cmdstan <- cmdstan_model(stan_file,
                     include_paths = "./rstanarm/src/stan_files",
                     force_recompile = TRUE,
                     cpp_options = list(CPPFLAGS = "-Wno-ignored-attributes"))

fit10.no.par.chain <- mod.cmdstan$sample(data = standata,
                         iter_sampling = 500, iter_warmup = 500,
                         adapt_delta = 0.9,
                         init = inits.jm,
                         chains = 4,
                         max_treedepth = 10L,
                         threads_per_chain = 1)

stanfit10.no.par.chain.cmdstan <- rstan::read_stan_csv(fit10.no.par.chain$output_files())
saveRDS(stanfit10.no.par.chain.cmdstan, file = "stanfit10.sin.cmdstan.rds")
}

## diagnostics on stanfit object
kable(get_elapsed_time(stanfit10.no.par.chain.cmdstan), "simple",
      caption = "Runtime in seconds of CmdStanR sampler with 500 iterations
                 (1 thread/chain)")
```

Table 3: Runtime in seconds of CmdStanR sampler with 500 iterations (1 thread/chain)

	warmup	sample
chain:1	144.798	174.923
chain:2	9324.200	17226.200
chain:3	139.170	162.015
chain:4	11509.800	2553.940

```
rstan::traceplot(stanfit10.no.par.chain.cmdstan, pars = "lp__")
```




```
sampler_params <- get_sampler_params(stanfit10.no.par.chain.cmdstan,
                                     inc_warmup = FALSE)
mean_energy_by_chain <- sapply(sampler_params, function(x) mean(x[, "energy__"]))
print(mean_energy_by_chain)
```

```
## [1] 8.289247e+123  1.626018e+05  1.622088e+22  1.625996e+05
```

```
print(summary(do.call(rbind, get_sampler_params(stanfit10.no.par.chain.cmdstan)),
              digits = 2))
```

```
##  accept_stat__    treedepth__      stepsize__      divergent__  n_leapfrog__
##  Min.   :0.67   Min.   : 1.0   Min.   :0.0000   Min.   :0    Min.   :   1
##  1st Qu.:0.97   1st Qu.: 2.0   1st Qu.:0.0000   1st Qu.:0    1st Qu.:   3
##  Median :1.00   Median : 7.0   Median :0.0022   Median :0    Median : 127
##  Mean   :0.98   Mean   : 5.2   Mean   :0.0151   Mean   :0    Mean   : 291
##  3rd Qu.:1.00   3rd Qu.:10.0   3rd Qu.:0.0172   3rd Qu.:0    3rd Qu.:1023
##  Max.   :1.00   Max.   :10.0   Max.   :0.0561   Max.   :0    Max.   :1023
##     energy__
##  Min.   : 1.6e+05
##  1st Qu.: 1.6e+05
##  Median : 6.9e+20
##  Mean   :2.1e+123
##  3rd Qu.:2.6e+122
##  Max.   :4.5e+124
```

```
sum.stats <- summary(stanfit10.no.par.chain.cmdstan)$summary[, c(1, 3, 4, 8:10)]
summary(sum.stats[, 6]) # Rhat
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.999    1.715    2.917    5.492    4.836 8599.027        2
```

```
keep.rows <- grep("^[almy]",  rownames(sum.stats))
```

```
kable(sum.stats[keep.rows, ], "simple",
      digits = c(4, 4, 4, 4, 4, 4),
      caption = "Continuous-time joint model fitted with CmdStanR without
      using within-chain parallelisation")
```

Table 4: Continuous-time joint model fitted with CmdStanR without using within-chain parallelisation

	mean	sd	2.5%	97.5%	n_eff	Rhat
yGamma1[1]	3.659040e+01	3.605140e+01	-2.85000e-01	73.1625	2.0021	163.5002
yAux1_unscaled[1]	1.615000e+00	2.219800e+00	1.05900e-01	5.4256	2.0020	8599.0272
a_z_beta[1]	-5.687000e-01	8.224000e-01	-1.79580e+00	0.2431	2.0037	33.1223
a_beta[1]	-8.430000e-02	1.219000e-01	-2.66200e-01	0.0360	2.0037	33.1223
yBeta1[1]	2.915100e+00	1.641600e+01	-1.86937e+01	27.5292	2.0032	40.4836
yBeta1[2]	2.096100e+00	2.349600e+00	-2.20200e-01	5.2137	2.0021	135.2098
yBeta1[3]	-2.618200e+00	2.689800e+00	-6.38680e+00	-0.0129	2.0021	130.9517
yBeta1[4]	9.018200e+00	9.603100e+00	-3.95900e-01	21.5867	2.0022	164.6894
yAux1[1]	1.511758e+02	2.077917e+02	9.91190e+00	507.8670	2.0020	8598.7697
yAuxMaximum	1.511758e+02	2.077917e+02	9.91190e+00	507.8670	2.0020	8598.7697
mean_PPD[1]	3.632670e+01	3.598710e+01	-3.23510e+00	72.3948	2.0041	30.1686
yAlpha1[1]	-3.831530e+01	1.312323e+02	-2.34604e+02	82.5193	2.0020	225.9296
lp__	-2.050368e+123	7.709868e+123	-3.88956e+124	-159833.9500	10.8984	1.3189

1.3 Using within-chain parallelisation feature of `CmdStanR`

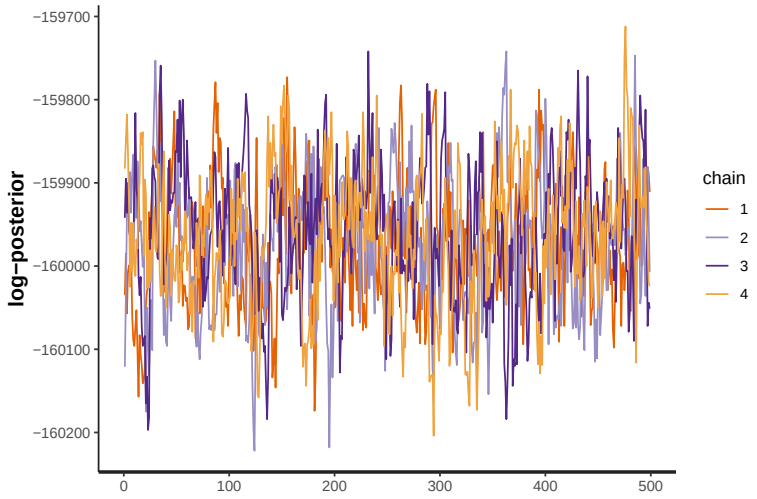
```
newcmdstan.threads2 <- FALSE
if(newcmdstan.threads2) {
load("~/jm.standata10.Rdata")

inits.jm <- get_inits(base10jm2000$stanfit)
stan_file <- file.path("/tmp/ioanna/mini-jm-stan-code-extracted.stan")


mod_threads2 <- cmdstan_model(stan_file,
                    include_paths = "/tmp/ioanna/rstanarm/src/stan_files",
                    force_recompile = TRUE,
                    cpp_options = list(CPPFLAGS = "-Wno-ignored-attributes",
                                       stan_threads = TRUE))
fit_threads2 <- mod_threads2$sample(data = standata,
                  iter_sampling = 500, iter_warmup = 500,
                  adapt_delta = 0.95,
                  init=inits.jm,
                  chains = 4,
                  max_treedepth = 10,
                  threads_per_chain = 2)

stanfit_threads2_cmdstan <- rstan::read_stan_csv(fit_threads2$output_files())
saveRDS(stanfit_threads2_cmdstan, file = "stanfit_threads2.cmdstan.rds")
}
```

Table 5: Runtime in seconds of CmdStanR sampler with 500 iterations (2 threads/chain)

	warmup	sample
chain:1	5853.15	1375.81
chain:2	4730.41	2048.94
chain:3	5147.23	2028.51
chain:4	4320.60	2265.03

```
stanfit_threads2_cmdstan <- readRDS("stanfit_threads2.cmdstan.rds")

kable(get_elapsed_time(stanfit_threads2_cmdstan),
      caption = "Runtime in seconds of CmdStanR sampler with 500 iterations (2 threads/chain)")
```

```
rstan::traceplot(stanfit_threads2_cmdstan, pars = "lp__")
```




```
sampler_params_threads2 <- get_sampler_params(stanfit_threads2_cmdstan,
                                              inc_warmup = FALSE)
mean_energy_by_chain_threads2 <- sapply(sampler_params_threads2,
                                        function(x) mean(x[, "energy__"]))
print(mean_energy_by_chain_threads2)
```

```
## [1] 162618.4 162618.0 162601.0 162603.1
```

```
print(summary(do.call(rbind, get_sampler_params(stanfit_threads2_cmdstan)),
              digits = 2))
```

```
##  accept_stat__   treedepth__    stepsize__     divergent__  n_leapfrog__
##  Min.   :0.67   Min.   :6.0   Min.   :0.051   Min.   :0    Min.   : 63
```

```
##  1st Qu.:0.92   1st Qu.:7.0   1st Qu.:0.052   1st Qu.:0    1st Qu.:127
##  Median :0.97   Median :7.0   Median :0.054   Median :0    Median :127
##  Mean   :0.95   Mean   :6.8   Mean   :0.056   Mean   :0    Mean   :117
##  3rd Qu.:0.99   3rd Qu.:7.0   3rd Qu.:0.057   3rd Qu.:0    3rd Qu.:127
##  Max.   :1.00   Max.   :7.0   Max.   :0.064   Max.   :0    Max.   :255
##     energy__
##  Min.   :162289
##  1st Qu.:162544
##  Median :162610
##  Mean   :162610
##  3rd Qu.:162673
##  Max.   :162899
```

```
sum.stats.threads2 <- summary(stanfit_threads2_cmdstan)$summary[, c(1, 3, 4, 8:10)]
summary(sum.stats.threads2[, 6]) # Rhat
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.9980  0.9988  0.9991  0.9993  0.9997  1.0181       2
```

```
keep.rows <- grep("^[almy]",  rownames(sum.stats.threads2))
```

```
kable(sum.stats.threads2[keep.rows, ], "simple",
      digits = c(4, 4, 4, 4, 4, 4),
      caption = "Continuous-time joint model fitted with CmdStanR using
      within-chain parallelisation (2 threads/chain)")
```

Table 6: Continuous-time joint model fitted with CmdStanR using within-chain parallelisation (2 threads/chain)

	mean	sd	2.5%	97.5%	n_eff	Rhat
yGamma1[1]	72.6634	0.3086	72.0797	73.2626	354.4502	1.0139
yAux1_unscaled[1]	0.1066	0.0004	0.1058	0.1074	1683.3623	0.9995
a_z_beta[1]	0.1830	0.0363	0.1091	0.2507	3310.2875	0.9982
a_beta[1]	0.0271	0.0054	0.0162	0.0372	3310.2830	0.9982
yBeta1[1]	1.3824	0.6259	0.1486	2.5844	363.3005	1.0038
yBeta1[2]	-0.1725	0.0282	-0.2262	-0.1163	278.0027	1.0027
yBeta1[3]	-0.0655	0.0333	-0.1292	0.0009	266.5422	1.0074
yBeta1[4]	-0.3053	0.0558	-0.4170	-0.1979	332.8192	1.0101
yAux1[1]	9.9763	0.0383	9.9008	10.0490	1683.2880	0.9995
yAuxMaximum	9.9763	0.0383	9.9008	10.0490	1683.2880	0.9995
mean_PPD[1]	72.2811	0.0687	72.1508	72.4161	1980.4671	0.9988
yAlpha1[1]	81.0261	0.9343	79.1556	82.8537	437.2903	1.0037
lp___	-159966.7915	78.4372	-160120.0250	-159813.9750	316.2811	1.0063

Furthermore, I have tested the options offered in `rstanarm` for variational inference, but the parameter space in this application is too complex for the VI algorithms to return meaningful results. Although VI is useful in some applications for linking a longitudinal component to a survival process, variational methods are not sufficient.

To fix convergence issues, I tried to increase the number of iterations and the value of the depth of the tree used by the NUTS sampler, as explained in 2.5.1. The step size gets adapted during the first iterations, the so-called adaptation phase (also known as burn-in in Gibbs sampling). Iterations had to increase twice from 1000 to 3000 and from 3000 to 4000 to get a sufficiently effective sample size, and the argument `adapt_delta` got from 0.8 to 0.99.

To diagnose issues with model fitting, I performed various checks using the `bayesplot` package and the ShinyStan app. The `bayesplot` package provides various plotting functions for graphical posterior predictive checking, creating graphical displays comparing observed data to synthetic data from the posterior predictive distribution (Gabry et al. 2017). The idea behind posterior predictive checking is that if a model is a good fit, then it should be able to generate data that look like the observed data.

In the next analysis step, I used the `posterior_traj()` function, which generates an estimated subject-specific longitudinal trajectory of the biomarker used in model fitting. These values are obtained from the generated posterior distribution. Since we know the true values of the biomarker from the data, we can assess whether these trajectories are realistic or not.

The `posterior_survfit()` function was used next to generate estimated survival probabilities based on draws from the posterior predictive distribution. The survival probabilities are conditional on an individual's random effects. Therefore for test data, the default behaviour is to sample new group-specific coefficients for the new individuals using a Monte Carlo scheme that conditions upon their longitudinal biomarker data. I, therefore, needed to restrict the testing set size in the interest of running time. Furthermore, I extended the functionality to generate survival estimates in arbitrary time points, e.g., survival probability at the end of

each year of follow-up, rather than estimating survival at times with received information only[5].

In light of the scaling issues I was having, I turned my focus to using `CmdStanR` for model fitting (the second approach featured in the notebook). `CmdStanR` is a lightweight shell interface to Stan for R, that utilises the C++ toolchain. `CmdStanR` (version 0.3.0.9) contains the `sample()` function for model fitting, which requires the input data to be given in a format readable by Stan. I extracted the Stan code used for the MCMC sampling within the `stan_jm()` function of `rstanarm` to use it as a base for modifying the model's parametrisation and prior choices. However, in the given state, the Stan code was not exceptionally readable due to include-statements and the large number of options it supports. Various compiling issues and getting familiar with the new toolchain gave me a better understanding of the inner workings of `rstanarm` and mainly how the model parameters are handled.

To extract the input data used by the `stan_jm()` model, I had to modify the `rstanarm` code itself to store a record of the data, and I subsequently, plug them into the data argument of the `sample()` function in order to compare the `sample()` and `stan_jm()` outputs. It was observed that the `CmdStanR` sampler ran very fast, giving the wrong impression that the sampling was performed correctly, but it was actually due to an insufficient exploration of the parametric space. Therefore, the resulting posterior density was not necessarily reliable and accurate.

Unfortunately, the `standata` object lacked essential initial values, such as prior means, which are crucial for the sampler to effectively specify the joint likelihood. In order to ensure the proper functionality of the sampler, it was necessary to provide initial values that closely approximated the region where the majority of the full posterior distribution was located. These initial values were retrieved from the corresponding `stan_jm()` model that had converged to a target distribution.

[5]https://github.com/IoannaThoma/PhD/tree/main/code.

On the one hand, the `stan_jm()` function has a built-in routine to supply initial values (default option init = "prefit") by fitting the longitudinal and event submodels separately and then using values from these to set initial values for sampling the joint likelihood.

However, the `CmdStanR` does not have a built-in routine to replicate this; thus, the initial values for each chain needed to be supplied manually. Thus, I extracted the initial values from the `stanfit` object created by `stan_jm()` to provide them as argument to the modelling method of `CmdStanR`. Since `sample()` and `stan_jm()` share the same Stan code, using the init = "prefit" option when generating the `standata` was also valid for `CmdStanR`.

Running `stan_jm()` with four chains for fifty iterations, I saved the final values of each chain and provided them as initial values to the `sample()` method of the `CmdStanR` model. Using the initial values computed from `stan_jm()`, the `CmdStanR` modelling function passed all diagnostics. Displayed below are two figures illustrating the diagnostic assessment of the fitted generalised linear mixed model.

In figure 4.1, the parameter `alpha` refers to the intercept of the regression model. It represents the average response, given that all other predictors are set to zero. Parameters `beta[1]` and `beta[2]` refer to the coefficients or parameter estimates associated with specific predictors in the model. The values of these coefficients indicate the magnitude and direction of the relationship between each predictor and the response variable. The parameter `sigma_indiv` represents the standard deviation associated with the random effects in the fitted linear mixed model. It quantifies the variability, i.e., dispersion of the individual-specific random effects, capturing variations among subjects. Lastly, `lp` refers to the log probability, and `energy` refers to the energy of the Hamiltonian system in HMC sampling.

Figure 4.2 shows the densities of the parameters and provides information about the likelihood of different values occurring for each variable. The mean of a parameter's density represents the expected value of that parameter. No issues have been observed with the parameters in this instance.

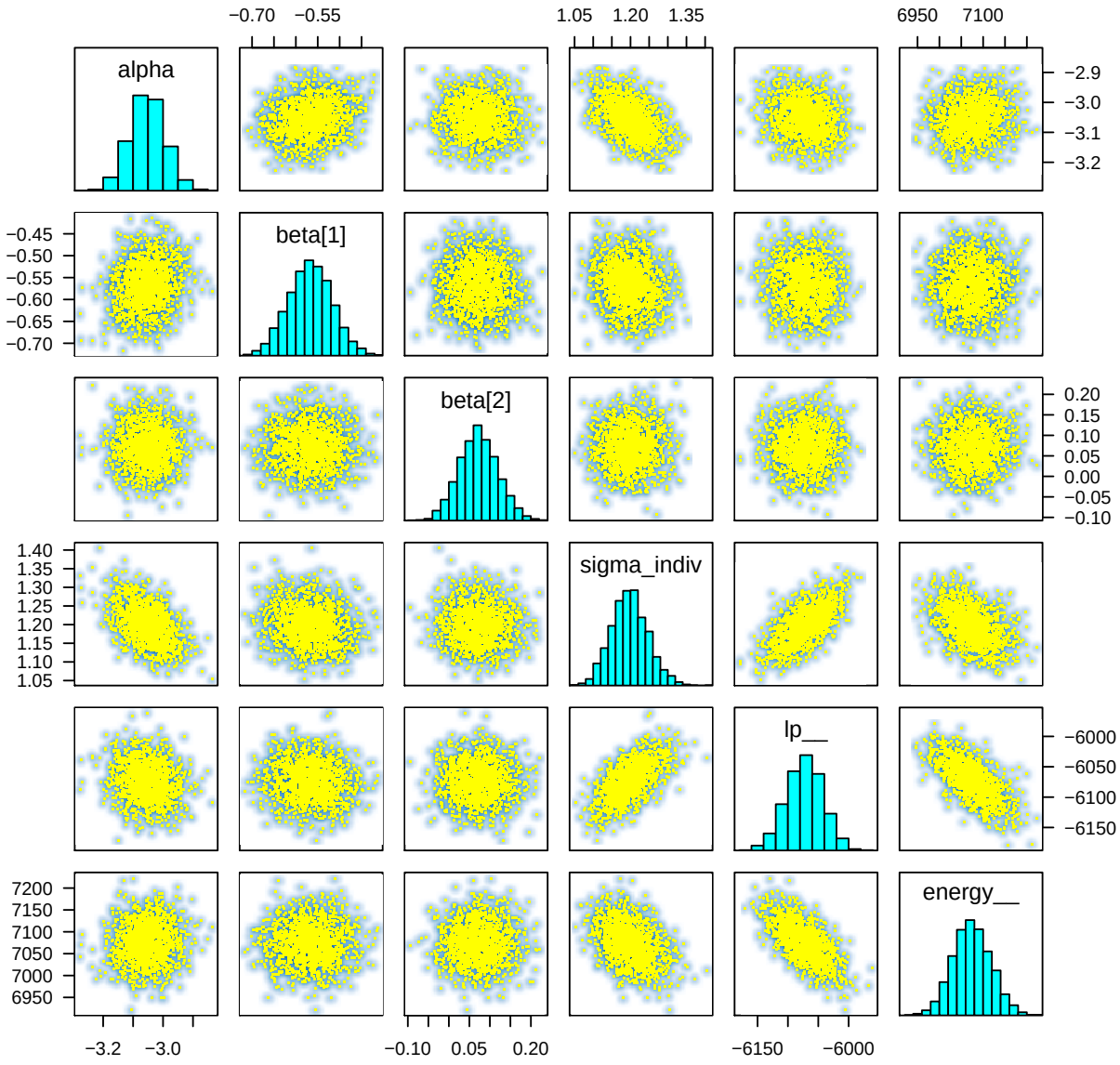


Figure 4.1: 6x6 grid of the parameters of the linear mixed model.

It appeared that the memory problem with the `rstanarm` package, in which memory usage spikes during the adaptation phase, was overcome by using `CmdStanR`. However, the limited sample size was prone to divergent transitions, and sampling was failing. On the other hand, the four chains do not mix at all and have very different average energy. A 500 warm-up and 500 sampling iterations were completed in 3 hours, meaning the sampler was not stuck in one parameter space as before.

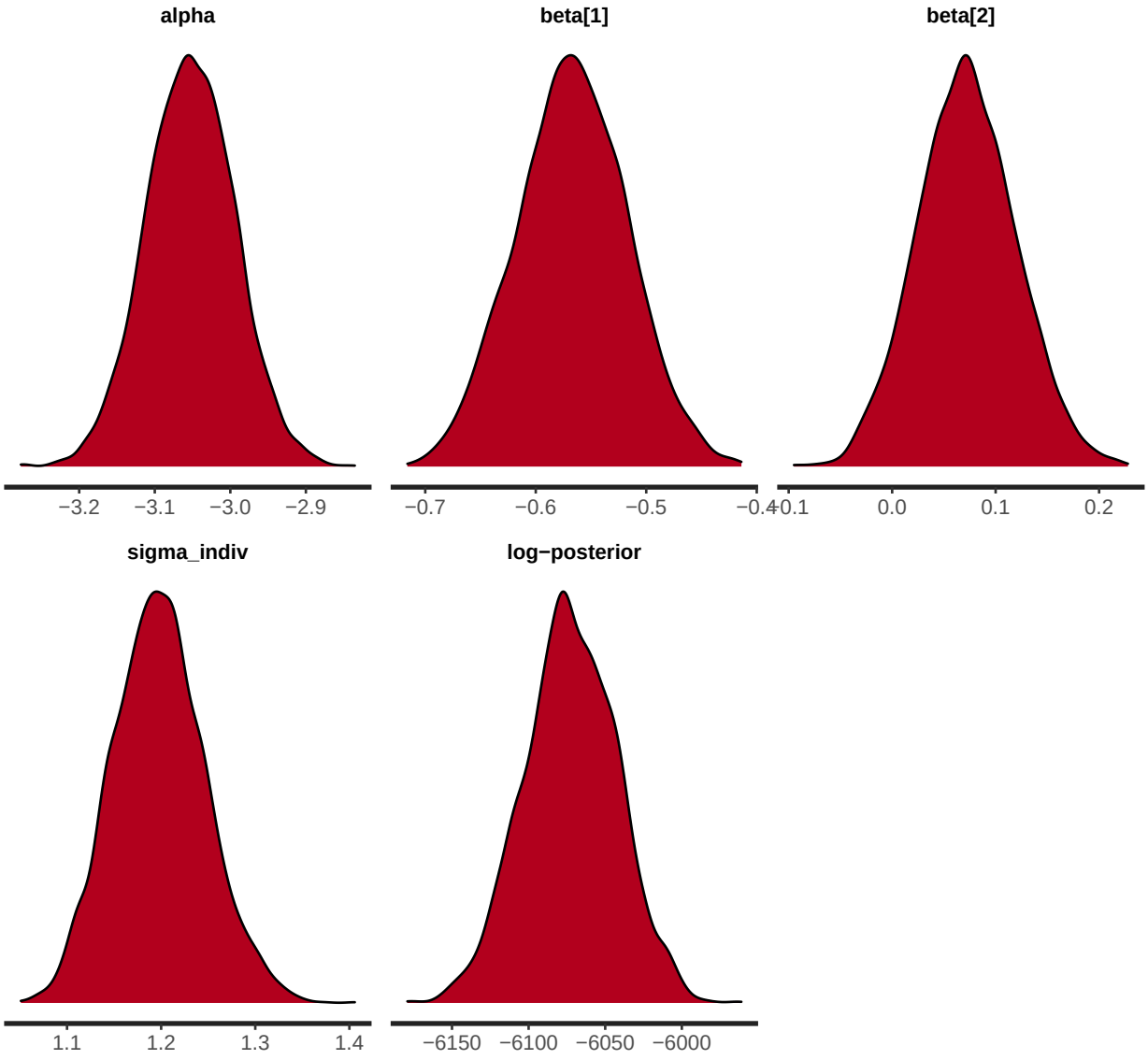


Figure 4.2: Probability density functions associated with each parameter in the linear mixed model.

To evaluate the meaningfulness of the posterior quantities, I needed to make the posterior summaries intelligible by relabelling the variable names. By comparing coefficients, I worked out which variables from one match the other. Not all parameters from `CmdStanR` would be relevant, as not all have a match. The interesting ones were the regression coefficients and those were relatively easy to reconcile.

With default settings on the example dataset PBC (Mayo Clinic Primary Biliary Cholangitis data) provided within `rstanarm`, the model takes about three minutes to run with `stan_jm()`. However, with `CmdStanR`, without using within-chain parallelisation, some chains complete in a few seconds, while others take about two minutes.

It is possible that the chains that run fast were not able to explore the parameter space efficiently as happened before, and therefore they were not exploring the entire posterior distribution. As a rule of thumb, in this situation, the $\hat{R}$ diagnostic (which compares the between- and within-chain estimates for model parameters) should help, and so should a trace plot of the log posterior.

Furthermore, the fitted model with `CmdStanR` could not be processed by the `posterior_survfit()` function for obtaining posterior predictions as this requires a `stan_jm()` object, and it does not work with a `stanfit` object. The current implementation would need to be extended in order to accommodate this requirement, which could be explored and addressed in future iterations[6].

As a temporary solution to enable the use of the existing functionality; `posterior_survfit()` from `rstanarm` for posterior predictions, I manually copied the `stanfit` object from `CmdstanR` into the object saved by `stan_jm()` to generate posterior predictions. However, that was not a viable solution since `CmdStanR` parallelisation for gradient calculation over many parallel computations gave unrealistic timings for larger datasets.

Through further trial and error, I established that there is no easy way to overcome the computational limitations of `stan_jm()`, and even porting it to `CmdStanR` did not speed it up at the time.

Upon inspecting the Stan code and identifying the key steps, it appears that matrix multiplication is the computationally intensive operation. Rather than looping over individuals, based on the nature of parallel operations typically performed on GPUs, compiling for GPU support

[6]Intermediate results are shown in embedded notebooks in Chapter 4 and the Appendix.

was deemed more likely to be successful than using the `reduce_sum()` chunking option. This is because GPUs are specifically designed to handle parallel computations efficiently, and leveraging GPU support can often lead to improved performance.

I, therefore, ran the `sample()` function with >1 thread/chain to see if there was any speedup from some undocumented parallelisation under the hood. Utilising within-chain parallelisation can only yield noticeable improvements if multithreading is effectively employed. Multithreading refers to the concurrent execution of multiple threads within a single process, allowing for parallel computation and potentially faster execution of matrix multiplication, which seems to be the barrier in this type of joint modelling.

`CmdStanR` was evaluated with 10 threads per chain, with the expectation that parallelised matrix algebra calculations would be performed noticeably faster. However, with the GPU support, the mixing of the sampler was regrettably unreliable ($N_{eff} = 2$ and $\hat{R}$ values > 100). It was probable that the retained samples, after discarding the burn-in period, still exhibited a significant correlation with their respective previous draws.

The presence of high correlation among retained samples in a MCMC simulation can have both positive and negative implications. On one hand, the correlation between samples indicates that the chain is exploring the parameter space effectively and not getting stuck in local regions. This suggests that the chain is mixing well and providing a representative sample from the target distribution.

On the other hand, high correlation can impact the efficiency of the MCMC algorithm. Highly correlated samples indicate that subsequent samples do not provide much new information beyond what has already been sampled. This can result in slower convergence, increased autocorrelation, and longer runtime.

Therefore, it is important to assess and diagnose the correlation structure of the MCMC samples to determine the impact on the analysis. In my situation, the trace plot of the log posterior and the very high $\hat{R}$ values indicated that the default `CmdStanR` settings have not

converged. Using GPU parallelisation might raise issues if there are errors or bugs in the GPU code or if the GPU hardware or software environment is not properly set up or compatible with the Stan program, causing the program to misbehave.

Hence, it is important to note that the specific implementation and configuration of GPU parallelisation can influence the behaviour of a Stan program in applications like this. Some of the potential issues that may arise with GPU parallelisation include:

- Compatibility: Ensuring that the GPU hardware and software environment is compatible with the Stan program and its requirements. This includes verifying that the GPU device is supported, drivers are up to date, and necessary libraries are installed.

- Implementation errors: Writing GPU code can be more complex than traditional CPU code, and errors in the implementation can lead to unexpected behaviour and unreliable results. Gaps in implementation were the most probable reason for failing at the time I attempted to use it.

- Memory management: GPU parallelisation requires careful management of memory resources. Incorrect memory allocation, deallocation, or data transfers between CPU and GPU can cause issues such as memory leaks, crashes, or performance degradation.

Another point of interest is that within-chain parallelisation is not given as an option to the user. In fact, the user needs to rewrite parts of the Stan code to use the `reduce_sum()` method. The next step was to parallelise the loop over individuals that increments the log posterior using `reduce_sum()` to split it into the necessary chunks.

Using `CmdStanR` for fitting the joint model resulted in a considerably slower sampling process; with an approximate rate of 100 iterations per hour when applied to the 10% dataset. Nonetheless, this method was advantageous as it avoided confinement to a limited region, as observed previously. Ideally, parallelisation of the sampling would be desirable to achieve a significant speedup in computation.

4.3.1 Implications

In summary, one should be able to run almost any model with a single biomarker with parallelisation.However, the quadrature step in this joint modelling implementation scales exponentially with the number of biomarkers, hampering modelling efforts that include more than two biomarkers.

The primary limitation of using the `CmdStanR` method for sampling the posterior distribution of the joint model is that the user needs to reformat the output from `CmdStanR` as a `rstanarm` object, in order to be able to use `rstanarm` functionality for posterior predictions of the longitudinal outcome(s) and of survival/time to event.

In summary, the state-of-the-art implementation of `stan_jm()` function of the `rstanarm` package does not perform computationally efficiently, and alternative of porting the code into the `CmdStanR` functionality for sampling, which would allow for GPU parallelisation, required a lot of code manipulation and made us lose the convenient functions already implemented in `rstanarm` for posterior predictions. Given that the time savings were neither guaranteed nor very promising, and `CmdStanR` was too limited to help improve the efficiency of `stan_jm()`, it was clear that the primary use of the `stan_jm()` joint model would be to establish a benchmark against which I would compare other predictive models.

To recapitulate, it was found that there were severe limitations in the application of joint modelling to this set of data, including:

- the computation of multivariate joint models is not that straightforward, albeit mathematically sound and comprehensive,
- the need to drop available data to satisfy the requirements of the specific implementation in case of a multivariate submodel,
- the variance-covariance matrix for random effects becomes more complicated as the number of longitudinal processes increases, and

- using a quadrature-based method is poorly suited to numerical computation, as computational cost then rises exponentially with the number of biomarkers.

It is important to consider these implications when fitting joint models, as the number and estimation of individual-level parameters relies on both the number of individuals (sample size) and the number of longitudinal responses. With a larger sample size, there is more information to estimate the parameters accurately. As the number of observations per individual increases, it provides more data points for estimating individual-specific trajectories. Consequently, the precision and reliability of estimating individual-level parameters tend to improve with an increased sample size.

Similarly, the number of longitudinal measurements per individual affects the number of individual-level parameters. If there are more longitudinal measurements available, it allows for a more detailed characterisation of individual trajectories. This may lead to a higher number of individual-level parameters needed to capture the variation and heterogeneity in longitudinal profiles accurately, which translates into an increased computational burden. Hence, there exist certain limitations and factors that necessitate careful consideration when employing joint modelling techniques. These drawbacks and issues include the following:

Choosing the appropriate model structure and selecting relevant covariates can be challenging in joint modelling. The selection process requires careful consideration of the relationships between the longitudinal and time to event processes, as well as consideration of potential confounding factors. Having a sufficient amount of data, to obtain reliable estimates. If data are sparse or subject to missingness, the performance of joint models may be compromised.

Furthermore, like any statistical model, joint models are based on certain assumptions about the input data. Violations of these assumptions, such as non-linearity or non-normality, can affect the model's validity and the quality of its results.

Therefore, extracting meaningful and actionable insights from joint modelling outputs requires careful consideration and interpretation. It is important to note that these disadvantages

are not inherent to all joint modelling implementations, and they can vary depending on the specific methods and datasets being used. Researchers should carefully evaluate these limitations and consider their implications before applying joint models in their analyses.

In light of the limited predictive performance observed in CVD risk models, I deemed it necessary to select an outcome that would be more conducive to developing a robust and high-performing joint model. Predicting CVD was proven challenging due to the complex interactions among various risk factors, however, this joint modelling approach is not reliable to handle and capture multiple longitudinal responses. Moreover, CVD often has a long latency period, during which individuals may remain asymptomatic. Hence, risk prediction models need to account for this extended period and accurately capture the early signs or risk indicators that may precede the onset of clinical symptoms. Therefore, I shifted to using a Bayesian sequential updating approach based on time-splitting. In the next chapter, I go on to describe the theory underpinning the model set up for the Bayesian time-splitting approach in more detail.

To that end, I opted for an outcome where the biomarkers are expected to demonstrate greater predictive power and a stronger prior association, namely predicting time to renal replacement therapy (RRT) from longitudinal eGFR data. This deliberate choice serves two purposes: first, it allows for a more pronounced demonstration of the development of a more effective joint model, and second, it leverages the inherent relationship between the selected outcome and the biomarker to enhance the model's predictive capabilities.

Chapter 5

Theory and development of the Bayesian updating time splitting approach with `ctsem`

The chapter is structured as follows:

1. Gaussian state-space models (section 5.1)
2. The Kalman filter, an efficient algorithm for updating the imputed biomarker values (section 5.2)
3. Fitting a joint model by sequential updating (section 5.3)
4. Hierarchical Gaussian state-space models, implemented in the R package `ctsem` (section 5.4)

This chapter describes a reformulation of the joint model based on sequential Bayesian updating, developed and evaluated as an alternative approach to the `stan_jm()` function included in the `rstanarm` package. Instead of trying to model the time to event by integrating over the unobserved trajectories of the biomarkers, this approach is based on splitting follow-up

time into short intervals and fitting a Poisson regression model where the covariates are the predicted values of the latent biomarkers at the start of each person-time interval, based on all observations up to that time point.

The motivation for considering this alternative approach is that it offers a wider functionality for specifying the longitudinal component and latent process of the joint model, and it is a computationally efficient method that can scale to large datasets with multiple biomarkers.

In the previous chapter, I showed the limitations of applying joint modelling to HbA_{1C} and CVD outcome data, using functionality implemented in `rstanarm`, where the standard joint modelling approach can only fit a generalised linear mixed model to the biomarker data. The most widely used class of models is a linear mixed model with individual-varying intercepts and slopes. With the sequential updating approach, we can fit a broader family of models known as Gaussian state space models or continuous-time structural equation models. These models generalise the linear mixed model to include terms for autoregressive drift and diffusion (Voelkle et al. 2012).

P. McKeigue (2022) has recently elaborated on the concept of using a sequential updating approach to jointly model biomarker and time to event outcomes, providing a demonstration of the theory using publicly available data, in particular, the so-called Mayo Clinic Primary Biliary Cholangitis (PBC) dataset (Dickson et al. 1989). The implementation he discusses is based upon recently emerged software, namely the package 'ctsem' (Driver, Oud, and Voelkle 2017). This work has been a major source of inspiration for my research and has expanded my research horizon providing valuable insight of survival modelling.

In the following sections, I am placing this methodological construction within the appropriate context for my setting, considering the relevant factors and conditions that are necessary for its comprehension and application to using the Kalman filter for the longitudinal component and then proceeding to the prediction of time to event within a sizeable dataset of individuals with T1D.

5.1 Continuous-time structural equation modelling

Structural equation modelling (SEM) represents the dependencies between multiple variables, usually including unobserved (latent) variables, by equations.

A continuous-time model is a model within which variables that evolve over time are modelled.

Consider a continuous-time model with observed variables X, latent variables L, and error terms E. The structural equation model can be represented using path coefficients (λ) (i.e., the connection strength) and error variances (θ) (Hair Jr et al. 2021). A simplified representation of a continuous-time model in SEM is given by the equations:

$$X = \Lambda_X L + E_X$$

$$L = \Lambda_L L + E_L$$

In this formulation, X represents the observed variables, L represents the latent variables, and E_X and E_L represent the corresponding error terms. Λ_X and Λ_L denote the path coefficients that represent the relationships between the underlying latent and observed states of a longitudinal process.

Although continuous-time models have a long history (Coleman and others 1964; Hannan and Tuma 1979), their use in the biological sciences has been sparse, in part due to a lack of suitable software to specify and estimate continuous-time models. Many applications of dynamic systems modelling (Izzo and Vecchio 2007) (Grijalva et al. 2007) are limited to discrete-time constructions, conditional on the assumption that the time intervals between measurements and/or subjects are constant. Although this is not common in observational studies, intervals of non-varying lengths are more commonly achievable, when desired, in

clinical trials.

In the broader literature on longitudinal data analysis, continuous-time modelling, which explicitly accounts for the timing of measurements, has several important advantages over discrete-time modelling, see, e.g., Oud and Delsing (2010); Voelkle et al. (2012); Deboeck and Preacher (2016). Continuous-time modelling can make full use of the information contained not only in the observations themselves, but also in the exact timing of the measurements. By using stochastic differential equations to estimate the underlying process, continuous-time models allow for any pattern of measurement occasions, e.g., irregularly spaced intervals.

The key idea of continuous-time structural equation modelling (CTSEM) is to model a stochastic process for how the latent variables evolve over time.

The growing availability of health records has fuelled the interest in continuous-time models, because they are inherently well-suited to handling asynchronous measurement occasions and enabling comparisons of studies with different time intervals between observations (Hecht and Zitzmann 2020). However, their use has yet to be widely adopted, and for the most part, this is attributable to the lack of suitable software to fit efficiently continuous-time state-space models. Although a range of R packages deals with stochastic differential equation modelling, most implementations are applied to a single subject applications. These include sde (Iacus 2007), yuima (Brouste et al. 2014), SIM.DiffProc (Guidoum and Boukhetala 2020), and POMP (King, Nguyen, and Ionides 2015).

Linear mixed models are a special case of this more general family of models for longitudinal data known as continuous-time dynamic models or state-space models. In the context of SEM, models can be expanded to allow for the estimation of measurement error through the use of multiple indicator latent factors. The generalised mixed model which is a combination of fixed and random effects can be considered as one where the latent factors underlying the repeated measures reflect the fixed and random effects associated with stability and change of the repeated measures over time (Curran 2003). Importantly, this family of models can be

extended to include autoregressive drift and diffusion, and this is implemented in `ctsem`.

5.1.1 Drift process

An autoregressive drift model can be represented as follows:

$$Y_t = \lambda Y_{t-1} + \epsilon_t + \delta_t$$

In this equation, Y_t represents the observed variable at time point t, Y_{t-1} represents a lagged value at time point $t-1$, λ is the autoregressive coefficient, ϵ_t is the error term, and δ_t represents the drift term. To estimate the autoregressive drift model within the SEM framework, we would like to specify the paths between the observed variables at different time points, including the autoregressive paths (λ) and the drift paths (δ). Additionally, we need to include the error terms (ϵ) and specify a constant variance structure based on the dependencies in the data (Draper and Smith 1998; Pardoe 2020).

Autoregressive drift can be seen as a mean-reverting process, as it depicts the overall trend of a family of random variables (a stochastic process) to drift back to their grand average. It primarily captures gradual, smooth changes in a process in which the properties might have been changed so that there is a tendency for the changing process to move back towards a central location (of high concentration), with a greater force when the process is away from the centre.

5.1.2 Diffusion process

In the context of SEM, a diffusion process refers to a specific type of model that captures the spread or diffusion of influence or information through a network of variables. Mathematically, a diffusion process can be represented using equations that describe the change in each variable over time, taking into account the influences from other variables. These equations often involve autoregressive terms and time-varying coefficients.

A *diffusion process* is a stochastic process that describes a trajectory that consists of a succession of random steps, such as a random walk. Such a process might be the solution to a stochastic differential equation. The one-dimensional random walk can be seen as a continuous-time Markov chain. In the diffusion process resulting from a random motion, there is movement from a region of high concentration to a region of low concentration. A familiar example is the scent of a flower that quickly permeates the still air of a room. The concept could be incorporated into the structural modelling of longitudinal biomarker data.

Over long follow-up times, models that allow for diffusion (random walk in continuous time) are more realistic than models that only allow a fixed slope with time. Such models have been shown to give a better fit for specifying the longitudinal eGFR (Diggle, Sousa, and Asar 2015).

This family of models can be specified with the stochastic differential equation (Voelkle et al. 2012):

$$d\eta(t) = (\mathbf{A}\eta(t) + \mathbf{b})dt + \mathbf{G}d\mathbf{W}(t)$$

where A is autoregressive effect (drift), b is slope, G scales the diffusion process $W(t)$

with measurement errors generated by

$$\mathbf{x} = \mathbf{\Lambda}\eta + \epsilon$$

The matrix $\mathbf{\Lambda}$ specifies the loadings of the observed biomarkers $\mathbf{x}$ on the latent variables η.

The $\mathbf{A}$ (autoregressive) matrix encodes autoregressive effects on the diagonal and cross-lagged effects (the directional effects between variables at different points in time) off the diagonal.

The vector $\mathbf{b}$ specifies the slopes of the biomarkers with time. In the stochastic differential equation, it appears as the intercept. With individual-specific random effects for the intercepts

and slope, we have a hierarchical state-space model.

The Cholesky factor matrix $\mathbf{G}$ scales the Wiener process $\mathbf{W}$. This diffusion process is the limiting form of a discrete-time random walk. $d\mathbf{W}(t)/dt$ is Brownian motion.

Several other models used for longitudinal data can be viewed as special cases of this hierarchical continuous-time dynamic model:

- With $\mathbf{A} = 0$, $\mathbf{G} = 0$, and $\mathbf{\Lambda} = \mathbf{I}$ we have a linear mixed model
- With $\mathbf{A} = 0$, $\mathbf{b} = 0$, and $\mathbf{\Lambda} = \mathbf{I}$ we have a diffusion-only model as used by (Diggle, Sousa, and Asar 2015) to model longitudinal measurements of kidney function.
- With $\mathbf{A} = \mathbf{I}$, $\mathbf{G} = 0$, $\mathbf{\Lambda} = \mathbf{I}$ and $\epsilon = 0$ (no measurement error), we have a last-observation-carried-forward model.

In summary, to fit a state-space model, the Kalman filter makes a forward pass through the data to compute the state probability distribution at each time point, conditional on all observations up to that time point. This technique allows us to estimate the imputations of the latent variables at the start of each time interval (Luo 2018) and thus implementing the Poisson time-split joint model.

5.2 Kalman filter: a message-passing algorithm

Sequential updating requires an efficient algorithm for imputing the values of the latent biomarkers at each time point recursively, conditional on all observations up to that time point.

For any Gaussian state-space model (i.e., `ctsem`), the Kalman filter algorithm can be used to impute the biomarkers. The Kalman filter takes as input some noisy measurements and tries to infer from those measurements of the possible state of a latent part of the stochastic process (MacKay, Mac Kay, and others 2003; Russell, Norvig, and Davis 2009). The Kalman

filter consists of two main steps: the prediction step and the updating step. The prediction step estimates the state of the system at the next time point, based on the previous state estimate and the system dynamics. The updating step incorporates new measurements into the state estimation, adjusting the estimate based on new data and the uncertainty associated with the measurements.

Kalman filter was first applied to a wide range of tracking and navigation problems and it is much applied in time series analysis (Durbin and Koopman 2012; Commandeur, Koopman, and Ooms 2011; Petris, Petrone, and Campagnoli 2009; Hyndman and Athanasopoulos 2018). Defining the filter in terms of state-space models simplifies the implementation of the filter in the discrete scale, another reason for its usefulness to our application, which depends on discrete person-time intervals (time split).

Message-passing algorithms help to construct a solution to a global problem by splitting the calculation into smaller parts that are easier to specify. Kalman filters are widely used to implement inferences for Gaussian state-space models. The Kalman filter provides a prediction about the next state being x, integrating over the error covariance matrix, p_x, representing the uncertainty of the state estimate. Therefore, the estimate of the next latent state is updated accordingly every time to reflect what is learnt. At every time point, the Kalman filter computes the probability distribution of the latent variables, conditional on all observations up to that point. At the next time point, this distribution is combined with the new observations to generate an updated distribution. We can obtain multiple imputations of the latent states from this distribution or a single best estimate. A graphical representation of updating the estimate of a parameter vector β if additional data become available is shown in figure 5.1:

An n-dimensional dynamic random process can be modeled as follows. A vector difference equation

$$x_{k+1} = \Phi_k x_k + u_k,\ k = 0, 1, 2, ...$$

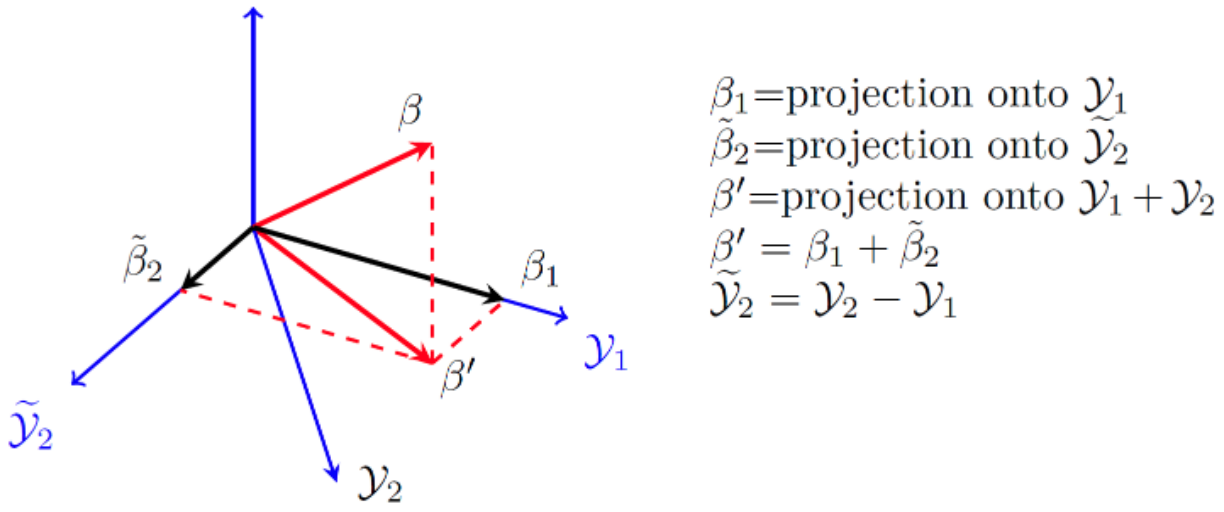


Figure 5.1: A visualisation of the parameter vector β being projected onto different subspaces of a Hilbert space $\mathcal{H}$. Intuition: Given new data, the updating is based on the part of the new data that is orthogonal to the old data.

which defines how the random vector x_k change over time (Masnadi-Shirazi, Masnadi-Shirazi, and Dastgheib 2019).

Here, x_k is an n-dimensional state vector where each component is a random variable, Φ_k is a known $n \times n$ matrix, u_k is an n-dimensional input random vector with zero mean such that there is zero correlation between present input at k and past input at l, i.e.,:

$$E[u_k u_l^T] = Q_k \delta_{kl} = \begin{cases} Q_k & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}$$

Taking a wide enough period, the rate at which the process changes might look reasonably constant, whereas when we zoom in, the rate of the changes may look rather substantial. There might be considerable changes in the short term, but the long-term trend is much smoother. Moreover, time intervals of finer granularity lead to a better approximation of the continuous time and enable more precise estimation of dynamic structure parameters, e.g., autoregressive effects, where the current value depends linearly on the past values.

As we build better and better representations incorporating the most recent data, what was once process noise and uncertainty about the parameter values is slowly turned into predictable elements, allowing for a whole range of possibilities for the current states. Although the state-space model will fail to predict, to some extent, genuine fluctuations and incorporate entirely the variance inherited in the latent process generation, the value of continuous-time models is rather high. They are flexible and retrospectively adapt to the gained knowledge from the errors acquired in the data generation, yielding more accurate predictions.

Let us consider data arriving from a process sequentially and wish to update inference on an unknown parameter θ. The joint distribution $f(x_1, x_2, \ldots, x_n|\theta)$ represent the probability distribution of X given the parameters θ, i.e., the likelihood of observing different values of X under specific parameter values θ. The Bayesian prior distribution $\pi(\theta)$ at time n leads to a density for data conditional on θ as

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta)f(x_2|x_1, \theta) \cdots f(x_n|\mathbf{x_{n-1}}, \theta)$$

where we let $\mathbf{x_i} = (x_1, \ldots, x_i)$. Note that we are not assuming $X_1, X_2, X_3, \ldots, X_n, \ldots$ to be independent conditionally on θ. At time n, we update the distribution of θ to its posterior

$$\pi_n(\theta) = f(\theta|\mathbf{x_n}) \propto \pi(\theta)f(\mathbf{x_n}|\theta),$$

where $\pi(\theta)$ is the prior distribution of the set of parameters in θ. From a dynamic perspective, when a new observation arrives, we claim that just before time $n + 1$, our knowledge of θ is summarised in the distribution $\pi_n(\theta)$ so we just use this as prior distribution for the new piece of observation and update the posterior as

$$\pi_{n+1}(\theta) \propto \pi_n(\theta) \times L_{n+1}(\theta) = \pi_n(\theta)f(x_{n+1}|\mathbf{x_n}, \theta),$$

where $L_{n+1}(\theta)$ is the likelihood to observe the data given these values of θ. Hence, we notice

that at time n we only need to keep a representation of π_n and ignore the past. The current π_n contains all the information required to revise knowledge i.e. the likelihood $L(\theta)$ when a new data point arrives becomes $L_{n+1}(\theta)$. This updating is known as recursive.

Hence, we use a Markovian model for the changing states of the parameter θ of the form:

$$f(\theta_0) = \pi(\theta_0), \quad f(\theta_{i+1}|\theta_{\mathbf{i}}) = f(\theta_{i+1}|\theta_i),$$

where $\theta_0, \theta_1, \ldots$ are latent states which are dynamically updated, as new data points arrive:

$$f(x_i|\theta_{\mathbf{i}}, \mathbf{x_{i-1}}) = f(x_i|\theta_i),$$

which suggests that the distribution of the random variable x_i given the parameters $\theta_{\mathbf{i}}$ and the previous observations $\mathbf{x_{i-1}}$ is independent of the previous observations $\mathbf{x_{i-1}}$ conditional on the parameter θ_i. Hence, the joint density of states and observations is given by the formula:

$$f(\mathbf{x_n}, \theta_{\mathbf{n}}) = \pi(\theta_0) \prod_{i=1}^{n} f(\theta_{i+1}|\theta_i) f(x_i|\theta_i).$$

The most common tasks associated with inference about the changing states of θ are

- Filtering: specify the current state, $f(\theta_n|\mathbf{x_n})$
- Prediction: specify the next state, $f(\theta_{n+1}|\mathbf{x_n})$
- Smoothing: specify the past state at time k, $f(\theta_k|\mathbf{x_n}), \quad k < n$

This approach is traditionally attributed to Kalman from a result in 1960 (Kalman 1960), but was, in fact, fully described by the Danish statistician T.N. Thiele in 1880 (Thiele 1880) (Hald 1981), where he modelled sequential data consisting of a sum of a regression component, a Brownian motion and a white noise, and derived the procedure known as Kalman filtering; a recursive construction that evaluates the regression component and estimates the Brownian

motion (Lauritzen 1981).

Therefore, the Kalman filter is an efficient algorithm for generating updated predictions of the hidden states (biomarkers) from a Gaussian state-space model. The fitting of a Gaussian state-space model and the use of the Kalman filter to generate imputations is implemented in the R package `ctsem`.

The next step involves using Poisson regression models, including the underlying true value of the biomarker as a time-varying covariate as generated for the Kalman filter at the start of each person-time interval.

Furthermore, a last-observation-carried forward model was evaluated as a baseline for comparison of predictive performance.

5.3 Fitting a joint model by Bayesian sequential updating

Two-stage joint modelling approaches in which a model for the biomarker trajectories is fitted first, and the results are plugged into a model for the time to event in the second stage have been shown to be incorrect, because the likelihood does not factor over these two stages: the likelihood of the parameters for biomarker trajectories depends on the observed events. This may introduce severe bias (Henderson, Diggle, and Dobson 2000; Sweeting and Thompson 2011). This type of bias in survival analysis is known as 'immortal time bias'. This bias can lead to distorted estimates of the association between exposure and outcome.

Immortal time is a time period during which the outcome of interest cannot occur due to the study design or the exposure definition. This can happen when the exposure is defined based on a time-dependent event, or when the follow-up time is not correctly recorded. During this immortal time, individuals may appear to have a lower risk of the outcome solely because they have not yet reached the point where the outcome could occur.

The bias arises because the immortal time is inadvertently included in the exposed group but excluded from those censored during the analysis. This creates an artificial difference in the follow-up times between the two groups, leading to biased estimates of the exposure-outcome association (Yadav and Lewis 2021).

Existing approaches to joint modelling of time to event and longitudinal biomarker observations in a single model are based on modelling time to event as the response variable. With a time-varying covariate like the trajectory of a biomarker, the likelihood of the model parameters given time to event has to be evaluated by computing the hazard rate at a set of time points chosen to give an approximation to the area under the curve: a procedure known as quadrature. As the number of biomarkers increases, the number of points required for quadrature scales exponentially and model fitting becomes computationally intractable (Mauff et al. 2020). I have been able to fit joint models of two biomarkers, so up to a point it is tractable, albeit computationally intensive.

Sequential updating is an alternative way to capture the coupling between the longitudinal and event submodels. In the early 1980s, it was shown that instead of modelling the time to event for a survival analysis, the events can be modelled as a Poisson arrival process, in which each individual is censored at the first event. Even though the observations in each person-time interval are not independent draws from a Poisson distribution, they can be modelled as if they were (Rodriguez 2007).

With time split into many short time intervals, the hazard rate can be modelled as constant within each interval: this is equivalent to a survival curve that is piecewise exponential. With time-updated covariates, any survival curve can be modelled using Poisson regression.

To eliminate immortal-time bias, at each person-time interval the likelihood must depend only on observations made up to the beginning of the interval. With sequential Bayesian updating at each time an observation arrives, the posterior distribution of parameters from the last update becomes the prior for the next update. The log-likelihood of the model is

accumulated as a sum over updates at the start of time intervals.

Let η_i denote the value of the *latent state* and x_i the observed value. At each update, the probability $p(\eta_i)$ is conditioned on no event or censoring having occurred up to the ith time point, as shown below

$$\text{revised distribution} \propto \text{current distribution} \times \text{new likelihood}$$

$$p(\eta_i \mid \eta_{i-1}) = p(\eta_i \mid \eta_{i-1}, y_{i-1} = 0)$$

In the i-th interval, we observe whether an event occurs, and compute the likelihood of the event occurrence $y \in \{0, 1\}$ as the average of $p(y_i \mid \eta_i)$ over the distribution given by $p(\eta_i)$. If no event occurs, we can update the longitudinal model using the biomarker observation that arrives at the $(i + 1)$-th time point.

$$p(\eta_{i+1} \mid x_{i+1}, \eta_i, y_i = 0) \propto p(\eta_{i+1} \mid \eta_i) \times p(x_{i+1} \mid \eta_{i+1})$$

Therefore, for this type of model (observed x_i conditional on unobserved η_i that evolve as a Markov process) there is an efficient algorithm for computing the likelihood of the model and the probability distribution of η_i at each time point, conditional on all observations up to that point, by a forward pass through the data.

The likelihood of the model is the probability of the observations $x_1, \ldots, x_n$ given the model (eliminating the η_i).

Each interval of a Poisson process may be perceived as a Bernoulli trial, which is either a success or a failure.

The latent state is assumed to evolve as a Markov process; in other words that probability distribution at time i depends only on value at time point $i - 1$. The observed values x_i depend only on the latent state η_i at that time point. We do not have to condition on the

event, because no event has occurred. As soon as an event occurs, the observations stop.

The Poisson distribution probability mass function gives the likelihood λ of observing k events in a time period, given the length of the exposure and the average events per unit of time. The rate parameter λ can be thought of as the expected number of events in an interval T.

If the rate of occurrence is μ per unit of time, then the number of incidents is Poisson with mean $\lambda = \mu \times T$. To get the hazard rate, we take the log $\mu = X \times \beta$, using the Poisson generalised model's coefficients β from 2.1.

The models used for prediction are based on the inclusion of a landmark time point, up to which individuals need to be observable and after which all biomarker observations are deliberately censored in a 2-fold cross-validation approach. I have set the landmark time to be five years since the subject's entry into the study. The interest lies in quantifying the increment in risk prediction when more rigorous models that explicitly account for the time of each measurement are considered for specifying the longitudinal trajectory of a biomarker, using it as a benchmark the last-observation-carried-forward model.

The performance of the constructed Poisson models is assessed on previously unseen data, provided that the imputed biomarker values are strictly taken at the start of the prediction interval, and these prediction intervals are not truncated, but they run to the end. The rationale for that is that the time of the event is unknown in the test data. Model discrimination is quantified using the area under the receiver operating characteristic curve (AUC), also known as C-statistic. Furthermore, model calibration, the accuracy of risk estimates concerning the agreement between the predicted and observed number of events, is evaluated thoroughly by decile of predicted risk.

5.4 Implementation of hierarchical state-space models

Continuous-time models involve a two-level data structure, repeatedly measured values nested within individuals (groups). Forward updating is conditional on all past values prior to a time point t^*. The Kalman filter algorithm propagates new estimates upon the arrival of new information and uses a time step specified by the user. Therefore, the time split is determined from the time step and the timing of the last observation before an arbitrary t^*. For example, given a 45-day time step and no arrival of new observations, the Kalman filter would update the latent variable every 45 days. By updating when new observations arrive, the intervals can be a maximum of 45 days long or shorter.

In this context of SEM, I have fitted models to specify the longitudinal biomarker data.

The following figures (7.5 and 5.4) depict the LMM specification for the longitudinal eGFR, in which the deterministic change (drift) **A** and the random change (diffusion) **G** are both 0, whereas, in the drift and diffusion LMM that follows, both components are defined and estimated by the system. The outputs are given from the function `ctModelLatex()` in the `ctsem` package. The ctsem-model is a `ctStanFit` object, and the linearised normal approximation for subject parameters and covariates effects is shown. In addition to the time-updated eGFR data, the covariates gender, baseline age and baseline diabetes duration have been included in the models.

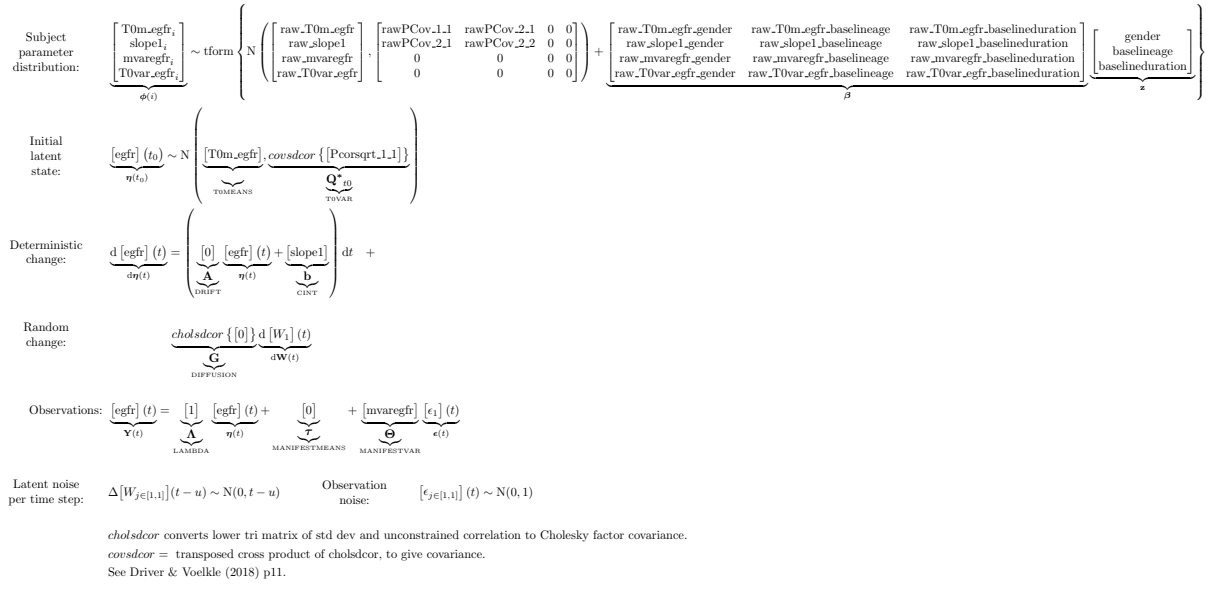


Figure 5.2: *Model fitting output.* Equation of subject-level structural equation model. Representation of a continuous-time linear mixed-effects model displaying matrix dimensions and equation structure for eGFR.

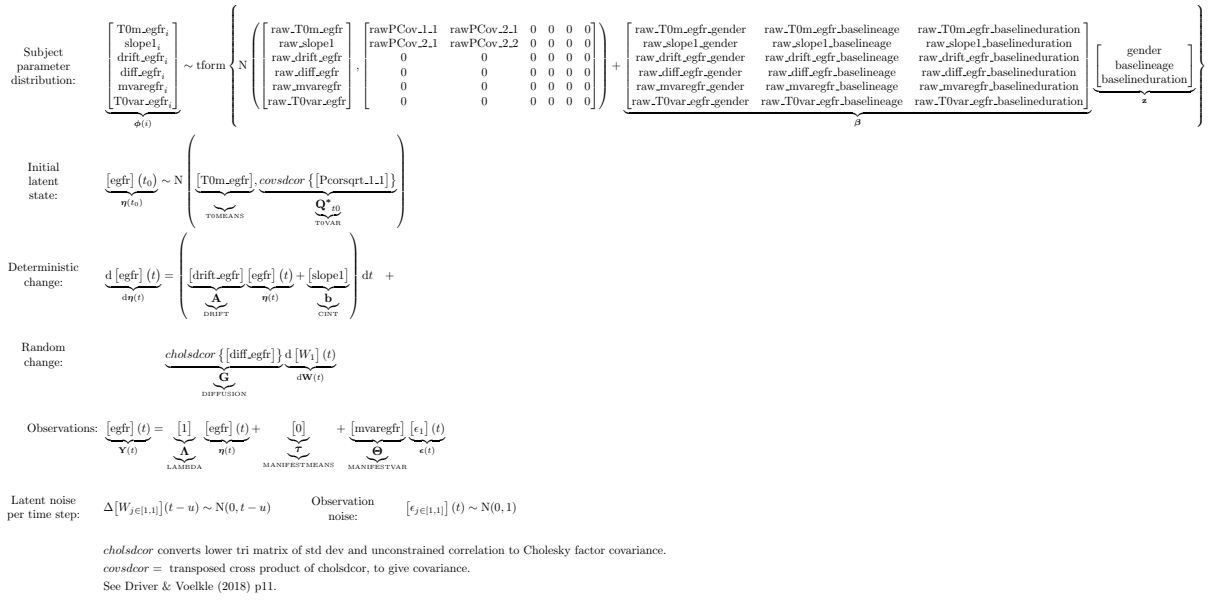


Figure 5.3: *Model fitting output.* Continuous-time representation of an LMM with drift and diffusion enabled, displaying matrix dimensions and equation structure. This comprises a mathematical extension of the simpler LMM to include drift and diffusion components for eGFR.

5.4.1 State-space model fitted to longitudinal data with `ctsem`

I have used the R package `ctsem` and Poisson regression models as a scalable alternative to `rstanarm` for the task of dynamic risk prediction. Firstly, for the longitudinal biomarker trajectories, a continuous-time model must be specified by creating an object of `ctsem` class, using the function `ctModel()`. Upon specification, the model is fitted to the longitudinal data using the function `ctFit()`, after which summary and plot methods may be used to examine parameter estimates, standard errors, and fit statistics.

Expectation matrices are then generated for each individual according to the specified input data and observed timing data. Optimisation using the Kalman filter is used to estimate the parameters, typically with a first pass using a penalty term (or prior) to find a region of high probability without extreme parameters, and then a second pass using the first as starting values.

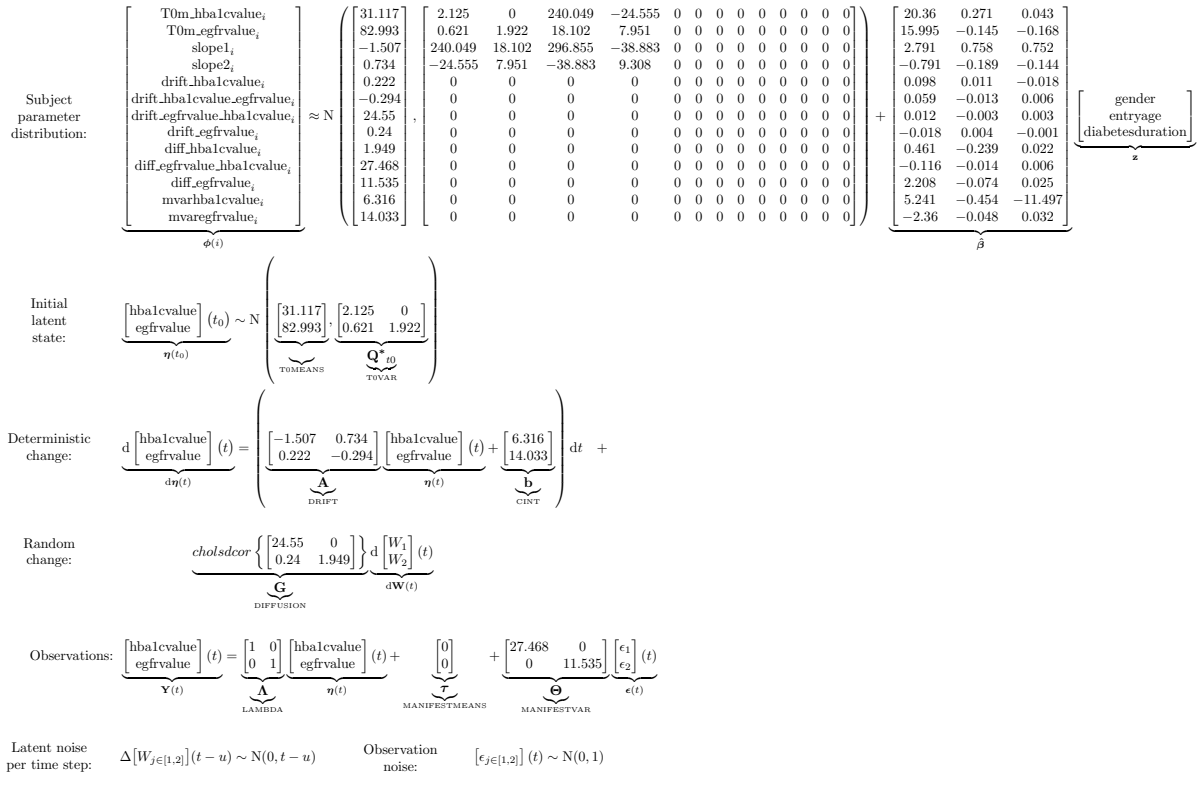


Figure 5.4: *Model fitting output.* Continuous-time representation of an LMM with drift and diffusion for two biomarkers.

5.4.1.1 Specifications of state-space models fitted to longitudinal data

To specify a linear mixed model with random intercepts and slopes but no drift and no diffusion, we specify options `DRIFT=0`, `DIFFUSION=0` and `CINT`="slope". The only individual-specific parameters are the intercepts and slopes for each biomarker. To impute time-split values from the ctsem model fitted to the training dataset, I have used the function `ctKalman()`, looping over all subjects.

I have first examined the fit of alternative models based on a multilevel bivariate Gaussian state-space model, using the R package `ctsem`.

The specification of each model is based on the following formula:

- $d\eta(t) = (A\eta(t) + b)dt + GdW(t)$ where A is autoregressive effect (drift), b is slope, G scales the diffusion process $W(t)$
  - With $A = 0, G = 0$, we have a linear mixed model,
  - With $A = 1, b = 0, G = 0$, we have a last-observation-carried-forward mode, which is used as baseline.

In the case of using two latent variables: To specify a model with drift and diffusion but no slopes, we comment out the lines `DRIFT=0`, `DIFFUSION=0` to allow these effects to be learned, and comment out the line `CINT=c('slope1','slope2')` so that the slope parameters are set to their default values of zero.

For a linear mixed model that allows diffusion but no drift, we uncomment the lines `DRIFT=0` and `CINT=c('slope1','slope2')`. For a linear mixed model that allows drift but no diffusion, we uncomment the lines `DIFFUSION=0` and `CINT=c('slope1','slope2')`. Finally, for a linear mixed model that allows drift and diffusion, we uncomment the lines `DIFFUSION=0`, `DRIFT=0` and `CINT=c('slope1','slope2')`.

The `ctsem` software package allows a hierarchical CTSEM to be fitted, including both

population-level and individual-level parameters. The original R package ctsem (Driver, Oud, and Voelkle 2017; Neale et al. 2016) has been rewritten to use the Bayesian package Stan (Driver and Voelkle 2018). Stan uses gradient-based algorithms to sample the posterior distribution or to fit a maximum likelihood model (Carpenter et al. 2017; Betancourt 2017). This makes it possible to average over the posterior distribution of unobserved variables.

Each individual trajectory is built by sequential Bayesian updating based on a Kalman filter run of the specified state-space models. Then the fit of the state-space models fitted to the biomarker data with the package `ctsem` is inspected and employed for forward prediction.

In this chapter, I described the theory of using a Kalman filter to update sequential data of biomarkers with the goal of including them as time-updated covariate data in Poisson regression models for time to event. In addition, I have highlighted the importance of landmarking, i.e., censoring all data coming after the start of the prediction window, to avoid biased extrapolations of individualised profiles. In the next chapter, I show how this theory is applied to predict the progression rate of renal failure in a T1D population, and I elaborate further on the strengths and limitations of such an approach.

Chapter 6

Progression to renal replacement therapy in a T1D population

This chapter discusses the application of the previously discussed methodology to dynamically model longitudinal biomarker data, obtained by SCI-diabetes, with a focus on eGFR and time to renal replacement therapy (RRT), as a proxy to predict end-stage renal disease (ESRD), using data from a national cohort of individuals with type 1 diabetes.

Chronic kidney disease (CKD) is defined as a reduction in kidney function or structural damage (or both) present for more than three months, with associated health implications (Webster et al. 2017). CKD is classified based on the underlying cause of the disease (e.g. hypertension, diabetes, glomerular disease), GFR/eGFR and the level of proteinuria (Transplant Work Group and others 2009). Patients are classified as G1-G5, based on the eGFR, and A1-A3 based on the ACR (albumin:creatinine ratio) as detailed in figure 6 (National Kidney Foundation webpage):

				Albuminuria categories		
				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				<3 mg/mmol	3 to 30 mg/mmol	>30 mg/mmol
GFR categories	G1	Normal	>90 mL/min/1.73m^2	Low risk	Moderate risk	High risk
	G2	Mild impairment	60-89 mL/min/1.73m^2	Low risk	Moderate risk	High risk
	G3a	Mild-moderate impairment	45-59 mL/min/1.73m^2	Moderate risk	High risk	Very high risk
	G3b	Moderate-severe impairment	30-44 mL/min/1.73m^2	High risk	Very high risk	Very high risk
	G4	Severe impairment	15-29 mL/min/1.73m^2	Very high risk	Very high risk	Very high risk
	G5	Established renal failure	≤15 mL/min/1.73m^2 or on renal replacement therapy (dialysis or transplant)	Very high risk	Very high risk	Very high risk

While the outcome of ESRD is usually defined as being in receipt of RRT or having an eGFR < 15 mL/min/1.73 m^2 (Colombo et al. 2020), the eGFR component of the definition of ESRD has deliberately been left out, as the primary objective of such a translational study is centered around improving the contribution of eGFR trajectories through the longitudinal biomarker submodel towards the accurate specification of time-to-event.

One argument for using a composite end-point of ESRD defined as start of RRT or eGFR < 15 is that using RRT only would exclude those who are evaluated as unlikely to benefit from RRT (for instance someone severely disabled by a stroke). It is also arguable that the composite end-point of ESRD is what clinicians want to be able to predict.

For this study, I have defined a composite outcome comprising (a) initiation of renal replace-

ment therapy (RRT) as a proxy for ESRD, (b) deaths with a mention of renal failure in the death certificate, assuming that censoring is independent or unrelated to the likelihood of progression to ESRD, also known as non-informative censoring. This means that the participants whose data are censored would have the same time to failure distribution if they were actually observed until the end of the study.

Previously, I attempted to predict cardiovascular disease from biomarkers, but found that although the fit to the biomarker submodel was much better with 'ctsem' than with a simple LMM, this did not appreciably improve prediction of time-to-event. For that reason, I intentionally chose to evaluate joint modelling in a situation where the biomarker (eGFR) and the event (start of RRT) are very tightly coupled. This aligns with other statistically sound joint modelling and precision medicine exemplars where a threshold value of the biomarker is included in the definition of the event. (Ilic et al. 2018; Sheikh et al. 2021; Parr, Hall, and Porta 2022).

The structure of this chapter is as follows:

1. I first describe the dataset used in this analysis and some details for the definition of the RRT outcome (section 6.1.1)

2. I briefly give some background pertinent to modelling ESRD via eGFR with a focus on the statistical models (section 6.2)

3. The objectives of the analysis are given next, followed by the methods (section 6.3.1)

4. Comparison of the fitting of the submodels for eGFR trajectories leveraging the various specifications described in chapter 5 (section 6.4)

5. Implications are drawn from the different modelling variations for the eGFR trajectories (section 6.4.2)

6.1 Data set up for modelling

In the following section, the reader can get some insight into the data setup, the definition of the outcome, the incident rate within the observed person-years and a table that summarises the baseline characteristics.

6.1.1 Population characteristics pertinent to the analysis

The analysis has been conducted using the information on individuals registered in Scottish Care Information (SCI) - Diabetes. Access to such data has generated many longitudinal studies in the last years (Wild et al. 2016; Walker et al. 2018; Captieux et al. 2021; Jeyam, Gibb, et al. 2021; Jeyam et al. 2022).

The study spans a decade: from 1-1-2008 to 1-1-2018. The T1D cohort with renal disease outcome data comprises 29121 individuals residing in Scotland who were closely followed by the healthcare system, with 449,330 eGFR measurements available.

I have defined a subgroup of 2633 individuals with *a baseline eGFR below 60 mL/min/1.73* m^2 at the study start for various sensitivity analyses. This subgroup does not necessarily include individuals that experience chronic kidney disease (defined as eGFR < 60 mL/min/1.73 m^2 that persists for 3 months or more). This *filtered* dataset represents individuals with at least mild kidney function loss, presumably due to T1D. This filtering aims to investigate whether this subgroup's rate of progressing to RRT is predicted more closely compared to individuals with T1D and no sign of kidney dysfunction at baseline.

Person-time is the total time contributed to the study by all subjects. The full cohort included 29121 individuals whose average age at baseline is 39.9 years, and average diabetes duration is 15.4 years. Mean follow-up length is 7.7 years. There are 16303 males and 12818 females, each of whom has 15.4 eGFR observations on average. The RRT incidence is 799 events, 2.7%. Table 6.1 summarises the population characteristics for the entire cohort and the filtered subgroup over the entire follow-up period. From those with baseline eGFR < 60, there was a

	Patient cohort					
Variable	Population	RRT	No RRT	eGFR<60	RRT	No RRT
Sex						
Male	16303	420	15883	1123	250	873
Female	12818	379	12439	1510	246	1264
Age (years)						
0-20	18.25 (0.52)	19.13(0.66)	18.25 (0.52)	18.49 (0.74)	19.8 (NA)	18.42 (0.69)
20-50	35.41 (8.56)	38.23 (7.93)	35.34 (8.57)	40.7 (6.9)	40.44 (6.8)	40.83 (6.96)
50+	61.62 (8.87)	65.54 (10.06)	61.41 (8.75)	67.58 (9.53)	66.79 (10.17)	67.71 (9.42)
Diabetes duration (years)						
1-5	1.21 (1.66)	1.39 (1.67)	1.21 (1.66)	1.14 (1.61)	1.22 (1.13)	1.14 (1.64)
6+	20.13 (11.22)	24.75 (12.04)	19.97 (11.16)	27.17 (12.99)	26.9 (11.79)	27.23 (13.27)
Mean eGFR (ml/min/1.73 m^2)	90.92 (27.84)	21.75 (20.93)	92.87 (25.41)	42.44 (22.29)	17.98 (13.52)	48.12 (19.97)
Follow-up (years)	8.18 (2.89)	5.52 (2.88)	8.26 (2.86)	6.95 (3.35)	4.81 (2.9)	7.45 (3.26)

Table 6.1: Demographics of individuals with renal failure outcome data.

single individual with eGFR < 30 at baseline. The overall distribution of eGFR measurements across the entire follow-up period is given by the table 6.1.1.

Table 6.2: Summary statistics of eGFR data at baseline

values
Min. : 3.00
1st Qu.: 76.00
Median : 96.00
Mean : 90.93
3rd Qu.:111.00
Max. :165.00

6.2 Previous work on predicting time to renal disease

Despite considerable improvements in the management of glucose levels in the last years, the mortality rate in patients with T1D is still high (Mameli et al. 2015). A finding confirmed by several studies in the United States and internationally, with respect to standardised mortality ratios revealing that patients with T1D have mortality rates that are 3–18 times higher than would be expected in their respective countries. There is marked geographic variation in mortality, and further notable differences between males and females, compared to the general population (Secrest, Washington, and Orchard 2021).

In addition, excess mortality in people with long T1D duration (over 30 years) stems partly from cardiovascular events (Deckert, Poulsen, and Larsen 1978). Interestingly, within the first 20 years from the onset of T1D, the most significant part of the excess mortality is attributed to renal failure (Dorman et al. 1984). Diabetic nephropathy turning into end-stage renal disease (ESRD), which may result in renal replacement therapy (RRT) contributes significantly to increased mortality (Orchard et al. 2010) (Stadler et al. 2006). Mortality rates for all major diabetes-related complications (acute, renal, cardiovascular, and infectious) have been summarised by Secrest et al. (2010) and can be reviewed analytically in the following link.

In the last decade, new technologies have emerged, e.g., insulin pumps, glucose sensors, etc., offering reasons to be hopeful that late T1D complications might be prevented through better monitoring. Moreover, drug development advancements for preventing or reversing moderately acute impaired renal function have been pivotal. Finally, accurate estimation of the rate of progression to renal failure is critical in lowering T1D mortality rates.

Creatinine is routinely used as a marker of renal function. Creatinine is a molecule that is processed by the kidney, and therefore the rate by which creatinine is removed, divided by the blood plasma concentration, indicates the rate of glomerular filtration (GFR) and the diluting capacity of kidneys, known as tubular function. Glomeruli are tiny filters in the

kidneys that help remove toxins from the blood.

The creatinine clearance is specified from serum creatinine (SCr) levels reported in mg/dL and 24-hour urine monitoring. The following is the Cockcroft and Gault formula, which was introduced in 1973 for assessing creatinine clearance:

$$C_{Cr} = \frac{(140 - \text{age}) \times \text{weight}}{72 \times \text{SCr}} \quad [\times 0.85 \text{ if female}].$$

The expected range of the creatinine clearance test is 100-130 ml/min in females and 110-150 ml/min in males (Gowda et al. 2010). Although creatinine is usually produced at a fairly constant rate by the body depending on muscle mass, it is also influenced by muscle function and composition, activity, diet and health status (Banfi and Del Fabbro 2006). In addition, fluctuations in the values of serum creatinine may indicate kidney dysfunction.

Serum creatinine has been the gold standard for assessing kidney function despite being an insensitive and unreliable predictive biomarker (Samra and Abcar 2012; Ostermann, Kashani, and Forni 2016; Swedko et al. 2003; Delanaye, Cavalier, and Pottel 2017; Bargnoux et al. 2018). However, accumulated experience and studies have led to discussion towards markers of kidney function, which would be non-invasive, accessible in blood or urine samples, cost-effective and capable of early detection, among other characteristics.

The gold standard measurement of GFR involves the injection of inulin and its clearance by the kidneys (Kampmann and Hansen 1981). However, the use of inulin is invasive, time-consuming, and an expensive procedure. Alternatively, the biochemical marker creatinine found in serum and urine is commonly used in the estimation of GFR (Gowda et al. 2010). Creatinine clearance (CrCl) is the volume of blood plasma cleared of creatinine per unit time (Shahbaz and Gupta 2023).

In order to avoid the invasive creatinine clearance test, clinicians routinely use the estimated glomerular filtration rate (eGFR) to measure how much blood the kidney filters clean every

minute based on body size. The eGFR is usually calculated from serum creatinine using an isotope dilution mass spectrometry (IDMS) traceable equation. The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula is one of the most widely used IDMS traceable equations for estimating GFR in patients aged 18 and over (Brand et al. 2011). The CKD-EPI formula uses a two-slope spline model to specify the relationship between GFR and serum creatinine, age, sex, and race. The CKD-EPI equation is given below (Raman et al. 2017):

$$\text{eGFR} = 141 \times \min(\frac{\text{SCr}}{\kappa}, 1)^{\alpha} \times \max(\frac{\text{SCr}}{\kappa}, 1)^{-1.209} \times 0.993^{\text{age}} \ [\times 1.018 \text{ if female}] \ [\times 1.159 \text{ if black}], \tag{6.1}$$

where κ is 0.7 for females, and 0.9 for males, and α is -0.329 for females and -0.411 for males. The equation does not take in weight or height data because the results are reported normalised to 1.73 m^2 body surface area, which is an accepted average adult surface area.

To sum up, the advantages of using eGFR are that serum creatinine is measured routinely and offers an accurate approximation of the filtration rate based on blood samples only, instead of the cumbersome and invasive screening of urine levels for 24 hours.

The consensus is that renal failure is defined as eGFR less than 15 mL/min/1.73 m^2, and it is therefore expected that eGFR trajectories would be the ultimate determinants of renal replacement therapy (RRT). However, a few patients with ESRD have preserved eGFR levels (Chang et al. 2013). Therefore, screening of eGFR might not be always ideal as an early clinical marker to early detect declining kidney function and must be complemented by spotting urine for albumin-to-creatinine ratio (ACR) to detect albuminuria (Nelson et al. 2006; Botev and Mallié 2008; Toffaletti 2010).

Scientists and clinicians have made great efforts in the past decades to discover and validate novel biomarkers for predicting the risk of renal disease. Recently emerged biomarkers for early detection of renal impairment have been discussed in Section 2.2.1.

Yang et al. (2021) in their attempt to build a machine learning risk prediction model of renal failure for a CKD population, confirmed the consensus view that the top risk factors associated with renal failure are serum creatinine, age, urine acid, systolic blood pressure, and blood urea nitrogen.

Tangri et al. (2017) developed a regression model for people with CKD stages 3 to 5, including age, sex, and urinary albumin-creatinine ratio at baseline as time-invariant covariates, and eGFR, serum albumin, phosphorus, calcium, and bicarbonate values as time-dependent covariates. In addition, they constructed a last-observation-carried-forward (LOCF) model for time-updated covariates. They have found that eGFR is more strongly associated with kidney failure in the LOCF model versus the baseline model (HR, 0.44 versus 0.65). They have additionally found that the LOCF regression model with eGFR as a time-dependent covariate incrementally improves risk prediction for kidney failure over a static model with only baseline eGFR.

For many decades, physicians have been taught that the slope of the decline in reciprocal serum creatinine versus time reflects quite accurately the underlying changes in creatinine clearance and allows an accurate assessment of the rate of progression to renal disease. Although a linear trend is more informative than assuming a simple random walk for the changing eGFR levels, likely, calculating a slope for eGFR is not as perfectly informative as it was considered back in the 1970s because the decline of kidney function is not strictly linear (Kassirer 1971; Levey, Perrone, and Madias 1988; Lacour 1992). In addition, the rate at which kidney function declines in different age groups and under specific preexisting conditions has not been conclusively determined. Therefore, the questions I am interested in answering are how informative of RRT the slope of longitudinal eGFR might be, and whether a better model for longitudinal eGFR improves the prediction of the rate of progression to renal failure.

6.3 Methods & Observations

6.3.1 Objectives of the analysis

The rate of progression to renal failure of individuals early in the course of renal disease cannot be reliably estimated. Hence, a question arises about whether employing models more sensitive to the biomarker changing could perform better than conventional approaches in fitting the longitudinal data like last-observation-carried-forward (LOCF). The LOCF approach is commonly used to assess the risk of progression to a disease when time-varying covariates are considered.

It is expected that the past trajectory of eGFR represents a quite informative individual slope given that loss of kidney function progresses with time. However, there is no conclusive evidence that this trend is linear. In contrast to HbA_{1c}, where fluctuations and higher values can be brought down with treatment and might follow a random walk over a while, a biomarker like eGFR, which is even more directly associated with organ function, is more likely to follow a monotonic trajectory.

The following sections describe our approach to:

- evaluate how eGFR trajectories evolve over time using a dynamic systems modelling approach based on differential equations of a continuous-time form,
- and produce risk predictions of time to renal failure, conditional on individual-level parameters of longitudinal eGFR, where various Kalman filter algorithms estimate latent states.

The study conducted by Diggle, Sousa, and Asar (2015) is the most closely related to the analysis described herein and has been a source of inspiration for us. They have developed a routine for real-time monitoring of progression to renal failure, employing a linear mixed-effects model extended by a random intercept and integrated Brownian motion, i.e., diffusion (section 2.2.2).

The overarching aim of this chapter is to evaluate how informative longitudinal eGFR is in assessing the development of renal disease in individuals with T1D and the contribution of using more dynamic models for the longitudinal biomarker to predict the rate of progression to ESRD.

Furthermore, the research question is extended to how long the time-updated eGFR trajectories of individuals with T1D must be to add to the prediction of the rate of progression to renal failure on top of clinical variables like sex, age at baseline and diabetes duration.

The choice of predicting time to RRT based on longitudinal eGFR observations was made on the belief that eGFR is a better predictor for ESRD (as it is used effectively in the definition of the outcome) compared to HbA_{1c}. Thus, the eGFR-RRT prediction problem would make a better exemplar for joint modelling than predicting time to CVD from HbA_{1c}. It is important to emphasise that the biomarker data used in my analyses belong to individuals diagnosed with a chronic disease, namely T1D, that might drift more than usual over time. In most instances, these longitudinal processes are highly dynamic; thus, we need to employ dynamic, Bayesian modelling to handle changing trends through time and identify potential proxies that contribute to the accurate specification of risk. For this reason, the development and evaluation of joint models that specify the correlation and error structures between noisy sequential biomarker measurements are essential to maximise dynamic risk prediction.

Furthermore, since the joint modelling approach specifies the entire trajectory of every biomarker-subject pair to compute the survival function, as opposed to conventional modelling approaches that only exploit biomarker observations at single time-points, we anticipate that there will be an increment in prediction of time-to-event, even for complicated relationships and biological pathways. Notwithstanding the speculated increment in risk prediction, joint modelling remains computationally intensive and not yet fully optimised. Hence, the most reasonable approach was to first leverage the most predictive/strongly related biomarker option in order to develop and demonstrate the modelling process.

As a first step, the performance of various state-space models and a traditional last-observation-carried-forward (LOCF) model for eGFR have been compared. In the second stage, the time-updated imputed eGFR data were provided into generalised linear models of Poisson likelihood with the intent of estimating the event rate.

In the following, I describe:

1. the use of time-splitting Bayesian models for predicting progression to renal failure in a T1D population
2. the results of the development of the biomarker submodel using `ctsem` and LOCF, and the way those specifications influence the prediction.

In chapter 5, I have examined the underlying theory of fitting the so-called time-splitting joint model, which is split into two components. The first piece of work involves the development of the biomarker trajectory, i.e., the submodel for the longitudinal component, which I constructed using the `ctsem` functionality.

Subsequently, I have evaluated the performance of the different submodels using metrics such as deviance and AIC, intending to evaluate the developed longitudinal models for eGFR and HbA_{1c} initially, and by comparison to deduce the number of parameters needed to fit the longitudinal trajectories of the biomarkers in the most optimal way regarding deviance and AIC. Finally, in the following results chapter, I describe the final fitting of the model for progression to RRT, comparing the performance characteristics of each joint model.

The main conclusions drawn so far from my developments have been that:

1. Extending the LMM also to include a term for drift and diffusion leads to the best fit of the biomarker data in this population of individuals with T1D.
2. A population-level parameter model needs to be revised to specify individual-level parameters of previously unseen data, especially if such individuals have not been part of the training.

6.3.2 Approaches to exploiting time-updated data

I have employed the `ctsem` package for the specifications of the longitudinal trajectories. This software produces projections of the biomarker trajectories based upon individual-level parameters for the rate of change of eGFR. A limitation of this approach is that we cannot learn the individual parameters for those who have not been included in the training.

Therefore, for the development of this methodology such software limitations have been a major factor. To work around this software limitation, as part of the training set, the setup includes a deliberate censoring of those individuals for whom we make predictions after a particular point of their follow-up, so-called *landmark time point*. This enables forward predictions based only on previous data, and eliminates the introduction of immortal-time bias, introduced in section 2.4.1.

Were the `ctsem` software designed for clinical prediction applications, optimally it would have been implemented in a way that uses the population-level model in order to learn the individual-level parameters and make predictions for entirely new patients, for whom we do not have available data. For instance, those previously unseen individuals may be homeless and not registered with a general practitioner.

Hence, the landmark time point works as a cutoff, after which we update the eGFR trajectories which are used for dynamic prediction, based on knowledge accumulated until the landmark point. The landmark point is set up to 5 years after the subjects' entry. During the first 2826.527 person-years (up to 5 years since entry for each subject), there were 290 RRT occurrences and 37070 out of the 62812 eGFR observations.

Put differently, the landmark point is calculated with respect to individuals' entry point to the study. Single landmark point to eliminate immortal-time bias was proposed by Anderson et al. (1983) in order to perform a survival analysis of tumour response. The paper suggested that when we compare people that responded to treatment with those that did not, we create bias in favour of responders, who have survived long enough to record their response. Therefore,

the artificial censoring I have introduced does not allow for information on eGFR arrival times to be considered in the specification of the survival function of those who technically are 'unseen' by the model during the prediction period.

6.3.3 Splitting the dataset into training and testing subsets

In this analysis, I have employed a k-fold cross-validation approach with an external validation on unseen data of T1D patients using a landmark time point.

The pipeline of the analysis goes as follows: First, the original sample is randomly partitioned into k equally-sized subsets. Next, each data partition is used to train a separate model, which is then validated on the withdrawn proportion of the data. Overall, this process is repeated k times, using a different partition of the data at each iteration. Therefore, each observation is used for validation exactly once. The advantage of this approach is that all observations are used for training and validating the fitted models. This allows us to validate the fitted models on the largest number of events possible.

The cross-validation approach requires refitting a generalised Poisson model k times to different subsets of the data. For the task of dynamic prediction, a different segment of the data (test individuals) is deliberately censored at five years within each training fold. Those individuals, who are censored on purpose at five years within the current training fold, comprise the complementary testing fold to be used with that model. In this case, the model is fitted using longitudinal data for N individuals for up to five years. After that point, biomarker information and event indicators are withheld from training. Using data from the future would notoriously give rise to immortal time bias, which in turn would yield biased estimations. Therefore, the training/test split concerns the individual follow-up.

The testing folds are created upon the condition that subjects must have contributed five years to the study. Dynamic predictions are based on imputed data, i.e., Kalman filter imputations of eGFR, which are agnostic to any observation of these individuals after five years. The

implementation of `ctsem` requires that within each training fold, there are at least some individuals who are still followed up five years after entry to be able to make predictions for those 'unseen' individuals in the five-year prediction window.

6.3.4 Kalman filter setup and presentation of training scenarios

At this point, someone might ask why we need to include those that have less than five years of follow-up to train the model. As explained later on, the fitting process has been adapted to test both scenarios (A. follow-up ≤ 5 years and B. follow-up ≥ 5 years), only to understand at a later stage, the intricacies of including only in training subjects who progressed to RRT in their first 5 years of follow-up.

The current section gives an overview of the steps that led us to two major statistical analysis designs. The estimation is based on coupling time-updated biomarker data generated from a Kalman filter algorithm with Poisson models for time to event (as described in chapter 5). In addition, I have specified a range of stochastic processes to determine the rate of change of longitudinal eGFR, and assess whether there were systematic differences between the entire cohort and the filtered subgroup.

The imputation algorithm of the Kalman filter depends on two parameters. Imputed eGFR data are generated conditional on past observations up to

1. pre-determined time points, as specified by a time step given by the user, and

2. the original arrival times of observed data for each subject.

A k-fold cross-validation design has been employed, and a landmark time point is determined, up to which an individual needs to be observable and free of an event in order to be assigned in one of the k testing folds. This landmark point has set up to five years, i.e., half of the longest possible follow-up. Recall that the average follow-up time for the filtered dataset is 6.95 years and respectively for the full cohort 8.18 years. Therefore, five years after the

baseline is a reasonable threshold that most participants have outlived.

The first design allows subjects that have been followed up for less than five years to contribute to the training phase. Although those individuals (with length of follow-up ≤ 5 years) are not included in the test set, they actively contribute to model fitting for as long as they are event-free. Subjects that are observable but contribute zero observations to the study during the first five person-years are excluded from the analysis. I call this statistical analysis design *analysis A*, i.e., inclusion of all individuals in model training, regardless of follow-up length.

Limitations of this study design: The subjects that are followed up for less than five years do not enter the cross-validation process. They are only included in the training folds, as they do not meet the criterion to enter the testing folds (having a follow-up length of at least five years). As a result, the event distributions between training and test subjects differ substantially since the testing folds are constructed on the basis that subjects are event-free for 5 person-years. On the contrary, the training data also include events that occurred within the first five person-years.

The predictive models, in this case, are less calibrated: they assign a very low probability of an event in each interval for most individuals because the test subjects were seen as low risk since they do not have an event within five years from entry. Therefore I refined the analysis to restrain the training folds so that they include subjects with follow-up ≥ 5 years to be compatible with the testing folds. I call this *analysis B*. When both training and testing folds are taken together in this scenario, they have equal numbers of individuals. Hence, in this case, the cross-validation concerns individual observation time before and after the landmark point.

By excluding from cross-validation those individuals whose follow-up was shorter than five years, I ensure that the training set does not include any individuals who had an event within the first five years (had an event occurred, they would not be part of any test fold). Hence, the training folds of analysis B include only individuals who were event-free during their

initial five years. As a result, the event distributions between train and test folds do not deviate significantly from each other, and model predictions are well-calibrated, as explained next in chapter 7.

To sum up, the number of events in the training folds under this specification matches the number of events that occurred after the landmark point. Analysis B discards those subjects with an event within the first five years. Therefore, the total number of events in the training phase equals the number of events occurring during the prediction period.

Note that in this refined design, the initial five years are all *control-years* (years that do not contain any event). The total length of follow-up between training and test data is now almost the same, with the difference being that the last interval of the training set may be truncated if an event occurs. On the other hand, the prediction intervals are all of equal, fixed length (I have studied various options such as one year, half year, etc., which is further discussed in chapter 9) since event time is a priori unknown, and the hazard rate must remain constant throughout the interval for the prediction to be accurate.

To demonstrate the designs and get the pipeline working, I have split individuals who are followed up for at least five years into two folds, in one of which their biomarker data and event status are withheld to allow for forward predictions of time to RRT. I then used the model agnostic to any data for these test individuals after the landmark time to make predictions. Analyses have also been performed using a 10-fold cross-validation, yielding similar results. However, using ten folds to replicate all different settings, i.e., eight model specifications (described in section 6.4), two designs: follow-up length restrained to 5 years or not, six different time steps to produce imputations, and two input samples (total and filtered cohort) has been relatively computationally intensive.

6.3.5 Format of input: eGFR data

I have log-transformed the eGFR data to reduce any skewness in the dataset and be more compatible with other studies. The log-transformation helps stabilise the variance of the data, particularly in this context where the variability increases with the magnitude of the data. Since more granular time steps than one year have also been used, most of the final 'made-up' datasets feature a substantial number of imputed sequential data, especially when individuals are followed up for long periods. I have preserved the original mean and standard deviation of eGFR to reverse the transformation if deemed necessary. The eGFR, age and diabetes duration data are also standardised.

To motivate more clearly the reason for transforming the data, the distributions of the raw eGFR values and the scaled and log-transformed eGFR values for the filtered dataset are shown respectively by figures 6.1 and 6.2.

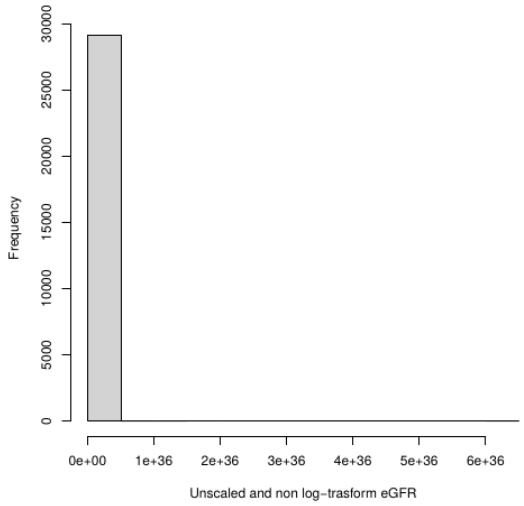


Figure 6.1: Distribution of original eGFR

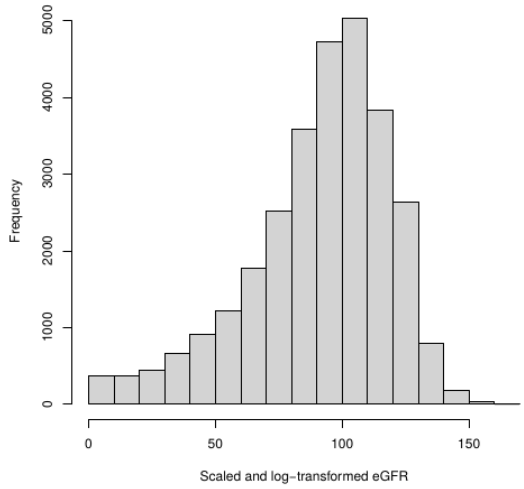


Figure 6.2: Distribution of transformed eGFR

The format of the input that goes into the model is one row per observation for each subject (*long format*), as shown in the following extract:

Table 6.3: Raw and standardised log-transformed eGFR given with time of measurement (lapsed since baseline), and event indicator. Ten first observations for a randomly selected patient who experienced no RRT event.

raw eGFR	log eGFR	time of obs	event
52.204	0.574	0.137	0
67.856	1.107	0.808	0
67.845	1.106	0.830	0
56.716	0.742	1.843	0
56.706	0.742	1.867	0
60.017	0.857	2.801	0
60.001	0.857	2.839	0
76.493	1.350	5.292	0
74.743	1.303	6.308	0
60.434	0.871	7.461	0

The imputation frequency depends on the user's time step and the observations' original arrival times.

Informative example

Suppose a subject had a biomarker value lastly measured at Time = 8.5 years since entry to the study and an event at Time = 10 years (1.5 years after the last biomarker update). In this thought experiment, we are using a time step for the Kalman filter equal to 1, i.e., to impute eGFR values every one-year *during the training period*. Additionally to the imputed values at the end of one-year-long intervals, we could also get imputations sooner, as long as new biomarker data become available within the one-year interval. Put differently, the arrival

of a data point triggers a Kalman filter imputation, whenever this comes.

An one year and half long interval, (starting at Time = 8.5, which is the last arrival of their observation, ending at Time = 10, which is the event time) would not be valid because this interval would run longer than the maximum interval length, set to be one year.

Since an interval such as [8.5, 10], i.e., one year and half long interval is prohibitive in this particular setup, the algorithm operates as follows: Counting backwards from the moment of event, the subject would have an interval spanning from 9 to 10 years (since entry), along with an interval starting at 8.5 (biomarker data arrival), and ending at Time 9. The flag for the event within the interval [9, 10] is 1, and elsewhere is 0, because in intervals prior to [9, 10], the subject has been event-free.

6.4 Development of the biomarker submodel using `ctsem`

6.4.1 Specification of state-space models

I compared a range of state-space models, also known as continuous-time structural equation models or dynamic Bayesian networks, to determine which model fits the eGFR data best. Continuous-time models involve a two-level data structure, sequentially measured values nested within subjects. The most challenging part of the model to estimate is individual-level parameters, i.e., random effects. Among the models I trained, I used continuous-time autoregressive effects models coupled with a stochastic diffusion process.

Among the benefits of this model class is that it scales favourably to high-dimensional datasets and is likely to explain highly variable biomarker trajectories among between people with T1D concerning the estimated variance of the distribution of the longitudinal data.

I have used hierarchical state-space models that rely on differential equations to capture the

underlying variance in observations and infer time to event. In addition, the initial LMM has been extended to allow for autoregressive effects that vary over time and random effects to account for the between-subject variability. In this case, the longitudinal process has been explicitly specified based on a multi-level univariate Gaussian state-space model.

Brownian motion, and Langevin processes are all instances of Gaussian processes (stochastic differential equations) that explain how a system evolves according to a set of deterministic and fluctuating (i.e., random) forces. In the Brownian motion of molecules, for example, a sample path of a diffusion process infers the trajectory of a molecule subjected to rearrangement due to collisions with other molecules. Equivalently, the Wiener process is a specific type of continuous-time stochastic process that integrates a Gaussian process that exhibits Brownian motion.

The following properties signify the equality between Wiener process and Brownian motion: Like Brownian motion, the Wiener process exhibits the Markov property, where the future behaviour depends only on the current state and not on the past. Gaussian increments signify the increments of the Wiener process, i.e., the changes in value between two time points, are normally distributed with mean zero and variance proportional to the time interval.

In linear mixed-effects models with autoregressive drift effect, the current state is regressed on the previous response along with the fixed and random effects, i.e., the current state depends on current values and past covariate trajectories (Diggle et al. 2002).

The developed models for eGFR were interfaced via the R package `ctsem` 3.7.6 running on R version 4.1.2 (2021-11-01). All the models for the longitudinal data apart from the LOCF model have been referred to as *ctsem models* for convenience as they have been developed using the particular software. Mainly, the model specification was carried out via the modelling function `ctModel()`, and fitting was derived via `ctStanFit()` that employs the powerful probabilistic language Stan as a backend.

The models under consideration are the following (in order of increasing complexity):

1. Regression model with individual slopes and intercepts, i.e., a linear mixed-effects model (LMM),
2. Regression with drift as an autoregressive effect, implying that the process tends to drift towards the areas of high concentration/average values over time,
3. Regression with diffusion to model sudden nudges/perturbations of the longitudinal process,
4. Regression with drift and diffusion but no slopes (i.e., no LMM),
5. LMM with diffusion,
6. LMM with autoregressive drift effect,
7. LMM with both drift and diffusion (full model).

6.4.2 Model comparison

For modelling the longitudinal eGFR data, I employ and demonstrate the `ctsem` functionality. The `ctKalman()` function uses the Kalman filter algorithm to output the predicted estimates of the latent states of the biomarker base d on the data fit with `ctStanFit()`. The `ctKalman()` output depends on the mode of the parameter distribution. Ideally, one would sample multiple imputations from the posterior distribution and take the average prediction. However, as a start, I used the posterior mode to obtain a single imputation of each biomarker state as the default specification in my attempts to sample the posterior distribution of the longitudinal data.

Furthermore, using the `ctKalman()` function, I generated various graphs that depict the original data y and the estimates of the 'true' value of y, given all observations. By replacing 'y' by η, the function returns the latent states of the so-called *manifest* variable, as shown in figures 6.3, 6.4 and 6.5.

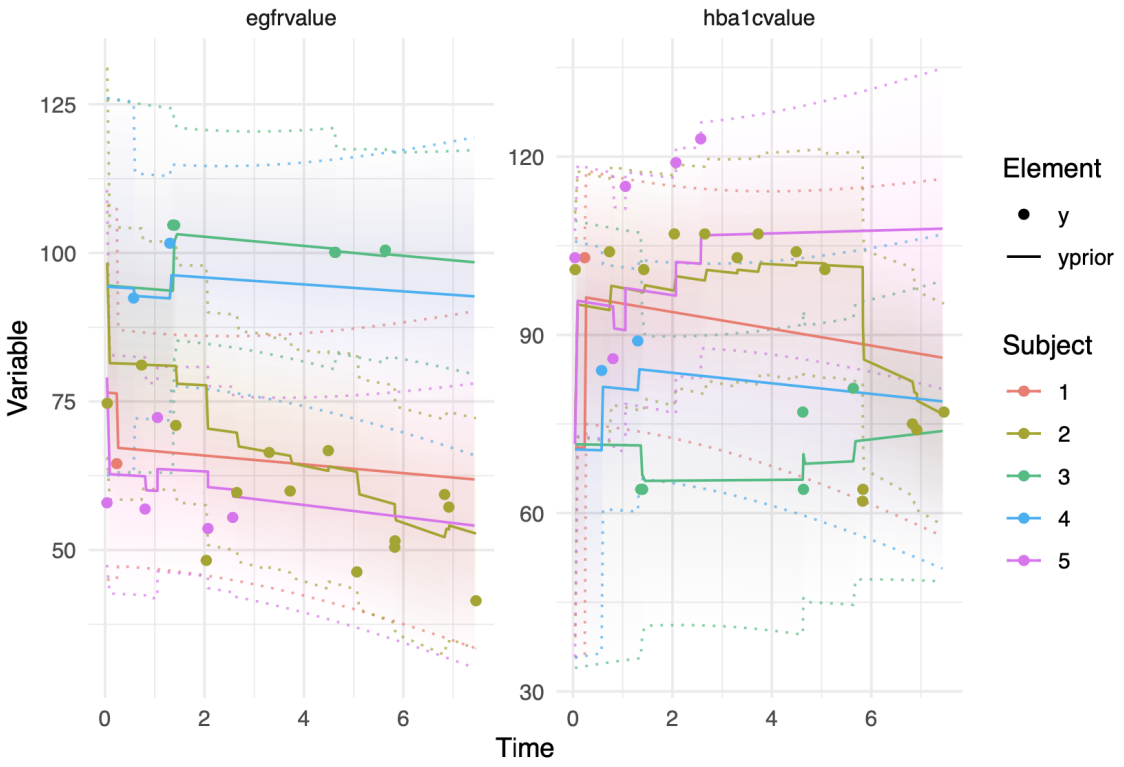


Figure 6.3: Linear mixed model with no diffusion or drift: updates to trajectory of five individuals by Kalman filter

More specifically,

- Filter: the first panel of figure 6.4 depicts the Kalman filter predictions for each time point conditional on the actual data of a randomly selected subject. Here, the model extrapolates forward in time any prior knowledge obtained from the data, determining the latent states. The model updates its predictions as new observations arrive.

- Smoother: the smoothed estimates depicted on the second panel of figure 6.4 are conditional on *all time points* in the data - past, present, and future.

In figure 6.4, the first panel depicts the *yprior* computation: original data y of *subject 118* and the prior distribution updating mechanism. The second panel depicts the *ysmooth* computation: the estimates of the *true* value of y given all observations for the same subject.

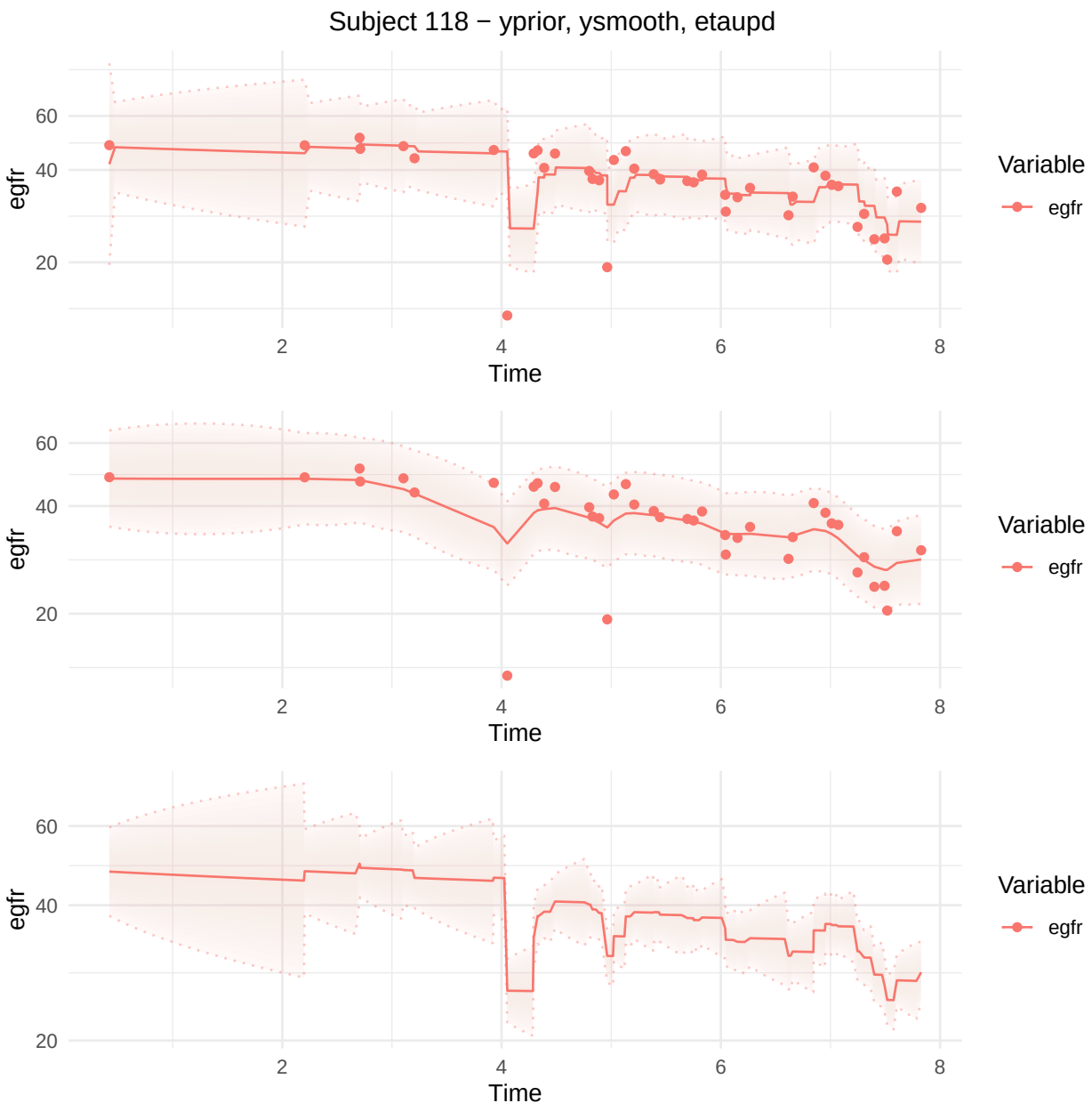


Figure 6.4: LMM state-space model specified for longitudinal eGFR of random subject

The smoothed estimates depicted in the plot are not included in modelling time to event to minimise algorithmic bias. The third panel depicts the *etaupd* computation: replacing the original values y by the predicted latent states η, the Kalman filter generates a complete trajectory of eGFR, which is in turn plugged into a Poisson regression model for time to event as time-updated predictor.

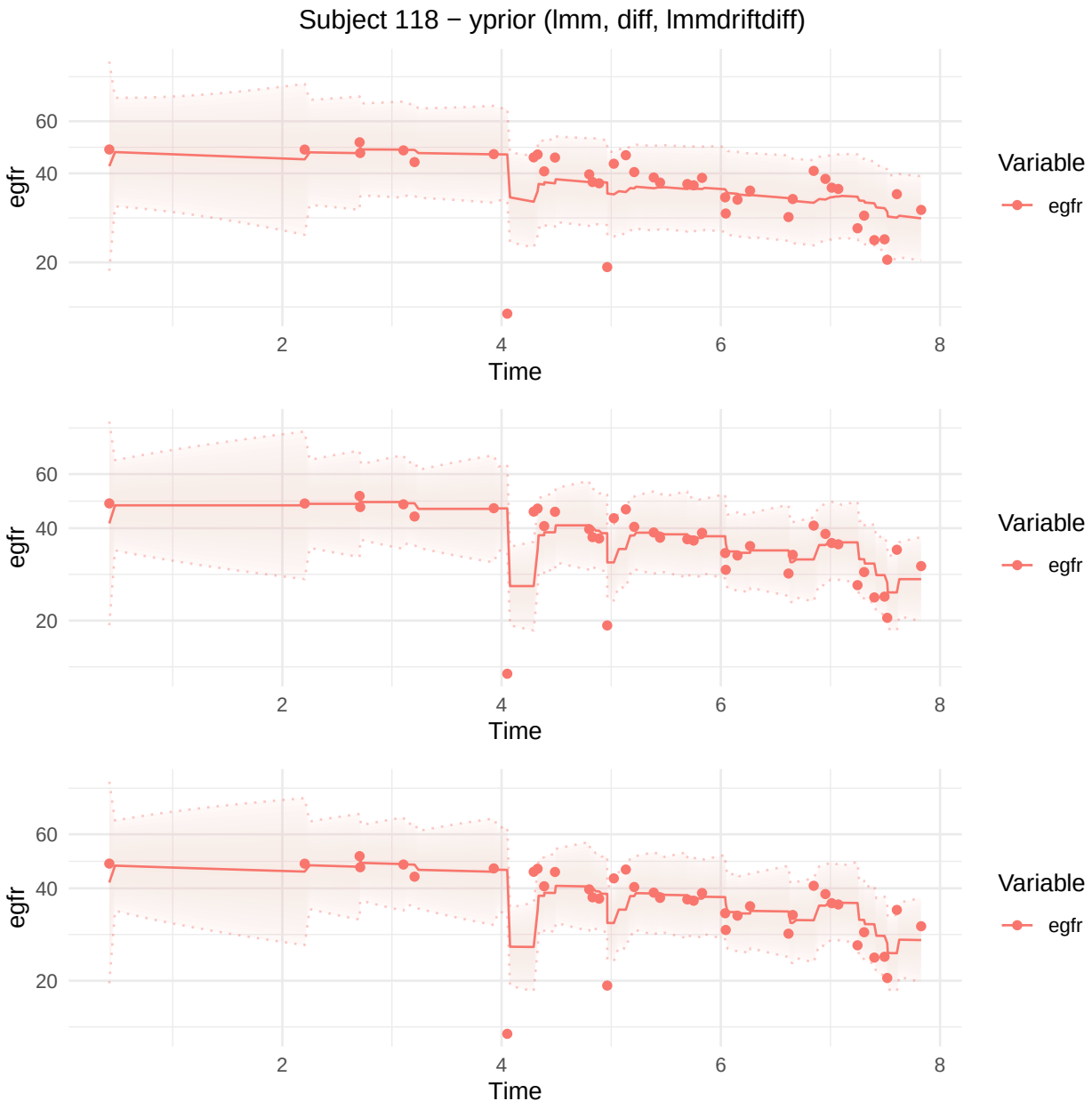


Figure 6.5: Three state-space model specifications under comparison

In figure 6.5, the first, second, and third panel depict an LMM, a model with diffusion, and an LMM with drift and diffusion, respectively. These diagrams show the Kalman filter imputations of the latent states, conditional on observations strictly up to the prediction time. The updating mechanism of the prior distribution is also shown. The imputations of latent states slightly differ depending on the model specification (whether the model includes slopes, drift or diffusion processes). The second and third panels arguably look quite similar. This is

because the diffusion process dominates in both modelling specifications.

However, the functionality for smoother estimates has been used for illustration purposes only. Including information from the future at the time of prediction would induce severe bias. Thus the smoothing algorithm has not been employed for the task of dynamic prediction. Lastly, the graphs are respectively based on the maximum likelihood estimate and posterior mean of the parameters.

The following two data extracts show the actual data of subject 118 at original arrival time points (table 6.5) and the respective Kalman filter predictions of latent states of eGFR based on the LMM (table 6.4.2). Not each time point shown in the second table is an original time point of arrival of an observation. The algorithm produces imputations for any arbitrary time point specified by the user. The filter uses the estimated prior distribution and the observed data up to that point to impute the states of eGFR.

Table 6.4: Actual observations times

arrival time of observation	original eGFR
0.05476	50.73734
0.10130	50.72075
0.38056	49.39172
0.50650	49.34804
0.76934	43.86357
0.81040	46.94766
0.81314	46.94675
0.93908	48.41158
0.94456	48.40972
1.03765	42.83192

Table 6.5: Extra biomarker values and times based on Bayesian updating

time of imputation	imputation of stand/sed log eGFR
0.00000	45.76539
0.05476	49.68982
0.10130	50.06144
0.38056	49.21433
0.50650	48.94547
0.76934	47.00294
0.81040	46.87303
0.81314	46.87749
0.93908	46.74716
0.94456	46.96477
1.00000	46.78285
1.03765	46.11076
1.28679	44.15129
1.63176	40.68697
1.76591	38.71778
1.80698	37.66357
1.99316	37.41287
2.00000	37.37475
2.06708	37.73202
2.09719	38.10346

Figures 6.6 and 6.7 give an overview of the two staged imputation-regression pipeline.

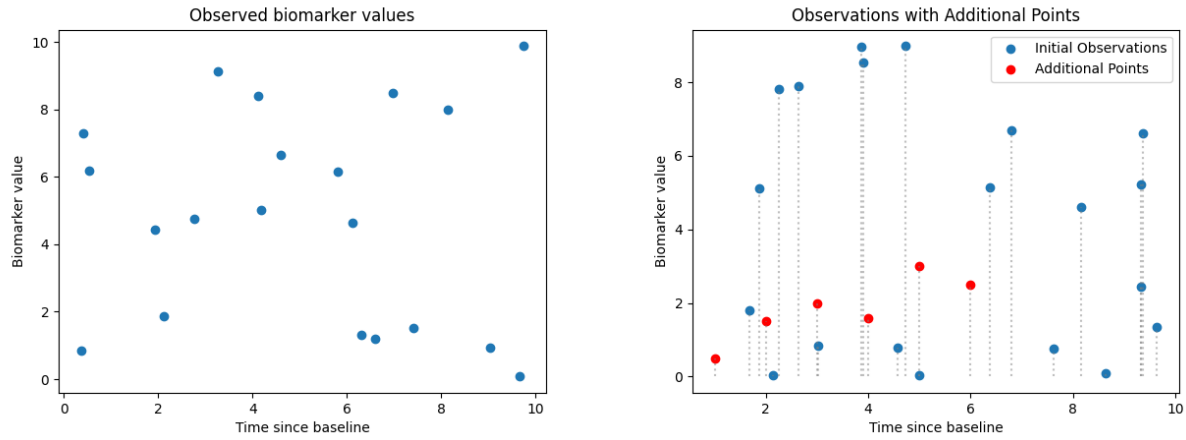


Figure 6.6: *Stage 1* Randomly generated initial data. Time-split and imputations generated via a Kalman filter at regular intervals including original time points.

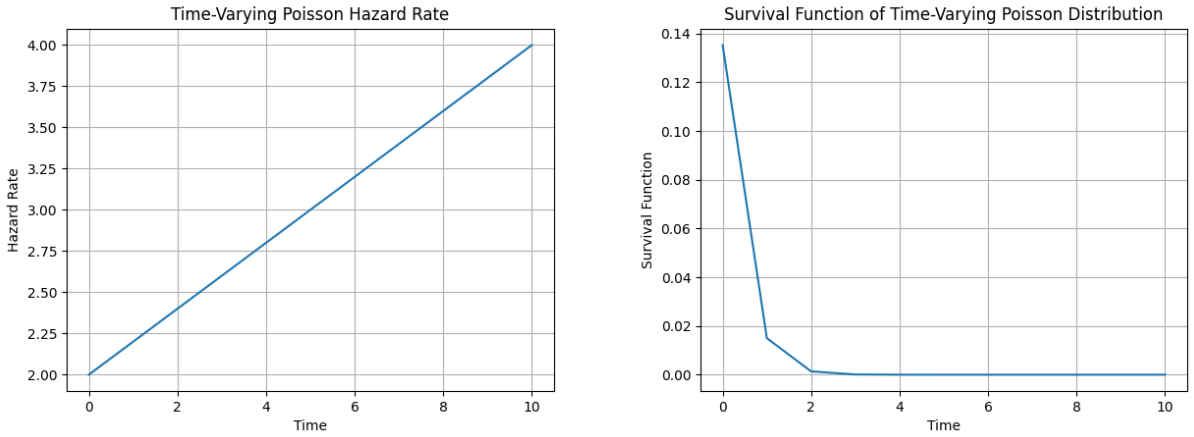


Figure 6.7: *Stage 2* Approximate the hazard function. Imputed values fed into a Poisson regression model as time-updated data.

Summarily, the `ctsem` software encodes functionality for estimating the biomarker's latent states using the Kalman filter algorithm. The Kalman filter estimates a distribution for the input data and the mode representing the latent state at an arbitrary time, see figure 6.3. Therefore, evaluating how the Kalman filter performs for different individuals is interesting. At the individual level, the Kalman filter updates the state each time there is a new observation available, i.e., the number of sequential observations might determine the frequency at which the updating happens. Additionally, the Kalman filter also performs with a time step, which evokes updates (to specify a latent state) without new data has arrived.

Note that the updates evoked by the time step all occur simultaneously for every individual. Therefore under this updating scheme, one could obtain synchronous biomarker data to be used by other joint modelling implementations that operate under this requirement, such as `rstanarm`.

6.4.3 Running times of joint model fitted using `stan_jm()` and of state-space models fitted to longitudinal data with `ctsem`

I have also fitted in parallel a univariate joint model per fold to use it as benchmark for the filtered subgroup, employing the modelling function `stan_jm()` from `rstanarm`. Table 6.6 gives the running times required to fit the joint model.

Table 6.6: Runtime in hours of fitting a `stan_jm()` model with 1000 iterations and 4 chains on 2673 subjects and one biomarker.

	warmup	sample	total
chain 1	4.255	2.015	6.270
chain 2	5.288	2.042	7.330
chain 3	7.670	7.162	14.832
chain 4	6.225	1.977	8.202

To highlight the contrast, the running times required for fitting five *ctsem* models in parallel (LMM, LLM with drift, LMM with diffusion, No LMM (a model with drift and diffusion but no slopes), LMM with drift and diffusion) for three study designs are given in table 6.7:

Table 6.7: Running times of fitting biomarker data using `ctsem`.

individuals	biomarkers	time	unit
29118	2	81.576	hours
29764	1	28.464	hours
2673	1	2.790	hours

6.4.4 Varying interval lengths

Individual time-splitting is determined by both the time step given by the user and the original time points that data have arrived at. Hence, each subject's follow-up time has been split into intervals of varying length.

Via the Kalman filter algorithm, I had an imputed value assigned at the beginning of each interval. The longest possible interval between two updates (in the absence of observed data) is 365.25 days. This choice is motivated by the observed frequency of the available data. each subject included in the analysis had at least three eGFR observation per year (90.2% of the full dataset).

With the Kalman filter, one can obtain updates dynamically as frequently as every one day, without substantially complicating the analysis. Since all imputed data comprise the input used in Poisson regression models, in which time is split into person-time intervals with a biomarker imputation being available at the beginning of each interval, we are interested in approximating the continuous hazard of the event by using a piecewise exponential survival function which assumes a constant hazard rate in each interval and can change between intervals. Therefore, I have applied the developed pipeline using a selection of interval lengths to assess how it affects the prediction of time to event.

By using gradually smaller interval lengths, the number of sequential updates, which are not introduced by the arrival of a new biomarker measurement, is increased. The shortest interval length I determined through the time step argument was 1.4 days (365.25 days/256). Note that all interval length choices divide 5 exactly to ensure that we always have an interval starting at the landmark point (Time = 5 years).

Although short interval lengths meet the assumption of a constant hazard within each interval, from an algorithmic point of view, frequent updates (determined by short interval lengths) might increase the error of the estimation of latent states if a value is carried forward for a long time until we have new information.

A peek into results: the closest imputations to actual data were estimated when the specified time step was the one closest to the original arrival rate of the repeated measurements. Although this varies between individuals, the most observed frequency was monthly to weekly.

In the following part, I elaborate more on the interval lengths and how this affects the calibration of predictions. I also present various visualisations of the imputed data to ease comprehension.

Chapter 7

Rationale of imputations before and after the landmark point

As indicated in the previous chapter, I outlined the theory and explained that I first fit a submodel to the longitudinal biomarker and then fit the Poisson model for time to event, including the submodel of the longitudinal component as a time-updated covariate. In this chapter, I describe the fitting of the biomarker submodel in more detail and evaluate the performance of that fitting.

As explained before, the time-split joint model formulation involves two stages: the imputed data obtained by various Kalman filters for the longitudinal eGFR as an extension to the LOCF model in the first step are used as time-varying covariates to fit a Poisson regression model on the training set. Follow-up of individuals who comprise the test set is deliberately censored at Time 5.

The Poisson model accounts for the varying intervals between observations (biomarker data arrive frequently but not periodically) by including an offset term for the logarithm of the interval length. The additional information I provide the model with is an eGFR value, sex, baseline age and baseline diabetes duration. All covariates are standardised, and eGFR has

also been log-transformed.

This method aims to emulate a realistic situation in which I am making forward predictions for individuals who have been followed up for a number of years. Therefore, if enough information is accumulated, it can enable forward prediction, based on a model already trained on a similar population. Unfortunately, due to software limitations bound to `ctsem`, we cannot use a model that has never seen any data for the individuals chosen for forward prediction. Therefore, the training set has to include participants who are fully observed and followed up to the end of the study period and then censor the subjects whom one wants to predict for.

For the task of forward prediction, I have defined a landmark time point, up to which the subject we wish to predict for must be event-free and observable. After landmark time, all data for the particular subject are deliberately censored to evaluate the predictive performance of the trained model on the test data, which are imputed data taken to reflect past trajectories.

To fit the Poisson model, the data of individuals comprising the testing population are not included *after* the landmark time point. Hence, a cross-validation approach is employed to split N event-free individuals up to the landmark point (analysis B is the dominated approach for reasons explained in section 6.3.4) into two folds. Data after five years are deliberately censored if those people are selected for forward prediction. Otherwise, the data are given to the Kalman filter as usual, and imputations rely on this new information. Each fold contains $\frac{N}{2}$ censored individuals. Hence, the model that has not seen data for the group of $\frac{N}{2}$ subjects is then used for forward prediction on this half.

Biomarker data imputed after landmark time have dual use and come from two imputation schemes. The first one yielded data spanning the entire course of follow-up in order to be used for training. For this task, imputed data are computed at (a) pre-determined time points (e.g., 1, 2, 3), which all these individuals share according to a user-chosen time step, and (b) the observed time points of observations for each subject, which apply to each one individually.

Furthermore, inferring a long trajectory from a limited number of observations may be

less reliable. To that effect, I have excluded a few individuals with less than three eGFR measurements throughout follow-up, to ascertain that a realistic long-term trend can be identified and fed back into the Poisson regression for the event. These individuals are 2855 out of 29121.

To evaluate the model's predictive performance, I use the `predict()` function on test data generated from the Kalman filter. However, the frequency of the updates this time only depends on the selected time step (how often the biomarker value needs to be updated). These imputations are unrelated to any observed information after Time 5. This imputation scheme inherently includes more uncertainty because the trajectory does not depend on new observations arriving after the landmark time, but it only depends on the built-up trajectory until that time. As a consequence, the censored follow-up, i.e., the prediction window for each individual who is intentionally censored, is split into equally-spaced intervals, the length of which is determined exclusively by the user's time step. The Kalman filter imputations for the testing folds are generated at selected time points, which do not exceed the maximum duration of the study.

The predicted values, i.e., test data generation process goes as follows: for the user-specified time step, the Kalman filter estimates the latent eGFR states. The prediction intervals span years five until the end of individual follow-up, rounded up to the closest integer[1].

For instance, if a subject is observed for 1965 days = 5.38 years in total, the last estimate will fall into the interval whose upper bound is 6. For significant splits of time, e.g., one-year versus one-week-long intervals, there will be a single or a few imputations that are included in an interval that goes beyond the observed end of follow-up. But the beginning of the last interval always starts before the end of follow-up.

Therefore, every subject in the testing fold has data imputed on the same data points for as long as they are observed. If they are removed from the study, their last interval will be

[1]The end of an interval may not be an integer when the timestep is a fraction of a year. Bottomline is that each prediction interval should always run to its designated end value.

rounded up to the next integer.

To predict the event status, I use the model fitted on the training set to compute the probability of an event in each *person-time* interval from 5 years until the personal end of follow-up. The constructed intervals for which we predict a failure probability represent person-time between imputed (synthetic) biomarker observations: the model is agnostic of any actual data after the landmark time. This scheme is followed for the generation of the test data.

The rationale for letting the last interval run slightly beyond the end of follow-up, instead of truncating at the moment of the event, is to prevent the model from miscalculating the predicted risk at the interval. The time of event is not included in these predicted data: the model has been outcome-agnostic for the individuals included in the test set. All prediction intervals are of equal length, with the last running to a predetermined end to allow for equal chances of an event happening within the interval of choice.

7.1 What is being projected forwards after five years?

7.1.1 How do the observed and predicted values compare?

The figures 7.1 to 7.4 included in this section show the original and imputed biomarker data of a randomly selected individual. They showcase the two imputation schemes: each Poisson model is fitted on eGFR imputations generated according to the true arrival times of new observations that fall into the first five years of follow-up, plus imputations at pre-determined time points retrieved by the chosen time step. This imputation scheme applies to the data generation of individuals used to *fit* the models. The imputed data of individuals used to *test* the fitted models, which cover years five to ten (at most), are generated differently. There is a value yielded for pre-defined, equally-spaced time points (same time step according to which I generated values before landmark time). The trajectory of the test individuals is updated when new data arrive strictly up to Time 5. After that, the trajectory is projected forward

based on what has been learned so far.

The following visualisations help to understand how the imputed trajectories look like before and after year five, and the data that go into training/testing, respectively.

Each Poisson model has been trained with imputed data from a Kalman filter up to Time 5 for test individuals. The test imputations are given by the smoothed segment (coloured either red or green, depending on the fold the individual belongs to), which starts at the dashed vertical line. From this point forward, the trajectory is not updated according to new biomarker data. Thus, the trajectory looks relatively flattened: it is converging to a long-term average. In the case of the LOCF model for the longitudinal eGFR data, the trajectory after Time 5 remains flat. In the alternative scenario where the subjects do not get intentionally censored at training in order to be used as test subjects afterwards (which happens in one of the two folds for a given individual each time), all data coming after Time 5 are taken into account as normal. In this case, the Kalman filter imputations follow the observed values (as shown in the following figures) for the entire length of follow-up, i.e., landmarking does not affect the data generation process.

Fig 7.1 depicts the observed and imputed trajectories of a random subject who did not progress to renal replacement therapy until the end of the study. The black data points give the original observations. The selected subject did not experience an event within the first five years of follow-up, thus, is eligible to get intentionally censored for forward prediction. Due to cross-validation, the subject will be fully observed in one of the two folds, shown by the (green line) and censored in the second one. The red line depicts the trajectory that stops being updated after five years (landmark time). The diagram shows all the imputed eGFR latent states up to five years (jagged red segment) and the projected trajectory (smoothed red segment) in the absence of new data, conditional solely on past values. The green and red paths (each one corresponds to one fold) are negligibly different in places before Time 5, because they come from two runs of the Kalman filter with an LMM with drift and diffusion

configuration.

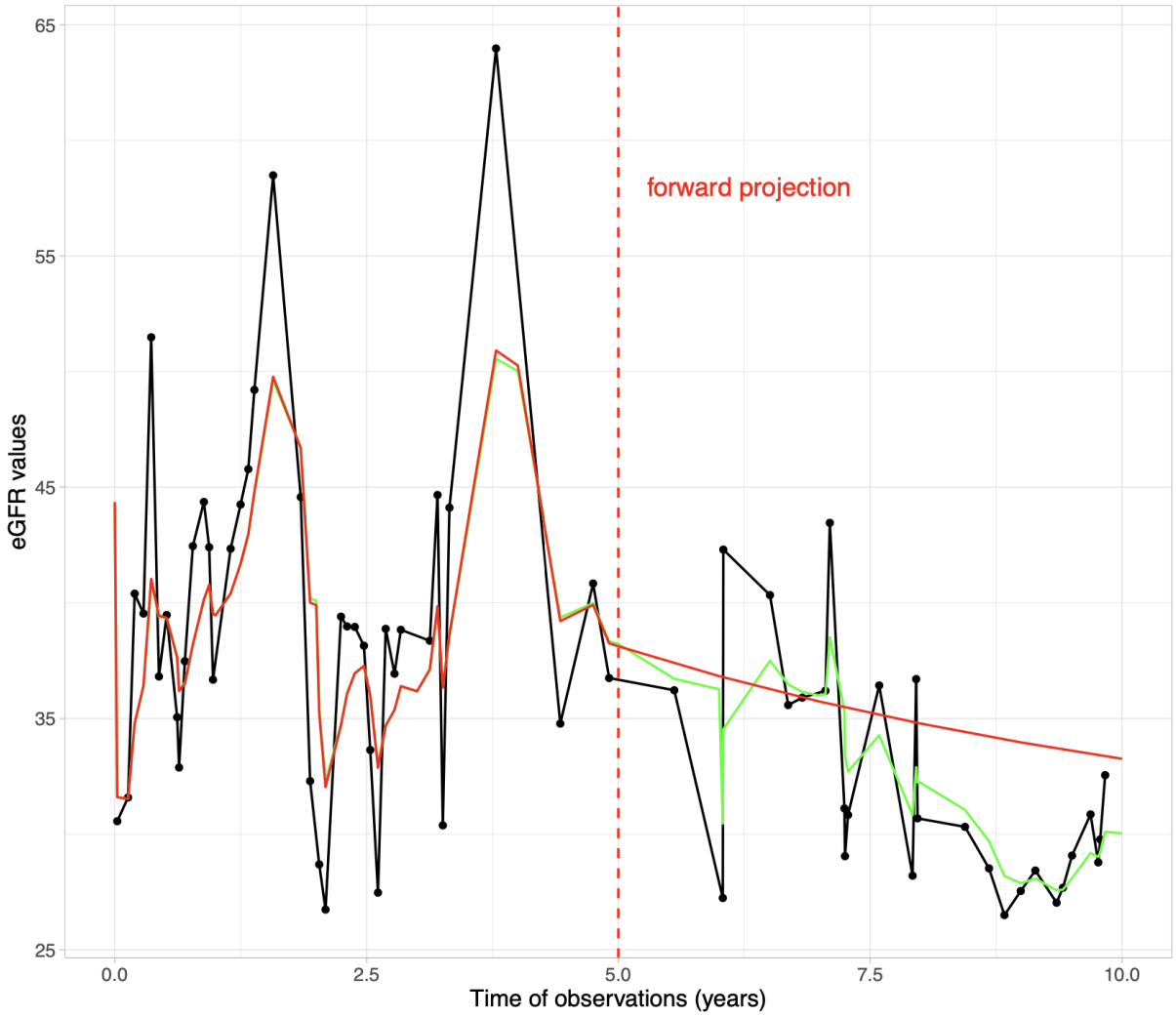


Figure 7.1: Observed trajectory (black). Full follow-up imputed eGFR based on LMM, drift and diffusion (green). Time-censored, imputations until 5 years are updated conditional on the real data (red). After landmark time, the real data are unavailable, and the learned trajectory is projected forwards (shown by the smoothed segment) to assess predictive performance.

Figure 7.2 depicts the same individual as figure 7.1 with baseline eGFR 30, who survives follow-up, where a range of state-space models is compared for eGFR imputations. The trajectories closely follow the distribution of observed data until censoring time. After five years, the observed data are withheld, and the path is extrapolated forward, conditional on previous Kalman filter updates (red smoothed segment). Whether the projected path is slope-free or not depends on the Kalman filter configuration. We observe that the diffusion

model's projection closely approximates the LMM, drift and diffusion trajectory for that individual. Therefore, it is seen that the diffusion component captures the fluctuations in changing eGFR, while the mixed-effects component conveys the underlying long-term trend. An LMM with just slopes or the drift model does not render an adequately realistic situation.

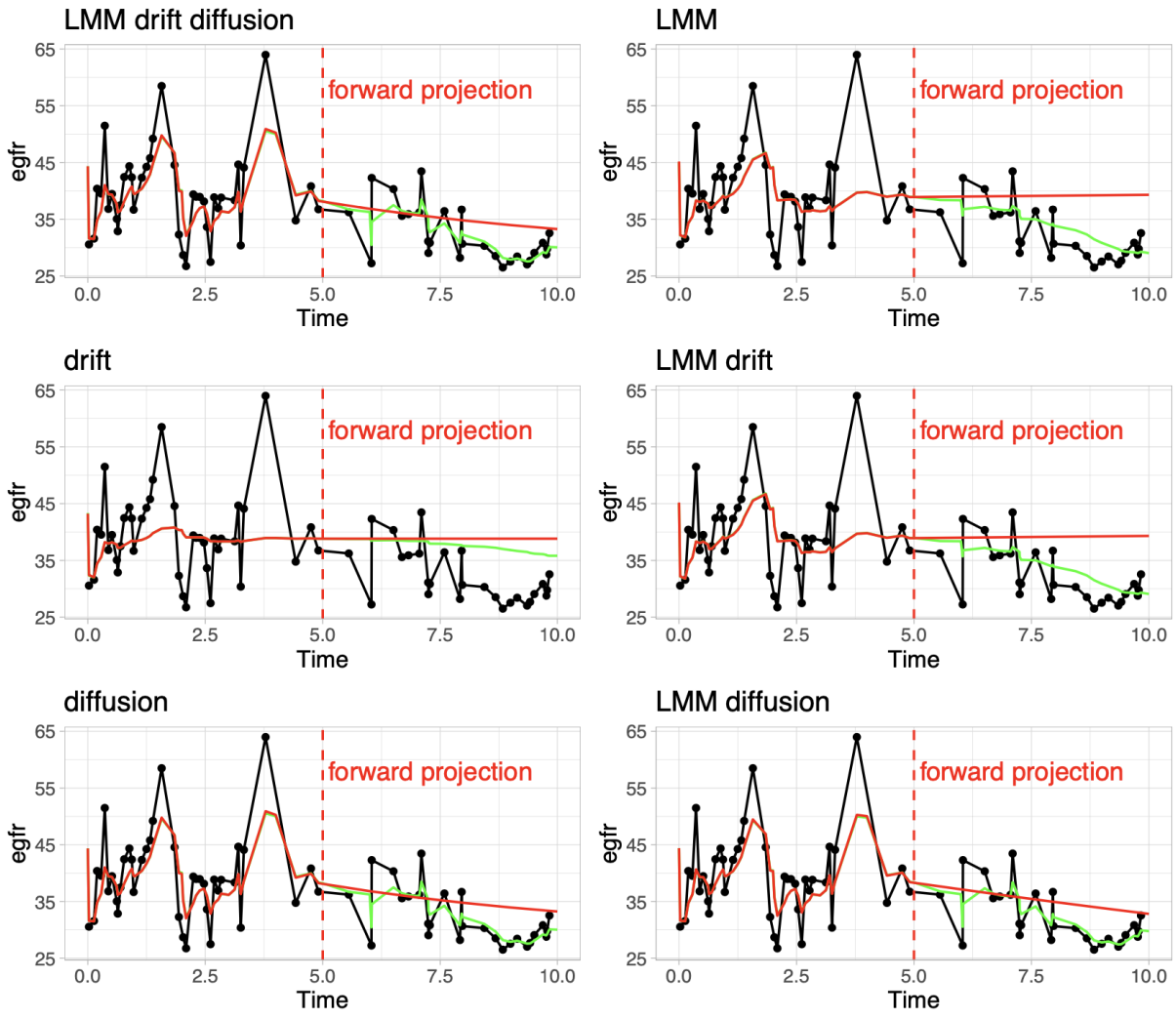


Figure 7.2: Subject 11, baseline eGFR 30. Observed values (black). Continuously-updated imputed trajectory (green). Kalman filter imputations are generated as new data arrive until year 5, comprising the training input (red). The predictive performance is evaluated based on imputations shown by the smoothed segment after year 5 every 21 days.

Figure 7.3 depicts a random individual with baseline eGFR 50 who progressed to renal replacement therapy during their 8th year of follow-up. A range of state-space models is compared for imputed eGFR. The subject's trajectory of eGFR has a downward trend, monotonically decreasing, which is best described by the LMM model (slopes) with drift. When the diffusion process is included in this model, it places the trajectory higher, which is not coinciding with reality.

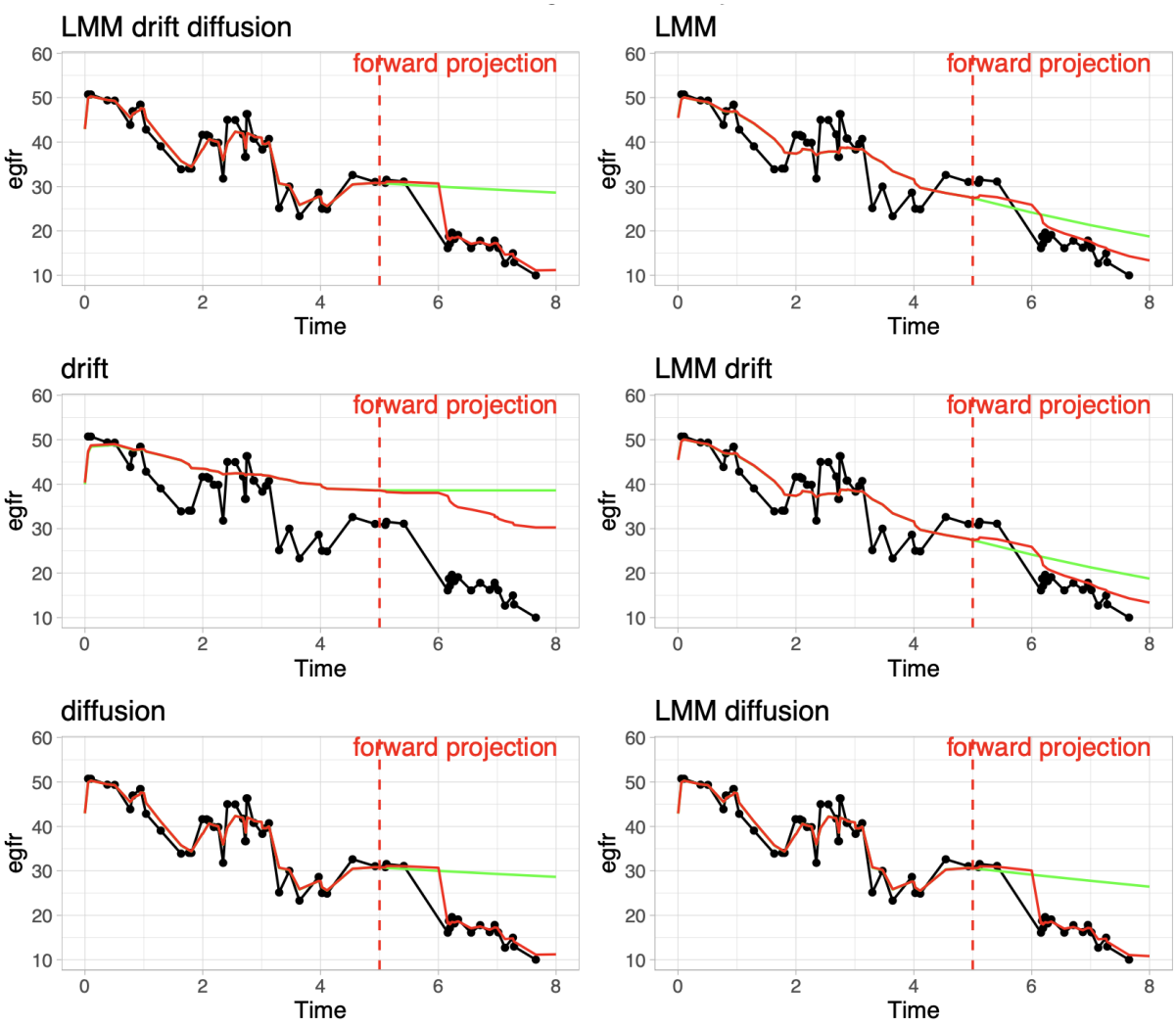


Figure 7.3: Subject 118, baseline eGFR 50. Observed values (black). Continuously-updated imputed trajectory (green). Kalman filter imputations are obtained as new data arrive until year 5, comprising the training input (red). The predictive performance is evaluated based on imputations shown by the smoothed segment after year 5 every 21 days.

Figure 7.4 depicts an individual with baseline eGFR 84 who survives follow-up. Fluctuating

eGFR, which falls from year 2 to year 7 and rises thereafter. The mixed-effects and diffusion model better depicts the rate of change of eGFR. On the other hand, the LMM with drift makes the projected trajectory look more optimistic than the actual situation.

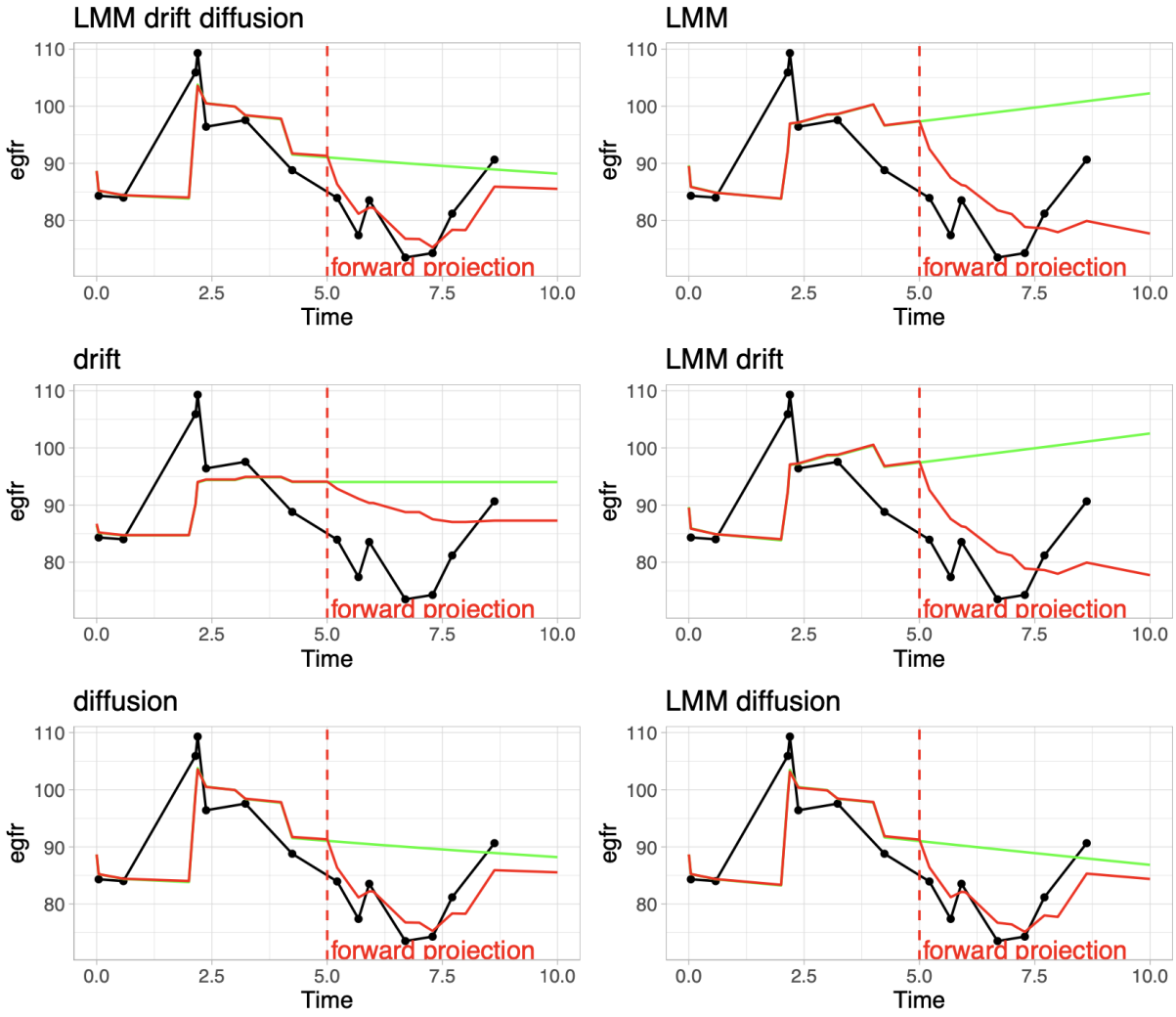


Figure 7.4: Subject 1, baseline eGFR 84. Observed values (black). Continuously-updated imputed trajectory (red). Kalman filter imputations generated after year 5 (green). The predictive performance is evaluated based on imputations shown by the smoothed segment after year 5 every 21 days.

Those imputations, therefore, may not be as accurate as the information used for training due to the lack of newly observed data after the threshold of five years. Imputations used in model training depend on new data coming throughout follow-up, whereas for forward predictions, all observed data are censored after landmark time (at the beginning of the prediction window). This absence of external information is necessary when testing the model's predictive performance to avoid biased estimations that stem from data from the future.

Hence, to guess future biomarker values and, subsequently, time to event, only the trend built up to the landmark point informs future trajectories. This difference in the generation of train and test biomarker data could translate into reduced model calibration, which is dealt with by ignoring the censoring of the interval containing the event, as explained further below.

7.2 Evaluation and comparison of the submodels of eGFR

Model comparison has been based on the log-likelihood given the training data with a penalty for the number of parameters. I have used the log-likelihood, the number of parameters and the Akaike information criterion (AIC) of each model as means for model ranking. The AIC is defined as $2k - 2L_{train}$ where k is the adequate number of parameters and L_{train} is the log-likelihood given the training data. As a rule of thumb, AIC can be used to compare models run with different parametric forms, with the lowest AIC indicative of the best fit. In fact, among the seven developed *ctsem* models, the LMM, which includes drift and diffusion components, presents the lowest deviance ($-2L_{train}$) and the lowest AIC, in every used sample of the filtered group (baseline eGFR < 60 mL/min/1.73 m^2). An extract of a single fold is shown in table 7.1.

The fit of each model is standardised. I have used the LMM as the reference model, and reported the differences from the reference model (L_{train} of model X - L_{train} of LMM, AIC of model X - AIC of LMM). The larger the deviation from the reference model, the better the fit of the model.

Therefore, the best Δ Loglik is the largest positive value, whereas the best Δ AIC is the lowest negative value. The LMM with drift and diffusion has the lowest deviance and the lowest AIC. It has been found that including a diffusion term in the longitudinal model outperforms the models without it in terms of fitting the biomarker data.

The measures used to compare the observed versus the predicted values on the test data after the five-years landmark have been discussed in the predictive evaluation chapters 8 and 9.

Table 7.1: Model comparison using the log-likelihood and Akaike Information Criterion (AIC)

Long/nal model	Δ Loglik	N of pars	Δ AIC
model.lmm	0.0	15	0.0
model.drift	-12502.9	13	25001.7
model.diff	5071.8	13	-10147.7
model.driftdiff	5078.0	17	-10152.1
model.lmmdiff	5250.5	19	-10493.1
model.lmmdrift	3.3	19	1.4
model.lmmdriftdiff	5285.6	23	-10555.3

A number of graphical illustrations is given next (figures 7.5 to 7.8) that help us compare the various model fits for a random group of subjects. The graphs feature a random sample of subjects, two of whom are censored at the landmark time (subjects 56 and 60) and utilised for prediction, and the remaining three are followed up until the end of the study. As before, eGFR is log-transformed and standardised to facilitate presentation.

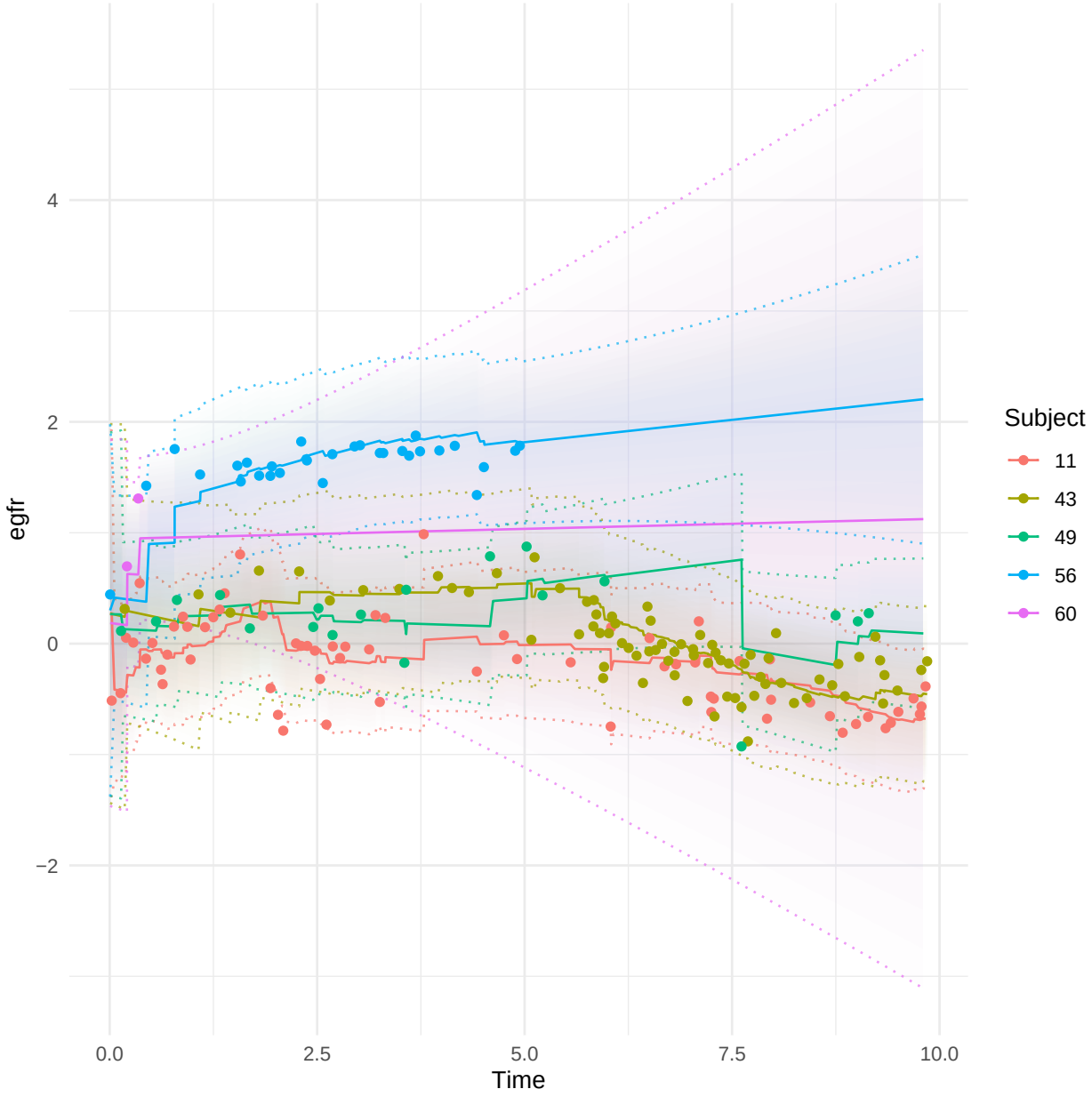


Figure 7.5: Here, we inspect trajectory plots developed based on the LMM. We see that the shrinkage towards the population mean could be better. The individual intercept and slope dominate in this example.

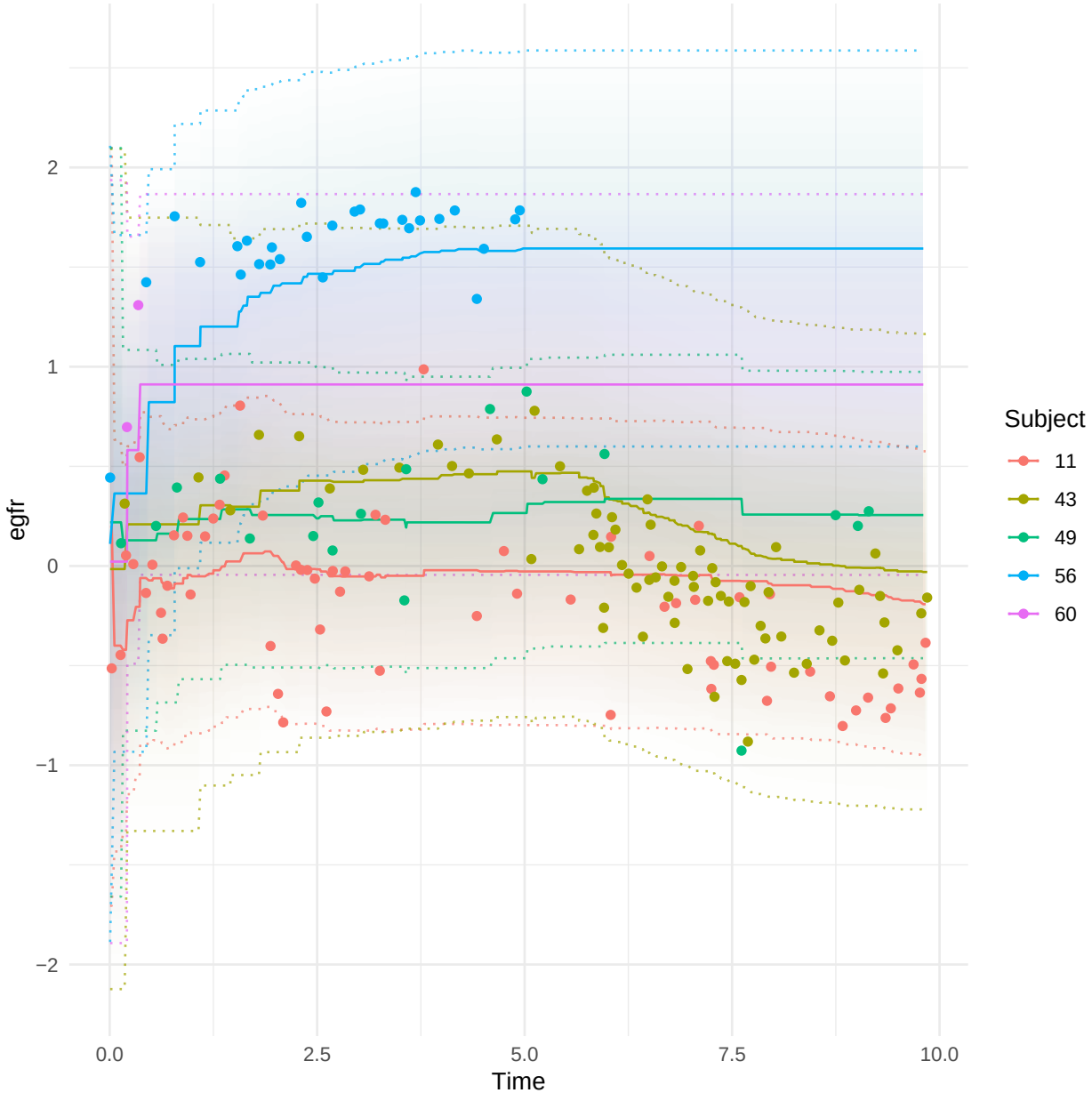


Figure 7.6: Inspecting trajectory plots developed based on the drift model.

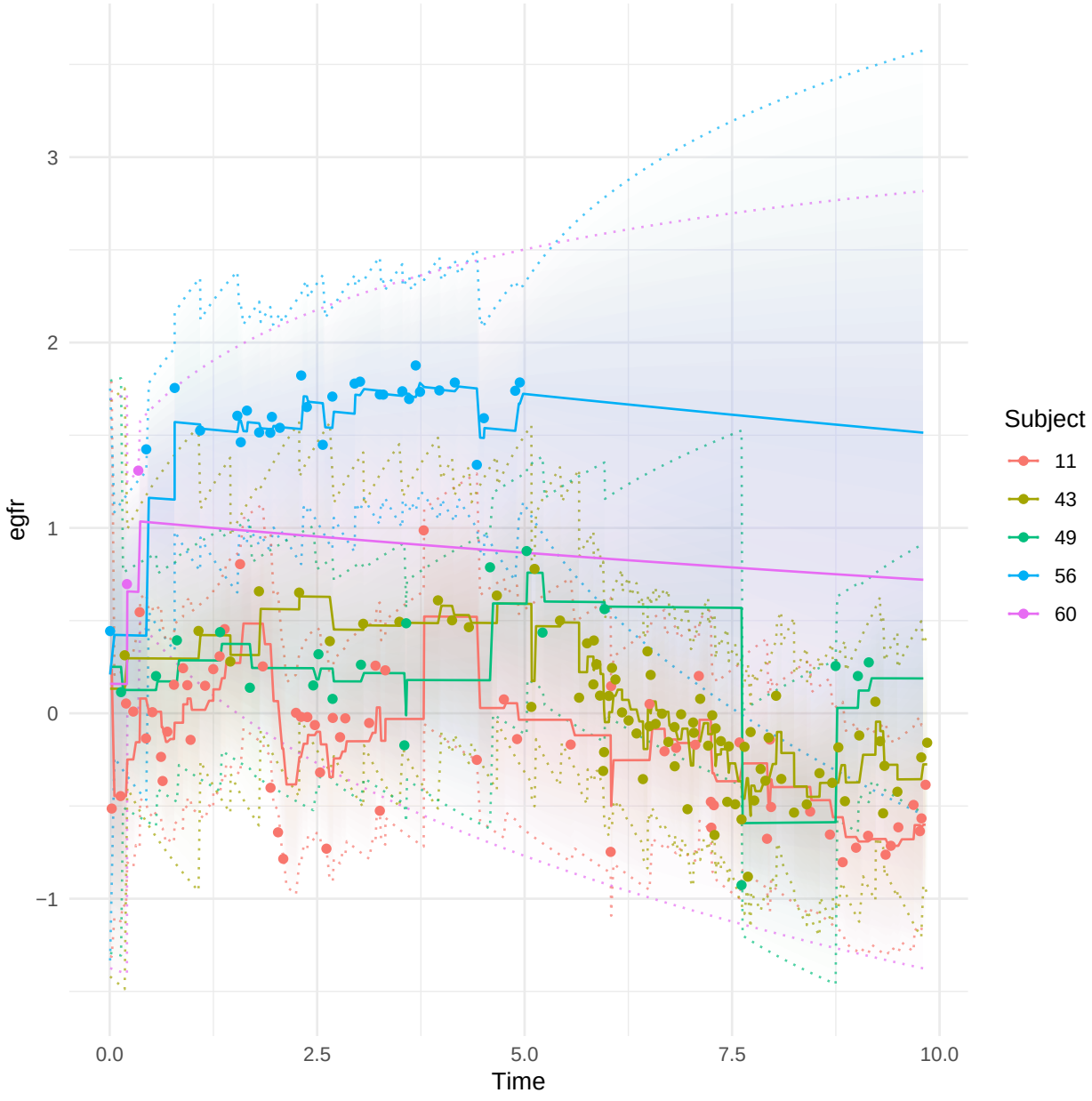


Figure 7.7: Inspecting trajectory plots developed based on the drift and diffusion model.

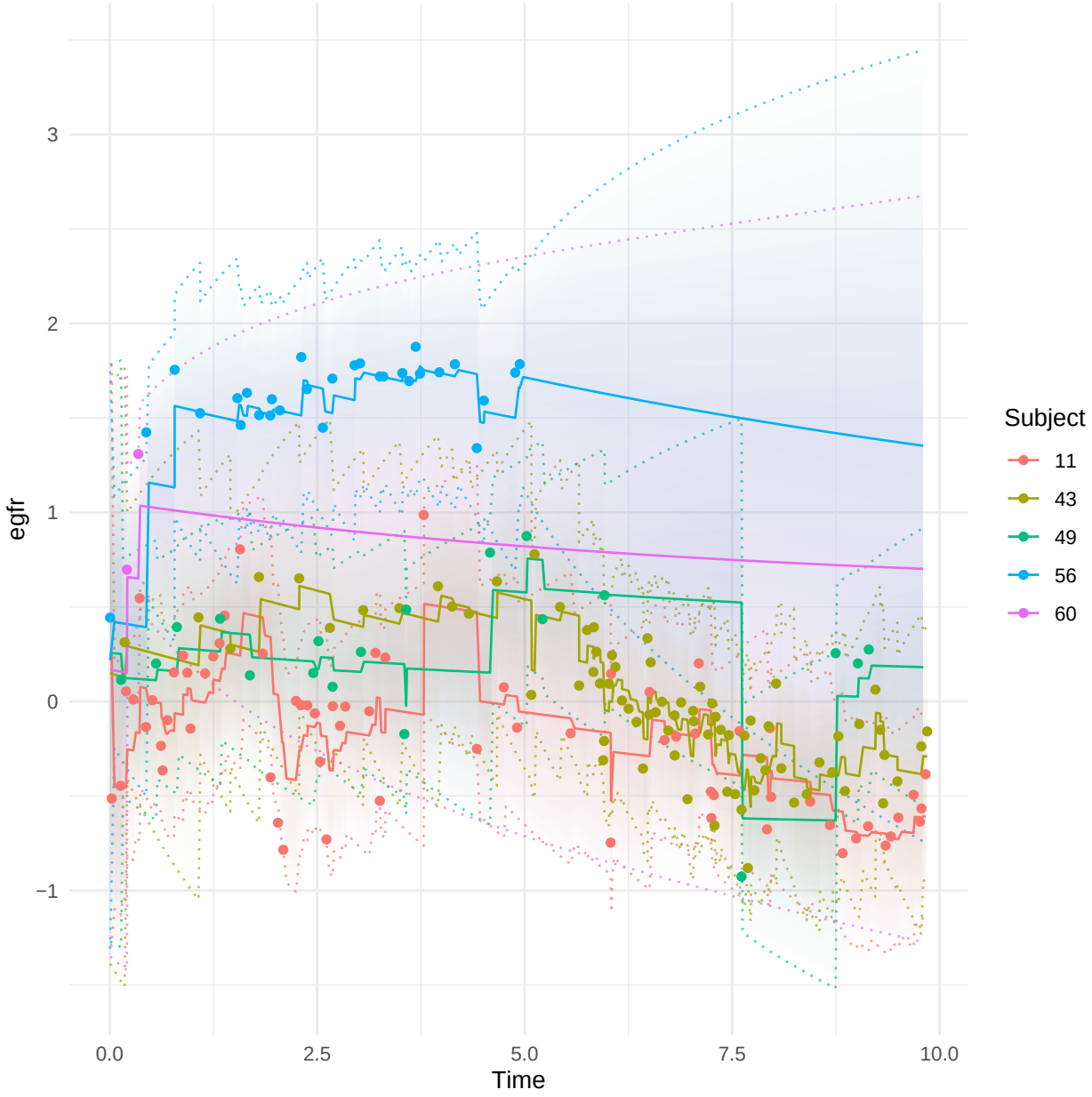


Figure 7.8: Inspecting trajectory plots developed based on the LMM with drift and diffusion. Extending the LMM to allow for drift and diffusion allows the trajectories to curve slightly away from a linear path towards the population mean.

7.3 Discussion

In this chapter, I have given some examples to ease comprehension regarding the different imputation configurations that apply to the developed method. The figures may have provided a better understanding of how the synthetic data are being used. When the updating mechanism is agnostic to the external information arriving after the landmark time, the generated imputations are then used as test data to evaluate the fitted Poisson models. When the updating mechanism generates imputations conditional on external information, the data are used to train the various Poisson models.

The more frequently (routinely) the biomarker data are updated in real life, the better the dynamic techniques approximate the rate of progression to renal complications. That, in turn, allows clinicians to make informed decisions about the course of treatment and time until the next assessment.

I have shown that differential equations to specify biomarker trajectories better fit the longitudinal data over a basic LMM. An additional strength of using a broader family of continuous-time state-space models via `ctsem` is that it facilitates the inclusion of all available biomarker data, without putting restrictions in terms of how sporadically or asynchronously (in the case of modelling more than one biomarker) the measurements have been. Time-varying intervals between observations are supported in `rstanarm` too, but `rstanarm` requires the biomarker data to have been measured synchronously, i.e., the observed time points to indicate all biomarker measurements, if more than one. This is particularly restrictive and resulted in the discarding of a considerable amount of unmatched biomarker observations (HbA_{1c}, eGFR). Moreover, the need to average data to represent a fixed period of one interval using a single value is circumvented by using `ctsem`.

Furthermore, every diagram within the chapter that shows how the models may compare proves that, different components of the longitudinal model are required to capture different aspects of the process, and different rates of change for unique trajectories and individuals.

For example, when the rate of change can be characterised as deterministic, the LMM coupled with a drift term outperforms the LMM models that include diffusion terms. However, in case of random fluctuations in individual levels of eGFR, the diffusion process is better suited for learning the underlying variance in the data.

It is observed that including autoregressive drift does not fit well with fluctuating eGFR data. Drift's tension to pull the new imputations towards the long-term average explains this poor fitting. It is thus resembling, in many cases, the LOCF imputation. In general, drift is helpful when fitting longitudinal data in which the rate of change is primarily monotonic.

The bottom line is that whether drift or diffusion improves the fit to the data depends on the overall trend of the biomarker of each individual, meaning the evolution and the derivatives of the process over time. Diffusion may capture nudges in the data but broadly converges to an average value when used for the trajectory projection. Conversely, an LMM coupled with drift effects predicts satisfactorily the future states, with a monotonic biomarker's trajectory.

Therefore, as a rule of thumb I support fitting many possible specifications for projecting longitudinal trajectories using the functionality available in 'ctsem' and, after assessing them based on their log-likelihood and using their AIC, choosing the model that provides the closest fit to the data.

In both datasets analysed (filtered and full cohorts), the LMM with drift and diffusion specification has the best fit to the longitudinal eGFR data. However, to avoid overfitting, I have utilised all the longitudinal submodels constructed for the time-updated eGFR, and employed a cross-validation approach to assessing the predictive performance of Poisson regression models with respect to time until a renal replacement therapy. In the following evaluation chapter, I present the fitting of the Poisson models when using time-updated data given by different Kalman filter specifications and evaluate their predictive performance compared to an LOCF-based Poisson model.

Chapter 8

Comparison of the use of `ctsem` with LOCF for fitting joint models for eGFR and RRT

In the previous chapter, I explained and evaluated the various models built for the longitudinal data. Before detailing that, I first gave more details about the content of the dataset used for this modelling, some elements of the dataset and the setup for the structural equation modelling, followed by details for model fitting. Here, I elaborate on the Poisson models fitting using the data generated from the Kalman filter as time-updated covariates. Then I compare those models to a simpler Poisson model that uses as time-updated eGFR a trajectory made by carrying the last available observation forward.

To recap, the Poisson time-splitting joint model has been implemented as follows: a Kalman filter configuration is used first to obtain imputations of latent states of eGFR at the beginning of each person-time interval. This imputation step is completed very fast compared to the alternative modelling within `rstanarm`. In the second step, the estimated trajectories of eGFR are fed as a time-updated covariate into the Poisson regression models for time to

event. I have used a cross-validation approach to improve the fitting of the Poisson regression models.

The current chapter concerns the fit and predictive performance of the developed Poisson models for time to RRT, conditional on time-updated eGFR data. Two separate analyses have been performed using the full cohort, one being a subset of the other, with baseline eGFR at most 60 mL/min/1.73 m^2. The main dataset is an extract of the national T1D population with linkage to renal outcome records from the national renal registry.

Finally, the chapter discusses the refinements made in the analysis to include B-splines for fitting the time-updated eGFR data. Then, I present the performance characteristics of the models I have fitted. The log-likelihood and AIC show how well the models fit the data for the number of parameters that have been specified. Using these metrics, one can evaluate the strength of the association between the parameters and the outcome, i.e., the coefficients of the covariates. This is how the probability of an event is estimated, given that the log odds for the event are a linear combination of the covariates included in the model. In this case, the function that converts the log odds to probability is the logistic function.[1]

Hence, the reader can expect to see the following in the coming sections:

1. Results of the model fitting procedure and model calibration
2. The impact of censoring and calibration per decile of predicted risk
3. Using B-splines for trajectories made up using LOCF and `ctsem`
4. Calibration of model performance with and without fitting B-splines for eGFR
5. Overall insights before proceeding to the final remarks in the last result chapter

[1]The sigmoid curve is given by the equation $f(x) = \frac{L}{1+exp(-k)(x-x_0)}$, where x_0 the x value of the sigmoid midpoint, L is the maximum value of the function and k is the logistic growth rate (steepness) of the curve.

8.1 Do we learn the same Poisson model regardless of how eGFR is modelled?

For the task of forward prediction of progression to end-stage renal disease, I have fitted Poisson models with imputed eGFR data as time-updated covariates. As part of a joint model construction, each Poisson model has been given a longitudinal eGFR submodel, one specification out of the eight available (seven longitudinal submodels obtained by using the Kalman filter algorithm and the last one by using the LOCF approach), and has been fitted using the `glm()` function with a Poisson likelihood for the rate of event. The Poisson equation for the current application is:

$$\text{RRT rate} = exp(\beta_0 + \beta_1 \text{baselineage} + \beta_2 \text{gender} + \beta_3 \text{baselineduration} + \beta_4 \text{egfr} + \text{offset(log(interval length))}$$

The interest lies in assessing how different the coefficients are estimated from the different fitted models. The outputs of the eight fitted Poisson models that employ the values of eGFR taken by the longitudinal models fitted with `ctsem` are given for comparison in Ttables 8.1, 8.2, 8.3, and 8.4. All models correspond to the filtered group (eGFR $<$ 60 mL/min/1.73 m^2), and apart from the time-updated eGFR, they include time-invariant information about subjects' baseline age, baseline duration of diabetes and sex. The last three covariates are standardised and given at the beginning of each person-time interval as years elapsed since entry. Finally, each model accounts for the duration of each person-time interval as an *offset* term (tstart).

Table 8.1: Poisson model fitted to latent biomarker values imputed by a Kalman filter based on an LMM (columns 1, 2, 3), and a drift effects model (columns 4, 5, 6)

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-3.792	0.087	0.000	-4.580	0.111	0.000
tstart	-0.082	0.023	0.000	0.112	0.019	0.000
gender	-0.096	0.052	0.063	-0.165	0.051	0.001
baselineage	0.132	0.051	0.010	-0.061	0.046	0.187
baselineduration	0.120	0.054	0.026	0.088	0.058	0.127
egfr	-1.035	0.033	0.000	-1.792	0.051	0.000

Table 8.2: Poisson model fitted to latent biomarker values imputed by a Kalman filter based on a diffusion model (columns 1, 2, 3), and a drift-diffusion model (columns 4, 5, 6)

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-4.752	0.116	0.000	-4.765	0.116	0.000
tstart	0.017	0.020	0.389	0.019	0.020	0.325
gender	-0.102	0.051	0.046	-0.115	0.051	0.024
baselineage	0.096	0.046	0.039	0.102	0.046	0.029
baselineduration	0.118	0.058	0.042	0.111	0.059	0.059
egfr	-1.689	0.044	0.000	-1.701	0.044	0.000

Table 8.3: Poisson model fitted to latent biomarker values imputed by a Kalman filter based on an LMM with diffusion model (columns 1, 2, 3), and an LMM drift model (columns 4, 5, 6)

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-4.195	0.096	0.000	-3.778	0.087	0.000
tstart	-0.061	0.021	0.004	-0.082	0.023	0.000
gender	-0.144	0.051	0.005	-0.095	0.052	0.067
baselineage	0.176	0.050	0.000	0.134	0.052	0.009
baselineduration	0.105	0.056	0.062	0.109	0.054	0.043
egfr	-1.350	0.037	0.000	-1.017	0.032	0.000

Table 8.4: Coefficient comparison of the full Poisson model based on a Kalman filter including LMM, drift and diffusion processes (columns 1, 2, 3), and an LOCF model (columns 4, 5, 6).

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-4.082	0.093	0.000	-4.921	0.117	0.000
tstart	-0.069	0.022	0.001	0.070	0.019	0.000
gender	-0.218	0.051	0.000	-0.091	0.051	0.074
baselineage	0.193	0.052	0.000	0.127	0.045	0.005
baselineduration	0.114	0.056	0.043	0.134	0.058	0.021
egfr	-1.248	0.035	0.000	-1.516	0.038	0.000

The ‘gender’ predictor was coded as numerical, although any linear transformation of its

values does not change either the shape of the distribution or the correlation of scores on that variable with those of any other variable. This does not make any difference in terms of fit, but it hides the label of the non-reference sex.

Typically the information that can be found in model summaries is the estimated Poisson regression coefficients with standard error, the estimated relative rates (i.e., the exponential of the coefficients), a Wald test statistic (testing the null hypothesis that the regression coefficient is zero or, equivalently, the relative rate of event associated with this explanatory variable is unit) and an associated P-value.

In a Poisson regression model, the coefficients represent the logarithm of the expected count of the outcome variable per unit change in the corresponding independent variable. Analogously, the coefficients in the Poisson models can be interpreted as the expected change in the log of the rate of event occurrence corresponding to a unit increase in the predictor of interest, holding constant all other predictors in the model.

Since the exponential function is used in the Poisson regression equation, a negative coefficient implies that as the value of the independent variable decreases, the expected count of the outcome variable increases. The outputs of the models are comparably close. Tables 8.1 to 8.4 show how the coefficients compare depending on the model specification for eGFR. In all Poisson models, among eGFR, baseline age, sex, and baseline diabetes duration, the strongest impact on the incidence rate of RRT stems from eGFR; as its value decreases over time, the expected rate of event increases.

The outputs suggest that while the drift and diffusion LMM has been shown to fit best the longitudinal data of the filtered group (lowest deviance and AIC), demonstrated at table 7.1, the fit of the Poisson regression models for the rate of event does not deviate significantly among those model specifications for the time-updated covariate; eGFR always has the largest effect on the rate of event. More particularly, the drift model identifies the strongest effect of eGFR on the rate of event (coeff: -1.792), and the LMM model and LMM drift estimate

the lowest effects (coeffs: -1.035 and -1.017, respectively). In this filtered population, the Poisson based on the LOCF model does quite well (coeff: -1.516), presumably due to the strong observed autoregressive effect in the data, the last observation is the most informative of the coming one.

I have also fitted a Poisson model using B-splines with six degrees of freedom to specify the effect of the arrival times of each biomarker measurement on time to event and reviewed the fitting. Table 8.5 shows the summary of such a model for reference and comparison of the coefficients derived from the Poisson models that do not employ splines for the timings. However, to simplify the analysis, I did not employ B-splines in the final construction, as observation times progress linearly with follow-up.

Table 8.5: Poisson time-splitting model fitted to latent biomarker values imputed by a Kalman filter specifying an LMM with diffusion and drift effects.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.02	0.63	-9.49	0.000
gender	-0.33	0.07	-4.76	0.000
diabetesduration	0.31	0.08	4.11	0.000
entryage	0.40	0.07	5.51	0.000
entryhba1cavg	0.24	0.08	2.86	0.004
hba1cvalue	0.45	0.09	4.93	0.000
egfrvalue	-0.31	0.08	-3.94	0.000
splines::bs(tstart, df = 6)1	1.94	0.97	2.00	0.045
splines::bs(tstart, df = 6)2	1.18	0.62	1.89	0.058
splines::bs(tstart, df = 6)3	2.19	0.78	2.79	0.005
splines::bs(tstart, df = 6)4	0.74	0.75	0.98	0.325
splines::bs(tstart, df = 6)5	1.82	0.84	2.18	0.029
splines::bs(tstart, df = 6)6	1.78	0.75	2.38	0.017

The following part summarises the coefficients of fitting the various Poisson models, including eGFR trajectories specified by LMM, drift and diffusion models with and without slopes, and an LOCF model, using biomarker data from the full cohort for visual comparison (tables 8.6, 8.7, 8.8, and 8.9). The effects of the covariates on the incidence rate are milder in this non-filtered population with T1D, with age at baseline (at study entry) being less statistically significant. The coefficients of eGFR remain larger and negative as well; decreasing values have a greater effect on the rate of event, which matches a clinician's intuition. Moreover, the

strongest eGFR effect is again found by using the autoregressive drift effects model (coeff: -1.960) and the lowest effect is estimated by the LOCF model (coef: -1.452). The remaining model specifications estimate a coefficient of around -1.60, which implies that both slopes and drift are important in order to capture the effect in this wider population with unfiltered eGFR at baseline.

Table 8.6: Poisson model fitted to latent biomarker values imputed by a Kalman filter, including an LMM (columns 1, 2, 3) and drift model (columns 4, 5, 6) as part of the time-splitting joint modelling approach applied to the full cohort.

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-7.762	0.372	0.000	-7.889	0.365	0.000
tstart	0.144	0.047	0.002	0.230	0.046	0.000
gender	-0.226	0.067	0.001	-0.177	0.066	0.008
baselineage	-0.128	0.073	0.078	-0.426	0.075	0.000
baselineduration	-0.224	0.070	0.001	-0.229	0.075	0.002
egfr	-1.600	0.038	0.000	-1.960	0.050	0.000

Table 8.7: Poisson model fitted to latent biomarker values imputed by a Kalman filter, including diffusion (columns 1, 2, 3) and drift-diffusion effects (columns 4, 5, 6), as part of a time-splitting joint modelling approach applied to the full cohort.

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-7.449	0.365	0.000	-7.427	0.365	0.000
tstart	0.088	0.047	0.060	0.085	0.047	0.070
gender	-0.199	0.067	0.003	-0.197	0.067	0.003
baselineage	-0.125	0.075	0.097	-0.121	0.076	0.109
baselineduration	-0.234	0.071	0.001	-0.247	0.072	0.001

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
egfr	-1.610	0.036	0.000	-1.617	0.036	0.000

Table 8.8: Poisson model fitted to latent biomarker values imputed by a Kalman filter, including LMM and diffusion (columns 1, 2, 3) and an LMM and drift effects (columns 4, 5, 6), as part of a time-splitting joint modelling approach applied to the full cohort.

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-7.453	0.366	0.000	-7.747	0.371	0.000
tstart	0.087	0.047	0.065	0.141	0.047	0.002
gender	-0.195	0.067	0.003	-0.231	0.067	0.001
baselineage	-0.124	0.076	0.101	-0.113	0.073	0.124
baselineduration	-0.229	0.071	0.001	-0.246	0.071	0.001
egfr	-1.606	0.036	0.000	-1.603	0.038	0.000

Table 8.9: Coefficient comparison of the full Poisson model based on a Kalman filter including LMM, drift and diffusion processes (columns 1, 2, 3), and an LOCF model (columns 4, 5, 6) for the full cohort

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-7.421	0.365	0.000	-7.674	0.373	0.000
tstart	0.084	0.047	0.073	0.124	0.047	0.009
gender	-0.170	0.066	0.010	-0.128	0.066	0.053

	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
baselineage	-0.146	0.076	0.053	-0.041	0.075	0.584
baselineduration	-0.258	0.073	0.000	-0.143	0.070	0.042
egfr	-1.611	0.036	0.000	-1.452	0.031	0.000

8.2 Impact of censoring on model calibration

As mentioned before, the offset term in the Poisson regression model allows for varying lengths of intervals, particularly for the last interval of each individual being truncated, if an event has occurred. Therefore, both censoring and unequal lengths of intervals are correctly accounted for, when computing the fitted values (of predicted events). I have confirmed that the observed and expected number of events perfectly agree within the training sets, verifying that the fitted models are perfectly calibrated, as described in section 2.6.2.

Subsequently, the fitted Poisson models have been used to estimate the expected number of events in the test datasets. The peculiarity of this case is that the time of event is not known until it is observed, thus the last intervals must run all the way to the end and not be truncated in order to return meaningful results. Furthermore, I have observed that model calibration is better when the trajectory projected forward depends on a Kalman filter estimation rather than the LOCF scheme for imputing the unknown test data. This was established by assessing how the hazard rates change as we progressively shorten the interval lengths. I have experimented with the following interval lengths for obtaining imputations:

1. 365.25 days,
2. 365.25/2 = 182.625 days,
3. 365.25/4 = 91.3125 days,
4. 365.25/8 = 45.65625 days,
5. 365.25/16 = 22.82812 days,
6. 365.25/32 = 11.41406 days,
7. 365.25/64 = 5.707031 days,
8. 365.25/128 = 2.853516 days,
9. 365.25/256 = 1.426758 days.

I have only used values that divide 5 years (5*365.25) exactly to ascertain that we always have an interval starting at the landmark point. However, working with truncated representations of floating numbers is better avoided, due to various implications that it can bring to the analysis, such as complicating comparisons when rounding to different decimal points, and reproducibility.

Furthermore, as a sanity check, I fitted a Poisson model on the test data like they were training data, and used this model to assess calibration and ensure that there was not an artefact hidden in the data. Correctly, the sum of fitted values agreed with the sum of the `predict()` output regarding the expected number of events.

Last but not least, I have investigated calibration patterns in each year of follow-up separately. That can be extended to all months and weeks that comprise follow-up. The rationale for this check was to evaluate how well the model performs in terms of predicting time to event accurately as we go further into the future. It would make sense that the further we go into the future, the less accurate the predictions become. However, this was not conclusively observed, as the quality of predictions seems independent of whether there is a long-term trend to be captured on an individual trajectory, or the direction of change is too arbitrary to be deduced.

8.2.1 Calibration plots per decile of predicted risk

The discrimination metric in statistical theory measures how often a model estimates a higher risk for individuals who experience the event of interest than those who do not. Similarly, calibration is the ability of a model to assign accurate probabilities of an event occurring at a particular time.

The person-time of the training set is censored, either at the first event or exit from the study due to other causes. On the contrary, the last interval of the test set is not censored to avoid biased results: time of event is not known a priori. As a result, the length of follow-up in

the test set gets slightly inflated, i.e., it may run negligibly beyond the individual's observed follow-up to meet the condition of having complete, non-truncated prediction intervals.

Some implications stemming from whether the last interval of an individual is censored at the time of event, or it runs all the way to the end in the *training* data this time are relevant to mention and demonstrate.

8.2.2 Implications

The fitted values generated by the function `glm()` with Poisson likelihood are the expected numbers of events up to the end of each person-time interval, allowing for censoring at time of event. The fitted values are obtained by transforming the linear predictors by the inverse of the link function used. On a general note, the link function generalises linear regression by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

An agreement between observed and fitted sums of the response variable is guaranteed when maximising the fitting of a model with a likelihood in the exponential family. However, unequally spaced person-time intervals cannot be considered exchangeable. Therefore the response variable (event/no event) on these intervals cannot be represented as a mixture of independent and identically-distributed Bernoulli variables.

Hence, the computation of the expected number of events in each person-time interval, while allowing for censoring at first event, does not naturally rank individuals by predicted risk. However, the fitting infers how many events to expect in each interval, primarily identifying high-risk individuals.

This phenomenon is further demonstrated by the calibration plots shown in figures 8.1 to 8.3, where I have grouped the fitted values obtained from fitting the Poisson model using the training data into deciles of predicted risk (from low to high). This was done for every Poisson model fitted and for both study groups. Figures 8.1, 8.2 demonstrate the second

training fold of the filtered subgroup, in which the sum of the observed events is 387 and the sum of the fitted values is 387 (383 events are contained in the first fold respectively).

I have plotted observed against the expected number of events by decile of predicted risk, depicted on the left-hand side. On the right-hand side, there is a scatter plot of the observed number of events which shows how these compare to the expected number of events, which remains constant among deciles. The horizontal dotted line depicts the expected number of events in each bin. The majority of predicted events is expected to fall within the interquartile range, which is depicted by the shaded area.

From a well-calibrated predictive model, it is expected that most predicted events fall within the 25th and 75th percentiles. However, because the person-time intervals which contain an event are right-censored, the model may miscalculate the number of expected events that fall into the very right of the interval. To address this issue, I have plotted the fitted values of a model in which last person-time intervals are indifferent to event occurrence and run to the end - the same concept as with the test set. This adjustment is necessary to correctly assess model performance and compare the calibration obtained using the fitted model against unseen data.

Furthermore, the observed events in each decile of predicted risk should be, in principle, evenly distributed above and below the expected number of events. This rule is adequately met when censoring in the training/test data is disregarded, as we see by comparing figures 8.1 and 8.2. This happens because, under this condition, the model assimilates the prediction intervals to the training intervals that contain an event.

In particular, figure 8.1 depicts the second fold of the training data, which contains 2633 individuals, of whom 387 advanced to renal replacement therapy. The person-time intervals are at most one-year long. In this scenario, the last person-time interval is truncated, if an event occurs. The fitted values shown are from a Poisson model, which assumes a linear relationship between eGFR and event risk. The time-updated eGFR is specified via a Kalman

filter based on an LMM-drift-diffusion model.

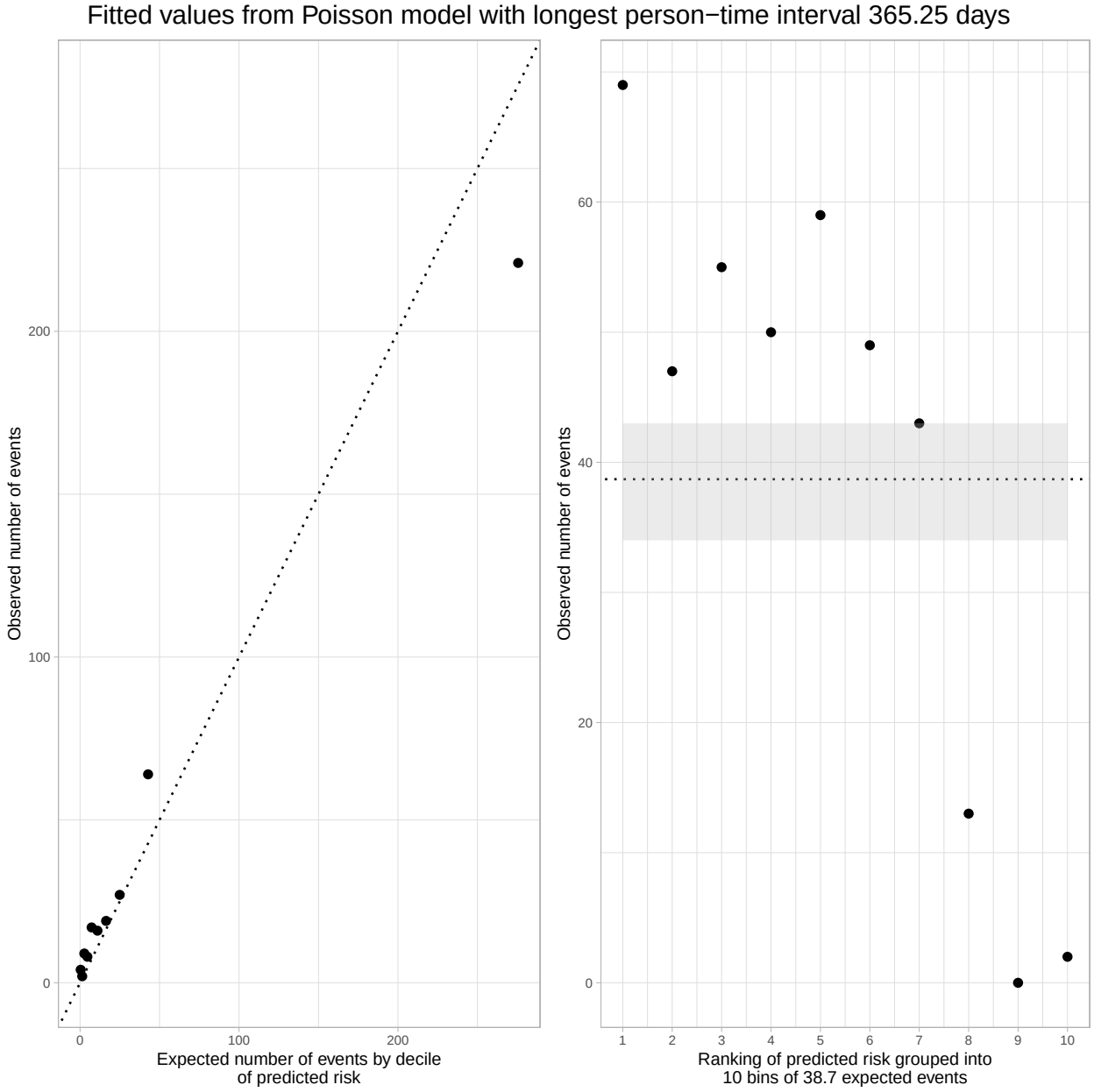


Figure 8.1: Half training data with some events being deliberately censored (landmarking) and some intervals being truncated.

Figure 8.2 depicts the same training data again. However, this time, the last interval of each subject is non-truncated, i.e., all intervals have a length equal to one (year). All last person-time intervals are indifferent to event occurrences and run to their designated end. Indeed, ignoring the event occurrence and running all intervals until the end improves calibration

because it treats the training person-time the same as prediction person-time.

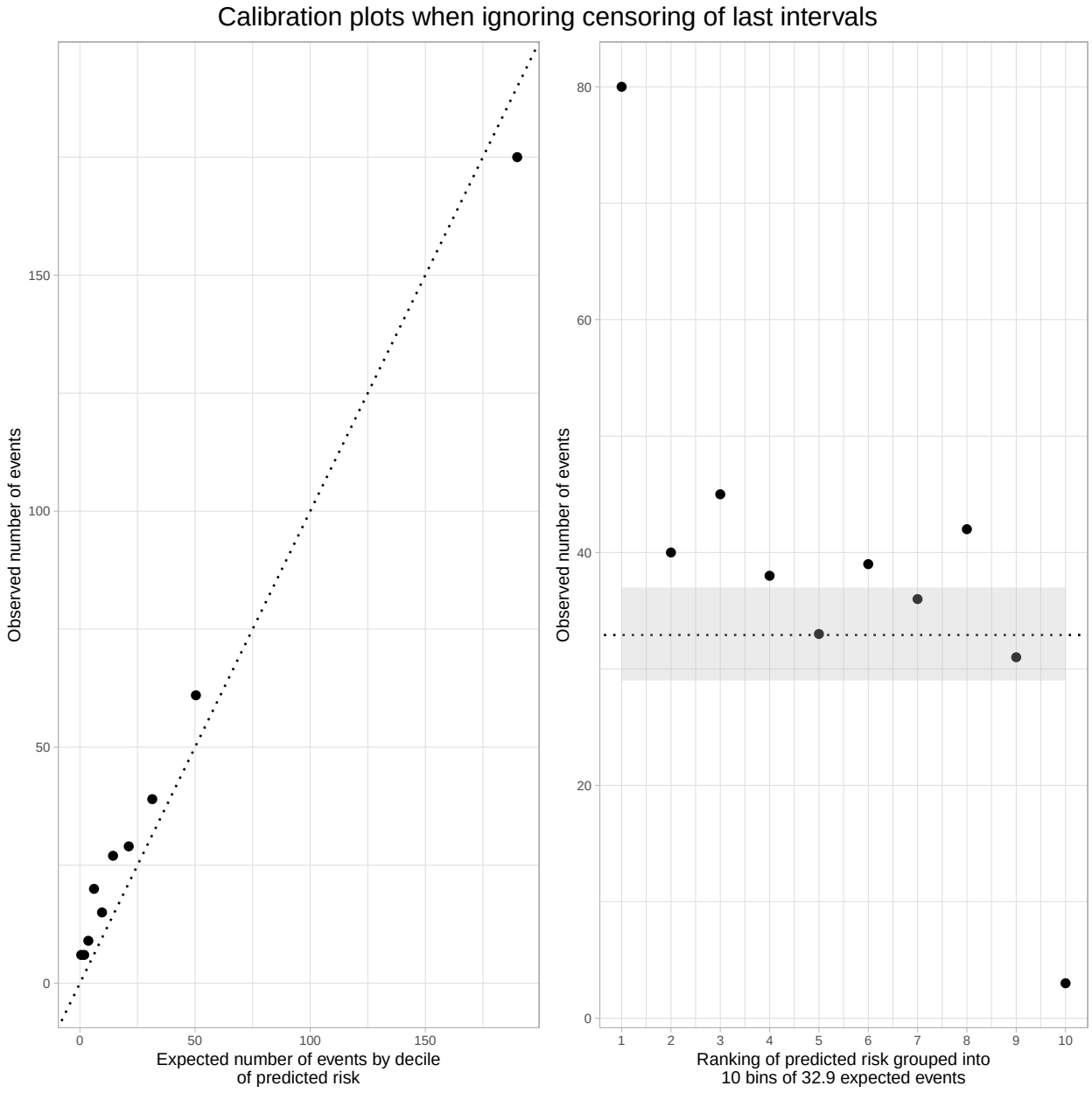


Figure 8.2: Half training data with some events being deliberately censored (landmarking), but all intervals have equal length, ignoring event occurrence.

Figure 8.3 shows the test data comprising 1725 individuals and 222 observed events. The length of intervals is set to one (year). The fitted values shown are derived from a Poisson model that takes in time-updated eGFR data specified by an LMM coupled with drift and diffusion processes. The testdata are generated by models that consider actual measurements

until a landmark time and then project the trajectories forward. It looks like the model underestimates the number of individuals being low-risk, and overestimates those at high-risk.

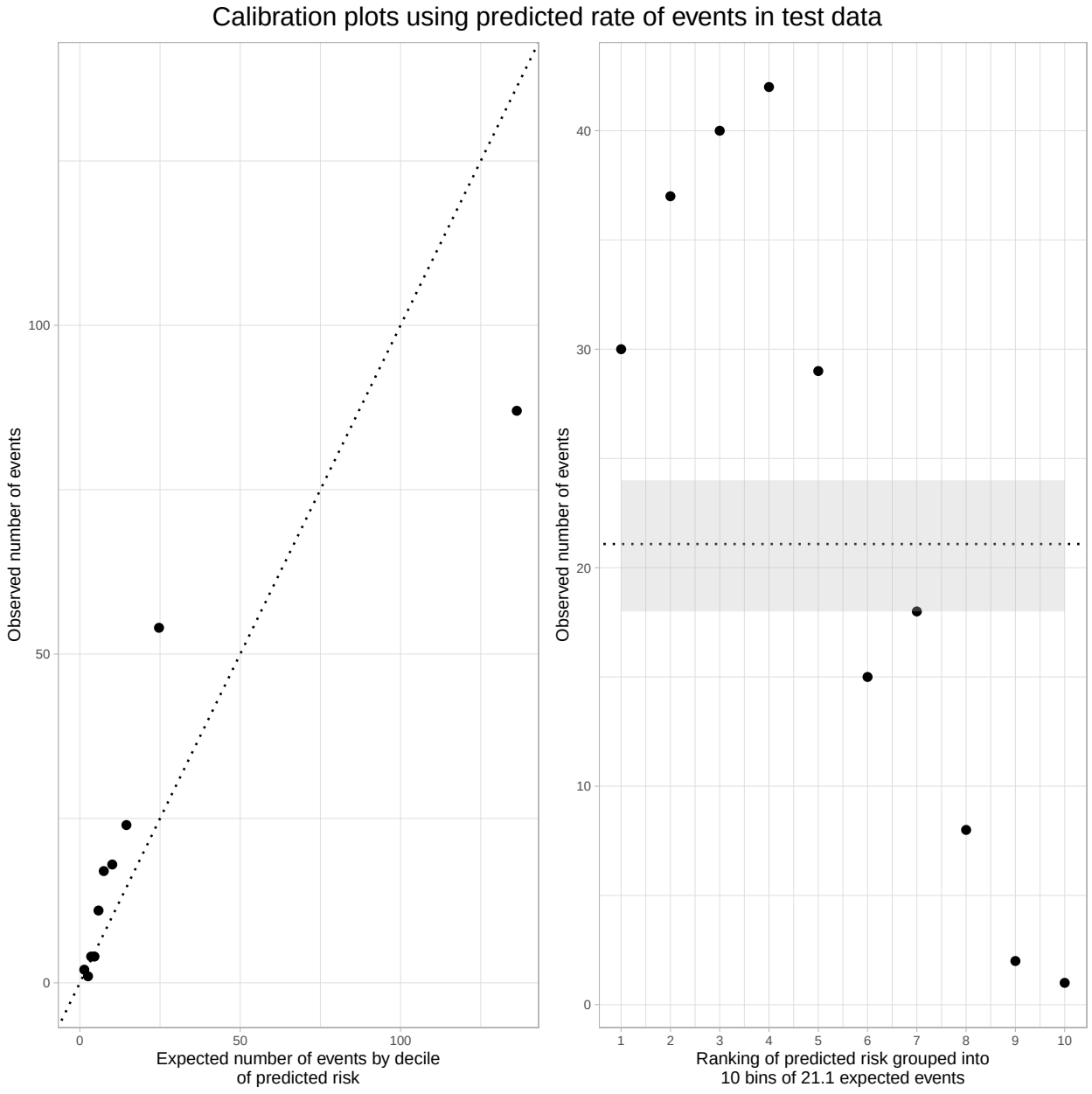


Figure 8.3: Event data corresponding to those individuals who are deliberately censored in the training data shown in figures 8.1 and 8.2.

8.3 Model refinement to include a B-spline function

In light of the aforementioned findings, the effect of longitudinal eGFR on renal failure progression might be non-linear for chunks of time. Since we do not make predictions that fall beyond the end of the study, we may use as well a B-spline function (as introduced in Section 2.3.1) to specify the complex relationship between longitudinal eGFR and time to event.

Therefore, I refined the model to specify a B-spline for log-eGFR starting with using six equidistant knots. Recall that I previously used six degrees of freedom with the `stan_jm()` modelling function in `rstanarm` as the default option. Therefore, reusing the default, given that the average follow-up duration of the full cohort has been 15.4 years, sounded reasonable. Furthermore, given the cross-validation approach, where within the training set, half subjects are followed all the way to the end of the study, and half subjects are intentionally censored to perform forward prediction, the model considers data only for half of the population for the years after the landmark point. This fact also justified the choice of six nodes. Moreover, the knots are placed at equally spaced percentiles of observed event times.

However, identifying the optimal knot number is not trivial for non-uniform spaces, e.g., curves with various turning points or being discontinued at places (Dung and Tjahjowidodo 2017). Hence, future work could investigate the minimum number of knots needed to specify the biomarker data on such a large dataset.

In addition, it is worth mentioning that the refined models with splines have been fitted using a different data format. The reasoning behind reconstructing the input data was to perform diagnostics with respect to having moderately calibrated predictions. Therefore, I carried out the time-splitting using the times of observations of every subject as times of measurements for every subject included in the training set. I used the Kalman filter imputations given at pre-determined time points for each individual to impute at a second step, the eGFR values of time points that were observed in some individuals but were missing in some others by carrying forward the last available Kalman filter imputation. Arguably, this approach inflates

the number of imputations plugged into the Poisson model and increases the computational time needed to generate such input. Therefore, I trained a limited number of Poisson models dependent on this data format. In particular, the LMM model with drift and diffusion for eGFR was used in the refined Poisson model.

Table 8.10 shows the Poisson coefficients of a model fitted on imputed eGFR data of 2633 individuals (filtered subgroup) using B-splines. The number of the measurements has been increased to 4799869, under the splitting and imputations settings described earlier.

Table 8.10: Poisson model fitted to latent biomarker values imputed by a Kalman filter, including LMM and drift effects, using B-splines for the eGFR data.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-33.379	5.855	-5.701	0.000
tstart	0.038	0.021	1.829	0.067
splines::bs(egfr, df = 6)1	46.685	8.011	5.828	0.000
splines::bs(egfr, df = 6)2	30.500	5.216	5.847	0.000
splines::bs(egfr, df = 6)3	28.444	5.973	4.762	0.000
splines::bs(egfr, df = 6)4	16.319	6.040	2.702	0.007
splines::bs(egfr, df = 6)5	108.518	64.375	1.686	0.092
splines::bs(egfr, df = 6)6	-1221.674	1460.507	-0.836	0.403
gender	-0.156	0.051	-3.066	0.002
baselineage	0.041	0.047	0.859	0.390
baselineduration	0.073	0.059	1.241	0.215

This alternative data format is worth further exploring. However, it inflates the input of the model and requires much more computational time than creating the previously used input format, in which all individuals share imputations at pre-determined time points according to user's time step, and apart from that, they have individual times of observations.

Of course, the smaller the time step, the more frequent the time points that all individuals have an imputation generated. Hence, the Poisson model that uses updated daily input data converges to the refined model that uses all times of observations observed in the dataset for every subject. For the latter, the input is created by carrying forward the last imputed value to fill in time points observed in other individuals.

Figure 8.4 depicts the same data as do figures 8.1 and 8.2. In this scenario, the fitted Poisson model includes a B-spline for log-eGFR, the trajectory of which is specified using an LMM with drift and diffusion. The plot depicts the scenario where interval truncation is not ignored on the left-hand side, while on the right-hand side, all intervals run to the end. We observe that calibration is further improved by using B-splines to determine the magnitude of the effect of eGFR on the rate of event.

Figure 8.4 shows an illustration based on the test data of 1725 individuals, of whom 222 had an event (same as in figure 8.3) and one year-long person-time intervals. The fitted Poisson model includes a B-spline for log-eGFR, the trajectory of which is based on an LMM with drift and diffusion. There is apparently some overfitting taking place when using B-splines for eGFR to train the model. This is primarily due to the fact that we need to generate observations for every observed arrival time of all individuals for everyone and this inflates the data points. As an ad-hoc solution, I have carried the last observation forward, between Kalman filter imputations, when necessary.

The calibration of the model with a B-spline for log-eGFR is improved compared to the calibration plots shown first (figures 8.1 and 8.2), which are based on fitted values of the Poisson models without splines for eGFR. The refined model has been fitted without restricting

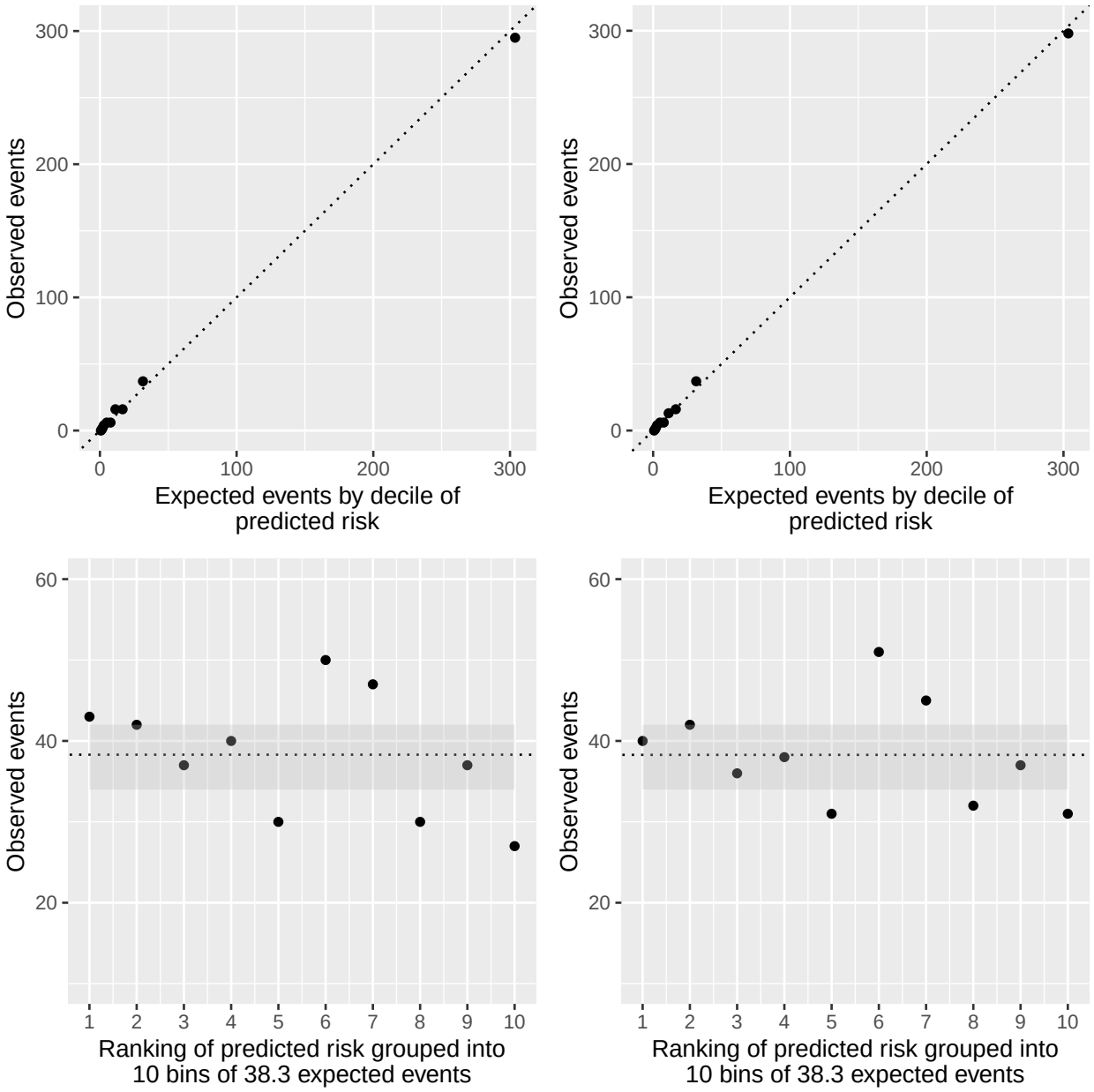


Figure 8.4: Analysis A, fitting on training data of model with splines. Interval truncation at event occurrence included (left column), intervals being complete ignoring event occurrence (right column).

follow-up length to be longer than five years. The calibration plots of the Poisson models with B-splines for log-eGFR of individuals with follow-up ≥ 5 years are further improved, which also improves prediction, as shown in figures 8.6 and 8.7.

Fig 8.6 shows the same training data as before. Similarly, the fitted Poisson model includes

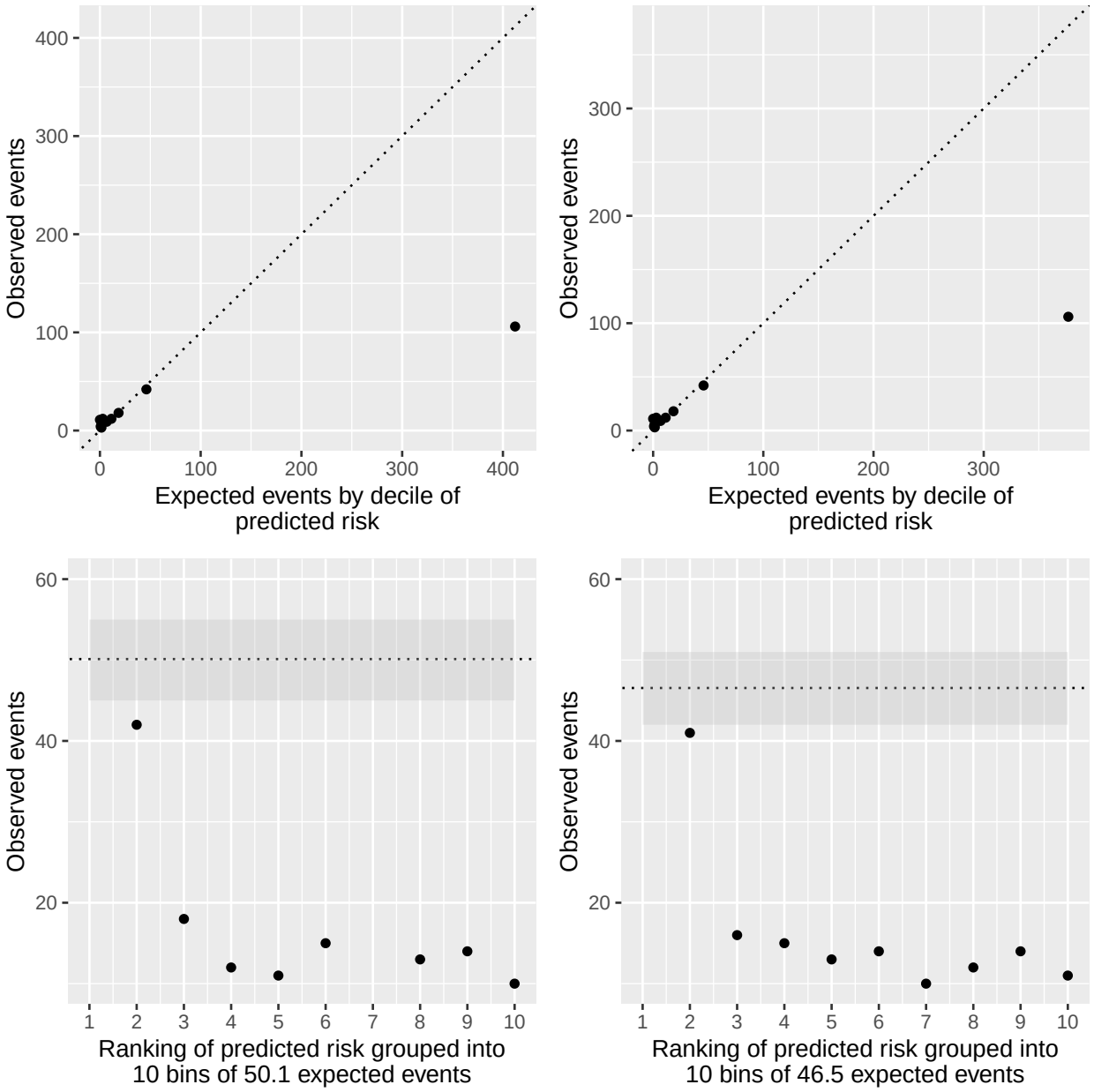


Figure 8.5: Analysis A, the prediction model underestimates the number of individuals being at risk in all deciles of predicted risk.

a B-spline for log-eGFR, the trajectory of which is specified using an LMM with drift and diffusion. The plot depicts the alternative scenario, where only subjects with length of follow-up ≥ 5 years are included in the models instead of the previous set of calibration plots 8.4, 8.5. This refinement shows the best model calibration.

Fig 8.7 shows an enhanced calibration on the predictions based on the test data of 1725

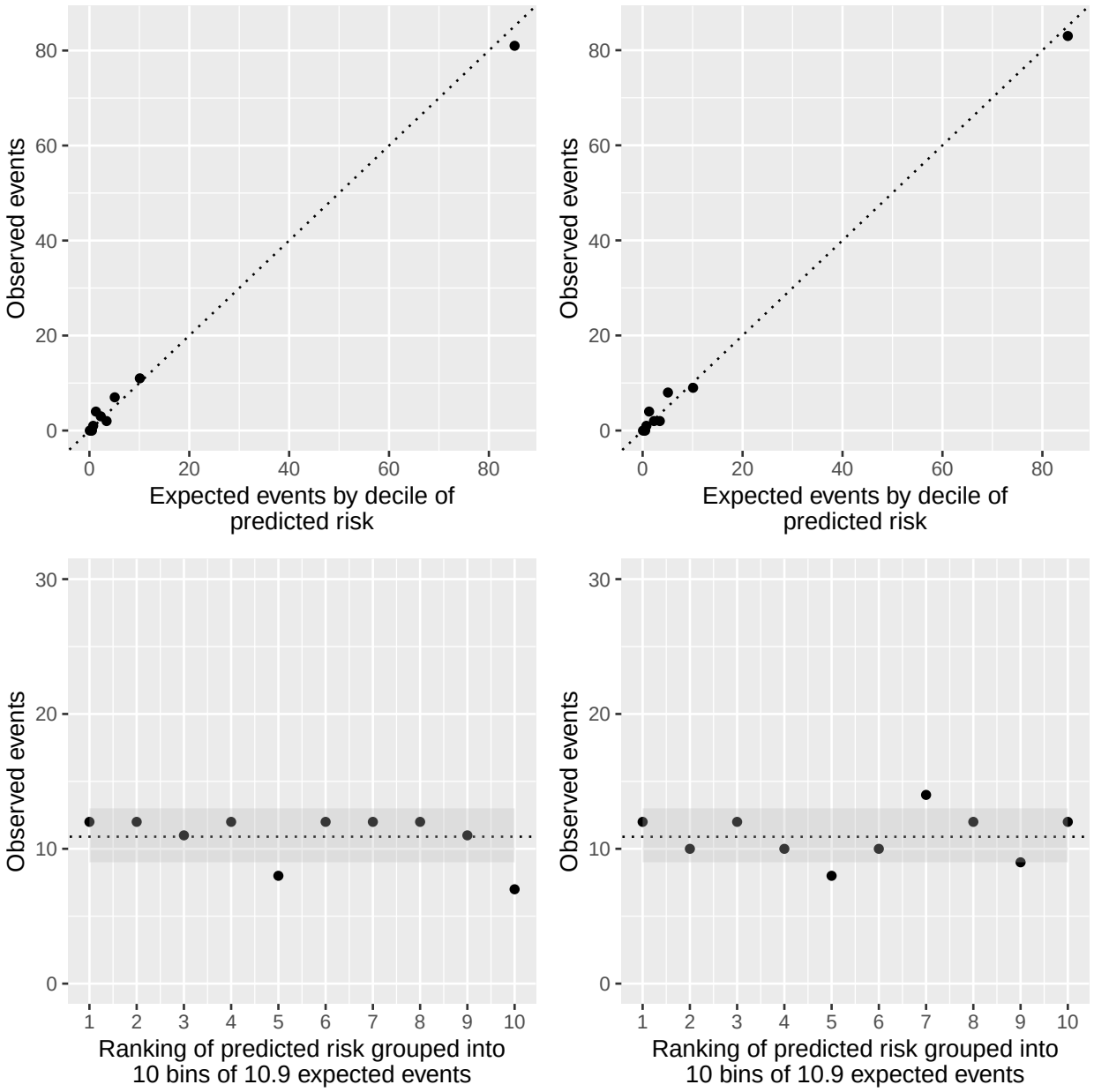


Figure 8.6: Analysis B fitting on training data of model with splines. Interval truncation at event occurrence included (left column), intervals being complete ignoring event occurrence (right column).

individuals with 222 RRT events within one-year-long intervals. Analysis B is shown to be the most calibrated scenario of them all.

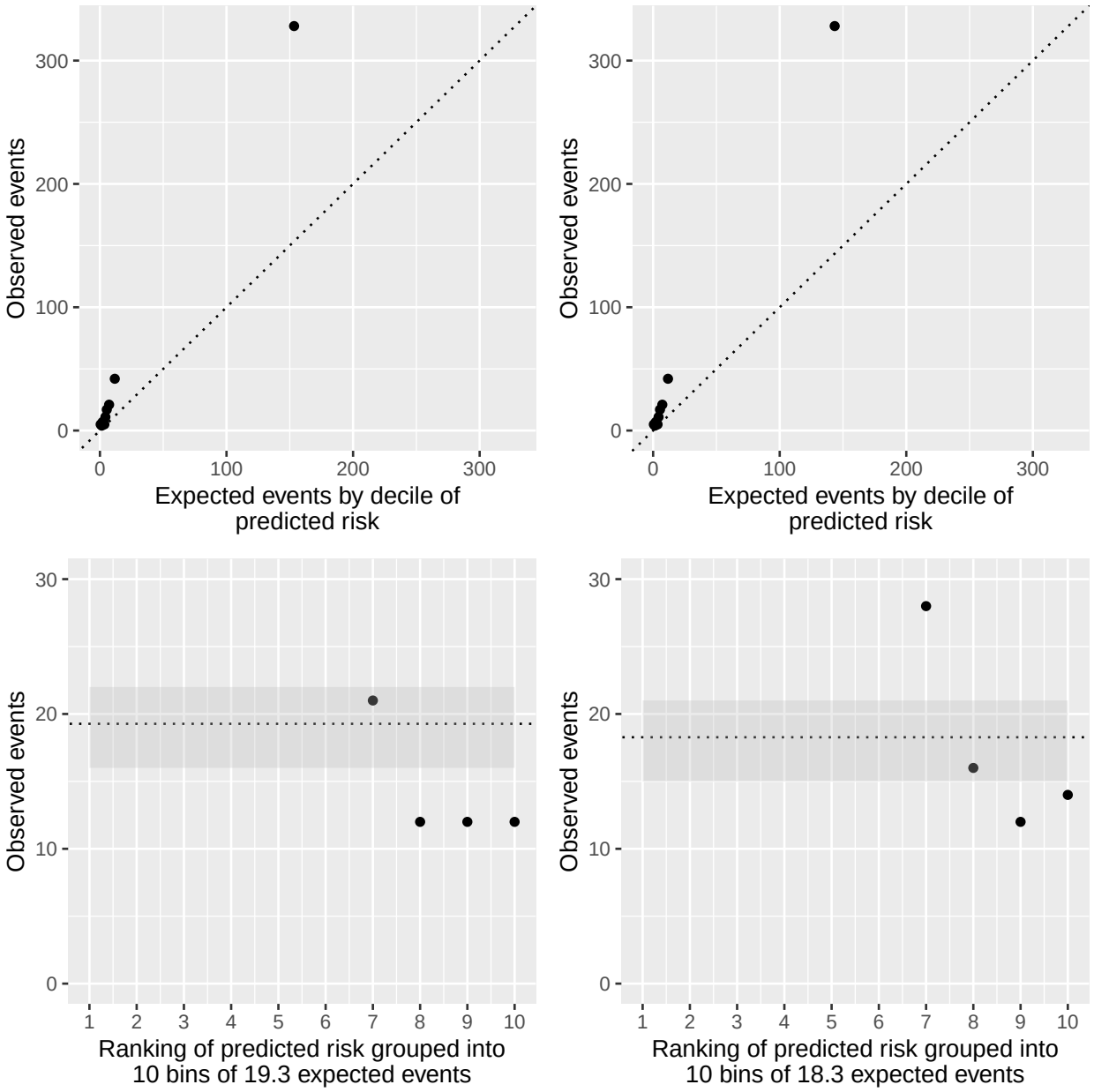


Figure 8.7: Analysis B, prediction with test data from the model with splines.

8.4 Evaluation of designated Poisson models for RRT risk

With a survival dataset reformatted as person-time intervals of fixed length, the predictive performance of the Poisson models can be evaluated by comparing *case* person-time intervals in which an event occurs with *non-case* person-time interval in which no event occurs.

To predict event status in the test data, I use the models fitted to the training dataset and compute the probability of an event in each person-time interval from the 5th year onwards. For any subject who is censored before the end of the study, we round the length of individual follow-up to the closest interval end (e.g., a follow-up of 264 days corresponds to 1 year, and not 0 years, for a time step of 365.25 days).

I have evaluated the predictive performance of the fitted models on withdrawn data using a 2-fold cross-validation approach to maximise the number of events used for assessing the fitted models, and compared the results obtained by models, including time-updated eGFR from stochastic models and the LOCF approach.

The probabilities of events from the 2 folds combined are computed conditionally on imputed biomarker data from five years onwards on all the 1725 subjects (filtered dataset) who have been followed up for at least five years and have been intentionally censored in one of the two folds: 863 individuals belong in fold one, and 862 individuals belong in fold two.

Table 8.13 shows what the test data of a random subject look like for intervals one year long. This individual has not experienced renal failure during a follow-up period of 3653 days (approximately ten years). This extract aims to show how the probabilities of event change over time conditional on the imputation method. Imputations shown are computed via

1. a Kalman filter configuration, including diffusion,
2. a Kalman filter based on LMM, drift and diffusion,
3. the LOCF approach.

Table 8.11: Biomarker imputations for a random subject based on a diffusion process. We observe that the values remain very close because they only depend on the most recent value due to a random walk assumption. This is quite similar to the LOCF process.

interval begins	interval ends	imputed log eGFR	prob. of event
5	6	-0.0361595	0.0101505
6	7	-0.0361594	0.0102216
7	8	-0.0361594	0.0102931
8	9	-0.0361593	0.0103652
9	10	-0.0361593	0.0104378

Table 8.12: Biomarker for the same random subject imputed by an LMM with drift and diffusion.

interval begins	interval ends	imputed log eGFR	prob. of event
5	6	-0.0646023	0.0136371
6	7	-0.1344659	0.0146557
7	8	-0.1962433	0.0155831
8	9	-0.2508704	0.0164135
9	10	-0.2991748	0.0171446

Table 8.13: Same subject, biomarker given based on LOCF.

interval begins	interval ends	imputed log eGFR	prob. of event
5	6	-0.1390048	0.0134864
6	7	-0.1390048	0.0142828
7	8	-0.1390048	0.0151259
8	9	-0.1390048	0.0160184
9	10	-0.1390048	0.0169631

The updates provided by each Kalman filter algorithm represent past data and estimations up to the beginning of the interval I am predicting for. The Kalman filter is agnostic to any data arriving after five years, i.e., the calculated trajectory only depends on data observed before Time 5, and after this point, it gives a projection of possible future values against which the fitted Poisson model is evaluated. Furthermore, the LOCF approach is also employed for the test data, which extrapolates forward the last imputation obtained right before the landmark time point.

8.5 Strengths and limitations

A salient point to mention is imputations used to validate the fitted Poisson models deviate from imputations used in the training process, due to the model being agnostic to any biomarker data arriving after the landmark time for all subjects comprising the test folds.

Personalised probabilities of risk of events span time five until end of individual follow-up. To avoid biased estimates of the hazard rate, the prediction intervals must run until their designated end (as if they do not contain an event), as opposed to the person-time intervals used in training that are explicitly truncated, when an event occurs. In addition, prediction

intervals must remain complete, so that the rate of event is distributed uniformly throughout the interval. Incorporating information about the time of an observed event, or biomarker arrival, would severely bias risk prediction.

The presented two-staged method is compatible with the state-of-the-art joint modelling approach implemented in `rstanarm`, which fits a mixed-effects model to the longitudinal data first and then uses the learned parameters to fit a time to event model.

So far in the thesis, I have demonstrated a scalable joint modelling method as an alternative to the `rstanarm` implementation, and I have applied this method to estimate the rate of progression to renal replacement therapy up to five years in the future, starting from a set landmark time, using the national T1D cohort of individuals with renal disease records in Scotland (30000 individuals). In addition, the demonstrated formulation obtains survival probabilities for any arbitrary future time point within the prediction window, as opposed to the `rstanarm` implementation that provides predictions only at the time points a biomarker measurement arrives.

The main limitation of `rstanarm` is the need to integrate the random effects out of the joint distribution calculation by computing the quadrature of the relevant area, which becomes intractable for more than a single frequently-measured biomarker.

Despite the scalability issue met, the `rstanarm` functionality remains a useful building block for future extensions, and can be successfully applied in smaller-scale datasets. This software was evaluated and slightly extended using a small subset of T1D individuals (N=2000) with cardiovascular disease as outcome data and time-updated HbA_{1c} and eGFR as manifest variables for the outcome.

I have also used this functionality as a benchmark against which I have compared the performance of the alternative methodological approach for joint modelling of longitudinal and time to event data . The sequential updating process via a Kalman filter, as a first step and the Poisson regression model for the rate of event based on the longitudinal trajectories

obtained by the Kalman filter may handle person-time intervals of fine granularity in a timely manner, and performs comparably well to the continuous-time Bayesian model estimated by `stan_jm()` for the rate of progression to RRT (C-statistic > 0.80).

The demonstrated joint modelling approach has been developed using the functionality offered by `ctsem` to specify longitudinal trajectories of biomarker data. The available information spans a decade the most (observational study), and I have assessed retrospectively whether the generated biomarker data and predictions of time to event are accurate and outperform the traditional LOCF for the biomarker trajectory. The strength of this research is its flexibility; I have explored many paths on how to maximise the information accrued on the biomarker data over time efficiently and estimate in turn, risk prediction robustly. The closer the model's projections are to reality, the greater their application in clinical settings.

Specifically for the rate of progression to RRT, proper specification of the rate of change of eGFR plays a critical role in predicting time to event robustly. Therefore, in the interest of time and given the large number of individuals with available outcome data, I stopped including HbA_{1c} as the second predictor in the survival model for RRT to assess the predictive power of eGFR explicitly, which counts as a limitation of this study. Having established that the procedure yields comparable estimates, extending it to more than one biomarkers is straightforward.

This modelling approach can help precision medicine in diabetes by predicting and preventing acute and long-term complications. Furthermore, obtaining frequent outcome updates promptly can identify the patients at the highest risk and, in turn, reduce emergency hospital admissions and mortality attributable to diabetes.

Last but not least, the time-splitting joint modelling approach serves a twofold purpose. First, I have investigated whether the LMM specification if extended to include drift and diffusion components performs better than simpler methods that do not infer a longitudinal pattern in the data. Extending the LMMs perform better in terms of predicting latent characteristics

of individual biomarker trajectories. Moreover, I have studied how the different Kalman filter configurations affect the Poisson model estimations and to what extent it affects the predictions yielded. This has been evaluated by projecting individualised profiles forward agnostic to real measurements and used as test data.

A limitation also worth mentioning is that while more granular intervals might approximate the Poisson hazard rate more accurately, the imputations based on the sequential updating approach may be less reliable as time progresses. Predicting forward for a relatively long period in the lack of actual new data might not be that trustworthy, depending also on the stochastic process configured. In my research, this phenomenon was not observed, because the prediction time was relatively short and the biomarker data were collected in high frequency; a fact that might implies that those individuals were monitored more closely due to actually being 'high-risk' (Glasziou, Irwig, and Mant 2005; Peyroteo et al. 2021).

Note that this implication concerns the imputation scheme used for generating the test data, against which I evaluated the fitted Poisson models, since those updates are data-agnostic. Hence, it is likely that a trade-off exists between the accuracy of predictions and the frequency of updating in the absence of new data.

Last but not least, the current implementation of `ctsem` has been employed to impute latent states of eGFR trajectories on those individuals with at least three observations in the study period. However, it can be restrictive in terms of out-of-sample predictions. The Kalman filter can only extrapolate a longitudinal trajectory for individuals who have been included in the train set. Note that a trajectory can be built up to an arbitrary time point, which does not need to be the same for all subjects, and this consists a major strength on the other hand. Therefore, the cross-validation split I have performed essentially concerns the individual follow-up length, not the people themselves. Each subject for whom I need to generate predictions should be observable for at least some duration, a feature that has resulted in the inclusion of a landmark time point, up to which people need to be event-free.

This constraint of `ctsem` implementation can be limiting, but landmarking appears to be a good workaround.

In the next and final evaluation chapter, I discuss the numerous variables embedded in the analyses, and how these parameters intertwine with each other with respect to the final increment in prediction.

Chapter 9

Reflection of the development and predictive accuracy of models for RRT risk

9.1 Interpretation of performance

To study the hazard rate of time to renal failure in (a) the national T1D cohort with renal outcome data and (b) a frail subset of the national cohort with baseline eGFR < 60 mL/min/1.73 m^2 (so-called filtered subgroup), I have split follow-up time, using a selection of time intervals ranging from 1 year to 1 day. In principle, the shorter the interval length, the better the approximation of the hazard of an event occurring within a person-time interval.

Furthermore, I have fitted models on different training subsets of data concerning their follow-up length and whether an event occurred before landmark time. Arguably, an individual who has an event before landmark time cannot be part of the test population because they are censored at the event occurrence and thus removed from the study. However, individuals who have an event within the first five years still contribute some time to model

training. Consequently, the first analysis design implemented included every individual without restricting whether they have an event within the first five years or later.

- Analysis A accounts for any follow-up length in model fitting, allowing the model to learn from the event times observed before landmark time.

- Analysis B restricts the model's training to those who have been observed for at least five years: thus, there are zero events included in training before landmark time.

Moving from analysis A to analysis B was required to correct for the fact that the distributions of events were dissimilar between training and test sets, resulting in poorly calibrated outcomes. In addition, since the test individuals have all outlived the landmark time in analysis A, they were considered low-risk for the event. Finally, the inclusion of the people who have been event-free for the first five years in model fitting was challenging to think of in advance because, in the real world, every patient with renal disease, followed up by a clinician, contributes to their clinician learning experience, despite the actual renal failure time, if any.

Consequently, each of the two configurations above yielded different results. They also demonstrate an artefact, a well-known epidemiological paradox worth mentioning. Studying a group of older patients who already have signs of renal dysfunction at baseline with eGFR < 60 mL/min/1.73 m^2 would make an expert expect that the probability of an event will be much higher in the years to come, as opposed to the rest of the cohort, which includes less frail individuals.

For connection, the filtered group has an average baseline age of 61 years and diabetes for 25 years, with the full dataset having an average baseline age of 40 years and diabetes duration of 15 years (table 6.1). However, highly frail subjects do not benefit from intervention or improvements in treatment plans as they have already substantially degraded and get dismissed from the study early on, quite earlier than the landmark time. Hence, survival later in the study improves and, counter-intuitively, is better than one would initially expect.

The Poisson models estimate a lower hazard rate when the pool is not restricted to a follow-up length of at least five years. This finding is demonstrated by figures 9.1 and 9.2, which compare the predicted events as estimated per each analysis design and for various interval lengths. Each plot of the grid represents the predictive performance of the respective Poisson model that utilises eGFR imputations based on a particular configuration, as explained already.

9.1.1 Increment in predictive performance moving from one-year to one-day intervals

I have evaluated the calibration of the refined Poisson model separately, as presented before. The refined model includes a B-spline to specify the effect of changing eGFR on the risk of an event.

The new model has been evaluated against a dataset formatted as imputed observations at all time points that people have a measurement within the original data plus some pre-determined time points common among all individuals. The respective model calibration is shown in figure 8.4. The predictive performance of the refined model favours the assumption that the rate of change of longitudinal eGFR varies non-linearly over time and heavily depends on individual-level random effects: thus, it is fair to say that the impact of eGFR on the risk of ESRD is not rigidly linear throughout follow-up.

The following figures represent the change in the calibration of the initial Poisson models that do not include a B-spline specification for the time-updated eGFR.

In particular, figure 9.1 show the Poisson models fitted using cross-validation of imputed eGFR data on the filtered group. The horizontal line depicts the observed (albeit unknown to the model) number of events within the test data. The featured subset contains 222 events between years 5 and 10. The LMM models are characterised by a tension to overestimate the risk of event due to assumptions of linearity, which, as discussed earlier, does not always hold for the relationship between changing eGFR and RRT occurrence. It might be that their relationship

is curved or has a threshold effect. Furthermore, it might be that the measurements of individuals are substantially correlated, i.e., the random effects are not entirely captured the correlation structure between observations. Last but not least, there might be some imbalance in the data or some influential observations, which have a disproportionate effect on model estimates.

On the other hand, the models with autoregressive drift and without linear mixed-effects are more likely to underestimate the number of events, because not all measurements and individuals are reverting to a long-term average within follow-up. That might have been true for a healthier group at baseline. Moreover, LMMs coupled with drift and diffusion processes exhibit the closest predictive performance to reality. It looks like the diffusion element (more like a random walk in the biomarker sequence) is essential to counterbalance the drift element, which fits moderately non-monotonic trajectories, as it does not incorporate all the random and sudden fluctuations in eGFR data. Finally, the model with diffusion components approximates the number of events in the test data well for all interval lengths, which overall implies that eGFR progression in people with T1D is quite irregular over time.

Figure 9.2 shows the predictive performance of all Poisson models for the test folds using the full dataset. In this broader population, the linear patterns are more obvious, ranking the LOCF model, the diffusion model and the diffusion-drift model last with respect to guessing the observed number of events. The full dataset is more balanced, and possible outliers are likely less ambiguous. Therefore, the mixed-effects models manage to specify sufficiently the rate of event. In addition to specifying linear slopes, the inclusion of a diffusion component appears to attain the best calibration. A likely interpretation of the numbers of predicted events being continuously on the rise (especially for intervals shorter than 45 days) might be that the fitted models do not assign a zero probability in an interval, which does not contain an event. As a consequence, the shorter the interval length, the more these non-zero probabilities add up, inflating the estimation of predicted events and the rate in the Poisson models, respectively.

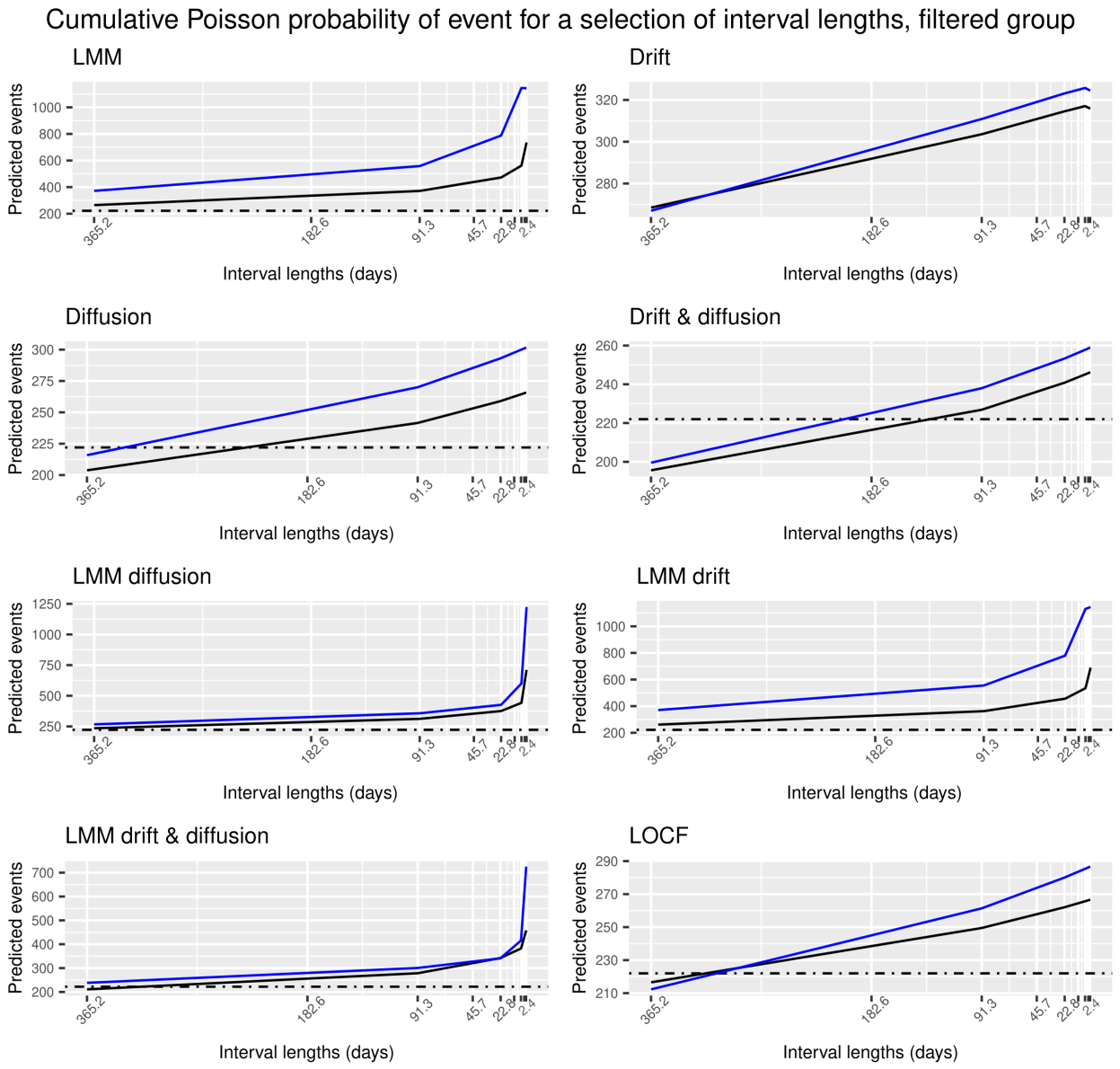


Figure 9.1: How does the specification of the longitudinal model for eGFR affect the calibration of Poisson models? Predictions regarding the filtered group for all interval lengths assessed.

As opposed to the complex Kalman filter configurations for the eGFR profiles, the simple LOCF approach extrapolates the last existing value until the end of the prediction period. This more naive imputation is not always sufficient, especially for individuals with a more linear, monotonic rate of change. For most individuals however, assuming a constant eGFR over the prediction period is broadly better than predicting a particular trend which is not

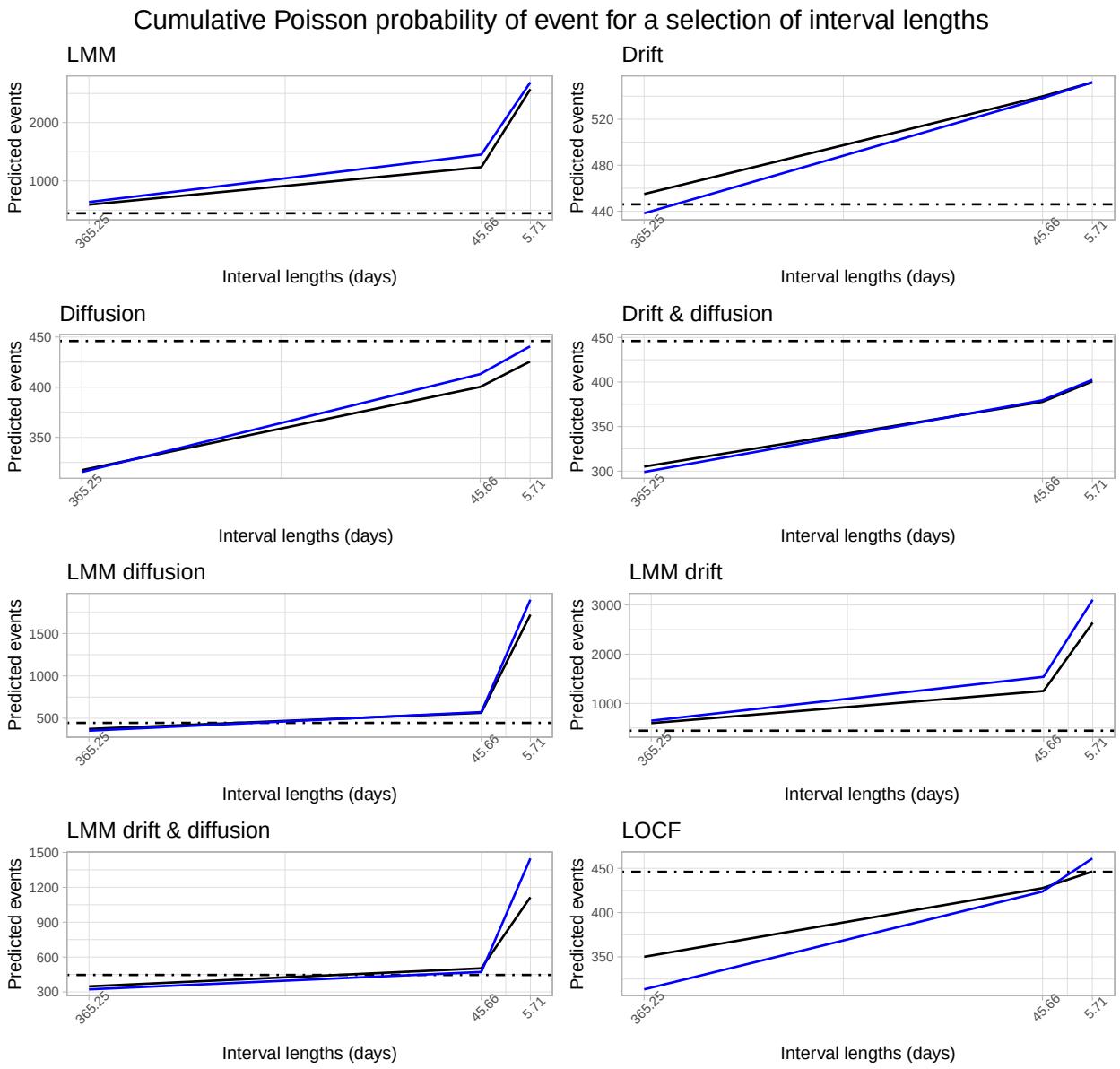


Figure 9.2: Predictions regarding the full cohort, all interval lengths assessed. The horizontal line depicts the observed, albeit unknown to the model, number of events within the testing folds; 446 events between years 5 and 10.

met in reality.

However, the LOCF approach is worth attempting, because it emulates the real-world scenario, where physicians do not have records for patients who have not interacted with the healthcare system because they are healthy, or cannot access treatment (Handy et al. 2017). Then,

physicians may project a simple trajectory based on a single measurement of a clinical variable, when this becomes available. Yet, they might estimate the rate of progression and make decisions about the course of treatment, and when that patient should be referred for evaluation next.

9.1.2 How does the LMM component add to the predictive performance?

It is expected that the models that include (a) random slopes and (b) autoregressive effects for building individualised eGFR profiles would outperform the more straightforward random walk and LOCF processes that do not depend on past trajectories.

Inspecting the trajectories obtained via a Kalman filter for a few individuals, it is observed that a drift and diffusion model which does not estimate slopes yields predicted trajectories that resemble a last-observation-carried-forward model. Extending the drift and diffusion model to also allow for slopes allows the trajectories to curve slightly away from a linear path towards the population mean.

Figure 9.1, which features the frail subgroup shows that the inclusion of random effects specified by the linear mixed-effects model is necessary for the correct estimation of the hazard rate over the test intervals, but not always enough. The Poisson models that have included time-updated eGFR measurements given by longitudinal models that include an LMM component provide an insightful slope regarding the individual trajectory, as opposed to the diffusion, drift and LOCF imputation schemes for eGFR, which cannot infer a gradient for the changing biomarker data. We recognise that the diffusion model resembles the performance of the LOCF model in the sense that these models feature a random walk and do not infer any kind of slope. Most importantly, the models that do not include slopes tend to infer a lower probability of event due to being agnostic to an average, long-term tendency; either increasing or decreasing trend.

Therefore, while including autoregressive drift when modelling the eGFR of the frail subgroup is not always performant, it still captures well the direction of change of the broader population with T1D. This is likely the case because someone with a normal range of eGFR can more easily revert to their long-term average. If eGFR levels suddenly deteriorate due to sickness or hospitalisation, it then stabilises back to the average level in most patients (Astor et al. 2008; Hemmelgarn et al. 2010).

In conclusion, using at least linear mixed-effects models for the longitudinal data, with or without drift and diffusion processes, enhances the prediction of the outcome as shown in figures 9.1 and 9.2; the predictions approximate more closely the observed number of effects compared to the other specifications.

The table 9.1 contains the measures of predictive performance for all models for a particular interval length, evaluated against the test data of the filtered subgroup. All models achieve adequately good discriminative performance (C-statistic > 0.8) on average. While it is helpful to assess performance in terms of C-statistic, the figures 9.1 and 9.2 give a more comprehensive view of models' performance for the entire range of interval lengths used.

Table 9.1: Predictive performance of Poisson models using the filtered group. Person-years equal to 8625.

model	observed	predicted	hazard rate	log-lik	C-statistic
lmm	222	526.727	264.677	-1113.916	0.801
drift	222	318.241	268.469	-933.738	0.825
diff	222	270.933	203.784	-999.707	0.807
driftdiff	222	252.504	195.602	-993.935	0.808
lmmdiff	222	410.688	235.177	-1050.421	0.817
lmmdrift	222	511.344	261.3	-1106.217	0.801
lmmdriftdiff	222	393.678	210.83	-1112.147	0.802

model	observed	predicted	hazard rate	log-lik	C-statistic
locf	222	268.689	216.554	-951.19	0.819

The performance of autoregressive drift effects when is coupled with a slope specification approximates closely the performance of just having an LMM alone, because especially the frail subgroup's eGFR trajectories tend to drop and does not usually revert to an overall trend. The LMM-diffusion and LMM-diffusion-drift models are the most calibrated in estimating the hazard rate of the frail subgroup, while the drift model is the most performant in terms of C-statistic (table 9.1).

The Poisson model is likely to associate lower risk of event with extended follow-up periods. The subjects in the test sets have been event-free within the first five years, and they are therefore, given a low probability of event over the assessed period, as the rate of change of eGFR appears to get more stable; in terms of observing less frequent updates (as opposed to the years before landmarking) and less informative updates due to the lack of new information.

9.1.3 Regarding the computations of the predicted number of events

The vector of hazard rates defines the cumulative hazard rate, i.e., the output of the `predict()` function multiplied by the fixed length of the prediction intervals (depending on the timestep each time) and our own `hazardrate.poisson()` function, shown here:

```
hazardrate.poisson <- function(model.poisson, Xmatrix) {
    ## returns hazard rates using coeffs from model.poisson,
    ## covariates from matrix Xmatrix, which has been generated by
    ## model.matrix to add a column of ones and to recode categorical
    ## variables as indicator variables
```

```
    beta <- matrix(model.poisson$coefficients, ncol=1)
    xbeta <- as.numeric(Xmatrix %*% beta) # linear predictor
    return(exp(xbeta))
}
```

In particular, the generalised linear Poisson model is used, with the argument `type` equal to 'response' to specify the rate of an event in each person-time interval. The output is obtained by evaluating the regression function of the new data provided (including imputed data and time-invariant covariates), specifying where to look for explanatory variables to be used for prediction.

Then we compute the probability of event occurrence in each person-time interval, as

$$1 - \exp(-\text{hrate} \times \text{interval length}).$$

In principle, we expect these two likelihoods of predicted events to get closer as we shorten the interval length. However, taking complete person-time intervals inflates the total person-time of observation so that the total expected events no longer equates to the observed events. I indeed observe a discrepancy between the observed and estimated number of events (the expected events are typically more than the observed) in the evaluation data with fixed length intervals.

It is worth pointing out that the hazard rates must be interpreted in the same units of time that are used to measure exposure. If a subject is observable at the beginning of an interval, we need to include them in the entire interval. Therefore, the exposure time always represents one unit of time, i.e. one complete interval: it is not allowed for an individual in the test set to get partially exposed. We can set the time unit to any length, e.g., half-year or quarter-year. This configuration leads to exchangeable person-time intervals, where the individuals can be ranked by predicted risk. Each subject contributes to the risk set of cases or non-cases with

respect to the time being exposed.

By including a time-varying covariate, like eGFR, into the piece-wise exponential survival function, I have assumed that the effect of the time-updated covariate on risk does not depend on time. The effect must be constant within each person-time interval. According to this, it is fully legitimate to represent the response variable on these complete (uncensored) intervals as a mixture of independent and identically distributed Bernoulli variables, as explained in 8.2.1.

The inclusion criterion for entering the testing folds introduces a survival bias, on the grounds that the models underrate the event probability of subjects who have outlived the landmark point, because they are supposed to be less frail and less likely to have an event.

The predictive performance of the time-splitting joint model based on updated biomarker trajectories at the start of each person-time interval can be considered as valid as the fully Bayesian continuous-time joint model, if not more, because of the use of a broader family of models supported by `ctsem` compared to `rstanarm`, which specifies the longitudinal component.

Both approaches build a linear mixed model with random intercepts and slopes, which is a minimal requirement to specify a longitudinal submodel. However, the LMM, coupled with drift and diffusion *fits the biomarker data best* and is rather performant in predicting time to event, with the drift model being a fair compromise between treating eGFR as constant during the prediction period (LOCF), and inferring a gradient which is not met in reality, due to arbitrary fluctuation in eGFR of individuals who are not surging ahead in the course for renal failure.

The main takeaway points are the following:

The better fit to the longitudinal data, obtained by using the `ctsem` functionality, as opposed to the LOCF model, translates into a moderate increment in C-statistic for the risk of event. There is no substantial increase in the computational load if more than one biomarker is used

as inferred by longitudinal models built for both eGFR and HbA_{1c} in parallel in a timely manner.

Although the C-statistic of the Poisson model employing the LOCF imputed eGFR is relatively high, it significantly underestimates the expected number of events (figure 9.2), even when taking shorter intervals. On the other hand, as fitting and evaluation specifically elucidate, eGFR requires a better stochastic model to be determined more closely. There is a trade-off between model performance and fitting when actually observing an eGFR trend over the training window, and despite that observation, constant values (via LOCF) are used for prediction. With LMM and LMM-drift estimates of hazard rate being close, but rather higher than the estimates of the other models, it can be seen how the imputations from the various configurations resemble. As before, including the random slope component extrapolates the trajectories rather well, while the drift or diffusion processes may render a fairly conservative guess.

Likewise, the models which include an LMM component tend to be better calibrated (due to better capturing the long-term trend), especially when the interval of prediction is shorter and the hazard rate may change faster between intervals, in contrast to Poisson models, that include imputed eGFR obtained by employing solely drift, diffusion or LOCF imputation schemes (as shown in figures 9.1 and 9.2). From this point of view, the performance of the LOCF model is significantly poorer, and it is not improving with shorter interval lengths. This is a valid justification for using more elaborate imputation schemes instead of plain LOCF.

Additionally, the models which appear to be best calibrated are not the ones with the best C-statistic. Thus, the models are good in differentiating between event occurrence or not, but the predicted probabilities does not accurately reflect the true likelihoods of the outcomes. Moreover, the differences in C-statistics among Poisson models are much smaller than the differences among predicted probabilities of events derived by each model. The reason, why

the best-calibrated models may not have the highest C-statistic, can vary depending on the specific features of the dataset and the modelling context. It could be due to factors such as group imbalances or model's complexity. Focusing on calibration is particularly important in scenarios where accurate probability estimation is crucial, such as risk prediction.

In particular, models that include diffusion broadly perform better in discriminating cases from non-cases, however, models that include random slopes along with diffusion are the best in terms of calibration. Besides, although the LOCF models are less calibrated, they do quite well at differentiating cases from non-cases. The relatively good differential performance is speculated to stem from the similarity of LOCF to the diffusion process in the sense that they both weigh more on the most recent observation of the subject in order to predict the risk.

9.2 Implications

The filtered dataset contains subjects who are progressing faster to renal failure, as opposed to the total cohort. Predicting time to renal disease is an excellent example to demonstrate that building more precise functions of the longitudinal biomarker data makes a difference in the prediction of time to event. In both datasets, the most performant model in terms of model fitting to the data is the Poisson model, which depends on random slopes, drift and diffusion for eGFR trajectories.

When analysing all individuals together, then most patients appear to be at a lower risk of progression to ESRD –maybe because they have less frequently collected measurements of eGFR since they are not experiencing a decline in renal function yet, which explains why the survival models with time-varying eGFR based on stochastic processes for the longitudinal data, are more likely to overestimate the risk of event, as opposed to what it was observed by using the frail subgroup's observations, where the difference between observed and expected numbers of events is smaller.

The repeated measurements of eGFR can provide insights into the progression or decline of

renal function. A circular relationship or feedback loop is likely to exist; eGFR is commonly used as a prognostic measure of renal function. When eGFR data are collected more frequently, it allows for monitoring changes in kidney function over time. Frequent measurements, however, serve as an indication of declining kidney function, which in turn prompts further monitoring and medical interventions.

The LOCF imputation technique is outperformed by smarter imputation schemes, despite performing well at identifying the most vulnerable population. Furthermore, the Poisson-LOCF models fail to assign a substantially increased probability to subjects with high risk, which is a crucial drawback.

To recap, when dynamic system modelling is employed to specify time-varying covariates, such as biomarker data that are used in risk prediction models, these predictive models, in turn, perform considerably well in determining what proportion of individuals will remain event-free for the ensuing period beyond the landmark time point. They also, yield more reasonable latent state predictions than the LOCF models for eGFR.

In this chapter, I have examined the predictive performance of a time-splitting joint modelling approach to assess the risk of progression to renal failure using two overlapping groups of a national T1D population. Exploiting the joint modelling concept reduces bias since it explicitly disregards data coming during the forecast period and instead bases the predictions exclusively on past information. The findings presented herein contribute to the literature that estimates the effect of a changing longitudinal process, such as a biomarker, on time to event.

Therefore, these results add to existing evidence that stochastic processes, such as autoregressive drift coupled with a linear mixed-effects model and diffusion components can approximate better any long-term trends in fluctuating eGFR data. It is also found that longitudinal eGFR is more likely not to be linearly related to the risk of event throughout the follow-up length and, hence, a piece-wise exponential survival function that includes a B-spline function

would perform better in estimating the changing rate of eGFR in individuals with T1D.

On the other hand, the conventional LOCF model, which has been broadly used in clinical modelling of time-updated data, is not robust enough when the hazard rate is changing fast, as in the frail subgroup with baseline eGFR less than 60 mL/min/1.73 m^2. Despite its good differential performance, the Poisson-LOCF model predicted risk in each person-time is more conservative than the alternative state-space models.

This statistical formulation approach might be helpful to identify patients who are progressing fast towards renal failure and thus having eGFR measurements updated frequently, and also mitigate overtreatment towards the proportion of the T1D population that experiences a regular rate of progression to renal failure. The significant strength of the Bayesian sequential updating approach is that it outperforms the LOCF models in settings with sparser collections of biomarker data, i.e., less frequently collected and updated.

Chapter 10

Discussion

10.1 Concluding summary

Today's great opportunity to maximise prediction stems from large-scale electronic health record collections containing longitudinal data on risk factors for important clinical outcomes.

To that end, I have harnessed data on a cohort with type 1 diabetes for whom there are available records on developing cardiovascular disease and renal failure outcomes, along with a variety of relevant biomarker data. Firstly, I applied joint modelling of longitudinal and time to event data, a flexible suite of models, which is the current state-of-the-art approach to survival analysis based on longitudinal data. The rationale behind this approach was to assess if using joint modelling improves the risk prediction of CVD in people with T1D, supported by the evidence that harnessing the information hidden in longitudinal data could better inform risk prediction.

However, having encountered substantial scaling problems using this method, I have turned my attention to an alternative two-staged joint modelling approach based on sequential Bayesian updating for the longitudinal data and Poisson time-splitting for time to event. I applied the latter joint modelling approach to the problem of predicting progression to end-stage renal

disease (ESRD), using collected data on renal replacement therapy (RRT) of the particular cohort.

The reasoning for demonstrating the method, modelling the rate of potential progression to RRT instead of CVD, using longitudinal eGFR data, was to develop a simplified and manageable exemplar of applying joint modelling to a more tractable prediction problem. The longitudinal eGFR profiles were extrapolated frequently enough and contained measurements to evaluate the joint model formulation; therefore, using a single biomarker for showcasing the development was the most natural choice. In that matter, I assessed the performance of multiple model specifications since the fitting was less computationally cumbersome for the purpose of demonstrating the method.

My research has made a significant contribution to addressing the identified scaling issues of the evaluated `rstanarm` implementation and using more flexible stochastic models to determine the underlying process of the changing biomarkers and connect natural processes to the rate of progression to developing complications of diabetes. With the aim of establishing a place for joint models in the statistician's toolkit, I conducted a thorough examination of two formulations that facilitate more flexible and efficient risk prediction.

The use of a broader class of continuous-time state-space models has been examined with the objective of specifying more robust individual biomarker trajectories able to inform time to event. Improving the prediction of time to event highly depends on the effect of the longitudinal process on the development of the clinical outcome. A more precise fitting to the underlying transition states of the biomarker data could translate to bigger increments in risk estimation. To establish the gain, I used as a benchmark against the join models the traditionally used last-observation-carried-forward approach for the specification of the biomarker trajectory.

The predictive analysis of the rate of progression to RRT suggests that longitudinal eGFR trajectories determined by using autoregressive drift effect and diffusion processes, fit the data

best, along with specifying random intercepts and slopes, and outperform the constant LOCF extrapolated trajectory when used as new data to predict future events. To effectively specify a longitudinal process, the choice of the method depends on whether the biomarker exhibits a monotonically changing pattern over time, fluctuates considerably, or lacks a clear direction.

In terms of improving public health, the scrutiny of promising implementations and the choice of appropriate methods for specifying longitudinal processes and biomarkers can have significant implications. By identifying effective models and methodologies, practitioners can gain deeper insights into the progression of health conditions and better predict individual risk.

This, in turn, can lead to more tailored and targeted interventions, early detection of health issues, and optimised allocation of resources to prevent or manage diseases. The application of joint models and appropriate longitudinal analysis contributes to evidence-based decision-making, which is crucial for designing effective public health strategies, enhancing healthcare outcomes, and ultimately improving the overall health of populations.

Within individual-level, the reviewed Bayesian joint modelling implementation requires all biomarkers to be measured synchronously; otherwise, they cannot be modelled jointly. This feature severely restricts the ability to specify multiple longitudinal processes in parallel, resulting in a significant proportion of the original observations being discarded when the dates of biomarker measurements do not match.

However, despite this limitation, these implementations serve adequately for demonstrative purposes and lay the foundation for further development and refinement. They offer valuable insights and initial evidence of the potential of joint modelling techniques in addressing complex longitudinal relationships between biomarkers. As research in this area grows, it is expected that more versatile approaches will be devised to accommodate asynchronous measurements and overcome the challenges associated with the joint modelling of multiple longitudinal processes, like the introduction of a Kalman filter to infer the unobserved latent

states of the true values of the biomarkers conditional on all past observed data up to an arbitrary time point.

An additional bottleneck concerning multivariate joint models is that likelihood estimation is computationally intense, and the computation time scales unfavourably, as the number of biomarkers increases due to complex *quadrature* calculations.

The statements made herein may be considered as post-hoc observations. The introduction of precision medicine and personalised profiles of a growing number of individuals with chronic diseases has highlighted the magnified heterogeneity in substantial collections of data, making it challenging to determine in advance how individual longitudinal trajectories will behave over time. Through a major overhaul of my work on predicting future states for individuals with diverse biomarker trajectories, I have come to realise the significance of improving computational efficiency, which currently serves as a major bottleneck in the concept of precision medicine. I aspire that joint modelling approaches will facilitate meaningful interventions based on individual characteristics.

The number of software implementations that compute the likelihood of clinical outcomes based on time-updated biomarker data is rising (Bohr and Memarzadeh 2020; Johnson et al. 2021; Subbiah 2023; Hadjichrysanthou et al. 2020). My work has been influenced by recent advancements in inference algorithms that specify the likelihood of the predictive distribution surface. However, due to the substantial number of individual-level parameters requiring estimation, the computation of the likelihood surface needs to be approximated. Achieving an accurate approximation, or model fitting, involves sampling the most probable values from the underlying distribution. However, this process can become computationally intensive, especially for large-scale datasets.

As such, there is a growing need for more efficient schemes for posterior predictive inference. Developing improved algorithms and computational approaches can help address the challenges posed by large datasets and complex models, allowing for faster and more precise inference in

the context of predictive modelling and precision medicine applications. These advancements are critical for leveraging the potential of coupling statistical machine learning and precision medicine and enhancing the understanding and specification of diverse health trajectories.

Arguably, predicting the time to renal failure is an important practical problem, and improving existing risk prediction models might mitigate undertreatment in patients early in the course of the disease. Moreover, as new information becomes more frequently available, the concept of providing real-time risk prediction becomes more attractive. For example, preventing acute ketoacidosis in people with diabetes would be a fitting application to showcase methods for dynamically updated prediction. To that end, the use of differential equations to study the dependencies between underlying latent processes offers valuable insights, particularly in cases where these relationships may not follow linear patterns for the majority of instances.

The joint model fitted within the R package `rstanarm` accounts for the correlation that is induced by clustered observations (e.g., subjects, hospital, country). Albeit mathematically well designed, in the presence of two or more biomarkers, the computational limitations of quadrature become apparent; this method is poorly suited to numerical computation. In particular, the computation time scales unfavourably -exponentially- as the number of biomarkers increases. The joint model implemented in `rstanarm` requires numeric integration of thousands of biomarker trajectories, which is prohibitively time-consuming and computationally intensive for even moderately-sized datasets.

On the other hand, the Poisson time-splitting approach replaces the evaluation of the likelihood as an integral over the hazard function by factoring the likelihood over many short intervals. As this does not require quadrature computation, which is much faster than the continuous-time multivariate joint model.

The main takeaway point is that `rstanarm` was the state-of-the-art package up until recently, using the most advanced algorithms available at the time for Bayesian analysis. However, the developers themselves also expressed that its current implementation does not really scale

to any reasonably big dataset, to the effect that they allow other developers to find a way to speed it up. This is arguably feasible, but not by the currently employed approach that requires random effects integration, which fundamentally depends on rather slow quadrature computations.

To sum up, I have demonstrated an efficient approach that enables the fitting of joint models in realistic computing times and does not require quadrature computation for the longitudinal biomarker data like the most modern approach of `stan_jm()` that severely hampers the inclusion of more than two biomarkers. I have specified longitudinal profiles of eGFR and HbA_{1c} of individuals with T1D using various state-space models as an extension to the LOCF model for the biomarker trajectory. The imputations of missing values using an LOCF model have been used as a benchmark. Finally, all imputations have been fed as time-updated covariates into Poisson regression models for time to event.

Understanding the intertwined nature of biological processes is crucial for enhancing our ability to predict disease outcomes, develop targeted interventions, and ultimately advance the field of precision medicine.

The results obtained from the renal failure study provide convincing evidence that robust models can effectively determine the progression to end-stage renal disease. Incorporating longitudinal eGFR along with the clinical variables of sex, age, and diabetes duration leads to relatively modest improvements in predicting the time to event, even though the predictions are already reasonably accurate. The C-statistic, which measures average discrimination performance, does not show substantial improvement, which is unsurprising.

From a clinical perspective, the association parameter may be of greater interest in understanding the relationships between the variables. Currently, only the current value of the biomarker data has been experimented with, but exploring whether considering the rate of change or the area under the curve could enhance individual profiles and predictions is an essential research direction worth pursuing.

Regarding data handling, the LOCF approach works well in this exemplar case, as there are frequently collected observations. This translates into adequate model training, which can, in turn, provide informative time to event predictions.

Overall, the study's findings underscore the potential of robust modelling in predicting renal and cardiovascular disease progression, and further investigations into various data aspects can contribute to the promotion of precision medicine in managing complex biological processes like progression to renal failure.

10.2 Final remarks

This thesis has the following salient messages:

1. Calculating survival functions when we have time-varying covariates is complicated because it needs to specify a trajectory for each variable. Recently implemented software enables us to apply more flexible dynamic models to specify latent states of time-varying biomarker data pertinent to the outcome of interest and quantify the impact of the time-varying effects on the shape of the hazard rate of event.

2. This work investigates an alternative joint modelling approach, based on Poisson time-splitting, in which the observed events are modelled as a counting process over many short person-time intervals. The biomarker values at the start of each interval are imputed based on Bayesian sequential updating, using only observations up to the start of that interval from a class of models, known as hierarchical continuous-time dynamic models, of which linear mixed models are a special case.

3. With my approach, the generated biomarker imputations that fall after the landmark point are held for testing via cross-validation and do not include any real data arriving after this time.

4. I have conducted a double-prediction task: A. the Kalman filter predicts the trajectory

after the landmark time, conditional on the actual observations up to this time, and these imputed values are subsequently treated as test data, B. The Poisson model is first fitted on the imputed biomarker data conditional on observations until the landmark time, and it then predicts time to event using the imputed test data.

5. I have tested the performance of the Poisson models using Kalman filter imputations about future states of the biomarker trajectory instead of taking the conventional approach, which is to test the Poisson models on actual withheld observations. This approach has two prediction stages and errors are likely to be reflected in the calibration of the fitted models.

6. The proposed approach constructs the longitudinal trajectories based on advanced equations and uses the time-updated trajectories to fit a model for time to event. Furthermore, using one of the simplest parametric models to estimate the event rates is advantageous because including time-updated biomarker data in each interval helps reweigh the baseline hazard function as necessary.

7. The risk of the event is evaluated at every person-time interval. With many short intervals, the approximation of a Poisson time-splitting approach to a continuous-time survival model can be made arbitrarily close. For each constructed person-time interval, I have predicted the latent state of the biomarker, according to a Gaussian state-space model. I have also carried out the last-observation-carried-forward method as a benchmark. The piece-wise exponential approach based on the Poisson likelihood is reasonably robust for settings where the hazard rate remains constant within each interval. To be helpful, the trajectory should be somewhat informative of the hazard rate.

8. To control for immortal-time bias when generating the biomarker path, it is crucial to ensure that we are not including any observations from future time intervals. To elaborate on this, when using the Kalman filter algorithm to estimate latent states, it is

not allowed to use smoothing functions but only to conduct a filtering pass to compute the posterior distribution, given the observed data up to the beginning of each interval.

9. Partial exposure is not allowed in a time-splitting approach: I thus allow for the time intervals of the testing set to run all the way to the end. Observed censoring in training data is accounted for by including an offset term of exposure time (interval length) in the Poisson model. This is a necessary condition for treating the predicted events as exchangeable observations among intervals.

10. From a clinical perspective, the small increments in the predictive performance of the rate of progression of renal failure reflect the complexity of the processes defining time to developing complications in T1D. Models that perform well at guessing patients who will experience the outcome of interest (yes/no decision) in the near future are not necessarily predicting the time to event accurately.

11. Using a stochastic variation to capture the changes in risk factors could be a major determinant of hazard heterogeneity. Variation in disease risk often goes far beyond what is captured by measured risk factors (Aalen et al. 2015). Even in the case of predicting renal failure from longitudinal eGFR, frailty remains a large component. In fact, a large number of people with renal problems do not progress to the later stages of the disease but die prematurely due to numerous comorbidities and complications.

12. Heterogeneity is inherent due to stochastic processes. Varying frailty between individuals may have several different explanations: environment, genetics etc. (or it may be a purely stochastic phenomenon). In addition, some randomness averages out at the population-level model. Therefore, individualised risk prediction may need to be more accurate than estimating the overall population trends.

13. For the biomarker model, diffusion and autoregressive drift effects are larger than the linear trend. Our models do not allow for any lag in the effect of the latent biomarker on the hazard rate, so if the latent biomarker rises, the modelled hazard rate will increase

simultaneously. There may be lag effects of the changing biomarker, but even in this large dataset, we do not have enough information to detect them.

14. Frailty (unmeasured heterogeneity in individual susceptibility) can cause rate ratios associated with risk factors to "wear off" with time, in a survival analysis. This is because those who have survived without developing the disease, the rate at which the most susceptible individuals are removed is higher among those who have high levels of the risk factor than among those who have has low levels of the risk factor. This is relevant mainly to cardiovascular disease, where the rate ratios associated with risk factors wear off with age. This is especially likely to affect people with type 1 diabetes because the risk of CVD is high even at young ages, so the most frail individuals are removed more rapidly.

10.3 Outlook & Further directions

Risk prediction at the individual level presents a challenging task, both mathematically and computationally, primarily due to the significant number of random effects that determine the longitudinal process and influence the outcome of interest. While joint models implemented in `rstanarm` are elegantly formulated, they become impractical for modelling large-scale observational study data and electronic health care records due to their reliance on quadrature.

To tackle the intractable parameter estimation, a possible solution would be a method proposed by Mauff et al. (2020), known as importance sampling. This technique improves the estimation of the posterior probability distributions. In addition, one would expect the performance to be further improved by using multiple imputations of the biomarkers followed by averaging over the models.

Bottom line is that the time-splitting joint modelling approach, based on `ctsem` and Poisson regression, provides a better fit to the longitudinal data, producing risk predictions comparable to those obtained by continuous-time Bayesian joint models. This approach offers a scalable

alternative to `rstanarm`, and its predictive performance is acceptable when the risk factors adequately relate to the outcome of interest. Additionally, its cautious design helps overcome common pitfalls encountered in survival analysis, leading to unbiased estimates.

Considering these advantages, it is strongly recommended to adopt the time-splitting joint modelling approach for fitting robust joint models efficiently when dealing with large datasets. By leveraging this approach, researchers can obtain timely and accurate risk predictions, making it particularly valuable for precision medicine applications where large amounts of data need to be analysed to gain insights into disease progression and individual biomarker trajectories.

References

Aalen, Odd O, Morten Valberg, Tom Grotmol, and Steinar Tretli. 2015. "Understanding Variation in Disease Risk: The Elusive Concept of Frailty." *International Journal of Epidemiology* 44 (4): 1408–21.

Abd ElHafeez, Samar, Graziella D'Arrigo, Daniela Leonardis, Maria Fusaro, Giovanni Tripepi, and Stefanos Roumeliotis. 2021. "Methods to Analyze Time-to-Event Data: The Cox Regression Analysis." *Oxidative Medicine and Cellular Longevity* 2021: 1–6.

Altman, Douglas G, and Patrick Royston. 2000. "What Do We Mean by Validating a Prognostic Model?" *Statistics in Medicine* 19 (4): 453–73.

Anderson, James R, Kevin C Cain, Richard D Gelber, and others. 1983. "Analysis of Survival by Tumor Response." *J Clin Oncol* 1 (11): 710–19.

Andrinopoulou, Eleni-Rosalina, and Dimitris Rizopoulos. 2016. "Bayesian Shrinkage Approach for a Joint Model of Longitudinal and Survival Outcomes Assuming Different Association Structures." *Statistics in Medicine* 35 (26): 4813–23.

Asar, Özgür, James Ritchie, Philip A Kalra, and Peter J Diggle. 2015. "Joint Modelling of Repeated Measurement and Time-to-Event Data: An Introductory Tutorial." *International Journal of Epidemiology* 44 (1): 334–44.

Association, American Diabetes, and others. 2006. "Diagnosis and Classification of Diabetes Mellitus." *Diabetes Care* 29 (1): S43.

Astor, Brad C, Stein I Hallan, Edgar R Miller III, Edwina Yeung, and Josef Coresh. 2008. "Glomerular Filtration Rate, Albuminuria, and Risk of Cardiovascular and All-Cause Mortality in the Us Population." *American Journal of Epidemiology* 167 (10): 1226–34.

Austin, Peter C, Ewout W Steyerberg, and Hein Putter. 2021. "Fine-Gray Subdistribution Hazard Models to Simultaneously Estimate the Absolute Risk of Different Event Types: Cumulative Total Failure Probability May Exceed 1." *Statistics in Medicine* 40 (19): 4200–4212.

Baade, Peter D, Patrick Royston, Philipa H Youl, Martin A Weinstock, Alan Geller, and Joanne F Aitken. 2015. "Prognostic Survival Model for People Diagnosed with Invasive Cutaneous Melanoma." *BMC Cancer* 15 (1): 1–13.

Baart, Sara J, Roel LF van der Palen, Hein Putter, Roula Tsonaka, Nico A Blom, Dimitris Rizopoulos, and Nan van Geloven. 2021. "Joint Modeling of Longitudinal Markers and Time-to-Event Outcomes: An Application and Tutorial in Patients After Surgical Repair of Transposition of the Great Arteries." *Circulation: Cardiovascular Quality and Outcomes* 14 (11): e007593.

Banfi, Giuseppe, and Massimo Del Fabbro. 2006. "Serum Creatinine Values in Elite Athletes Competing in 8 Different Sports: Comparison with Sedentary People." *Clinical Chemistry* 52 (2): 330–31.

Barbieri, Sebastiano, Suneela Mehta, Billy Wu, Chrianna Bharat, Katrina Poppe, Louisa Jorm, and Rod Jackson. 2022. "Predicting Cardiovascular Risk from National Administrative Databases Using a Combined Survival Analysis and Deep Learning Approach." *International Journal of Epidemiology* 51 (3): 931.

Bargnoux, Anne-Sophie, Nils Kuster, Etienne Cavalier, Laurence Piéroni, Jean-Sébastien Souweine, Pierre Delanaye, and Jean-Paul Cristol. 2018. "Serum Creatinine: Advantages and Pitfalls." *J Lab Precis Med* 3 (8): 71.

Beisswenger, Paul J. 2012. "Glycation and Biomarkers of Vascular Complications of Diabetes." *Amino Acids* 42 (4): 1171–83.

Bellera, Carine A, Gaëtan MacGrogan, Marc Debled, Christine Tunon De Lara, Véronique Brouste, and Simone Mathoulin-Pélissier. 2010. "Variables with Time-Varying Effects and the Cox Model: Some Statistical Concepts Illustrated with a Prognostic Factor Study in Breast Cancer." *BMC Medical Research Methodology* 10 (1): 1–12.

Berger, James O, Brunero Liseo, and Robert L Wolpert. 1999. "Integrated Likelihood Methods for Eliminating Nuisance Parameters." *Statistical Science*, 1–22.

Bermingham, Mairead L, Marco Colombo, Stuart J McGurnaghan, Luke AK Blackbourn, Frano Vučković, Maja Pučić Baković, Irena Trbojević-Akmačić, et al. 2018. "N-Glycan Profile and Kidney Disease in Type 1 Diabetes." *Diabetes Care* 41 (1): 79–87.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv Preprint arXiv:1701.02434*.

Betancourt, MJ, and M Girolami. 2013. "Hamiltonian Monte Carlo for Hierarchical Models. ArXiv. Org."

Bianconcini, Silvia. 2014. "Asymptotic Properties of Adaptive Maximum Likelihood Estimators in Latent Variable Models." *Bernoulli* 20 (3): 1507–31.

Bohr, Adam, and Kaveh Memarzadeh. 2020. "The Rise of Artificial Intelligence in Healthcare Applications." In *Artificial Intelligence in Healthcare*, 25–60. Elsevier.

Botev, Rossini, and Jean-Pierre Mallié. 2008. "Reporting the eGFR and Its Implication for Ckd Diagnosis." *Clinical Journal of the American Society of Nephrology* 3 (6): 1606–7.

Brand, Jan AJG van den, Gerben AJ van Boekel, Hans L Willems, Lambertus ALM Kiemeney, Martin den Heijer, and Jack FM Wetzels. 2011. "Introduction of the Ckd-Epi Equation to Estimate Glomerular Filtration Rate in a Caucasian Population." *Nephrology Dialysis*

Transplantation 26 (10): 3176–81.

Brilleman, Sam. 2022. "Estimating Joint Models for Longitudinal and Time-to-Event Data with Rstanarm."

Brilleman, Sam, Michael Crowther, Margarita Moreno-Betancur, J Buros Novik, and Rory Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." *Proceedings of StanCon 2018*, 1–18.

Brouste, Alexandre, Masaaki Fukasawa, Hideitsu Hino, Stefano Iacus, Kengo Kamatani, Yuta Koike, Hiroki Masuda, et al. 2014. "The Yuima Project: A Computational Framework for Simulation and Inference of Stochastic Differential Equations." *Journal of Statistical Software* 57: 1–51.

Brown, Elizabeth R, Joseph G Ibrahim, and Victor DeGruttola. 2005. "A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival." *Biometrics* 61 (1): 64–73.

Brown, ER. 2003. "A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data. Biometrics 59: 221228Brown Er, Ibrahim Jg (2003) Bayesian Approaches to Joint Cure Rate and Longitudinal Models with Applications to Cancer Vaccine Trials. Biometrics 59: 686693Brown Er, Ibrahim Jg, Degruttola V (2005) a Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival. Biometrics 61: 6473Carlin Bp, Polson Ng (1991) an Expected Utility Approach to Influence Diagnostics." *J Am Stat Assoc*.

Buse, Maria G. 2006. "Hexosamines, Insulin Resistance, and the Complications of Diabetes: Current Status." *American Journal of Physiology-Endocrinology and Metabolism* 290 (1): E1–E8.

Campbell, Kristen R, Rui Martins, Scott Davis, and Elizabeth Juarez-Colunga. 2021. "Dynamic Prediction Based on Variability of a Longitudinal Biomarker." *BMC Medical Research Methodology* 21 (1): 104.

Captieux, Mireille, Kelly Fleetwood, Brian Kennon, Naveed Sattar, Robert Lindsay, Bruce Guthrie, Sarah H Wild, and Scottish Diabetes Research Network Epidemiology Group. 2021. "Epidemiology of Type 2 Diabetes Remission in Scotland in 2019: A Cross-Sectional Population-Based Study." *PLoS Medicine* 18 (11): e1003828.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Carpenter, James R, and Michael G Kenward. 2007. "Missing Data in Randomised Controlled Trials: A Practical Guide." Health Technology Assessment Methodology Programme.

Chang, Tara I, Suying Li, Shu-Cheng Chen, Carmen A Peralta, Michael G Shlipak, Linda F Fried, Adam T Whaley-Connell, et al. 2013. "Risk Factors for Esrd in Individuals with Preserved Estimated Gfr with and Without Albuminuria: Results from the Kidney Early Evaluation Program (Keep)." *American Journal of Kidney Diseases* 61 (4): S4–S11.

Coleman, James Samuel, and others. 1964. "Introduction to Mathematical Sociology." *Introduction to Mathematical Sociology.*

Collins, Gary S, Joris A de Groot, Susan Dutton, Omar Omar, Milensu Shanyinde, Abdelouahid Tajar, Merryn Voysey, et al. 2014. "External Validation of Multivariable Prediction Models: A Systematic Review of Methodological Conduct and Reporting." *BMC Medical Research Methodology* 14 (1): 1–11.

Colombo, Marco, Akram Asadi Shehni, Ioanna Thoma, Stuart J McGurnaghan, Luke AK Blackbourn, Hayden Wilkinson, Andrew Collier, et al. 2021. "Quantitative Levels of Serum N-Glycans in Type 1 Diabetes and Their Association with Kidney Disease." *Glycobiology* 31 (5): 613–23.

Colombo, Marco, Stuart J McGurnaghan, Samira Bell, Finlay MacKenzie, Alan W Patrick, John R Petrie, John A McKnight, et al. 2020. "Predicting Renal Disease Progression in a

Large Contemporary Cohort with Type 1 Diabetes Mellitus." *Diabetologia* 63: 636–47.

Colombo, Marco, Erkka Valo, Stuart J McGurnaghan, Niina Sandholm, Luke AK Blackbourn, R Neil Dalton, David Dunger, et al. 2019. "Biomarker Panels Associated with Progression of Renal Disease in Type 1 Diabetes." *Diabetologia* 62: 1616–27.

Commandeur, Jacques JF, Siem Jan Koopman, and Marius Ooms. 2011. "Statistical Software for State Space Methods." *Journal of Statistical Software* 41: 1–18.

Condrat, Carmen Elena, Dana Claudia Thompson, Madalina Gabriela Barbu, Oana Larisa Bugnar, Andreea Boboc, Dragos Cretoiu, Nicolae Suciu, Sanda Maria Cretoiu, and Silviu Cristian Voinea. 2020. "MiRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis." *Cells* 9 (2): 276.

Cooper, Hannah, Sue Wells, and Suneela Mehta. 2022. "Are Competing-Risk Models Superior to Standard Cox Models for Predicting Cardiovascular Risk in Older Adults? Analysis of a Whole-of-Country Primary Prevention Cohort Aged over 65 Years." *International Journal of Epidemiology* 51 (2): 604–14.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187–202.

Crowther, Michael J, Paul C Lambert, and Keith R Abrams. 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): 1–8.

Curran, Patrick J. 2003. "Have Multilevel Models Been Structural Equation Models All Along?" *Multivariate Behavioral Research* 38 (4): 529–69.

Dafni, Urania. 2011. "Landmark Analysis at the 25-Year Landmark Point." *Circulation: Cardiovascular Quality and Outcomes* 4 (3): 363–71.

Deboeck, Pascal R, and Kristopher J Preacher. 2016. "No Need to Be Discrete: A Method for

Continuous Time Mediation Analysis." *Structural Equation Modeling: A Multidisciplinary Journal* 23 (1): 61–75.

Deckert, T, JE Poulsen, and M Larsen. 1978. "Prognosis of Diabetics with Diabetes Onset Before the Age of Thirtyone." *Diabetologia* 14 (6): 371–77.

Delanaye, Pierre, Etienne Cavalier, and Hans Pottel. 2017. "Serum Creatinine: Not so Simple!" *Nephron* 136 (4): 302–8.

Depaoli, Sarah, James P Clifton, and Patrice R Cobb. 2016. "Just Another Gibbs Sampler (Jags) Flexible Software for Mcmc Implementation." *Journal of Educational and Behavioral Statistics* 41 (6): 628–49.

Deshpande, Anjali D, Marcie Harris-Hayes, and Mario Schootman. 2008. "Epidemiology of Diabetes and Diabetes-Related Complications." *Physical Therapy* 88 (11): 1254–64.

Dickman, Paul W, Andy Sloggett, Michael Hills, and Timo Hakulinen. 2004. "Regression Models for Relative Survival." *Statistics in Medicine* 23 (1): 51–64.

Dickson, E Rolland, Patricia M Grambsch, Thomas R Fleming, Lloyd D Fisher, and Alice Langworthy. 1989. "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making." *Hepatology* 10 (1): 1–7.

Diggle, Peter, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, and others. 2002. *Analysis of Longitudinal Data*. Oxford university press.

Diggle, Peter J, Inês Sousa, and Özgür Asar. 2015. "Real-Time Monitoring of Progression Towards Renal Failure in Primary Care Patients." *Biostatistics* 16 (3): 522–36.

Dorman, JS, RE Laporte, LH Kuller, KJ Cruickshanks, TJ Orchard, DK Wagener, DJ Becker, DE Cavender, and AL Drash. 1984. "The Pittsburgh Insulin-Dependent Diabetes Mellitus (Iddm) Morbidity and Mortality Study: Mortality Results." *Diabetes* 33 (3): 271–76.

Draper, Norman R, and Harry Smith. 1998. *Applied Regression Analysis*. Vol. 326. John

Wiley & Sons.

Driver, Charles C, Johan HL Oud, and Manuel C Voelkle. 2017. "Continuous Time Structural Equation Modeling with R Package Ctsem." *Journal of Statistical Software* 77: 1–35.

Driver, Charles C, and Manuel C Voelkle. 2018. "Hierarchical Bayesian Continuous Time Dynamic Modeling." *Psychological Methods* 23 (4): 774.

Duane, Simon, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. 1987. "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–22.

Dung, Van Than, and Tegoeh Tjahjowidodo. 2017. "A Direct Method to Solve Optimal Knots of B-Spline Curves: An Application for Non-Uniform B-Spline Curves Fitting." *PloS One* 12 (3): e0173857.

Durbin, James, and Siem Jan Koopman. 2012. *Time Series Analysis by State Space Methods*. Vol. 38. OUP Oxford.

Elliott, Jackie, Solomon Tesfaye, Nish Chaturvedi, Rajiv A Gandhi, Lynda K Stevens, Celia Emery, John H Fuller, EURODIAB Prospective Complications Study Group, and others. 2009. "Large-Fiber Dysfunction in Diabetic Peripheral Neuropathy Is Predicted by Cardiovascular Risk Factors." *Diabetes Care* 32 (10): 1896–1900.

Ezzati, Majid, Stephen Vander Hoorn, Carlene M M Lawes, Rachel Leach, W Philip T James, Alan D Lopez, Anthony Rodgers, and Christopher J L Murray. 2005. "Rethinking the 'Diseases of Affluence' Paradigm: Global Patterns of Nutritional Risks in Relation to Economic Development." *PLoS Medicine* 2 (5): e133.

Faucett, Cheryl L, and Duncan C Thomas. 1996. "Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach." *Statistics in Medicine* 15 (15): 1663–85.

Fauvernier, Mathieu, Laurent Remontet, Zoé Uhry, Nadine Bossard, and Laurent Roche.

2019. " "SurvPen": An** R** Package for Hazard and Excess Hazard Modelling with Multidimensional Penalized Splines." *Journal of Open Source Software* 4 (40): 1434.

Ferranti, SD de, IH de Boer, V Fonseca, CS Fox, SH Golden, CJ Lavie, SN Magge, et al. n.d. "Type 1 Diabetes Mellitus and Cardiovascular Disease: A Scientific Statement from the American Heart Association and American Diabetes Association [Published Online Ahead of Print August 11, 2014]." *Circulation. Doi* 10.

Food, US, Drug Administration, and others. 2020. "About Biomarkers and Qualification."

Frost, Jim, and others. 2017. "Multicollinearity in Regression Analysis: Problems, Detection, and Solutions." *Statistics by Jim* 2.

Furgal, Allison KC, Ananda Sen, and Jeremy MG Taylor. 2019. "Review and Comparison of Computational Approaches for Joint Longitudinal and Time-to-Event Models." *International Statistical Review* 87 (2): 393–418.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2017. "Visualization in Bayesian Workflow." *arXiv Preprint arXiv:1709.01449*.

Gelman, Andrew, Daniel Lee, and Jiqiang Guo. 2015. "Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization." *Journal of Educational and Behavioral Statistics* 40 (5): 530–43.

Glasziou, Paul, Les Irwig, and David Mant. 2005. "Monitoring in Chronic Disease: A Rational Approach." *Bmj* 330 (7492): 644–48.

Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore. 2010. "Understanding Survival Analysis: Kaplan-Meier Estimate." *International Journal of Ayurveda Research* 1 (4): 274.

Gohda, Tomohito, Nozomu Kamei, Takeo Koshida, Mitsunobu Kubota, Kanako Tanaka, Yoshinori Yamashita, Eri Adachi, et al. 2020. "Circulating Kidney Injury Molecule-1 as a Biomarker of Renal Parameters in Diabetic Kidney Disease." *Journal of Diabetes*

Investigation 11 (2): 435–40.

Goldstein, Benjamin A, Gina Maria Pomann, Wolfgang C Winkelmayer, and Michael J Pencina. 2017. "A Comparison of Risk Prediction Methods Using Repeated Observations: An Application to Electronic Health Records for Hemodialysis." *Statistics in Medicine* 36 (17): 2750–63.

Goudie, Robert JB, and Sach Mukherjee. 2016. "A Gibbs Sampler for Learning Dags."

Gould, A Lawrence, Mark Ernest Boye, Michael J Crowther, Joseph G Ibrahim, George Quartey, Sandrine Micallef, and Frederic Y Bois. 2015. "Responses to Discussants of '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): 2202.

Gowda, Shivaraj, Prakash B Desai, Shruthi S Kulkarni, Vinayak V Hull, Avinash AK Math, and Sonal N Vernekar. 2010. "Markers of Renal Function Tests." *North American Journal of Medical Sciences* 2 (4): 170.

Gratton, Serge, Monserrat Rincon-Camacho, Ehouarn Simon, and Philippe L Toint. 2015. "Observation Thinning in Data Assimilation Computations." *EURO Journal on Computational Optimization* 3 (1): 31–51.

Gray, Robert J. 1992. "Flexible Methods for Analyzing Survival Data Using Splines, with Applications to Breast Cancer Prognosis." *Journal of the American Statistical Association* 87 (420): 942–51.

Grijalva, Carlos G, J Pekka Nuorti, Patrick G Arbogast, Stacey W Martin, Kathryn M Edwards, and Marie R Griffin. 2007. "Decline in Pneumonia Admissions After Routine Childhood Immunisation with Pneumococcal Conjugate Vaccine in the Usa: A Time-Series Analysis." *The Lancet* 369 (9568): 1179–86.

Gu, Harvest F. 2019. "Genetic and Epigenetic Studies in Diabetic Kidney Disease." *Frontiers*

in Genetics 10: 507.

Gubitosi-Klug, Rose A, Barbara H Braffett, Susan Hitt, Valerie Arends, Diane Uschner, Kimberly Jones, Lisa Diminick, et al. 2021. "Residual β Cell Function in Long-Term Type 1 Diabetes Associates with Reduced Incidence of Hypoglycemia." *The Journal of Clinical Investigation* 131 (3).

Guidoum, Arsalane Chouaib, and Kamal Boukhetala. 2020. "Performing Parallel Monte Carlo and Moment Equations Methods for Itô and Stratonovich Stochastic Differential Systems: R Package Sim. DiffProc." *Journal of Statistical Software* 96: 1–82.

Habbema, JDF, and J Hilden. 1981. "The Measurement of Performance in Probabilistic Diagnosis Iv. Utility Considerations in Therapeutics and Prognostics." *Methods of Information in Medicine* 20 (02): 80–96.

Hadjichrysanthou, Christoforos, Stephanie Evans, Sumali Bajaj, Loizos C Siakallis, Kevin McRae-McKee, Frank De Wolf, Roy M Anderson, and Alzheimer's Disease Neuroimaging Initiative. 2020. "The Dynamics of Biomarkers Across the Clinical Spectrum of Alzheimer's Disease." *Alzheimer's Research & Therapy* 12: 1–16.

Hair Jr, Joseph F, G Tomas M Hult, Christian M Ringle, Marko Sarstedt, Nicholas P Danks, Soumya Ray, Joseph F Hair, et al. 2021. "Evaluation of the Structural Model." *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*, 115–38.

Hald, Anders. 1981. "TN Thiele's Contributions to Statistics." *International Statistical Review/Revue Internationale de Statistique*, 1–20.

Handy, Lori K, Stefania Maroudi, Maura Powell, Bakanuki Nfila, Charlotte Moser, Ingrid Japa, Ndibo Monyatsi, et al. 2017. "The Impact of Access to Immunization Information on Vaccine Acceptance in Three Countries." *PloS One* 12 (8): e0180759.

Hanley, James A, and Barbara J McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (Roc) Curve." *Radiology* 143 (1): 29–36.

Hannan, Michael T, and Nancy Brandon Tuma. 1979. "Methods for Temporal Analysis." *Annual Review of Sociology*, 303–28.

Harjutsalo, Valma, and Per-Henrik Groop. 2014. "Epidemiology and Risk Factors for Diabetic Kidney Disease." *Advances in Chronic Kidney Disease* 21 (3): 260–66.

Harley, K, and C Jones. 1996. "Quality of Scottish Morbidity Record (Smr) Data." *Health Bulletin* 54 (5): 410–17.

Harrell Jr, Frank E, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. 1984. "Regression Modelling Strategies for Improved Prognostic Prediction." *Statistics in Medicine* 3 (2): 143–52.

Hatfield, Laura A, Mark E Boye, and Bradley P Carlin. 2011. "Joint Modeling of Multiple Longitudinal Patient-Reported Outcomes and Survival." *Journal of Biopharmaceutical Statistics* 21 (5): 971–91.

Health, National Institute for, and Care Excellence. 2021. "Chronic Kidney Disease: Assessment and Management Nice Guideline [Ng203]."

Hecht, Martin, and Steffen Zitzmann. 2020. "A Computationally More Efficient Bayesian Approach for Estimating Continuous-Time Models." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (6): 829–40.

Hemmelgarn, Brenda R, Braden J Manns, Anita Lloyd, Matthew T James, Scott Klarenbach, Robert R Quinn, Natasha Wiebe, Marcello Tonelli, Alberta Kidney Disease Network, and others. 2010. "Relation Between Kidney Function, Proteinuria, and Adverse Outcomes." *Jama* 303 (5): 423–29.

Henderson, Robin, Peter Diggle, and Angela Dobson. 2000. "Joint Modelling of Longitudinal Measurements and Event Time Data." *Biostatistics* 1 (4): 465–80.

Hickey, Graeme L, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona.

2016. "Joint Modelling of Time-to-Event and Multivariate Longitudinal Outcomes: Recent Developments and Issues." *BMC Medical Research Methodology* 16 (1): 1–15.

Ho, AM-H, PW Dion, CSH Ng, and MK Karmakar. 2013. "Understanding Immortal Time Bias in Observational Cohort Studies." *Anaesthesia*. Wiley Online Library.

Hoffman, Matthew D, Andrew Gelman, and others. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–1623.

Höhn, Andreas, Stuart J McGurnaghan, Thomas M Caparrotta, Anita Jeyam, Joseph E O'Reilly, Luke AK Blackbourn, Sara Hatam, et al. 2022. "Large Socioeconomic Gap in Period Life Expectancy and Life Years Spent with Complications of Diabetes in the Scottish Population with Type 1 Diabetes, 2013–2018." *Plos One* 17 (8): e0271110.

Holford, Theodore R. 1980. "The Analysis of Rates and of Survivorship Using Log-Linear Models." *Biometrics*, 299–305.

Hoyle, Rick H. 2012. *Handbook of Structural Equation Modeling*. Guilford press.

Hyndman, Rob J, and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. OTexts.

Iacus, Stefano Maria. 2007. "Sde: Simulation and Inference for Stochastic Differential Equations."

Ibrahim, Joseph G, Haitao Chu, and Liddy M Chen. 2010. "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data." *Journal of Clinical Oncology* 28 (16): 2796.

Ilic, Dragan, Mia Djulbegovic, Jae Hung Jung, Eu Chang Hwang, Qi Zhou, Anne Cleves, Thomas Agoritsas, and Philipp Dahm. 2018. "Prostate Cancer Screening with Prostate-Specific Antigen (Psa) Test: A Systematic Review and Meta-Analysis." *Bmj* 362.

Izzo, Giuseppe, and Antonia Vecchio. 2007. "A Discrete Time Version for Models of Population Dynamics in the Presence of an Infection." *Journal of Computational and Applied Mathematics* 210 (1-2): 210–21.

Jeyam, Anita, Helen Colhoun, Stuart McGurnaghan, Luke Blackbourn, Timothy J McDonald, Colin NA Palmer, John A McKnight, et al. 2021. "Clinical Impact of Residual c-Peptide Secretion in Type 1 Diabetes on Glycemia and Microvascular Complications." *Diabetes Care* 44 (2): 390–98.

Jeyam, Anita, Fraser W Gibb, John A McKnight, Brian Kennon, Joseph E O'Reilly, Thomas M Caparrotta, Andreas Höhn, et al. 2021. "Marked Improvements in Glycaemic Outcomes Following Insulin Pump Therapy Initiation in People with Type 1 Diabetes: A Nationwide Observational Study in Scotland." *Diabetologia* 64 (6): 1320–31.

Jeyam, Anita, Fraser W Gibb, John A McKnight, Joseph E O'Reilly, Thomas M Caparrotta, Andreas Höhn, Stuart J McGurnaghan, et al. 2022. "Flash Monitor Initiation Is Associated with Improvements in Hba1c Levels and Dka Rates Among People with Type 1 Diabetes in Scotland: A Retrospective Nationwide Observational Study." *Diabetologia* 65 (1): 159–72.

Jia, Xiaona, Mirza Mansoor Baig, Farhaan Mirza, and Hamid GholamHosseini. 2019. "A Cox-Based Risk Prediction Model for Early Detection of Cardiovascular Disease: Identification of Key Risk Factors for the Development of a 10-Year Cvd Risk Prediction." *Advances in Preventive Medicine* 2019.

Johnson, Kevin B, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. 2021. "Precision Medicine, Ai, and the Future of Personalized Health Care." *Clinical and Translational Science* 14 (1): 86–93.

Kalman, Rudolph Emil. 1960. "A New Approach to Linear Filtering and Prediction Problems." *Transactions of the ASME–Journal of Basic Engineering* 82 (Series D): 35–45.

Kampmann, JP, and J Mølholm Hansen. 1981. "Glomerular Filtration Rate and Creatinine

Clearance." *British Journal of Clinical Pharmacology* 12 (1): 7.

Kannel, William B, Thomas R Dawber, Abraham Kagan, Nicholas Revotskie, and JOSEPH STOKES III. 1961. "Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up Experience: The Framingham Study." *Annals of Internal Medicine* 55 (1): 33–50.

Kaplan, Edward L, and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81.

Kassirer, Jerome P. 1971. "Clinical Evaluation of Kidney Function: Glomerular Function." *New England Journal of Medicine* 285 (7): 385–89.

Katsarou, Anastasia, Soffia Gudbjörnsdottir, Araz Rawshani, Dana Dabelea, Ezio Bonifacio, Barbara J Anderson, Laura M Jacobsen, Desmond A Schatz, and Åke Lernmark. 2017. "Type 1 Diabetes Mellitus." *Nature Reviews Disease Primers* 3 (1): 1–17.

Khambhati, Jay, Marc Allard-Ratick, Devinder Dhindsa, Suegene Lee, John Chen, Pratik B Sandesara, Wesley O'Neal, et al. 2018. "The Art of Cardiovascular Risk Assessment." *Clinical Cardiology* 41 (5): 677–84.

King, Aaron A, Dao Nguyen, and Edward L Ionides. 2015. "Statistical Inference for Partially Observed Markov Processes via the R Package Pomp." *arXiv Preprint arXiv:1509.00503*.

Krzyszczyk, Paulina, Alison Acevedo, Erika J Davidoff, Lauren M Timmins, Ileana Marrero-Berrios, Misaal Patel, Corina White, et al. 2018. "The Growing Role of Precision and Personalized Medicine for Cancer Treatment." *Technology* 6 (03n04): 79–100.

Kucukelbir, Alp, Rajesh Ranganath, Andrew Gelman, and David Blei. 2015. "Automatic Variational Inference in Stan." *Advances in Neural Information Processing Systems* 28.

Kuijk, Sander MJ van, Frank JWM Dankers, Alberto Traverso, and Leonard Wee. 2019. "Preparing Data for Predictive Modelling." *Fundamentals of Clinical Data Science*, 75–84.

Kumar, Dhananjay, and Bengt Klefsjö. 1994. "Proportional Hazards Model: A Review." *Reliability Engineering & System Safety* 44 (2): 177–88.

Lachin, John M. 2016. "Fallacies of Last Observation Carried Forward Analyses." *Clinical Trials* 13 (2): 161–68.

Lacour, B. 1992. "Creatinine and Renal Function." *Nephrologie* 13 (2): 73–81.

Laird, Nan, and Donald Olivier. 1981. "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques." *Journal of the American Statistical Association* 76 (374): 231–40.

Lammert, Eckhard, Eckhard Lammert, Martin Zeeb, and Martin Zeeb. 2014. *Metabolism of Human Diseases: Organ Physiology and Pathophysiology*. Springer.

Lauritzen, Steffen L. 1981. "Time Series Analysis in 1880: A Discussion of Contributions Made by Tn Thiele." *International Statistical Review/Revue Internationale de Statistique*, 319–31.

"Lecture Notes on Generalized Linear Models." n.d. https://grodri.github.io/glms/notes/c7.pdf.

Lévesque, Linda E, James A Hanley, Abbas Kezouh, and Samy Suissa. 2010. "Problem of Immortal Time Bias in Cohort Studies: Example Using Statins for Preventing Progression of Diabetes." *Bmj* 340.

Levey, Andrew S, Ronald D Perrone, and Nicolaos E Madias. 1988. "Serum Creatinine and Renal Function." *Annual Review of Medicine* 39 (1): 465–90.

Li, Yan, Orestis A Panagiotou, Amanda Black, Dandan Liao, and Sholom Wacholder. 2016. "Multivariate Piecewise Exponential Survival Modeling." *Biometrics* 72 (2): 546–53.

Lim, David KE, James H Boyd, Elizabeth Thomas, Aron Chakera, Sawitchaya Tippaya, Ashley Irish, Justin Manuel, Kim Betts, and Suzanne Robinson. 2022. "Prediction Models

Used in the Progression of Chronic Kidney Disease: A Scoping Review." *PloS One* 17 (7): e0271619.

Lin, Xihong, Jeremy MG Taylor, and Wen Ye. 2008. "A Penalized Likelihood Approach to Joint Modeling of Longitudinal Measurements and Time-to-Event Data." *Statistics and Its Interface* 1 (1): 33–45.

Livingstone, Shona J, Helen C Looker, Eleanor J Hothersall, Sarah H Wild, Robert S Lindsay, John Chalmers, Stephen Cleland, et al. 2012. "Risk of Cardiovascular Disease and Total Mortality in Adults with Type 1 Diabetes: Scottish Registry Linkage Study."

Long, Jeffrey D, and James A Mills. 2018. "Joint Modeling of Multivariate Longitudinal Data and Survival Data in Several Observational Studies of Huntington's Disease." *BMC Medical Research Methodology* 18 (1): 1–15.

Lopez, Joseph. 2015. "Carl a. Burtis and David E. Bruns: Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics, Elsevier, Amsterdam, 1075 Pp, Isbn 978-1-4557-4165-6." Springer.

Lorenz, Max O. 1905. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association* 9 (70): 209–19.

Lunn, David J, Andrew Thomas, Nicky Best, and David Spiegelhalter. 2000. "WinBUGS-a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10 (4): 325–37.

Luo, Ya. 2018. "Joint Modeling of Longitudinal and Survival Data via Multivariate Mixed Effects State Space Model." PhD thesis, UC Santa Barbara.

MacKay, David JC, David JC Mac Kay, and others. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge university press.

Mameli, Chiara, Sara Mazzantini, Moufida Ben Nasr, Paolo Fiorina, Andrea E Scaramuzza,

and Gian Vincenzo Zuccotti. 2015. "Explaining the Increased Mortality in Type 1 Diabetes." *World Journal of Diabetes* 6 (7): 889.

Maser, Raelene E, Dorothy J Becker, Allan L Drash, Demetrius Ellis, Lewis H Kuller, Douglas A Greene, and Trevor J Orchard. 1992. "Pittsburgh Epidemiology of Diabetes Complications Study: Measuring Diabetic Neuropathy Follow-up Study Results." *Diabetes Care* 15 (4): 525–27.

Masnadi-Shirazi, Hamed, Alireza Masnadi-Shirazi, and Mohammad-Amir Dastgheib. 2019. "A Step by Step Mathematical Derivation and Tutorial on Kalman Filters." *arXiv Preprint arXiv:1910.03558*.

Mauff, Katya, Ewout Steyerberg, Isabella Kardys, Eric Boersma, and Dimitris Rizopoulos. 2020. "Joint Models with Multiple Longitudinal Outcomes and a Time-to-Event Outcome: A Corrected Two-Stage Approach." *Statistics and Computing* 30 (4): 999–1014.

Mauff, Katya, Ewout W Steyerberg, Giel Nijpels, Amber AWA van der Heijden, and Dimitris Rizopoulos. 2017. "Extension of the Association Structure in Joint Models to Include Weighted Cumulative Effects." *Statistics in Medicine* 36 (23): 3746–59.

McGurnaghan, Stuart J, Luke AK Blackbourn, Thomas M Caparrotta, Joseph Mellor, Anna Barnett, Andy Collier, Naveed Sattar, et al. 2022. "Cohort Profile: The Scottish Diabetes Research Network National Diabetes Cohort–a Population-Based Cohort of People with Diabetes in Scotland." *BMJ Open* 12 (10): e063046.

McGurnaghan, Stuart J, Paul M McKeigue, Stephanie H Read, Stefan Franzen, Ann-Marie Svensson, Marco Colombo, Shona Livingstone, et al. 2021. "Development and Validation of a Cardiovascular Risk Prediction Model in Type 1 Diabetes." *Diabetologia* 64 (9): 2001–11.

McKeigue, Paul. 2022. "Fitting Joint Models of Longitudinal Observations and Time to Event by Sequential Bayesian Updating." *Statistical Methods in Medical Research* 31 (10):

1934–41.

McKeigue, Paul M, Stuart McGurnaghan, Luke Blackbourn, Louise E Bath, David A McAllister, Thomas M Caparrotta, Sarah H Wild, Simon N Wood, Diane Stockton, and Helen M Colhoun. 2022. "Relation of Incident Type 1 Diabetes to Recent Covid-19 Infection: Cohort Study Using E-Health Record Linkage in Scotland." *medRxiv*.

Molnar, Frank J, Brian Hutton, and Dean Fergusson. 2008. "Does Analysis Using 'Last Observation Carried Forward' Introduce Bias in Dementia Research?" *Cmaj* 179 (8): 751–53.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. "Layer-Wise Relevance Propagation: An Overview." *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 193–209.

Moolgavkar, Suresh H. 2015. "Commentary: Frailty and Heterogeneity in Epidemiological Studies." *International Journal of Epidemiology* 44 (4): 1425–6.

Nathan, David M, and DCCT/Edic Research Group. 2014. "The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study at 30 Years: Overview." *Diabetes Care* 37 (1): 9–16.

Neal, Radford M. 2012. "MCMC Using Hamiltonian Dynamics. ArXiv E-Prints, Page." *arXiv Preprint arXiv:1206.1901*.

Neale, Michael C, Michael D Hunter, Joshua N Pritikin, Mahsa Zahery, Timothy R Brick, Robert M Kirkpatrick, Ryne Estabrook, Timothy C Bates, Hermine H Maes, and Steven M Boker. 2016. "OpenMx 2.0: Extended Structural Equation and Statistical Modeling." *Psychometrika* 81 (2): 535–49.

Nelson, AW, B Mackinnon, J Traynor, and CC Geddes. 2006. "The Relationship Between Serum Creatinine and Estimated Glomerular Filtration Rate: Implications for Clinical Practice." *Scottish Medical Journal* 51 (4): 5–9.

Ng, Ryan, Kathy Kornas, Rinku Sutradhar, Walter P Wodchis, and Laura C Rosella. 2018. "The Current Application of the Royston-Parmar Model for Prognostic Modeling in Health Research: A Scoping Review." *Diagnostic and Prognostic Research* 2: 1–15.

Orchard, TJ, AM Secrest, RG Miller, and T Costacou. 2010. "In the Absence of Renal Disease, 20 Year Mortality Risk in Type 1 Diabetes Is Comparable to That of the General Population: A Report from the Pittsburgh Epidemiology of Diabetes Complications Study." *Diabetologia* 53 (11): 2312–9.

O'Reilly, Joseph E, Anita Jeyam, Thomas M Caparrotta, Joseph Mellor, Andreas Hohn, Paul M McKeigue, Stuart J McGurnaghan, et al. 2021. "Rising Rates and Widening Socioeconomic Disparities in Diabetic Ketoacidosis in Type 1 Diabetes in Scotland: A Nationwide Retrospective Cohort Observational Study." *Diabetes Care* 44 (9): 2010–7.

Orenes-Pinero, Esteban, Sergio Manzano-Fernandez, Angel Lopez-Cuenca, Francisco Marin, Mariano Valdes, and James L Januzzi. 2013. "β-Trace Protein: From Gfr Marker to Cardiovascular Risk Predictor." *Clinical Journal of the American Society of Nephrology* 8 (5): 873–81.

Ostermann, Marlies, Kianoush Kashani, and Lui G Forni. 2016. "The Two Sides of Creatinine: Both as Bad as Each Other?" *Journal of Thoracic Disease* 8 (7): E628.

Oud, Johan HL, and Marc JMH Delsing. 2010. "Continuous Time Modeling of Panel Data by Means of Sem." In *Longitudinal Research with Latent Variables*, 201–44. Springer.

Pardoe, Iain. 2020. *Applied Regression Modeling*. John Wiley & Sons.

Parr, Harry, Emma Hall, and Nuria Porta. 2022. "Joint Models for Dynamic Prediction in Localised Prostate Cancer: A Literature Review." *BMC Medical Research Methodology* 22 (1): 1–19.

Pencina, Michael J, Ralph B D'Agostino Sr, and Ewout W Steyerberg. 2011. "Extensions of Net Reclassification Improvement Calculations to Measure Usefulness of New Biomarkers."

Statistics in Medicine 30 (1): 11–21.

Petris, Giovanni, Sonia Petrone, and Patrizia Campagnoli. 2009. *Dynamic Linear Models with R*. Springer Science & Business Media.

Peyroteo, Mariana, Inês Augusto Ferreira, Luis Brito Elvas, João Carlos Ferreira, and Luis Velez Lapão. 2021. "Remote Monitoring Systems for Patients with Chronic Diseases in Primary Health Care: Systematic Review." *JMIR mHealth and uHealth* 9 (12): e28285.

Pinheiro, José C, and Douglas M Bates. 1995. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics* 4 (1): 12–35.

Prentice, Ross L. 1982. "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model." *Biometrika* 69 (2): 331–42.

Press, William H, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. 2007. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge university press.

Prigge, Regina, John A McKnight, Sarah H Wild, Aveni Haynes, Timothy W Jones, Elizabeth A Davis, Birgit Rami-Merhar, et al. 2022. "International Comparison of Glycaemic Control in People with Type 1 Diabetes: An Update and Extension." *Diabetic Medicine* 39 (5): e14766.

Raman, Maharajan, Rachel J Middleton, Philip A Kalra, and Darren Green. 2017. "Estimating Renal Function in Old People: An in-Depth Review." *International Urology and Nephrology* 49: 1979–88.

Randers, Else, and Erland J Erlandsen. 1999. "Serum Cystatin c as an Endogenous Marker of the Renal Function–a Review."

Rawshani, Aidin, Araz Rawshani, Stefan Franzén, Björn Eliasson, Ann-Marie Svensson,

Mervete Miftaraj, Darren K McGuire, Naveed Sattar, Annika Rosengren, and Soffia Gudbjörnsdottir. 2017. "Mortality and Cardiovascular Disease in Type 1 and Type 2 Diabetes." *New England Journal of Medicine* 376 (15): 1407–18.

Rich, Jason T, J Gail Neely, Randal C Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W Wang. 2010. "A Practical Guide to Understanding Kaplan-Meier Curves." *Otolaryngology—Head and Neck Surgery* 143 (3): 331–36.

Rizopoulos, D. 2014. "JMbayes: Joint Modeling of Longitudinal and Time-to-Event Data Under a Bayesian Approach." *R Package Version 0.5-3*.

Rizopoulos, Dimitris. 2010. "JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data." *Journal of Statistical Software* 35: 1–33.

———. 2012. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC press.

Rizopoulos, Dimitris, and Pulak Ghosh. 2011. "A Bayesian Semiparametric Multivariate Joint Model for Multiple Longitudinal Outcomes and a Time-to-Event." *Statistics in Medicine* 30 (12): 1366–80.

Rizopoulos, Dimitris, Laura A Hatfield, Bradley P Carlin, and Johanna JM Takkenberg. 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): 1385–97.

Roberts, Gareth O, and Jeffrey S Rosenthal. 2004. "General State Space Markov Chains and Mcmc Algorithms."

Rodriguez, German. 2007. "Lecture Notes on Generalized Linear Models." *URL: Http://Data. Princeton. Edu/Wws509/Notes/C4. Pdf*.

Rudman, Najda, Olga Gornik, and Gordan Lauc. 2019. "Altered N-Glycosylation Profiles as

Potential Biomarkers and Drug Targets in Diabetes." *FEBS Letters* 593 (13): 1598–1615.

Russell, Stuart J, Peter Norvig, and E Davis. 2009. "Upper Saddle River." *Artificial Intelligence: A Modern Approach. 3rd Ed. Prentice Hall: NJ.*

Samra, Manpreet, and Antoine C Abcar. 2012. "False Estimates of Elevated Creatinine." *The Permanente Journal* 16 (2): 51.

Schluchter, Mark D. 1992. "Methods for the Analysis of Informatively Censored Longitudinal Data." *Statistics in Medicine* 11 (14-15): 1861–70.

Schofield, Jonathan, Jan Ho, and Handrean Soran. 2019. "Cardiovascular Risk in Type 1 Diabetes Mellitus." *Diabetes Therapy* 10 (3): 773–89.

Secrest, Aaron M, Dorothy J Becker, Sheryl F Kelsey, Ronald E LaPorte, and Trevor J Orchard. 2010. "Cause-Specific Mortality Trends in a Large Population-Based Cohort with Long-Standing Childhood-Onset Type 1 Diabetes." *Diabetes* 59 (12): 3216–22.

Secrest, Aaron M, Raynard E Washington, and Trevor J Orchard. 2021. "Mortality in Type 1 Diabetes."

Shah, Manasi S, and Michael Brownlee. 2016. "Molecular and Cellular Mechanisms of Cardiovascular Disorders in Diabetes." *Circulation Research* 118 (11): 1808–29.

Shahbaz, Hassan, and Mohit Gupta. 2023. "Creatinine Clearance." *StatPearls*.

Sheikh, Md Tuhin, Joseph G Ibrahim, Jonathan A Gelfond, Wei Sun, and Ming-Hui Chen. 2021. "Joint Modelling of Longitudinal and Survival Data in the Presence of Competing Risks with Applications to Prostate Cancer Data." *Statistical Modelling* 21 (1-2): 72–94.

Sherstinsky, Alex. 2020. "Fundamentals of Recurrent Neural Network (Rnn) and Long Short-Term Memory (Lstm) Network." *Physica D: Nonlinear Phenomena* 404: 132306.

Shlipak, Michael G, Ronit Katz, Bryan Kestenbaum, Linda F Fried, Anne B Newman, David S Siscovick, Lesley Stevens, and Mark J Sarnak. 2009. "Rate of Kidney Function Decline

in Older Adults: A Comparison Using Creatinine and Cystatin c." *American Journal of Nephrology* 30 (3): 171–78.

Shortliffe, Edward H, Edward H Shortliffe, James J Cimino, and James J Cimino. 2014. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer.

Simpson, Keith. 1993. "The Scottish Renal Registry." *Scottish Medical Journal* 38 (4): 107–9.

Soedamah-Muthu, Sabita S, Nish Chaturvedi, Daniel R Witte, Lynda K Stevens, Massimo Porta, and John H Fuller. 2008. "Relationship Between Risk Factors and Mortality in Type 1 Diabetic Patients in Europe: The Eurodiab Prospective Complications Study (Pcs)." *Diabetes Care* 31 (7): 1360–6.

Sperrin, Matthew, Emily Petherick, and Ellena Badrick. 2017. "Informative Observation in Health Data: Association of Past Level and Trend with Time to Next Measurement." *Stud Health Technol Inform* 235: 261–65.

Stadler, M, M Auinger, C Anderwald, T Kastenbauer, R Kramar, C Feinbock, K Irsigler, F Kronenberg, and R Prager. 2006. "Long-Term Mortality and Incidence of Renal Dialysis and Transplantation in Type 1 Diabetes Mellitus." *The Journal of Clinical Endocrinology & Metabolism* 91 (10): 3814–20.

Stevens, David, Deirdre A Lane, Stephanie L Harrison, Gregory YH Lip, and Ruwanthi Kolamunnage-Dona. 2021. "Modelling of Longitudinal Data to Predict Cardiovascular Disease Risk: A Methodological Review." *BMC Medical Research Methodology* 21 (1): 1–24.

Subbiah, Vivek. 2023. "The Next Generation of Evidence-Based Medicine." *Nature Medicine* 29 (1): 49–58.

Suissa, Samy. 2007. "Immortal Time Bias in Observational Studies of Drug Effects." *Pharmacoepidemiology and Drug Safety* 16 (3): 241–49.

Sullivan, Gail M, and Richard Feinn. 2012. "Using Effect Size—or Why the P Value Is Not Enough." *Journal of Graduate Medical Education* 4 (3): 279–82.

Sung, Ji Min, In-Jeong Cho, David Sung, Sunhee Kim, Hyeon Chang Kim, Myeong-Hun Chae, Maryam Kavousi, et al. 2019. "Development and Verification of Prediction Models for Preventing Cardiovascular Diseases." *PloS One* 14 (9): e0222809.

Swedko, Peter J, Heather D Clark, Koushi Paramsothy, and Ayub Akbari. 2003. "Serum Creatinine Is an Inadequate Screening Test for Renal Failure in Elderly Patients." *Archives of Internal Medicine* 163 (3): 356–60.

Sweeting, Michael J, and Simon G Thompson. 2011. "Joint Modelling of Longitudinal and Time-to-Event Data with Application to Predicting Abdominal Aortic Aneurysm Growth and Rupture." *Biometrical Journal* 53 (5): 750–63.

Tangri, Navdeep, Lesley A Inker, Brett Hiebert, Jenna Wong, David Naimark, David Kent, and Andrew S Levey. 2017. "A Dynamic Predictive Model for Progression of Ckd." *American Journal of Kidney Diseases* 69 (4): 514–20.

Tannock, Ian F, and John A et al Hickman. 2016. "Limits to Personalized Cancer Medicine." *N Engl J Med* 375 (13): 1289–94.

Team, NR. 2013. "Of Sw National Records of Scotland." *National Records of Scotland*.

Thiele, Thorvald Nicolai. 1880. *Om Anvendelse Af Mindste Kvadraters Methode I Nogle Tilfaelde, Hvor En Komplikation Af Visse Slags Uensartede Tilfaeldige Fejlkilder Giver Fejlene En" Systematisk" Karakter, Af Tn Thiele*. B. Lunos Kgl. Hof.-Bogtrykkeri.

Toffaletti, John G. 2010. "Improving the Clinical Value of Estimating Glomerular Filtration Rate by Reporting All Values: A Contrarian Viewpoint." *Nephron Clinical Practice* 115 (3): c177–c181.

Transplant Work Group, Kidney Disease: Improving Global Outcomes (KDIGO), and others.

2009. "KDIGO Clinical Practice Guideline for the Care of Kidney Transplant Recipients." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 9: S1–S155.

Tsiatis, Anastasios A, and Marie Davidian. 2004. "Joint Modeling of Longitudinal and Time-to-Event Data: An Overview." *Statistica Sinica*, 809–34.

Tsiatis, Anastasios A, Victor Degruttola, and Michael S Wulfsohn. 1995. "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and Cd4 Counts in Patients with Aids." *Journal of the American Statistical Association* 90 (429): 27–37.

Uno, Hajime, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. 2011. "On the c-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data." *Statistics in Medicine* 30 (10): 1105–17.

Uno, Hajime, Lu Tian, Tianxi Cai, Isaac S Kohane, and LJ Wei. 2013. "A Unified Inference Procedure for a Class of Measures to Assess Improvement in Risk Prediction Systems with Survival Data." *Statistics in Medicine* 32 (14): 2430–42.

Vachon, Celine M, Carla H Van Gils, Thomas A Sellers, Karthik Ghosh, Sandhya Pruthi, Kathleen R Brandt, and V Shane Pankratz. 2007. "Mammographic Density, Breast Cancer Risk and Risk Prediction." *Breast Cancer Research* 9 (6): 1–9.

Van Calster, Ben, Ewout W Steyerberg, Ralph B D'Agostino Sr, and Michael J Pencina. 2014. "Sensitivity and Specificity Can Change in Opposite Directions When New Predictive Markers Are Added to Risk Models." *Medical Decision Making* 34 (4): 513–22.

Vaupel, James W, Kenneth G Manton, and Eric Stallard. 1979. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography* 16 (3): 439–54.

Vistisen, Dorte, Gregers Stig Andersen, Christian Stevns Hansen, Adam Hulman, Jan Erik Henriksen, Henning Bech-Nielsen, and Marit Eika Jørgensen. 2016. "Prediction of First

Cardiovascular Disease Event in Type 1 Diabetes Mellitus: The Steno Type 1 Risk Engine." *Circulation* 133 (11): 1058–66.

Voelkle, Manuel C, Johan HL Oud, Eldad Davidov, and Peter Schmidt. 2012. "An Sem Approach to Continuous Time Modeling of Panel Data: Relating Authoritarianism and Anomia." *Psychological Methods* 17 (2): 176.

Walker, Jeremy, Helen Colhoun, Shona Livingstone, Rory McCrimmon, John Petrie, Naveed Sattar, and Sarah Wild. 2018. "Type 2 Diabetes, Socioeconomic Status and Life Expectancy in Scotland (2012–2014): A Population-Based Observational Study." *Diabetologia* 61 (1): 108–16.

Webster, Angela C, Evi V Nagler, Rachael L Morton, and Philip Masson. 2017. "Chronic Kidney Disease." *The Lancet* 389 (10075): 1238–52.

Wei, Lee-Jen. 1992. "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis." *Statistics in Medicine* 11 (14-15): 1871–9.

White, Christine A, Sassan Ghazan-Shahi, and Michael A Adams. 2015. "β-Trace Protein: A Marker of Gfr and Other Biological Pathways." *American Journal of Kidney Diseases* 65 (1): 131–46.

Whittemore, Alice S, and Joseph B Keller. 1986. "Survival Estimation Using Splines." *Biometrics*, 495–506.

"Who Needs the Cox Model Anyway?" 2018. https://bendixcarstensen.com/WntCma.pdf.

Wild, Sarah, Colin Fischbacher, John McKnight, and Scottish Diabetes Research Network Epidemiology Group). 2016. "Using Large Diabetes Databases for Research." *Journal of Diabetes Science and Technology* 10 (5): 1073–8.

Williams, Katherine V, Dorothy J Becker, Trevor J Orchard, and Tina Costacou. 2019. "Persistent c-Peptide Levels and Microvascular Complications in Childhood Onset Type 1

Diabetes of Long Duration." *Journal of Diabetes and Its Complications* 33 (9): 657–61.

Wilson, David B. 2021. "The Relative Incident Rate Ratio Effect Size for Count-Based Impact Evaluations: When an Odds Ratio Is Not an Odds Ratio." *Journal of Quantitative Criminology*, 1–19.

Wolfinger, Russell D. 1999. "Fitting Nonlinear Mixed Models with the New Nlmixed Procedure." In *Proceedings of the 24th Annual Sas Users Group International Conference (Sugi 24)*, 278–84.

Wu, Margaret C, and Raymond J Carroll. 1988. "Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process." *Biometrics*, 175–88.

Wulfsohn, Michael S, and Anastasios A Tsiatis. 1997. "A Joint Model for Survival and Longitudinal Data Measured with Error." *Biometrics*, 330–39.

Yadav, Kabir, and Roger J Lewis. 2021. "Immortal Time Bias in Observational Studies." *Jama* 325 (7): 686–87.

Yang, Songchun, Yuting Han, Canqing Yu, Yu Guo, Yuanjie Pang, Dianjianyi Sun, Pei Pei, et al. 2022. "Development of a Model to Predict 10-Year Risk of Ischemic and Hemorrhagic Stroke and Ischemic Heart Disease Using the China Kadoorie Biobank." *Neurology* 98 (23): e2307–e2317.

Yang, Yujie, Ye Li, Runge Chen, Jing Zheng, Yunpeng Cai, and Giancarlo Fortino. 2021. "Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data." *Big Data Research* 25: 100234.

Ye, Wen, Xihong Lin, and Jeremy MG Taylor. 2008. "Semiparametric Modeling of Longitudinal Measurements and Time-to-Event Data–a Two-Stage Regression Calibration Approach." *Biometrics* 64 (4): 1238–46.

Yu, Marc Gregory, Hillary A Keenan, Hetal S Shah, Scott G Frodsham, David Pober, Zhiheng He, Emily A Wolfson, et al. 2019. "Residual β Cell Function and Monogenic Variants in Long-Duration Type 1 Diabetes Patients." *The Journal of Clinical Investigation* 129 (8): 3252–63.

Zarulli, Virginia. 2016. "Unobserved Heterogeneity of Frailty in the Analysis of Socioeconomic Differences in Health and Mortality." *European Journal of Population* 32: 55–72.

Zhang, Zhongheng, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E Pieterse, and Catharina GM Groothuis-Oudshoorn. 2018. "Time-Varying Covariates and Coefficients in Cox Regression Models." *Annals of Translational Medicine* 6 (7).

Zhou, Amy, Maya Sabatello, Gil Eyal, Sandra Soo-Jin Lee, John W Rowe, Deborah F Stiles, Ashley Swanson, and Paul S Appelbaum. 2021. "Is Precision Medicine Relevant in the Age of Covid-19?" *Genetics in Medicine* 23 (6): 999–1000.

Appendix

ICD & OPCS codes used for defining CVD

Prior CVD ICD codes used to exclude people who have had an event in a period of 10 years prior to the study start date ICD10: I20, I21, I22, I23, I24, I25, I61, I63, I64, G45, I70.2, I70.8, I70.9, I73.9, E10.5, E11.5, E13.5, E14.5 ICD9: 410, 411, 412, 413, 414, 431, 433, 434, 435, 443.9, 440.2, 440.3, 440.4, 440.9, 250.7 Study CVD codes used to determine an event during the study period ICD10: I20, I21, I22, I23, I24, I25, I61, I63, I64, G45, I70.2, I70.8, I70.9, I73.9, E10.5, E11.5, E13.5, E14.5 ICD9: 410, 411, 412, 413, 414, 431, 433, 434, 435, 443.9, 440.2, 440.3, 440.4, 440.9, 250.7 OPCS4: K40, K41, K42, K43, K44, K45, K46, K48, K49, K50, K75, L29, L30, L31.0, L31.1, L31.3, L31.4, L31.5, L31.6, L31.7, L31.8, L31.9, L33, L34, L35, X09.1, X09.2, X09.3, X09.4, X09.5 OPCS3: 304, 082.8, 871.2, 873 CHD related death during the study period ICD10: I20, I21, I22, I23, I24, I25, I46 ICD9: 427.5, 410, 412, 413, 414

Intermediate outputs

Dynamic prediction of time to event given longitudinal observations: Evaluation of forward prediction using a joint model of longitudinal and time to event data. Evaluation of a two-stage approach based on Kalman filter updates and a time-split counting process as a scalable alternative to joint modelling.

Ioanna Thoma et al.

2021-05-17

Background

We present a real world data application to study the association between longitudinal biomarkers, namely HbA_{1c} and eGFR, and time to CVD event in individuals with type 1 diabetes.

We first consider a multivariate shared parameter joint model. For each biomarker a linear mixed model is specified with an individual-specific intercept and slope. We then estimate the hazard of the event using a parametric proportional hazards regression model. The probability of individual i still being event-free at time t is often referred to as the 'survival probability'. The log hazard of death at time t is modelled as dependent on the current values of the biomarkers.

The joint modelling approach specifies the likelihood of event by computing the hazard rate at a set of time points chosen to give an approximation to the area under the curve: a procedure known as quadrature. As the number of biomarkers increases, the number of points required for quadrature scales exponentially and model fitting becomes computationally intractable. This limitation has hampered the application of multivariate joint models with real data in practice.

Another approach that enables the fitting of such models with more realistic computational times* is the use of a multivariate mixed effects state space model. In this case the values of the longitudinal variables are modelled by the latent state of the state space model and then these are used as predictors in a Poisson regression model for the survival probability. Details about the model formulation are given below.

State space models are powerful in modelling dynamic processes due to their flexibility. Population effects and subject random deviations of any variable can be described by different stochastic processes. However, the expensive computational cost is a major hindrance to the application of the mixed effects state space model to healthcare data with large numbers of individuals.

Lastly, we compare these continuous-time methods to modelling time to event as a counting process over many short person-time intervals. The major assumption here is that if the intervals are short, the hazard rate within intervals can be treated as constant and the counts of events can be modelled by Poisson regression. With many short interval lengths, the approximation of a Poisson time-splitting model to a continuous-time survival model can be made arbitrarily close.

Study population

The study started on 1/1/2008 and ended on 1/1/2018. The study population includes 26327 individuals with HbA_{1c} data, where each has 22.09 observations on average. 97.3% of them (25616 individuals) also have eGFR data, with each having 26.21 observations on average.

To reduce computational complexity we took the observations of eGFR measured on the same date as HbA_{1c}. This resulted in a reduced number of participants and observation totals per individual. Consequently, 24779 individuals are included in the dataset and the average number of observations per individual now adds up to 14.13 measurements. Over 50% of the patients in the study have had their HbA_{1c} and eGFR measured at least once per year.

Baseline characteristics

The cohort was initially consisted of 14124 males and 10655 females. The average age was 36.09 and the average diabetes duration was 15.08 years. The number of CVD incidents occurred in the study period was 2235 (9.02%).

By scaling down the original dataset we make the model fitting process faster. We select 1000 controls and 100 cases such that the ratio of events is maintained. We select the subjects who were not censored within 5 years from their entry. This total corresponds to 795 individuals. We then drop any individuals who are followed up but have no observation within these first 5 years and this leaves us with 780 subjects. Then we censor half of the individuals who are followed up for at least 5 year, so that these 398 individuals, followed from 5 years till the end of study, comprise the test set.

The aim here is emulate a realistic situation in which we are trying to make forward predictions for individuals who have been followed for several years based on a model trained on the entire population. Therefore, we need to include in the training set some people who are followed to the end of the study period. This allows the program to fit a spline model for the baseline hazard rate that covers the entire study period.

Methods and results

We demonstrate the continuous-time approach to joint modelling by using the function stan_jm() from the `rstanarm` package to fit a multivariate joint model to the two longitudinal biomarkers and time to CVD event.

For each biomarker a linear mixed model is specified with an individual-specific intercept and slope. The event model includes `gender`, `age`, `diabetes duration` and `median HbA1c` as baseline covariates. The log hazard of death at time t is modelled as dependent on the current values of the biomarkers.

Table 1: Runtime in minutes of fitting a stan_jm() model with 1000 iterations

	warmup	sample	total
chain:1	90.2	197.6	287.8
chain:2	132.5	13.7	146.2
chain:3	82.2	17.6	99.8
chain:4	158.5	176.2	334.7

Table 2: Continuous-time bivariate joint model fitted with stan_jm()

	mean	sd	n_eff
Event|(Intercept)	-6.09	1.13	885
Event|gender	-0.42	0.21	1202
Event|entryage	0.04	0.01	1222
Event|diabetesduration	0.01	0.01	1492
Event|entryhba1cavg	0.00	0.01	1094
Event|b-splines-coef1	-0.55	0.52	680
Event|b-splines-coef2	-0.76	0.54	495
Event|b-splines-coef3	-0.75	0.73	375
Event|b-splines-coef4	-2.03	0.86	357
Event|b-splines-coef5	0.46	0.74	446
Event|b-splines-coef6	-1.40	1.02	861
Assoc|Long1|etavalue	0.04	0.01	668
Assoc|Long2|etavalue	-0.02	0.01	1029

```
## Warning in bs(x = c(`42` = 9.99041752224504, `42` = 0.042682205037857, `42`
## = 0.254216602137268, : some 'x' values beyond boundary knots may cause ill-
## conditioned bases

## Warning in bs(x = c(`45` = 9.99041752224504, `45` = 0.042682205037857, `45`
## = 0.254216602137268, : some 'x' values beyond boundary knots may cause ill-
## conditioned bases

## Warning in bs(x = c(`70` = 9.99041752224504, `70` = 0.042682205037857, `70`
## = 0.254216602137268, : some 'x' values beyond boundary knots may cause ill-
## conditioned bases
```

Comparison of state space models fitted to longitudinal data

We first examine the fit of alternative models based on a multilevel bivariate Gaussian state space model, using the R package `ctsem`.

To specify a model with drift and diffusion but no slopes (i.e. no LMM), we comment out the lines `DRIFT=0`, `DIFFUSION=0` to allow these effects to be learned, and comment out the line `CINT=c('slope1','slope2')` so that the slope parameters are set to their default values of zero.

For a linear mixed model that allows diffusion but no drift, we uncomment the lines `DRIFT=0` and `CINT=c('slope1','slope2')`. For a linear mixed model that allows drift but no diffusion, we uncomment the lines `DIFFUSION=0` and `CINT=c('slope1','slope2')`. For a linear mixed model that allows drift and diffusion, we uncomment the lines `DIFFUSION=0`, `DRIFT=0` and `CINT=c('slope1','slope2')`.

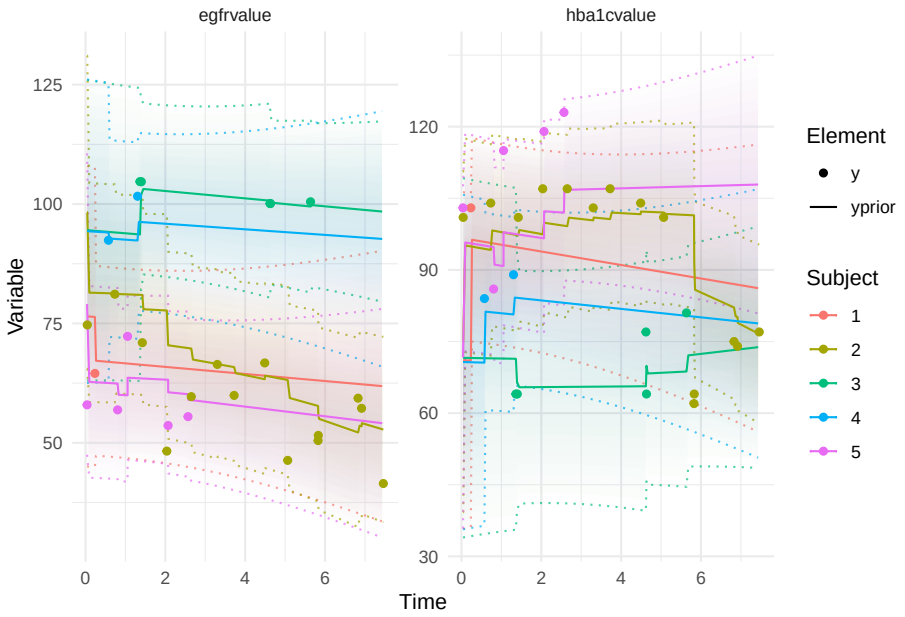


Figure 1: Linear mixed model with no diffusion or drift: updates to trajectory of five individuals by Kalman filter

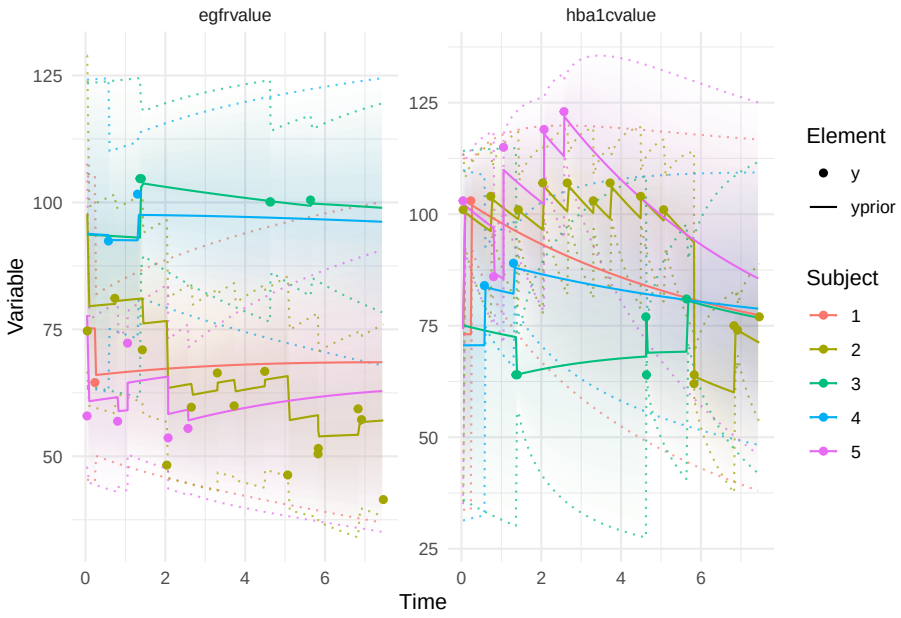


Figure 2: Model with diffusion and drift only: updates to trajectory of five individuals by Kalman filter

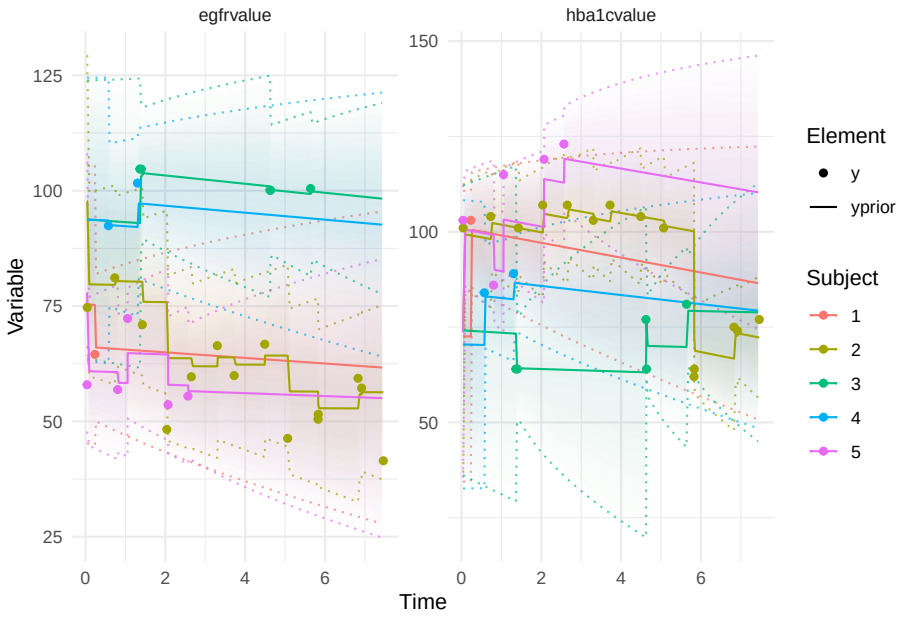


Figure 3: Linear mixed model with diffusion but no drift: updates to trajectory of five individuals by Kalman filter

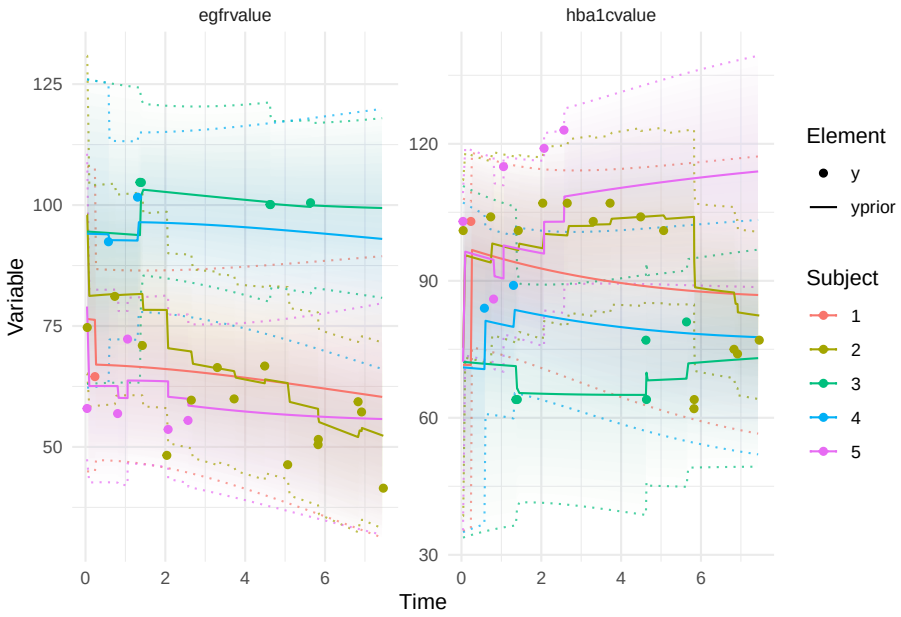


Figure 4: Linear mixed model with drift but no diffusion: updates to trajectory of five individuals by Kalman filter

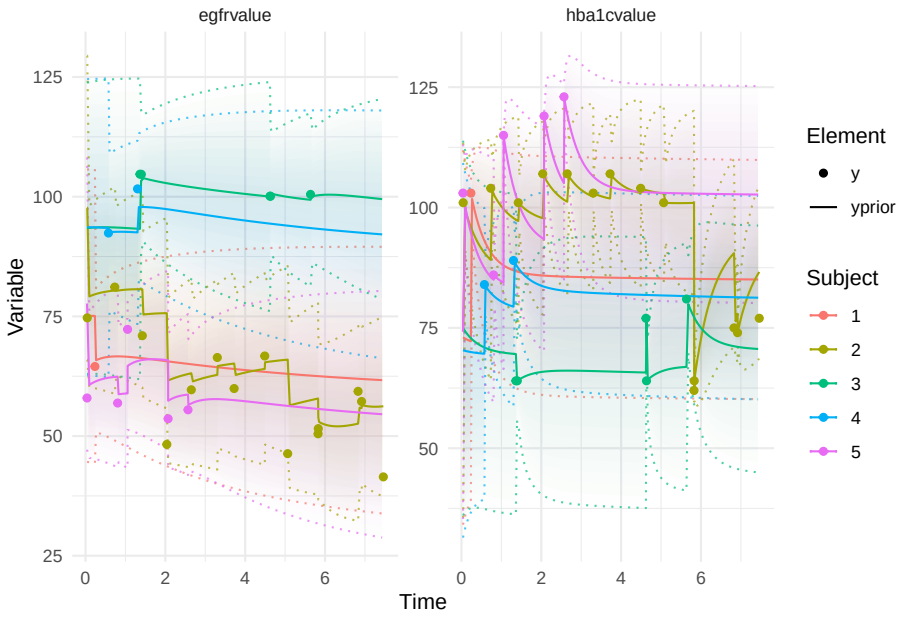


Figure 5: Linear mixed model with drift and diffusion: updates to trajectory of five individuals by Kalman filter

Comparison of state space models when used to impute biomarker values for time-splitting model of events

The imputed values are used to train a Poisson regression model on the training dataset, in which follow-up of those individuals who appear in the test dataset is censored at five years.

Table 3: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with diffusion and drift

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.41	1.10	-6.73	0.000
gender	-0.31	0.21	-1.51	0.131
entryage	0.06	0.01	6.49	0.000
diabetesduration	0.01	0.01	1.55	0.122
entryhba1cavg	0.02	0.01	4.17	0.000
hba1cvalue	0.02	0.01	2.17	0.030
egfrvalue	-0.03	0.00	-7.83	0.000
splines::bs(tstart, df = 6)1	2.22	1.09	2.04	0.042
splines::bs(tstart, df = 6)2	-1.50	0.78	-1.92	0.055
splines::bs(tstart, df = 6)3	2.25	0.90	2.51	0.012
splines::bs(tstart, df = 6)4	-1.62	1.01	-1.60	0.109
splines::bs(tstart, df = 6)5	1.80	1.16	1.55	0.121
splines::bs(tstart, df = 6)6	0.58	1.21	0.48	0.631

Comparison of fit to training data

Deviance	numparams	AIC	Model
0.0	34	0.0	LMM
-2831.6	47	-2805.6	Drift and diffusion, no LMM
-2374.5	46	-2350.5	LMM with diffusion
-148.3	50	-116.3	LMM with drift
-3931.1	62	-3875.1	LMM with drift and diffusion

Of the five models, the LMM with drift and diffusion has the lowest deviance and the lowest AIC.

Comparison of predictive performance on test data, using joint model as benchmark

Table 5: Comparison of predictive performance on test dataset

Model	Observed	Predicted	Person_years	C_statistic
lmm.fit	31	22.6	952.6	0.754
nolmm.fit	31	23.4	952.6	0.751
lmmdiff.fit	31	22.1	952.6	0.751
lmmdrift.fit	31	22.2	952.6	0.753
lmmdriftdiff.fit	31	23.0	952.6	0.752
stanjm	12	22.9	1337.5	0.688

Even with only a single imputation, the predictive performance obtained by the time-splitting approach with biomarker trajectories predicted at the start of follow-up beyond 5 years is as good as that obtained with the fully Bayesian continuous-time joint model. In both approaches, the longitudinal submodel is a linear mixed model with random intercepts and slopes.

Studying time to renal disease in people with type 1 diabetes

Ioanna Thoma

03 September 2021

Abstract

make dynamic prediction of time-to-event (renal replacement therapy in people with type 1 diabetes) given longitudinal data (eGFR, HbA1c)

Contents

1 Background

Long-term complications of type 1 diabetes (T1D) include nephropathy, retinopathy, neuropathy and vascular disease. Learning to predict these complications is important for clinical practice and decision making. In addition, for the management of chronic diseases such as diabetes where biomarkers are measured repeatedly over the course of the disease, clinicians need to be able to make dynamic risk predictions that are updated as new biomarker observations arrive. This is a key requirement for precision medicine to realise its potential.

It is imperative to understand why some individuals with T1D face a number of long-term, potentially life-threatening complications, while others may relish an uncomplicated long lifespan. Despite large improvements in the management of glucose levels in the last years, the mortality rate in patients with T1D is still high (Mameli et al. 2015). After a long disease duration, most of the excess mortality in people with T1D is considered to result largely from cardiovascular events (Deckert, Poulsen, and Larsen 1978).

Interestingly, during the first 20 years of T1D, most of the excess mortality is attributed to renal failure (Dorman et al. 1984). Diabetic nephropathy causing end-stage renal disease (ESRD) which may result in renal replacement therapy (RRT) contributes in a major way to increased mortality (Orchard et al. 2010) (Stadler et al. 2006). Studies provide evidence that life expectancy in haemodialysis patients is reduced fourfold on average versus healthy age-matched individuals (Cheema 2008).

In the last decade, new technologies have been emerged, e.g. insulin pumps, glucose sensors, etc that provide cause for optimism that late T1D-related complications may be prevented or at least be delayed. In addition, developing drugs to prevent or reverse impaired renal function is an important goal. However, there is wide variation in the rate of renal function decline among those with T1D with some people being much more susceptible than others.

The aim of the current work is to investigate incidence rates and time-to-renal disease using electronic healthcare record linkage of a Scottish T1D registry, and identify biomarkers that improve prediction of renal outcomes on top of clinical records. A key question is how longitudinal eGFR and HbA1c data can contribute usefully to prediction of progression of renal disease in individuals with T1D.

2 Methods

2.1 Characteristics of study population

This analysis was conducted using information on patients who are registered in SCI-Diabetes (Scottish Care Information - Diabetes) to provide their personal data for research purposes. This network tracks real-time clinical information on all 300,000 people with type 1 and type 2 diabetes in Scotland. In particular, the current analysis was conducted on 29126 individuals with type 1 diabetes and the study period starts on 1/1/2008 and ends on 1/1/2018.

The renal replacement therapy prevalence in the dataset is 466 events, 0.016%. Within the study period, each subject has 9.96 HbA_{1c} and eGFR observations on average. For greater clarity in their modelling, we have discarded unmatched biomarkers measurements and used only eGFR measurements that were taken on the same date as HbA1c.

The total contribution of the participants in this study is 253955 person-years. Person-time is the sum of total time contributed by all subjects. Furthermore, the dataset contains 16345 male and 12781 female participants aged 18 and above, with average entry age and diabetes duration 39.78 and 15.45 years respectively. The mean follow-up time is 6.9 years. During the first 5 years of follow-up, there were 200 RRT occurrences and 180258 of the 294043 biomarker observations. Biomarker data and covariates gender, baselineage, baselineduration and currentage are scaled and centred for the statistical analysis.

To speed up the training process, we selected 2000 subjects, maintaining the rate of events, 32 of whom had an RRT occurrence. To construct a training subset to be used for dynamic predictions, we censor at 5 years half the individuals who were still being followed up at 5 years (1652 people).

We evaluated the predictive performance of the biomarker models on withdrawn data using 2-fold cross-validation, and compared models including longitudinal biomarker information to baseline models that only contain clinical covariates. The first fold includes 2000 individuals where 826 had their biomarker data censored, and the second fold includes 2000 individuals of whom 826 are censored after Time 5.

The aim is emulate a realistic situation in which we are trying to make forward predictions for individuals who have been followed up for several years, based on a model trained on the entire population. Therefore, we need to include in the training set some people who are followed to the end of the study period. This allows the program to fit a spline model for the baseline hazard rate that covers the entire study period.

3 Dynamic models

We apply a continuous-time dynamic modelling approach to investigate the association of the biomarker trajectories and time-to-RRT. The growing popularity of real longitudinal data has fuelled the interest in continuous-time models because they are inherently well suited to handling unequally spaced measurement occasions and individualised assessment designs (Hecht and Zitzmann 2020). Furthermore, continuous-time models enable comparisons of studies with different time intervals between observations.

A number of attractive software solutions for estimating continuous-time models have been introduced, such as the R package `ctsem`, which includes both a frequentist (Driver, Oud, and Voelkle 2017) and a Bayesian estimation module (Driver and Voelkle 2018). Whereas frequentist estimation in `ctsem` seems appropriately fast, model run times in the Bayesian `ctsem` are rather high when using real life data.

Continuous-time models involve a two-level data structure (repeatedly measured values nested within persons). We perform a Bayesian analysis as we wish to benefit from the ability to include previous knowledge and the potential to estimate otherwise intractable models e.g., Van De Schoot et al. (2017). Therefore, we chose to use the Bayesian software STAN for its flexibility and stability.

3.1 Rationale for time-splitting approach

In survival analysis, it is common to model occurrence of event as a binary variable over many person-time intervals. The event cannot occur in interval i, if an event has occurred in any interval prior to i. Thus, the probability of an event occurring in interval i is not independent from whether an event has occurred in interval $i-1$. However, if the intervals are short, the hazard rate within intervals can be treated as constant and the counts of events can be modelled by Poisson regression, using biomarker values at the start of each interval. With many short interval lengths, the approximation of a Poisson time-splitting model to a continuous-time model can be made arbitrarily close.

3.2 Statistical analysis

The most challenging part of the model we try to estimate is individual level parameters (random effects). We fit a range of dynamic bivariate continuous-time models to the data, including parameters to capture continuous-time auto-effect and diffusion variance, person-specific random effects, their discrete-time counterparts and an error term for the process.

To stabilise the estimation, we express the discrete-time process intercepts in terms of the long-range process means, an approach also employed by Driver, Oud, and Voelkle (2017). All models were interfaced via the R package `ctsem` 3.5.3 running on R version 3.6.0 (Ripley and others 2001). For more explanations, examples, and illustrations of this (and other) continuous-time models see Hecht, Hardt, et al. (2019), Hecht and Voelkle (2019), Driver et al. (2017), Driver and Voelkle (2018), and Voelkle, Oud, Davidov, and Schmidt (2012). TODO: Add references

It is widely accepted that some information will be lost when using a model to represent a dynamic process. To assess the quality of the continuous-time models fitted to given sets of data, we use the log likelihood, number of parameters and Akaike information criterion (AIC) of each model (for each fold) as means for model selection. The Akaike Information Criterion (AIC) is defined as $2k - 2L_{train}$ where k is the effective number of parameters and L_{train} is the log-likelihood given the training data. Of the six `ctsem` models, the linear mixed model (LMM) which includes drift and diffusion components has the lowest deviance and the lowest AIC. Since a LMM model with drift and diffusion specifies appropriately the longitudinal trajectories of the biomarkers, it can be employed for obtaining imputations of the biomarkers at the beginning of each interval, in order to implement the time-splitting approach.

Increasing the size of the already big vector of observations might feel counter-intuitive, however so-called message-passing algorithms can be very effective at imputing latent biomarker values in between existing measurement occasions. The Kalman filter is such an algorithm that makes a forward pass through the data to compute the state probability distribution at each time point, conditional on all observations up to that time point. This can be used to impute the predicted values of the latent variables at the start of each time interval (Zhu, DeSantis, and Luo 2018).

We generate imputations from each training fold within each `ctsem` model. The imputed values are then used to train a Poisson regression model on the training dataset, in which follow-up of those individuals who appear in the test dataset is censored at 5 years. For this purpose the dataset is reformatted as person-time intervals of fixed length (one year long).

Table 1: Runtime in minutes of fitting a stan_jm() model with 1000 iterations

	warmup	sample	total
chain:1	459.4	71.6	531.0
chain:2	510.8	66.7	577.5
chain:3	255.1	81.9	337.0
chain:4	247.0	471.1	718.1

Table 2: Continuous-time bivariate joint model fitted with stan_jm() - Fold 1

	mean	sd	n_eff
Event|(Intercept)	-5.89	0.11	2454
Event|gender	-0.59	0.28	3192
Event|baselineage	-0.68	0.32	3071
Event|baselineduration	0.05	0.29	3602
Event|b-splines-coef1	-12.31	5.21	1834
Event|b-splines-coef2	-4.26	2.72	1897
Event|b-splines-coef3	-5.23	2.01	2535
Event|b-splines-coef4	-5.13	1.94	1539
Event|b-splines-coef5	-3.14	2.16	2561
Event|b-splines-coef6	-8.75	3.03	1802
Assoc|Long1|etavalue	0.54	0.32	2265
Assoc|Long2|etavalue	-3.25	0.44	1808

Table 3: Runtime in minutes of fitting a stan_jm() model with 1000 iterations

	warmup	sample	total
chain:1	307.5	65.9	373.4
chain:2	215.0	90.1	305.1
chain:3	217.3	90.3	307.6
chain:4	243.2	84.2	327.4

Table 4: Continuous-time bivariate joint model fitted with stan_jm() - Fold 2

	mean	sd	n_eff
Event|(Intercept)	-5.98	0.13	1443
Event|gender	-0.29	0.28	3406
Event|baselineage	-0.77	0.31	3753
Event|baselineduration	-0.21	0.39	4412
Event|b-splines-coef1	-16.44	6.91	2053
Event|b-splines-coef2	-3.46	3.14	2627
Event|b-splines-coef3	-8.13	2.49	1410
Event|b-splines-coef4	-4.02	2.20	1676
Event|b-splines-coef5	-10.60	3.02	1704
Event|b-splines-coef6	-5.41	1.95	1312
Assoc|Long1|etavalue	0.29	0.36	2826
Assoc|Long2|etavalue	-3.89	0.59	1181

Table 5: Model fits

model	loglik	npars	aic
model.lmm	-24243.0	34	48554.0
model.lmm	-24546.6	34	49161.3
model.lmm1	-25789.7	26	51631.4
model.lmm1	-26026.8	26	52105.7
model.nolmm	-21651.4	47	43396.7
model.nolmm	-21732.8	47	43559.6

model	loglik	npars	aic
model.lmmdiff	-22316.6	46	44725.1
model.lmmdiff	-22396.5	46	44885.0
model.lmmdrift	-24101.2	50	48302.4
model.lmmdrift	-24395.9	50	48891.8
model.lmmdriftdiff	-20738.1	62	41600.3
model.lmmdriftdiff	-20803.0	62	41730.1

Table 6: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with diffusion and drift - Fold 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.66	1.40	-9.06	0.000
tstart	0.19	0.08	2.37	0.018
gender	-0.43	0.24	-1.79	0.074
baselineage	-0.78	0.30	-2.62	0.009
baselineduration	0.04	0.27	0.14	0.891
hba1c	0.33	0.22	1.45	0.147
egfr	-3.73	0.48	-7.79	0.000

Table 7: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with diffusion and drift - Fold 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.65	1.54	-8.23	0.000
tstart	0.14	0.09	1.58	0.114
gender	-0.01	0.24	-0.02	0.981
baselineage	-0.86	0.26	-3.26	0.001
baselineduration	-0.14	0.35	-0.39	0.696
hba1c	0.17	0.25	0.67	0.505
egfr	-3.82	0.53	-7.15	0.000

Table 8: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with drift - Fold 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.97	1.13	-9.73	0.000
tstart	0.10	0.09	1.11	0.267
gender	-0.48	0.24	-1.97	0.049
baselineage	-0.74	0.30	-2.45	0.014
baselineduration	-0.04	0.29	-0.15	0.879
hba1c	0.60	0.26	2.33	0.020
egfr	-3.22	0.41	-7.92	0.000

Table 9: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with drift - Fold 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.53	1.07	-9.83	0.000
tstart	0.00	0.10	-0.03	0.973
gender	0.04	0.23	0.18	0.857
baselineage	-0.85	0.27	-3.13	0.002
baselineduration	-0.28	0.34	-0.84	0.402
hba1c	0.47	0.29	1.63	0.104
egfr	-3.27	0.41	-8.04	0.000

Table 10: Time-splitting Poisson model with biomarker values imputed by last observation carried forward

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-27.28	128533.80	0	1
gender	0.03	59189.04	0	1
baselineage	-0.01	87790.12	0	1
baselineduration	-0.05	105489.85	0	1
hba1c	0.01	40356.43	0	1
egfr	-0.03	87875.18	0	1

4 Evaluation of predictive performance

To predict event status in the test dataset, we use the model fitted to the training dataset, and compute the probability of an event in each person-time interval from 5 years to the end of study.

Table 11: Comparison of predictive performance on test dataset

Model	Observed	Predicted	Person-years	Log score	C-statistic
model.lmm	22	27.37921	8752	-145.4463	0.8970478
model.lmm1	22	23.24346	8752	-144.8489	0.8511455
model.nolmm	22	7.971904	8752	-138.611	0.893804
model.lmmdiff	22	9.453759	8752	-132.6318	0.8966937
model.lmmdrift	22	27.0383	8752	-139.9215	0.9054931
model.lmmdriftdiff	22	7.027217	8752	-157.2087	0.8683484
stanjm	22	15.215	7298.349	NaN	0.8833518

5 References

Cheema, Birinder Singh Bobby. 2008. "Tackling the Survival Issue in End-Stage Renal Disease: Time to Get Physical on Haemodialysis." *Nephrology* 13 (7). Wiley Online Library: 560–69.

Deckert, T, JE Poulsen, and M Larsen. 1978. "Prognosis of Diabetics with Diabetes Onset Before the Age of Thirtyone." *Diabetologia* 14 (6). Springer: 371–77.

Dorman, JS, RE Laporte, LH Kuller, KJ Cruickshanks, TJ Orchard, DK Wagener, DJ Becker, DE Cavender, and AL Drash. 1984. "The Pittsburgh Insulin-Dependent Diabetes Mellitus (Iddm) Morbidity and Mortality Study:

Mortality Results." *Diabetes* 33 (3). Am Diabetes Assoc: 271–76.

Driver, Charles C, and Manuel C Voelkle. 2018. "Hierarchical Bayesian Continuous Time Dynamic Modeling." *Psychological Methods* 23 (4). American Psychological Association: 774.

Driver, Charles C, Johan HL Oud, and Manuel C Voelkle. 2017. "Continuous Time Structural Equation Modeling with R Package Ctsem." *Journal of Statistical Software* 77 (5).

Hecht, Martin, and Steffen Zitzmann. 2020. "A Computationally More Efficient Bayesian Approach for Estimating Continuous-Time Models." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (6). Taylor & Francis: 829–40.

Mameli, Chiara, Sara Mazzantini, Moufida Ben Nasr, Paolo Fiorina, Andrea E Scaramuzza, and Gian Vincenzo Zuccotti. 2015. "Explaining the Increased Mortality in Type 1 Diabetes." *World Journal of Diabetes* 6 (7). Baishideng Publishing Group Inc: 889.

Orchard, TJ, AM Secrest, RG Miller, and T Costacou. 2010. "In the Absence of Renal Disease, 20 Year Mortality Risk in Type 1 Diabetes Is Comparable to That of the General Population: A Report from the Pittsburgh Epidemiology of Diabetes Complications Study." *Diabetologia* 53 (11). Springer: 2312–9.

Ripley, Brian D, and others. 2001. "The R Project in Statistical Computing." *MSOR Connections. The Newsletter of the LTSN Maths, Stats & OR Network* 1 (1). Citeseer: 23–25.

Stadler, M, M Auinger, C Anderwald, T Kastenbauer, R Kramar, C Feinbock, K Irsigler, F Kronenberg, and R Prager. 2006. "Long-Term Mortality and Incidence of Renal Dialysis and Transplantation in Type 1 Diabetes Mellitus." *The Journal of Clinical Endocrinology & Metabolism* 91 (10). Oxford University Press: 3814–20.

Van De Schoot, Rens, Sonja D Winter, Oisín Ryan, Mariëlle Zondervan-Zwijnenburg, and Sarah Depaoli. 2017. "A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years." *Psychological Methods* 22 (2). American Psychological Association: 217.

Zhu, Huirong, Stacia M DeSantis, and Sheng Luo. 2018. "Joint Modeling of Longitudinal Zero-Inflated Count and Time-to-Event Data: A Bayesian Perspective." *Statistical Methods in Medical Research* 27 (4). SAGE Publications Sage UK: London, England: 1258–70.

Studying time to renal disease in people with type 1 diabetes

Ioanna Thoma

24 September 2021

Abstract

make dynamic prediction of time-to-event (renal replacement therapy in people with type 1 diabetes) given longitudinal eGFR data

Contents

1 Methods

1.1 Characteristics of study population

This analysis was conducted using information on patients who are registered in SCI-Diabetes (Scottish Care Information - Diabetes) to provide their personal data for research purposes. This network tracks real-time clinical information on all 300,000 people with type 1 and type 2 diabetes in Scotland. In particular, the current analysis was conducted on 29764 individuals with type 1 diabetes and the study period starts on 1/1/2008 and ends on 1/1/2018.

The renal replacement therapy prevalence in the dataset is 487 events, 0.0164%. Within the study period, each subject has 15.17 eGFR observations on average.

The total contribution of the participants in this study is 256256 person-years. Person-time is the sum of total time contributed by all subjects. Furthermore, the dataset contains 16671 male and 13093 female participants aged 18 and above, with average entry age and diabetes duration 39.78 and 15.41 years respectively. The mean follow-up time is 7.1 years. During the first 5 years of follow-up, there were 205 RRT occurrences and 260052 of the 456867 biomarker observations. Biomarker data and covariates gender, baselineage, baselineduration and currentage are scaled and centred for the statistical analysis. To construct a training subset to be used for dynamic predictions, we censor at 5 years half the individuals who were still being followed up at 5 years (24471 people).

We evaluated the predictive performance of the fitted models on withdrawn data using 10-fold cross-validation, and compared the `ctsem` models to a Poisson model that uses the 'Last Observation Carried Forward' technique (LOCF) to extrapolate the biomarker data at each person-time interval. All 10 folds include full data on 27316 individuals, meaning that about 2448 subjects are censored after the landmark time of 5 years.

The aim is emulate a realistic situation in which we are trying to make forward predictions for individuals who have been followed up for several years, based on a model trained on the entire population. Therefore, we need to include

in the training set some people who are followed to the end of the study period. This allows the program to fit a spline model for the baseline hazard rate that covers the entire study period.

2 Dynamic models

We apply a continuous-time dynamic modelling approach to investigate the association of the biomarker trajectories and time-to-RRT. The growing popularity of real longitudinal data has fuelled the interest in continuous-time models because they are inherently well suited to handling unequally spaced measurement occasions and individualised assessment designs (Hecht and Zitzmann 2020). Furthermore, continuous-time models enable comparisons of studies with different time intervals between observations.

A number of attractive software solutions for estimating continuous-time models have been introduced, such as the R package `ctsem`, which includes both a frequentist (Driver, Oud, and Voelkle 2017) and a Bayesian estimation module (Driver and Voelkle 2018). Whereas frequentist estimation in `ctsem` seems appropriately fast, model run times in the Bayesian `ctsem` are rather high when using real life data.

Continuous-time models involve a two-level data structure (repeatedly measured values nested within persons). We perform a Bayesian analysis as we wish to benefit from the ability to include previous knowledge and the potential to estimate otherwise intractable models e.g., Van De Schoot et al. (2017). Therefore, we chose to use the Bayesian software STAN for its flexibility and stability.

2.1 Rationale for time-splitting approach

In survival analysis, it is common to model occurrence of event as a binary variable over many person-time intervals. The event cannot occur in interval i, if an event has occurred in any interval prior to i. Thus, the probability of an event occurring in interval i is not independent from whether an event has occurred in interval $i-1$. However, if the intervals are short, the hazard rate within intervals can be treated as constant and the counts of events can be modelled by Poisson regression, using biomarker values at the start of each interval. With many short interval lengths, the approximation of a Poisson time-splitting model to a continuous-time model can be made arbitrarily close.

2.2 Statistical analysis

The most challenging part of the model we try to estimate is individual level parameters (random effects). We fit a range of dynamic bivariate continuous-time models to the data, including parameters to capture continuous-time auto-effect and diffusion variance, person-specific random effects, their discrete-time counterparts and an error term for the process.

To stabilise the estimation, we express the discrete-time process intercepts in terms of the long-range process means, an approach also employed by Driver, Oud, and Voelkle (2017). All models were interfaced via the R package `ctsem` 3.5.3 running on R version 3.6.0 (Ripley and others 2001). For more explanations, examples, and illustrations of this (and other) continuous-time models see Hecht, Hardt, et al. (2019), Hecht and Voelkle (2019), Driver et al. (2017), Driver and Voelkle (2018), and Voelkle, Oud, Davidov, and Schmidt (2012). TODO: Add references

It is widely accepted that some information will be lost when using a model to represent a dynamic process. To assess the quality of the continuous-time models fitted to given sets of data, we use the log likelihood, number of parameters and Akaike information criterion (AIC) of each model (for each fold) as means for model selection. The Akaike Information Criterion (AIC) is defined as $2k - 2L_{train}$ where k is the effective number of parameters and L_{train} is the log-likelihood given the training data. Of the 5 `ctsem` models, the linear mixed model (LMM) which includes drift and diffusion components has the lowest deviance and the lowest AIC. Since a LMM model with drift and diffusion specifies appropriately the longitudinal trajectories of the biomarkers, it can be employed for obtaining imputations of the biomarkers at the beginning of each interval, in order to implement the time-splitting approach.

Increasing the size of the already big vector of observations might feel counter-intuitive, however so-called message-passing algorithms can be very effective at imputing latent biomarker values in between existing measurement occasions. The Kalman filter is such an algorithm that makes a forward pass through the data to compute the state

probability distribution at each time point, conditional on all observations up to that time point. This can be used to impute the predicted values of the latent variables at the start of each time interval (Zhu, DeSantis, and Luo 2018).

We generate imputations from each training fold within each `ctsem` model. The imputed values are then used to train a Poisson regression model on the training dataset, in which follow-up of those individuals who appear in the test dataset is censored at 5 years. For this purpose the dataset is reformatted as person-time intervals of fixed length (one year long).

Table 1: Model fits

model	loglik	npars	aic
model.lmm	0.0	15	0.0
model.lmm	35.3	15	-70.6
model.lmm	581.8	15	-1163.6
model.lmm	448.5	15	-896.9
model.lmm	650.0	15	-1300.0
model.lmm	1025.7	15	-2051.4
model.lmm	393.7	15	-787.4
model.lmm	-248.6	15	497.3
model.lmm	238.6	15	-477.1
model.lmm	828.8	15	-1657.6
model.nolmm	25245.6	17	-50487.1
model.nolmm	25197.6	17	-50391.3
model.nolmm	25538.0	17	-51072.0
model.nolmm	25558.1	17	-51112.2
model.nolmm	25737.8	17	-51471.5
model.nolmm	25829.1	17	-51654.1
model.nolmm	25593.4	17	-51182.7
model.nolmm	25011.8	17	-50019.7
model.nolmm	25181.6	17	-50359.3
model.nolmm	25583.2	17	-51162.4
model.lmmdiff	25034.7	19	-50061.5
model.lmmdiff	24979.6	19	-49951.1
model.lmmdiff	25324.7	19	-50641.3
model.lmmdiff	25354.5	19	-50701.1
model.lmmdiff	25537.2	19	-51066.5
model.lmmdiff	25624.4	19	-51240.9
model.lmmdiff	25392.8	19	-50777.7
model.lmmdiff	24786.2	19	-49564.3
model.lmmdiff	24986.9	19	-49965.8
model.lmmdiff	25378.4	19	-50748.8
model.lmmdrift	77.1	19	-146.2
model.lmmdrift	116.0	19	-224.1
model.lmmdrift	660.9	19	-1313.8
model.lmmdrift	530.7	19	-1053.5
model.lmmdrift	739.9	19	-1471.8
model.lmmdrift	1118.3	19	-2228.6
model.lmmdrift	480.0	19	-952.0
model.lmmdrift	-170.7	19	349.4
model.lmmdrift	320.4	19	-632.7
model.lmmdrift	917.9	19	-1827.8
model.lmmdriftdiff	27009.7	23	-54003.3
model.lmmdriftdiff	26916.5	23	-53817.0
model.lmmdriftdiff	27315.1	23	-54614.1
model.lmmdriftdiff	27310.8	23	-54605.7
model.lmmdriftdiff	27468.3	23	-54920.7
model.lmmdriftdiff	27587.8	23	-55159.5

model	loglik	npars	aic
model.lmmdriftdiff	27335.4	23	-54654.8
model.lmmdriftdiff	26770.0	23	-53524.1
model.lmmdriftdiff	26870.8	23	-53725.5
model.lmmdriftdiff	27330.3	23	-54644.6

Table 2: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with diffusion and drift - Fold 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.74	0.52	-32.18	0.000
tstart	0.02	0.02	1.22	0.221
gender	-0.15	0.05	-3.20	0.001
baselineage	-0.82	0.06	-14.36	0.000
baselineduration	-0.30	0.06	-4.91	0.000
egfr	-6.57	0.21	-30.98	0.000

Table 3: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with diffusion and drift - Fold 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.45	0.51	-32.09	0.000
tstart	0.03	0.02	1.56	0.118
gender	-0.15	0.05	-3.24	0.001
baselineage	-0.83	0.06	-14.43	0.000
baselineduration	-0.30	0.06	-4.88	0.000
egfr	-6.45	0.21	-30.83	0.000

Table 4: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with drift - Fold 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.56	0.26	-40.88	0.000
tstart	-0.10	0.02	-5.25	0.000
gender	-0.18	0.05	-3.84	0.000
baselineage	-0.67	0.05	-12.43	0.000
baselineduration	-0.13	0.06	-2.25	0.024
egfr	-4.02	0.11	-36.12	0.000

Table 5: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from linear mixed model with drift - Fold 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.48	0.26	-40.91	0.00
tstart	-0.10	0.02	-5.42	0.00
gender	-0.20	0.05	-4.32	0.00

	Estimate	Std. Error	z value	Pr(>|z|)
baselineage	-0.69	0.05	-12.70	0.00
baselineduration	-0.13	0.06	-2.33	0.02
egfr	-4.00	0.11	-36.08	0.00

Table 6: Time-splitting Poisson model with biomarker values imputed by last observation carried forward

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.97	0.24	-45.27	0.000
gender	-0.09	0.05	-1.95	0.051
baselineage	-0.53	0.05	-10.12	0.000
baselineduration	0.04	0.05	0.81	0.418
egfr	-2.80	0.07	-39.59	0.000
splines::bs(tstart.year, df = 6)1	0.95	0.42	2.28	0.022
splines::bs(tstart.year, df = 6)2	2.20	0.35	6.27	0.000
splines::bs(tstart.year, df = 6)3	2.11	0.33	6.30	0.000
splines::bs(tstart.year, df = 6)4	3.00	0.39	7.69	0.000
splines::bs(tstart.year, df = 6)5	2.37	0.46	5.10	0.000
splines::bs(tstart.year, df = 6)6	2.94	0.82	3.58	0.000

3 Evaluation of predictive performance

To predict event status in the test dataset, we use the model fitted to the training dataset, and compute the probability of an event in each person-time interval from 5 years to the end of study.

Table 7: Comparison of predictive performance on test dataset

Model	Observed	Predicted	Person-years	Log score	C-statistic
model.lmm	293	364.4608	128844	-8208.457	0.89159
model.nolmm	293	58.88546	128844	-2513.691	0.88987
model.lmmdiff	293	85.10565	128844	-2373.343	0.88988
model.lmmdrift	293	361.9266	128844	-8170.763	0.89121
model.lmmdriftdiff	293	52.83163	128844	-2709.261	0.8723

References

Driver, Charles C, and Manuel C Voelkle. 2018. "Hierarchical Bayesian Continuous Time Dynamic Modeling." *Psychological Methods* 23 (4). American Psychological Association: 774.

Driver, Charles C, Johan HL Oud, and Manuel C Voelkle. 2017. "Continuous Time Structural Equation Modeling with R Package Ctsem." *Journal of Statistical Software* 77 (5).

Hecht, Martin, and Steffen Zitzmann. 2020. "A Computationally More Efficient Bayesian Approach for Estimating Continuous-Time Models." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (6). Taylor & Francis: 829–40.

Ripley, Brian D, and others. 2001. "The R Project in Statistical Computing." *MSOR Connections. The Newsletter of the LTSN Maths, Stats & OR Network* 1 (1). Citeseer: 23–25.

Van De Schoot, Rens, Sonja D Winter, Oisín Ryan, Mariëlle Zondervan-Zwijnenburg, and Sarah Depaoli. 2017. "A

Systematic Review of Bayesian Articles in Psychology: The Last 25 Years." *Psychological Methods* 22 (2). American Psychological Association: 217.

Zhu, Huirong, Stacia M DeSantis, and Sheng Luo. 2018. "Joint Modeling of Longitudinal Zero-Inflated Count and Time-to-Event Data: A Bayesian Perspective." *Statistical Methods in Medical Research* 27 (4). SAGE Publications Sage UK: London, England: 1258–70.

RRT analyses

Ioanna Thoma

2021-11-17

Introduction

We model time-to-renal replacement therapy (RRT) via hierarchical dynamical models. Three different RRT analyses are presented herein.

Evaluation of the predictive accuracy of the developed risk prediction models is done by obtaining *forward predictions* for individuals who have been followed up for several years (observational study). This is implemented by censoring part of the training dataset after a landmark time point. For these analyses the landmark point is set to year 5.

RRT outcome definition

The RRT outcome is a composite outcome. Cases are all those individuals who initiated renal replacement therapy within the study period due to loss of kidney function or have died with a mention of renal failure in their death certificate (N17-N19). An event equal to 1 for an individual means any of these three scenarios: (death=1, rrt=1) or (death=0, rrt=1) or (death=1, rrt=0).

An individual might have not been observable at the beginning of the study on 1-1-2008, and have entered the study subsequently. If a subject experiences an event or stops being observable for any other reason, they get censored and do not contribute to the study any further.

The multivariate (two biomarkers) RRT analysis includes 29118 subjects of whom 767 have experienced an RRT event and 394 events occurred after year 5. For this analysis we have used HbA1c and eGFR data measured on the same date. For brevity we call this analysis as analysis A.

We subsequently used only eGFR data to study time-to-RRT on the full dataset as longitudinal eGFR trajectories are more informative of kidney function. The univariate RRT analysis (called analysis B for brevity) includes 29764 subjects of whom 816 have experienced an RRT event and 434 events occurred after year 5.

Last, an extra analysis has been conducted that restricts baseline eGFR to 60 or below, in order to study explicitly those individuals that are experiencing mild loss of kidney function, thus are more likely to have an RRT event. For brevity, we will refer to this restricted analysis as analysis C.

There are 2673 individuals included in analysis C. Among them 504 individuals have experienced an RRT event throughout the decade 2008-2018 and 208 events occurred after year 5.

Methods

A 10-fold cross-validation is performed to enhance training and ensure that the number of cases in the testing set is sufficent to yield accurate predictions. We use the package `ctsem`, a recent software implementation that allows the fitting of a wide range of continuous time state space models that rely on differential equations.

In conventional joint models (i.e. `stan_jm()`) a linear mixed model (LMM) is used for each longitudinal process. The simple LMM is a special case of continuous time state space models, which can include drift and diffusion processes in addition to a LLM.

After specifying five different continuous time state space models (we call these *ctsem models* for simplicity's sake), we fit them to the longitudinal observations and compare quality of fitting by assessing their log likelihood, number of parameters and using the Akaike information criterion (AIC). When a statistical model is used to represent the process that generated the data, the representation will almost never be exact, so some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model. The less information a model loses, the higher the quality of that model. Furthemore, the log likelihood describes the joint probability of the observed data as a function of the parameters of the specified statistical model. The likelihood generally encapsulates both the data-generating process as well as the missing-data mechanism that produced the observed sample.

The next step was to combine the *ctsem* longitudinal models with Poisson regression models for the time-to-event process, splitting the follow-up period into equally spaced person-time intervals. For each person-time interval, an unobserved (latent) variable representing the true value of the biomarker at the time is imputed from the longitudinal model and used as input in the time-to-event model. To generate the unobserved values prior to fitting the Poisson regressions, we use the Kalman filter, a sequential updating algorithm which takes into account all the available information only up to the beginning of each person-time interval. Using data from the future would notoriously give rise to immortal time bias. This more advanced method of imputation is compared to the conventional 'last observation carried forward' (LOCF) approach.

We have also fitted in parallel a univariate joint model per fold to use it as benchmark for analysis C, employing the state-of-the-art modelling function `stan_jm()` from `rstanarm`. Joint models are computationally intensive, therefore we only use it for the univariate restricted analysis where there are fewer subjects. The table below gives the running times required to fit the joint model.

Table 1: Runtime in minutes of fitting a `stan_jm()` model with 1000 iterations and 4 chains

	warmup	sample	total
chain:1	255.3	120.9	376.2
chain:2	317.3	122.5	439.8
chain:3	460.2	429.7	889.9
chain:4	373.5	118.6	492.1

State space models fitted to longitudinal data with `ctsem`

Running times required for fitting the 5 *ctsem models* in parallel for each analysis:

Table 2: Running times of fitting biomarker data

	no. of people	no. of biomarkers	duration	unit
A	29118	2	3.399	days
B	29764	1	1.186	days
C	2673	1	2.790	hours

Comparison of fit to training data

Table 3: Model fits - A

model	loglik	npars	aic
hba1cegfr.model.lmm	0.0	34	0.0
hba1cegfr.model.nolmm	46571.5	47	-93117.0
hba1cegfr.model.lmmdiff	37414.5	46	-74805.0
hba1cegfr.model.lmmdrift	1942.4	50	-3852.7
hba1cegfr.model.lmmdriftdiff	59127.4	62	-118198.8

Table 4: Model fits - B

model	loglik	npars	aic
model.lmm	0.0	15	0.0
model.nolmm	32826.1	17	-65648.3
model.lmmdiff	33703.9	19	-67399.8
model.lmmdrift	142.2	19	-276.4
model.lmmdriftdiff	34230.1	23	-68444.2

Table 5: Model fits - C

model	loglik	npars	aic
model.lmm	0.0	15	0.0
model.nolmm	7090.2	17	-14176.4
model.lmmdiff	7333.9	19	-14659.9
model.lmmdrift	2.9	19	2.3
model.lmmdriftdiff	7371.6	23	-14727.3

Comparison of state space models when used to impute biomarker values for time-splitting model of events

The imputed values are used to train a Poisson regression model on the training dataset, in which follow-up of those individuals who appear in the test dataset is censored at five years.

Table 6: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from LMM with drift - A

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.270	0.117	2.311	0.021
tstart	0.045	0.014	3.186	0.001
gender	-0.155	0.037	-4.136	0.000
baselineage	-0.079	0.039	-2.006	0.045
baselineduration	-0.109	0.040	-2.732	0.006
hba1c	0.348	0.046	7.608	0.000
egfr	-0.103	0.002	-48.083	0.000

Table 7: Time-splitting model with biomarker values imputed by last observation carried forward - A

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.698	0.058	-116.455	0.000
gender	-0.063	0.036	-1.720	0.085
baselineage	0.298	0.037	7.999	0.000
baselineduration	0.057	0.035	1.602	0.109
hba1c	0.298	0.032	9.284	0.000
egfr	-0.906	0.013	-67.712	0.000

Table 8: Poisson time-splitting model only fitted to time invariant covariates - A

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.639	0.084	-79.260	0.000
tstart	0.143	0.013	10.838	0.000
gender	0.046	0.037	1.233	0.218
baselineage	0.650	0.041	15.989	0.000
baselineduration	0.198	0.036	5.531	0.000

Table 9: Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from LMM with drift - B

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.507	0.107	4.745	0.000
tstart	0.027	0.014	2.012	0.044
gender	-0.144	0.036	-3.985	0.000
baselineage	-0.063	0.037	-1.684	0.092
baselineduration	-0.119	0.039	-3.058	0.002

	Estimate	Std. Error	z value	Pr(>|z|)
egfr	-0.112	0.002	-48.611	0.000

Table 10: Time-splitting model with biomarker values imputed by last observation carried forward - B

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.461	0.054	-118.976	0.000
gender	-0.035	0.035	-0.992	0.321
baselineage	0.318	0.036	8.791	0.000
baselineduration	0.074	0.035	2.115	0.034
egfr	-1.083	0.016	-68.898	0.000

Table 11: Poisson time-splitting model only fitted to time invariant covariates - B

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.471	0.079	-81.776	0.000
tstart	0.136	0.013	10.757	0.000
gender	0.042	0.036	1.166	0.243
baselineage	0.672	0.040	16.802	0.000
baselineduration	0.223	0.035	6.315	0.000

Table 12: Continuous-time univariate joint model fitted with `stan_jm()` - Poisson time-splitting model fitted to latent biomarker values imputed by Kalman filter from LMM with drift - C

	JM mean	JM SD	Poisson est.	Poisson SE
(Intercept)	-3.500	0.013	-3.847	0.086
gender	-0.216	0.059	-0.146	0.045
baselineage	0.069	0.056	0.126	0.044
baselineduration	0.070	0.079	0.017	0.046
egfr	-1.361	0.217	-0.721	0.020

Table 13: Time-splitting model with biomarker values imputed by last observation carried forward - C

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.998	0.059	-67.934	0.000
gender	-0.151	0.044	-3.434	0.001
baselineage	-0.119	0.039	-3.037	0.002
baselineduration	0.040	0.046	0.862	0.389
egfr	-0.899	0.030	-29.952	0.000

Table 14: Poisson time-splitting model only fitted to time invariant covariates - C

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.952	0.086	-46.222	0.000
tstart	0.076	0.016	4.790	0.000
gender	-0.147	0.045	-3.283	0.001
baselineage	-0.207	0.042	-4.896	0.000
baselineduration	0.092	0.046	1.996	0.046

Comparison of predictive performance on testing data

Table 15: Comparison of predictive performance on test dataset - A

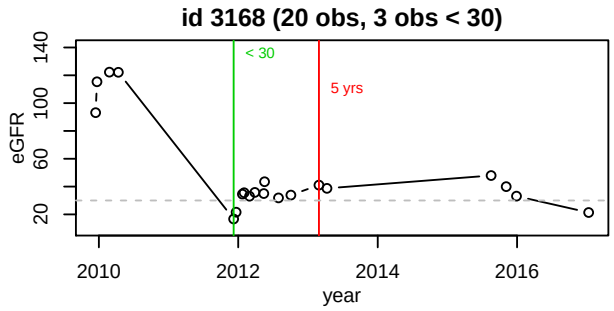
Model	Observed	Predicted	Person-years	Log score	C-statistic
hba1cegfr.model.lmm	458	517.271	127339	-2342.411	0.90636
hba1cegfr.model.nolmm	458	313.5247	127339	-2399.202	0.8888
hba1cegfr.model.lmmdiff	458	384.3632	127339	-2329.494	0.90223
hba1cegfr.model.lmmdrift	458	522.9768	127339	-2333.65	0.90826
hba1cegfr.model.lmmdriftdiff	458	388.7403	127339	-2317.053	0.90746
model.locf	458	259.9159	127339	-2339.564	0.88989
model.invariant	458	594.9419	127339	-2991.269	0.64961

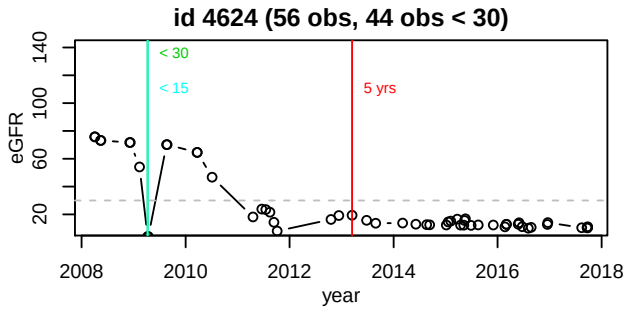
Table 16: Comparison of predictive performance on test dataset - B

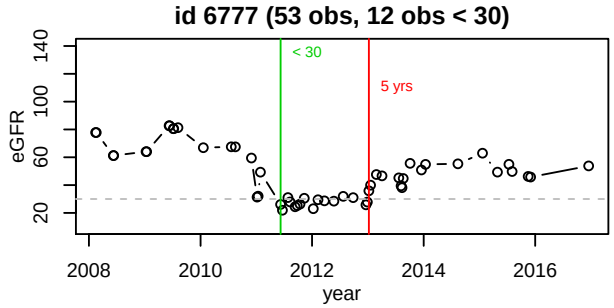
Model	Observed	Predicted	Person-years	Log score	C-statistic
model.lmm	464	519.7593	128844	-2528.531	0.89373
model.nolmm	464	277.3873	128844	-2568.091	0.89163
model.lmmdiff	464	351.0722	128844	-2523.315	0.89197
model.lmmdrift	464	519.5929	128844	-2531.289	0.89331
model.lmmdriftdiff	464	347.4431	128844	-2546.072	0.88944
model.locf	464	266.9784	128844	-2301.496	0.90304
model.invariant	464	612.1535	128844	-3029.036	0.65248

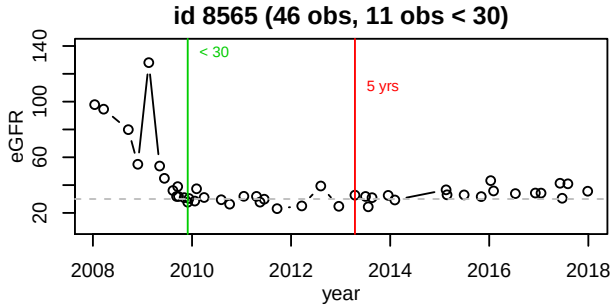
Table 17: Comparison of predictive performance on test dataset - C

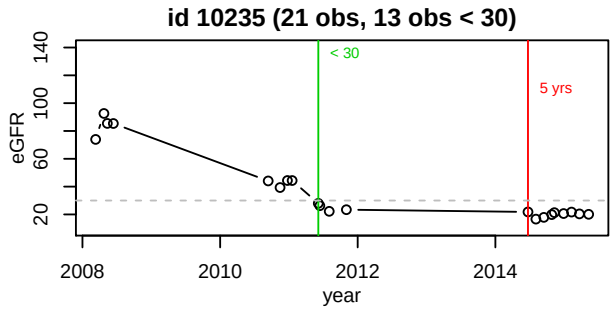
Model	Observed	Predicted	Person-years	Log score	C-statistic
model.lmm	230	229.6886	8768	-1787.442	0.81082
model.nolmm	230	147.9616	8768	-950.9871	0.82687
model.lmmdiff	230	200.8635	8768	-1042.986	0.8245
model.lmmdrift	230	229.374	8768	-1788.35	0.81109
model.lmmdriftdiff	230	196.1925	8768	-1076.55	0.8203
model.locf	230	158.0865	8768	-1011.661	0.72112
model.invariant	230	292.3499	8768	-1069.162	0.5669


id 3168 (20 obs, 3 obs < 30)
eGFR
20
60
100
140
< 30
5 yrs
2010
2012
2014
2016
year


id 4624 (56 obs, 44 obs < 30)
eGFR
20
60
100
140
< 30
< 15
5 yrs
2008
2010
2012
2014
2016
2018
year


id 6777 (53 obs, 12 obs < 30)
eGFR
20
60
100
140
< 30
5 yrs
2008
2010
2012
2014
2016
year


id 8565 (46 obs, 11 obs < 30)
eGFR
20
60
100
140
< 30
5 yrs
2008
2010
2012
2014
2016
2018
year


id 10235 (21 obs, 13 obs < 30)
eGFR
20
60
100
140
< 30
5 yrs
2008
2010
2012
2014
year