Joint Modelling of Competing Risks and Time-Dependent Covariates

Xinyi Liu

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

Department of Statistical Science

University College London

July 16, 2023

I, Xinyi Liu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In this thesis we propose a joint model for competing risks and longitudinal data. Our joint model provides a flexible approach to handle longitudinal data with complicated structures. Our model consists of a multi-state model for the competing risks and a general mixed model for the longitudinal outcomes, linked together by some latent random effects. For the joint model of one longitudinal outcome, we obtain the estimates of the parameters by maximising the marginal likelihood. We also extend the joint model to take into account multiple longitudinal outcomes simultaneously. To alleviate the 'curse of dimensionality' in integration, we propose to use Bayesian inference and use the posterior means as the estimates of the parameters. The joint models are applied to two datasets, the English Longitudinal Study of Ageing (ELSA) and the clinical data from the PhysioNet/Computing in Cardiology Challenge 2019. For the second dataset, we also propose a two-stage framework for disease early diagnosis. We construct a time-dependent loss function, and make diagnosis by minimising the expected loss.

Impact Statement

Longitudinal data and time-to-event data are frequently collected in various medical studies, and the joint modelling of longitudinal and time-to-event data is a topic of increasing importance. In the standard analysis of time-to-event data, it is assumed that there is only one event of interest. However, in practice, there usually exists events that compete with each other, and approaches specifically developed for modelling competing risks are of great interest. Methods developed in this thesis are suitable for the joint analysis of competing risks and longitudinal data with a great variety of structures. Moreover, the disease early detection framework proposed in this thesis provides a novel approach for dynamic prediction of disease. Compared with the approaches developed based on machine learning and deep learning algorithms, our method has better interpretability and can be extended and applied to various diseases.

Longitudinal data provide information on how the quantity of interest changes over time. The analysis of longitudinal data is a classic topic and has played an important role in research across different fields. Time-to-event data have information on whether an event occurs and the time till the event occurs. Joint models provide tools to study the two processes simultaneously when they are correlated. A joint model consists of a sub-model for longitudinal data i.e. time-dependent covariates, and a sub-model for time-to-event data. In this thesis, we model the time-dependent covariates by extending the linear mixed model. Instead of assuming a normal distribution for the error term, we relax this assumption and employ a skewed-normal distribution. Furthermore, we take into account the temporal variation in the variance of the error term. The model we proposed offers a flexible

way to characterize the time-dependent covariates with non-normal error terms or asymmetric distribution shapes. In addition, we discussed the extensions of the joint model to incorporate multiple time-dependent covariates. To alleviate the numerical problems when computing the marginal likelihood in high-dimensional space, we propose to use simulation-based estimation to approximate the marginal likelihood. The methods proposed in this thesis provide useful extensions for the joint analysis of time-dependent covariates and competing risks.

Disease progression prediction is a research topic of huge potential. In this thesis, a new framework for the early detection of disease was proposed based on joint modelling of competing risks and time-dependent covariates. The disease progression was described using different states. The multi-state model was used as the sub-model for competing risks. To achieve early detection of disease we proposed to make predictions by minimizing the expected loss. The expected loss was defined using a time-dependent loss function and the transition probabilities, which can be calculated from the joint model. This framework provides a novel approach for disease dynamic prediction.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor Professor Ardo van den Hout, for his invaluable guidance and consistent support throughout the past three years. Without his support and encouragement, the completion of this thesis would have been impossible.

I sincerely thank the Departmentally-funded Teaching Assistantship for funding my PhD study. Furthermore, I would like to convey my appreciation to the department for their assistance and support.

I would like to thank Shengning Pan and Wenyu Wang for generously sharing their knowledge and experience without hesitation. I am particularly grateful to Dr Guan Luo for his insightful research suggestions and warm assistance when I was visiting the Institute of Automation, Chinese Academy of Sciences.

I sincerely appreciate Helen Hu, Mengyuan Chen, Marta Campi, Yurong Ling, Xiaoou Lu, and Qibin Zhu. Their support and companionship meant the world to me during the most challenging moments of my life, providing me with the strength to persevere.

Finally, I want to express my gratitude to my family and friends for their consistent love and support.

Contents

Chapter 1

Introduction

Joint models provide a valuable method for statistical analysis of biomedical studies with longitudinal and time-to-event data. A joint model consists of a model describing the trajectories of longitudinal outcomes and a survival model for time-to-event data, both models are linked by random effects. In this way, joint models are able to take into account the association between the two processes.

In this thesis, we use the joint model to study the relationship between longitudinal outcomes and disease progression. For disease progression, we consider different health statuses as discrete states and model them using a multi-state model. For the modelling of the longitudinal data, in order to improve the model's capability to capture complex structures in the data, we propose to use the mixed model. Our approach extends the linear mixed model by relaxing the normal assumption of the error term and employs a skew normal distribution instead. Additionally, we take into account the temporal variation in the variance of the error term. We refer to this model as the general mixed model. The model parameters are estimated using maximum marginal likelihood.

We apply the joint models to analyse the English Longitudinal Study of Ageing (ELSA) dataset, which consists of longitudinal and time-to-event data related to cognitive impairment in the elderly population. We also apply the joint models to the dataset from the PhysioNet/Computing in Cardiology Challenge 2019, which consists of the clinical data of sepsis disease.

Furthermore, we aim to explore joint models of competing risks and multiple

longitudinal outcomes. By including more than one longitudinal outcomes, we anticipate enhanced predictions for disease progression. To alleviate the numerical problems when computing the marginal likelihood in high-dimensional space, we propose to utilise simulation-based estimation to approximate the marginal likelihood. However, it should be noted that this approach can introduce instability to the optimisation process. In order to overcome this issue, we introduce Bayesian inference and specify weakly informative priors. Therefore the posteriors converge towards the maximum likelihood estimates (MLE). The details of this approach will be presented in Chapter 6. It is important to note that this thesis remains focused on the frequentist approach and does not involve a transition to the Bayesian paradigm.

In the study of disease progression, disease diagnosis is an important task. In this thesis, we propose a framework for disease early diagnosis based on the joint model. We construct a time-dependent loss function for the sepsis early diagnosis problem, then perform diagnosis by minimising the expected loss. We apply the joint model of competing risks and time-dependent covariates to the clinical data of sepsis. Then we illustrate the disease early diagnosis framework using this model.

The rest of this chapter is organised as follows: In Section 1.1, we will first introduce the concept of longitudinal data and time-to-event data in medical studies. Then we will briefly discuss the disease diagnostic problem, with a specific focus on the sepsis diagnostic problem. In Section 1.2, we will provide a brief introduction to the joint model used in this study. In Sections 1.3 and 1.4, we will introduce competing risks and informative censoring. In Section 1.5, we will provide an overview of various methods employed in disease diagnosis. Finally, in Section 1.6, we will introduce the scope of the research. We will also briefly describe the joint model used in this study and our proposed disease early diagnosis framework.

1.1 Background

Longitudinal data consists of repeated measurements of the quantity of interest, therefore it has the information of how the quantity changes over time. In medical studies, it is common to collect longitudinal outcomes that are associated with disease progression. Changes in the longitudinal outcomes can reflect the development of disease over time. Time-to-event data tracks the time till the occurrence of the event of interest, e.g. the time till death. It also provides information on the disease progression. In some medical studies, both longitudinal and time-to-event data are available, and there is growing interest in characterising the relationship between both processes. For instance, in the English Longitudinal Study of Ageing (ELSA), the cognitive functions are tested at each follow-up. The scores of the cognitive tests are recorded, along with the time to death. With the information provided by both longitudinal data and the time-to-event data, it can be investigated which cognitive test is most relevant to cognitive functions. The trajectories of the cognitive tests can also be taken into account with the presence of death, allowing for dynamic predictions of an individual's cognitive function based on their past observations.

Apart from the study of chronic diseases, many clinical datasets of acute illness also consist of longitudinal data and time-to-event data. The dataset provided by the PhysioNet/Computing in Cardiology Challenge 2019 consists of clinical data of patients who have sepsis. Sepsis is a life-threatening condition caused by the body's overreaction to infections. When the body reacts to infection, white cells travel to the infection site, and this will trigger inflammation. However, the inflammation can become widespread in the body if the infection is severe or the immune system is weak. It might interrupt the blood flow, resulting in a decrease in blood pressure and stopping oxygen from reaching the organs and tissues, and it might further cause severe organ damage and death. Sepsis is a global health crisis. The mortality of sepsis is between 20% to 50%, recorded in Nasir et al. (2015). According to the World Health Organization et al. (2020), it kills 11 million people around the world each year. Everyone can get sepsis, but certain groups of people are at higher risk, especially children, pregnant women, and people with chronic diseases. Sepsis can

present itself with various symptoms, for example, fever or low temperature and shivering, muscle pain, increased heart rate, weak pulse and low blood pressure, etc.

Sepsis is often diagnosed by testing temperature, heart rate, breathing rate and blood. Early detection of sepsis is very important. Studies found that during the mandated emergency care for sepsis, each hour of delayed treatment is associated with a 4% to 8% increase in mortality. To help diagnose sepsis as early as possible, new clinical criteria were proposed. It was recommended to diagnose sepsis based on the score named Sequential Organ Failure Assessment (SOFA), which assigns a grade of abnormality to each organ system. More details about this score can be found in Vincent et al. (1996) and Lambden et al. (2019). Other organ failure scoring systems have also been developed and evaluated in the context of sepsis. In a study conducted by Khwannimit et al. (2018), the authors compare the SOFA score with two other sepsis criteria. The computation of the SOFA score requires various clinical and laboratory variables, including but not limited to Pao2 (arterial oxygen partial pressure), platelet count, creatinine level, and bilirubin level. However, despite the availability of scoring systems like SOFA, there is still a need for early sepsis detection. Statistical models and machine learning methods can play a crucial role in this regard, as they can simultaneously consider multiple variables and provide a quantitative risk assessment for patients suspected of having sepsis.

In this study, our aim is to develop methods for analysing disease progression using longitudinal data. We focus on developing a predictive model that can estimate an individual's health status and assess the potential occurrence of the disease of interest based on their health records. Additionally, we aim for the model to capture the correlation between the longitudinal data of interest and the disease outcome, thereby providing explanatory insights.

We propose to use the joint model to analyse longitudinal data and time-to-event data. We use discrete states to describe an individual's health status and utilise a multi-state model to capture the transitions between different states. For the longitudinal data, we extend the linear mixed model by relaxing the normal assumption

of the error term. Instead, we use a skew normal distribution to model the error term. Additionally, our model allows for variance of the error term changes with respect to time. These extensions enable our model for longitudinal data to capture more complex structures and offer broader applicability. We present the joint models to the cognitive function data (ELSA) and the clinical data of sepsis.

For the problem of early diagnosis of sepsis, we propose a two-stage framework. We consider the diagnosis of 'sepsis' and 'non-sepsis' as competing events. In the first stage, we fit a joint model of competing risks and time-dependent covariates. To handle the early diagnosis, we propose a time-dependent loss function, which encourages early diagnosis and penalises late diagnosis. In the second stage, we predict the labels of sepsis by minimising the expected loss.

The joint model and sepsis early detection framework we proposed can be applied to other datasets with both longitudinal and time-to-event data. Different sub-models for the time-dependent covariates can be selected, and different time-dependent loss functions can be designed based on different needs.

1.2 Joint Modelling of Time-dependent Covariates and Time-to-Event Outcomes

Joint model, also known as shared-parameter model, is a popular statistical approach used in the analysis of longitudinal and time-to-event data. The main advantage of joint models is their ability to utilise longitudinal data to improve the prediction or estimation of event occurrence. At the same time, the longitudinal data can provide additional insights into the underlying disease progression or risk factors that influence the event of interest.

The standard Cox model can be extended to take into account time-dependent covariates when modelling the time-to-event data, but it has some limitations. One of the main challenges is the assumption that the time-dependent covariates are external, which are covariates that develop or exist independently of a patient's survival. For example, age and gender are external covariates. However, the time-

dependent covariates that we are of interest in the disease diagnosis problem violates this assumption. They are generated by the stochastic process of the individual, i.e. they are internal covariates. In addition, it is unavoidable to have informative dropouts, and this process should not be ignored (Rizopoulos, 2012). For above situations, joint models provide a desirable solution.

The basic idea behind joint models is to model the relationship between longitudinal and time-to-event outcomes simultaneously, taking into account the fact that these outcomes may be correlated. The joint modelling consists of using two sub-models: a hazard model for the time-to-event outcome and a mixed model for the time-dependent covariates. The two sub-models are linked by shared random effects, see Rizopoulos (2012) for a detailed introduction to joint models.

The advantage of applying the joint model is two-fold. First of all, it provides an approach to take into account a model for time-dependent covariates when the primary interest is modelling the survival time. Secondly, it is able to model the longitudinal data more accurately since it can take into account the effect of the associated events. The use of joint modelling reduces the bias in the treatment effects estimates (Ibrahim et al., 2010).

The earliest applications of joint models was in the field of HIV research, such as works by De Gruttola and Tu (1994) and Faucett and Thomas (1996). Since then, joint models have been applied in a wide variety of fields, including cancer research, cardiovascular disease, and mental health. There is an extensive collection of work on joint modelling, and there are several reviews on recent developments, e.g. Papageorgiou et al. (2019) and Lawrence Gould et al. (2015). For the modelling of time-dependent covariate, Diggle et al. (2002) summarised and discussed the common models for different types of longitudinal data, for example, the linear models and the generalised linear models.

Various extensions of the basic joint model are proposed. For example, the extension of longitudinal and survival components to be multi-dimensional allows for more information to be taken into account simultaneously, see Chi and Ibrahim (2006) and Lin et al. (2002). The extension of the association structures between

the time-to-event data and the time-dependent, see work by Mauff et al. (2017) and Andrinopoulou et al. (2014). The extension of the time-to-event outcome with a single event type to multiple recurrent or competing events, see work by Huang et al. (2011) and Elashoff et al. (2008).

The extensions of the mixed-effects model of the time-dependent covariates are of interest as well. Gibbons et al. (2010) explored the linear and nonlinear generalised mixed-effects models and their alternatives for the longitudinal data. The generalised linear model assumes that the outcome follows a distribution from the exponential family. However, in practice, the outcome often has complicated structures, and the distributions from the exponential family might not suit the data well.

In addition, it is common to assume that the residuals have constant variance in linear mixed models. However, in practice, this assumption can be violated. When examining the trajectory of an outcome over time, we may observe that both the mean and variance of the outcome vary with certain predictor variables. For example, in the ELSA data, we find that the mean number of animal names people can recall decreases as people age, while its variance also decreases. This suggests that the variance of the error term in the outcome may not be constant over time and may be influenced by the predictor variables of interest. Therefore, we take into account the changes in variance over time to improve the generalisability of the model we proposed for disease diagnosis. Accounting for changes in the variance over time can lead to more robust and accurate diagnostic models because it allows the model to capture the inherent variability in the disease progression over time. By explicitly modelling the changes in the variance over time, the proposed method can provide more accurate predictions for new patients or populations. It may also help identify changes in disease progression that are not captured by diagnostic models that assume constant variance.

1.3 Competing Risks

In standard survival analysis, individuals under study are assumed to be at risk of a single event. However, in practice, usually more than one event may occur, and the occurrence of one of the events prevents other events from occurring. We refer to these events as competing events. For example, in medical research, the deaths of individuals due to different causes can be considered as competing events, and out of these events, only the first is observable.

Models that consider the dependencies between competing risks are of great interest. One approach proposed is to directly model the dependence between event times by specifying a joint distribution using a copula. The details can be found in Zheng and Klein (1995).

Another approach is to model the joint distribution of the event times and the corresponding event using the cause-specific hazard function, see Austin et al. (2016). In particular, this approach can be considered as a subclass of multi-state models. In a multi-state model, an individual can move between several distinct states over time, and the transitions can be affected by one or more predictor variables. With the presence of competing risks, an individual can experience different types of events that prevent him or her from transitioning to other states. Therefore, the competing risks can be modelled with a process that starts from an initial state and transit to several different endpoints. In this study, we use the multi-state model framework to fit a parametric competing risks model. This framework provides clear methods for selecting and estimating parametric models that describe the transition probabilities between different states as a function of time and covariates.

Cox regression model can also be extended to accommodate dependent censoring, using the Inverse Probability of Censoring Weighted (IPCW) approach, see Collett and Kimber (2014). IPCW involves assigning weights to the censored observations based on the probability of the censoring time given the observed data. These weights are used to adjust the analysis of the Cox regression. However, in this thesis, we focus on modelling competing risks using multi-state models, which have more flexibility than the Cox regression model, as they can handle interval-

censored data. Additionally, our proposed method can be illustrated using a three-state graph, which facilitates better understanding and interpretation by clinicians and other audiences.

Ongoing research is currently being conducted in the field of developing survival models for competing risks, such as work by Ishwaran et al. (2014), Monterrubio-Gómez et al. (2022), and Sparapani et al. (2020), focuses on incorporating flexibilities in various aspects, such as non-linear effects of covariates, time-dependent covariates, variable selection, and missing data.

1.4 Informative Censoring

Informative censoring is a common issue in survival analysis, where the censoring process is related to the outcome of interest. There are multiple reasons why informative censoring can arise, such as selective dropout from a study or competing risks. Selective dropout can occur when certain participants are more likely to drop out of a study in a non-random manner, such as those with more severe symptoms. Competing risks, on the other hand, refer to events that prevent the observation of the outcome of interest, such as death or other illnesses, and can lead to informative censoring if they are related to the unobserved outcome. For example, in a study examining the effect of cancer treatment on the time to cancer recurrence, patients who die before experiencing a recurrence may be censored, which can result in informative censoring. Addressing informative censoring is crucial to obtain accurate and reliable estimates in survival analysis, and various statistical methods have been developed to account for it. For a concise explanation of how dependent censoring can be incorporated into survival data modelling, see Collett and Kimber (2014). Additionally, Schluchter (1992) provides a comprehensive review of approaches proposed for analysing longitudinal data with informative censoring.

The sepsis diagnosis problem involves analysing the time of onset for sepsis during a patient's hospital stay, using data extracted from electronic medical records. While it is possible for patients to develop sepsis after leaving the hospital,

our main interest lies in detecting it as early as possible during their hospitalization. Moreover, since sepsis is an acute disease, individuals in the sepsis problem are observed on an hourly basis. It is unlikely for them to develop sepsis within the timeframe of this study. The competing event of hospital discharge or change in medical care can prevent the event of interest (onset of sepsis) from occurring as the first event, resulting in informative censoring. To address the sepsis diagnosis problem, we propose to use a competing risk model in the multi-state model framework. This approach is consistent with the statistical methods reviewed by Putter et al. (2007). This tutorial also covers practical scenarios in the medical field where competing risks may occur.

1.5 Disease Diagnosis

Disease diagnosis has been an active research area in medical and healthcare communities for decades. In recent years, there has been a growing interest in developing accurate and efficient models for disease diagnosis, leveraging the increasing availability of electronic health records (EHRs), medical imaging data, and genomic data.

Since disease diagnosis can be simplified as a classification problem, i.e. given the historical observations, whether a patient has a certain disease or not, all classification methods have the potential to be used for disease diagnosis.

Using time series classification with EHRs is a commonly employed method for disease diagnosis. EHRs contain a wide variety of medical data, such as patient demographics, clinical notes, laboratory results, medication history, and diagnostic codes. These data are typically recorded over time, which makes EHRs well-suited for time-series analysis. Disease diagnosis can be achieved using time-series classification techniques. Papers that show the utilisation of time-series classification for disease diagnosis include works such as Mansour et al. (2021) and Al-Hadeethi et al. (2020).

It is possible to address the challenge of diagnosing sepsis using the time series

classification method. This involves modelling the repeated observations of patients over time as a time series. Relevant information, such as mean, variance, and autocorrelation, can be extracted from the data to aid in the classification process. The hourly diagnosis provided by clinicians can be used for classification at each corresponding time point. One important benefit of this approach is its ability to capture complex temporal dependencies within the data.

In the context of binary classification, individuals are categorised into two groups: those who have sepsis and those who do not have sepsis. However, there may be some individuals who are still under observation, and it is not yet clear whether they have sepsis or not. These individuals have the potential to experience the onset of sepsis in the future and should be treated differently from those who do not have sepsis. In survival analysis, this is referred to as "right censored" since the event of interest (i.e., the onset of sepsis) has not occurred at the time of classification. Unfortunately, the time-series classification approach overlooks this aspect. Moreover, time series classification methods focus on providing predictions at predetermined time points, therefore the status of individuals between those time intervals are neglected. However, in practice, it is common for the exact time observation assumption to be violated, meaning that the exact occurrence time of the event is unknown. This can introduce measurement errors, and the predictive performance of time series classification methods may be affected.

We propose a solution to the sepsis diagnosis problem by using survival analysis. Instead of treating each diagnosis of an individual at each time point as an independent binary outcome, we focus on predicting the time of onset of sepsis. The sequence of diagnosis is transformed into time-to-event data, which reflects the time of sepsis onset for each individual. Individuals who are non-septic are considered to be at risk of sepsis until their last diagnosis indicates that the individual does not have sepsis. This approach takes into account the dependencies of diagnosis within each individual, as well as the presence of right-censoring. Furthermore, our proposed joint modelling framework accommodates interval censoring, eliminating the need for exact transition times.

Different from the time series classification approach, the proposed approach in this thesis is to utilise a joint model for modelling both the time-dependent covariates and the time-to-event data. The joint model is able to model multiple correlated time-dependent covariates jointly with other individual-level covariates, such as demographic information, clinical history, and other clinical measurements, to predict the likelihood of a disease or event occurring. This approach allows for the capturing of complex relationships between different covariates and the disease outcome, accounting for individual-level variability and correlation structures within and between different time-dependent covariates.

Leveraging the benefits of modelling the longitudinal data as time-series, the joint model can be further enhanced. For example, Yue and Kontar (2021) extend the joint model by combining the Cox model and the multivariate Gaussian convolution process for time series modelling. Additionally, Recent developments of applying joint models to personalised predictive modelling, such as the work by Papageorgiou et al. (2018) and Rizopoulos (2011) shows that the joint models have the potential to be used for disease early diagnosis.

Landmarking is another technique used in survival analysis for dynamically predicting the risk of an event for an individual. The basic idea is to divide the follow-up time into a series of landmark times and then compare the survival distributions of individuals who have experienced the exposure of interest before a given landmark time with those who have not. Examples of studies that demonstrate the use of the landmarking approach for dynamic prediction include: Van Houwelingen and Putter (2008); Van Houwelingen (2007). The landmarking approach relies on exact observation of the transition time, but this may not always be feasible in practice. In contrast, the joint modelling approach proposed in this thesis allows for interval censoring and does not have strict requirements for exact transition time, therefore offering more flexibility.

Apart from aforementioned methods, the application of machine learning offers advanced algorithms capable of analysing complex biomedical data. Kononenko (2001) provided a thorough overview of the history of applying ma-

chine learning methods to medical diagnosis. There are several machine learning methods that are commonly used for disease diagnosis: Support Vector Machines (SVM) is a binary classification algorithm that finds the optimal hyperplane in a high-dimensional space to separate different classes. SVM has been successfully applied to diagnose diseases such as breast cancer, diabetes, and heart disease, see works such as Sartakhti et al. (2012) and Orru et al. (2012). Artificial neural network (ANN) is a biologically inspired network of interconnected nodes that can learn and make predictions on data. It is capable of learning non-linear relationships between variables and can handle large datasets, see works such as Baxt (1995) and Abiodun et al. (2018). In the context of sepsis early diagnosis, similar methods can be employed. For instance, studies like Kok et al. (2020) and (Li et al., 2019) utilise neural networks to address this problem.

Rajula et al. (2020) discussed and compared the use of machine learning algorithms and statistical methods in medical problems. Compared with statistical methods, machine learning algorithms are more data-driven. In order to use traditional statistical methods, strong assumptions are usually required, for instance, the distribution of the error terms, the structure of the linear predictor, and the proportional hazards. Machine learning algorithms can utilise all information, while statistical methods usually require a priori selection of the variables.

In summary, both machine learning and statistical methods have been widely used for disease diagnosis. Kononenko (2001) discussed the requirement for a machine learning algorithm to be a satisfying medical diagnosis system. Including good performance, dealing with missing data, reduction of the number of tests, transparency of diagnostic knowledge and explanation ability. While machine learning models can identify patterns and relationships that are not easily visible to human experts, the lack of interpretability can limit their usefulness in certain domains, such as healthcare, where interpretability and explainability are critical for decision-making and trust-building. Joint modelling is a promising approach for disease diagnosis, particularly for predicting the risk of dementia using longitudinal data. However, further research is needed to evaluate the performance of joint

modelling in other disease diagnosis scenarios.

1.6 Scope of Research

In this thesis, we aim to develop methods for analysing disease progression using longitudinal data. We explore joint models of competing risks and multiple longitudinal outcomes. In addition, we present a novel framework for disease diagnosis utilising the joint model. In this section we provide a brief overview of our work, focusing on explaining and clarifying key methodological choices inherent to our research.

Based on GAMLSS proposed by Stasinopoulos and Rigby (2008), we extend the sub-model for the longitudinal data by extending the distributions of response variables to distributions controlled by three parameters: the location, scale, and shape parameters. And we allow the parameters to be linked with other covariates in any shape of the links. We further extend the joint model for competing risks and one time-dependent covariate to multiple covariates by assuming that the random effects of different response variables are correlated.

In this study, we focus on parametric models for the random effects model because they are easy to specify and estimate. While nonparametric models could be chosen to model more complicated structures, we limit ourselves to parametric models for simplicity. As for the choice of error terms, we opt for the skew normal distribution. This distribution encompasses the normal distribution, allowing for easy comparison with normal alternatives. In our analysis, we utilise two random effects: the random intercept and random slope. These random effects have relatively simple structures and can be easily interpreted. However, for future studies, more complex random effects models can be explored. For example, when analysing clinical data on sepsis, we only consider individuals from a single hospital. In the Physionet Challenge 2019, data from different hospitals are available. To increase the generalisability of the disease diagnosis model, we could incorporate random effects to capture the variation at the hospital level.

We demonstrate the application of a joint model incorporating a general mixed sub-model, using both the ELSA data and clinical data for sepsis. Furthermore, we illustrate the extension of joint models to accommodate multiple time-dependent covariates using the clinical data for sepsis.

Given the continuous nature of disease progression, we utilised the continuous-time multi-state model as the sub-model within the joint model framework. We assume the transition intensities change continuously with respect to time. Due to this assumption and the incorporation of time-dependent covariates in defining these intensities, the Markovian assumption is violated. However, within the general MSM framework, there is currently no established method to directly derive transition probabilities $P(t_1, t_2)$ for a changing transition intensity in time interval (t_1, t_2). Therefore we adopt piecewise constant approximation to compute the transition probabilities. The piecewise constant approximation method involves dividing a continuous time interval into smaller discrete time intervals, then it assumes the transition intensities are constant within the time intervals. The approximation of transition probabilities over a continuous time interval is achieved by multiplying the transition probabilities corresponding to each time interval.

For the joint model with one time-dependent covariate using a general mixed model, the parameter estimates are obtained by maximising the marginal likelihood. The marginal likelihood is approximated using the Gaussian-Hermite quadrature method. However, if we want to extend it to incorporate additional time-dependent covariates, the number of random effects increases, and this requires the calculation of integrals in a high-dimensional space. This is where the 'curse of dimensionality' comes in. We propose to utilise simulation-based estimation to approximate the marginal likelihood. However, this introduces instability to the computation of likelihood, leading to unreliable results of the optimisation process. To alleviate this problem, we introduce Bayesian inference and specify weakly informative priors. Ideally, the posterior distributions should converge towards the maximum likelihood estimates (MLE). The estimates of parameters can be obtained by calculating the posterior mean. The data analysis is performed using the software WinBUGS

1.4.

For the data pre-processing technique employed in this thesis, we applied the rescaling technique to avoid potential numerical problems. In addition, we performed forward imputation to the clinical data of sepsis. In the analysis of sepsis data, we have access to seven distinct time-dependent covariates: *heart rate* (HR), *systolic blood pressure* (SBP), *mean arterial pressure* (MAP), *diastolic blood pressure* (DBP), *respiration rate* (Resp), *pulse oximetry* (O2Sat) and *temperature* (Temp). However, it is important to note that these covariates exhibit varying degrees of missingness. The observations were recorded at an hourly frequency, but there are missing records for some of the covariates. This could be due to constraints in medical resources, limitations in the measuring technique, or clinical decisions not to monitor certain covariates for specific individuals on an hourly basis. The following table presents the missingness rate for each covariate. The missingness rate is calculated by dividing the count of missing values by the total number of observations in the dataset.

Covariates	HR	SBP	MAP	DBP	Resp	O2Sat	Temp
Missingness Rate	0.077	0.152	0.102	0.481	0.097	0.120	0.662

In the variable selection/model selection process, we employ the Akaike Information Criterion (AIC) to identify the optimal pair of time-dependent covariates for prediction purposes. But it is important to acknowledge that AIC is based on the likelihood function, therefore, is influenced by the sample size. Consequently, comparing AIC values across different sample sizes may lead to unfair comparisons.

To address this issue and maintain equivalent sample sizes, we performed forward imputation for each covariate. However, it is crucial to recognise that this imputation method introduces bias into the estimates. In order to minimise the impact of the bias introduced, we conducted the data analysis solely on the covariates *heart rate* (HR), *mean arterial pressure* (MAP), and *respiration rate* (Resp). These specific covariates exhibit relatively lower missingness rates in comparison to the other variables.

1.7 Outline

The rest of this report is organised as follows. In Chapter 2, we will describe the English Longitudinal Study of Ageing (ELSA) data and the clinical data of sepsis provided by the PhysioNet/Computing in Cardiology Challenge 2019. We will also introduce the task and the evaluation method used in sepsis early diagnosis problem. Chapter 3 presents the competing risks, the multi-state model, the mixed model, and the joint model. In Chapter 4, we present our joint model where the longitudinal outcome is modelled by the general mixed model with skew normal distribution, and allows for the time-dependent variances. We apply this model to the English Longitudinal Study of Aging (ELSA) data. We show that this model outperforms the model which assumes the error term follows normal distribution. In Chapter 5, we introduce the statistical decision theory. We define the time-dependent loss function, and show that the diagnosis of sepsis can be made by minimising the expectation of this loss function. Then we show that to calculate the expected loss, the joint model of the competing risks and the time-dependent covariates is needed. Chapter 6 presents the fundamental aspects of Bayesian inference. We illustrate the joint model with the sub-model for multi-variate time-dependent covariates on the clinical data of sepsis. Furthermore, we illustrate the early sepsis diagnosis framework on the clinical data. Finally, in Chapter 7, we discuss challenges in this study and potential future developments that can improve the predictive capabilities.

Chapter 2

Data

The extension of joint models has received considerable attention and has been subject to extensive exploration. By relaxing the assumption of the error term of the mixed model, the extension opens up new possibilities and potential for the joint model. The combination of minimizing expected loss and survival analysis represents a novel and innovative approach, combining two distinct methodologies into one framework.

For clinicians, the extended joint model provides a valuable tool for studying associations between covariates of interest and the specific disease under investigation. Furthermore, the proposed two-step sepsis early diagnosis framework can also be applied to the diagnosis of various other diseases.

In this thesis, we demonstrate our method by applying it to two datasets: the English Longitudinal Study of Ageing (ELSA) and the clinical data of sepsis from Physionet Challenge 2019. With the ELSA data, We demonstrate the effectiveness of the proposed joint model by illustrating its application with one time-dependent covariate. Compared with the sepsis data, the ELSA study exhibits a lower frequency of observations, resulting in reduced computational costs. Initially, we apply the joint models we have developed to the ELSA data and conduct simulation studies to evaluate the model's performance. Building upon this, we extend the joint model to accommodate two time-dependent covariates simultaneously, and we illustrate this on the clinical data of sepsis.

In this chapter, we will introduce each of the two datasets respectively.

2.1 English Longitudinal Study of Ageing (ELSA)

The English Longitudinal Study of Ageing (ELSA) is a UK-based panel study of the population aged ≥ 50 years. Individuals were followed up every two years, and information on economic status, cognitive function, and mental health was collected. The dataset can be accessed online (www.esds.ac.uk). More information about this study was introduced in Steptoe et al. (2013).

Wave 1 of the ELSA study involved recruiting 12,099 individuals, marking the initial phase of data collection. This phase focused on gathering baseline information from participants, encompassing demographic data, health status, lifestyle factors, social networks, economic indicators, and psychosocial measures. Because of the computational challenges of our methods, we use a random subsample of 1000 individuals for our current analysis. The same dataset was used in Van den Hout and Muniz-Terrera (2016) as well.

In our study, the sample size was determined considering the limitation of computation time. The data analysis on the ELSA data is conducted on a High-Performance Computing (HPC) system, and as the model complexity increases, the time required for model fitting also increases. To ensure that the model fitting process can be completed within the allocated 72-hour usage limit, a sample size of 1000 was chosen. This sample size strikes a balance between obtaining a sufficiently large sample for meaningful analysis and accommodating the computational constraints associated with more complex models.

The sample we used consists of 540 women and 460 men. The age of individuals ranges from 50 to 89 years. Table 2.1 shows the observation frequencies in the provided sample. The number of individuals who have been observed 1, 2, 3, 4, 5, 6 times are 179, 141, 127, 467, 5 respectively. Table 2.2 displays the time intervals between observations and their respective frequencies. It is obvious that the majority of individuals have a follow-up time of two years.

In the ELSA study, cognitive performance was measured in several different

ways. In this study, we have an interest in the outcomes of two tests: the number of animal names the individual can write out within 1 minute, and the number of words the individual can recall from a list of 10. The two graphs in Figure 2.1 demonstrate the trajectories of both scores of 30 randomly selected individuals. The trajectories show that the process of recalling animal names and words is subject to noise., but it is obvious that both scores decline as the age of people increases, i.e. the cognitive function declines when people grow old.

Our analysis focuses on investigating the association of a cognitive test with the onset of cognitive impairment in the presence of death as a competing risk. We define the states of cognitive impairment based on the number of words recalled: if the number of words is greater than 0, then the state is $s = 1$ (healthy); if the number of words is 0, then the state is $s = 3$ (severe cognitive impairment). The state $s = 2$ is defined as the death state. The three-state model is illustrated in Figure 2.2. In this application, we consider the death state and the severe cognitive impairment state as competing risks, and we model the number of animal names recalled as a numerical longitudinal outcome. Table 2.3 displays the frequencies of transitions from one state to another. From this table, it is clear that the number of individuals who have experienced death is 131, while the number of individuals classified as having severe cognitive impairment, according to our definition, is 111.

In the data pre-processing, we excluded 81 individuals who were interviewed only once. As a result, we illustrate the joint models on a sample with size $N = 919$. To mitigate numerical issues, in the data analysis, we rescale the age variable by subtracting 49 years and dividing it by 10.

Observation Times	1	2	3	4	5	6
Number of Individuals	81	179	141	127	467	5

Table 2.1: The observation frequencies in the ELSA data, indicating the number of individuals who have been observed 1, 2, 3, 4, 5, and 6 times.

Observation Time Interval (Years)	1	2	3	4	5	6	7	8	9
Frequency	226	1987	384	65	18	17	13	6	4

Table 2.2: The time intervals between observations and their respective frequencies in the ELSA data.

State	Transition to State 1	Transition to State 2	Transition to state 3
1	3412	131	111

Table 2.3: Frequencies of transitions from one state to another in the ELSA data.

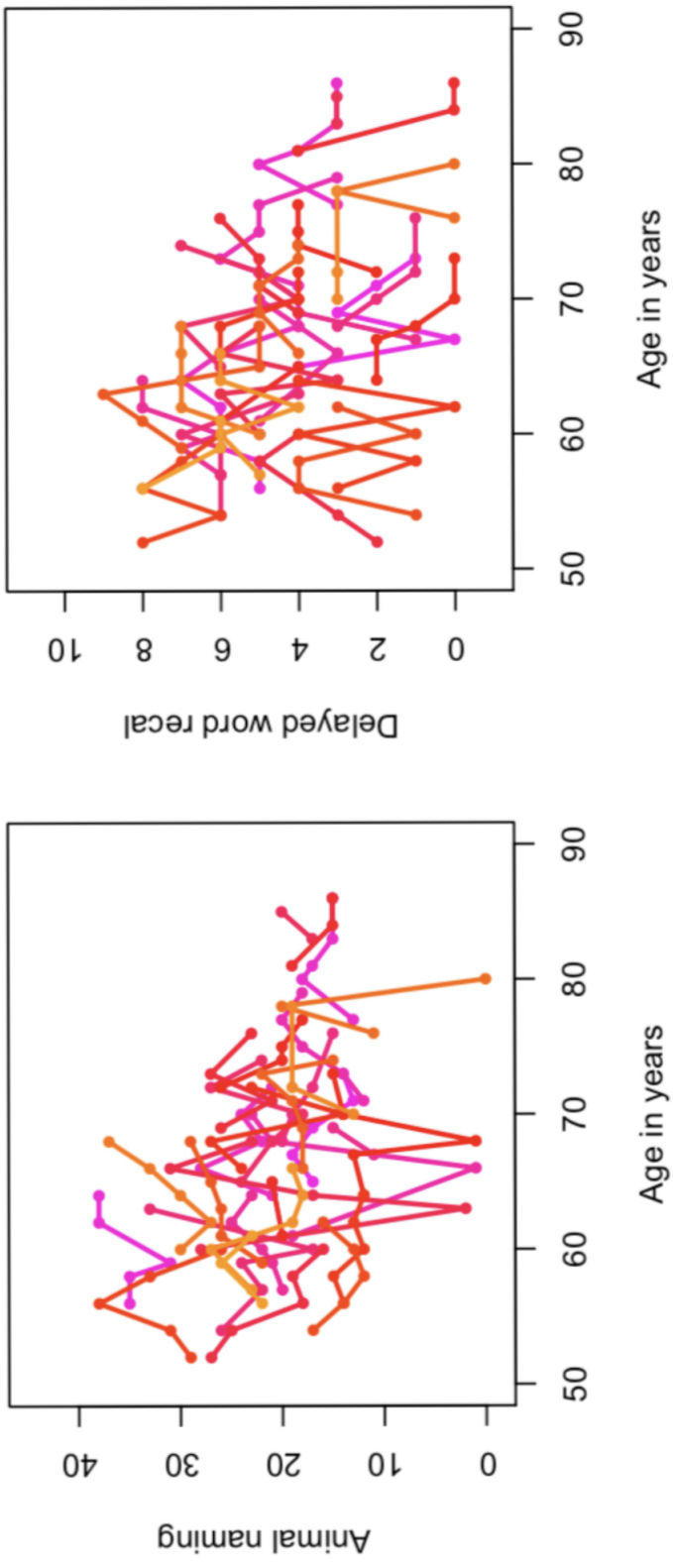


Figure 2.1: Left graph: Trajectories of number of animal names recalled by 30 randomly selected individuals; Right graph: Trajectory of number of words remembered by 30 randomly selected individuals.

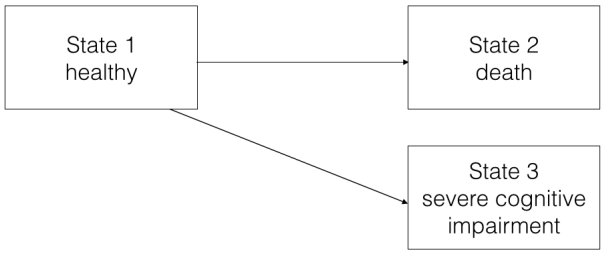


Figure 2.2: Three-state model for competing risks of death and severe cognitive impairment.

2.2 Clinical Data for Early Prediction of Sepsis

In this section, we will describe the clinical data we used in the study of early prediction of sepsis and the assessment of the performance of the models. In addition, we will introduce the task of early sepsis diagnosis.

2.2.1 Data Description

The dataset used in this study is from the Physionet Challenge 2019, a competition focused on developing algorithms for the early detection of sepsis. The data were collected over the decades in the United States. The dataset consists of the clinical records of 20,336 and 20,000 ICU patients from two distinct hospitals (hospital A and B respectively). For all patients, the demographic data, vital signs and laboratory values were recorded hourly. More details about this competition and dtatset can be found in Reyna et al. (2019a) and Reyna et al. (2019b). The dataset is accessible online (`https://physionet.org/content/challenge-2019/1.0.0/`).

The clinical data consists of 6 demographic variables, including age, gender, the types of ICU and the length of time staying in the ICU etc. The 20,336 patients from hospital A are from age 18 to age 89. Out of all the individuals, 8952 are female and 11384 are male. The individuals in the dataset were followed up for 27 hours on average.

For each individual, 7 vital signs were recorded respectively. The vital signs include Heart rate (HR), Pulse oximetry (O2Sat), Temperature (Temp), Systolic Blood Pressure (SBP), Mean arterial pressure (MAP), Diastolic Blood Pressure (DBP),

and Respiration rate (Resp). The trajectories of these vital signs are shown in Figure 2.4 and Figure 2.4. We present the trajectories of 5 individuals randomly chosen from the dataset. Each of the vital signs characterize different features of the body, therefore the ranges of each variables are different. In addition, the trajectories are varied in the behaviours. For example, in general the heart rates of individuals increase with time, while the pulse oximetry decrease with time. Furthermore, their variances changes with respect to time in various ways.

To explore the relationships and dependencies among various vital signs, a pairwise scatter plot matrix of the initial recordings of 7 vital signs for 500 randomly selected individuals is shown in Figure. This visualisation highlights the strong correlation between *systolic blood pressure* (SBP), *diastolic blood pressure* (DBP), and *mean arterial pressure* (MAP).

For each individual, 26 laboratory values were recorded as well. However, the measurement of these values are much more complicated than vital signs, and the low-frequency of observations made it difficult to treat these values in the same way as the vital signs. In Figure 2.6 and 2.7, we present the trajectories of 8 laboratory signs of 5 randomly chosen individuals. It is obvious that compared with the vital signs, the frequencies of observations of lab signs is much lower. There also exits a lot of individuals that have no records of some of the laboratory values.

For each individual the time point t_{sepsis} represent the onset time of sepsis. And for septic individuals, there is sequence of binary labels is used to indicate whether an individual has sepsis at a corresponding time point ($SepsisLabel = 1$) or does not have sepsis at that time point ($SepsisLabel = 0$). For non-sepsic individuals, the label sequence will always be 0. It is important to note that the status of whether an individual has sepsis or not before the observation of the onset time of sepsis (t_{sepsis}) is unknown. Therefore, the definition of the *SepsisLabel* is artificially designed to represent this information. Here we adopt the definition used in the PhysioNet Challenge 2019, where an individual is considered to have sepsis starting from 6 hours before the onset of sepsis. We present the relationship between *SepsisLabel* and t_{sepsis} as follows:

$$SepsisLabel = 1, \quad \text{if} \quad t \geq t_{sepsis} - 6$$
$$SepsisLabel = 0, \quad \text{if} \quad t < t_{sepsis} - 6$$

The aim of the study is to diagnose septic individuals as early as possible, and it is proposed that 6 hours earlier than the current diagnose time is the optimal time (see Reyna et al. (2019a) and Reyna et al. (2019b)). We pre-possessed the labels given in the dataset and the time to the onset of sepsis used to fit our joint model is 6 hours earlier than given in the dataset.

In this thesis we propose a new framework for the sepsis early diagnosis. We investigate the associations between the value of covariates and the onset time of sepsis with the presence of non-sepsis as the competing event. We define the disease progression states according to the diagnosis provided by the sequences of labels in the dataset: for observations during the follow-ups, all individuals with label 0 are at the risk of having sepsis, then the state is $s = 1$ (under risk); if the label is 1, then the state of the individual becomes $s = 2$ (sepsis onset). The state $s = 3$ (non-sepsis) is defined for the last observation for individuals diagnosed to be non-sepsis in the end (See Figure 2.3).

The data analysis in this study was only conducted using the clinical data collected from hospital A. In particular, a randomly selected subsample of 1000 individuals is used for our analysis. The decision to limit the sample size is explained in Section 5.5. The histogram in Figure 2.9 shows the observation frequencies of the provided sample. It is obvious that the follow-up time for most of the individ-

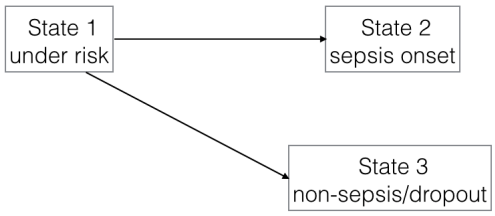


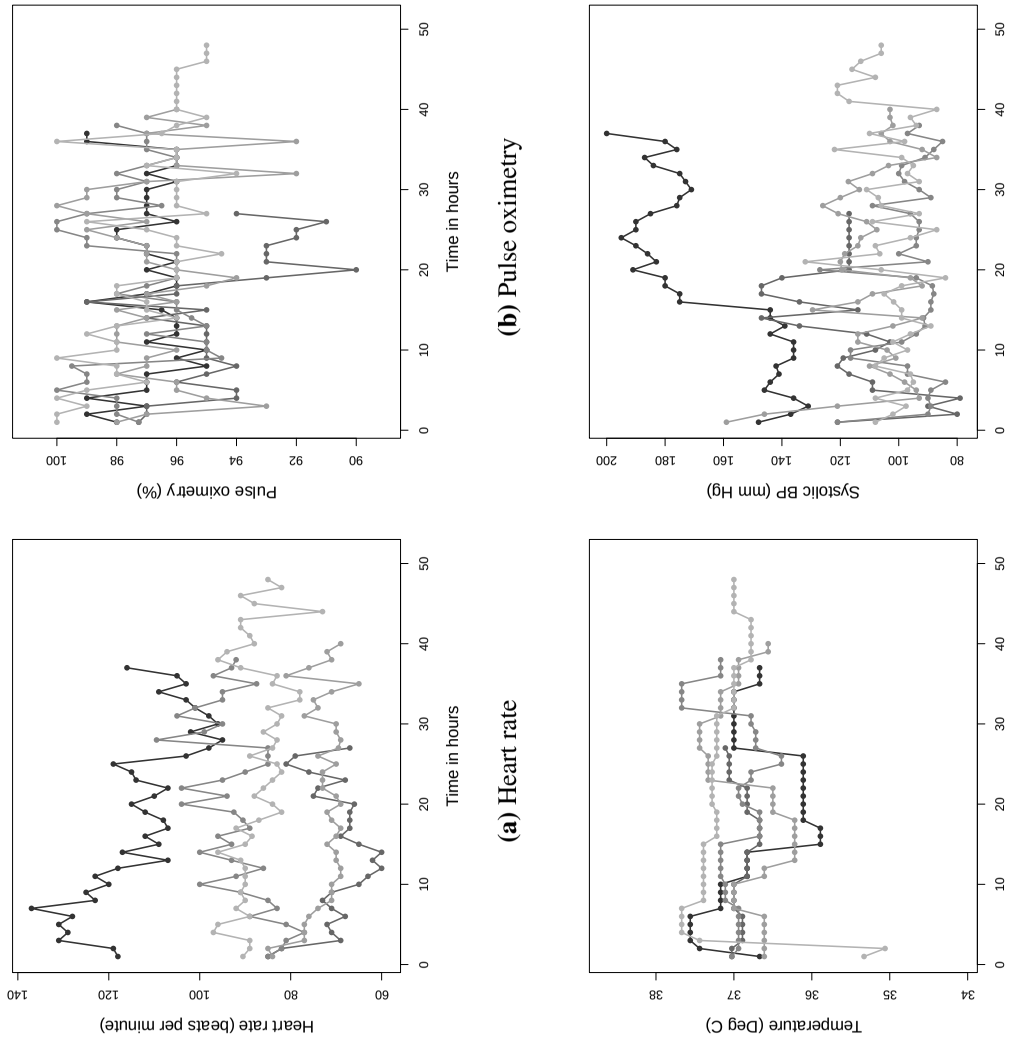
Figure 2.3: Transitions in the three-state model for the sepsis early diagnosis.

State	Transition to State 1	Transition to State 2	Transition to state 3
1	36848	78	922

Table 2.4: Frequencies of transitions from one state to another in the sepsis data.

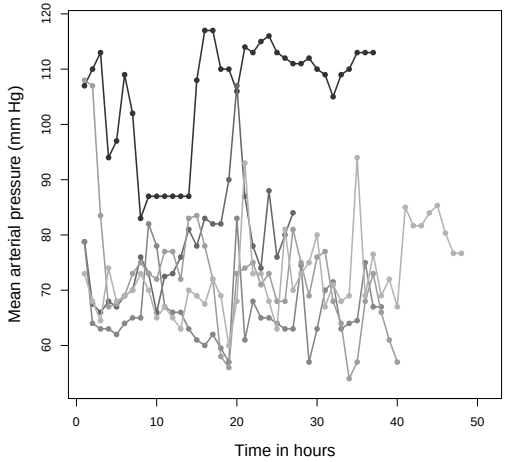
uals is around 25 hours. Table 2.4 displays the frequencies of transitions from one state to another. This table shows that there are 78 individuals who experienced the onset of sepsis, while 922 individuals did not have sepsis. The sample size was determined based on the high requirement for computing the marginal likelihood. In future studies, employing a larger sample size is recommended to enhance the analysis further.

In this thesis, we consider the sepsis state and the non-sepsis state as competing risks, and we model the competing risks using a three-state model. We model these vital signs as time-dependent covariates using the joint model. In Chapter 5, we fit joint models to the vital sign *respiration rate*. Different distributions and common variance structures are used. In Chapter 6, we fit joint models for bivariate covariates to the vital signs *heart rate* (HR), *mean arterial pressure* (MAP), and *respiration rate* (Resp). We select the pair of vitals signs that has the best performance, then illustrate the framework of sepsis early diagnosis based on joint model for this specific pair of vital signs.

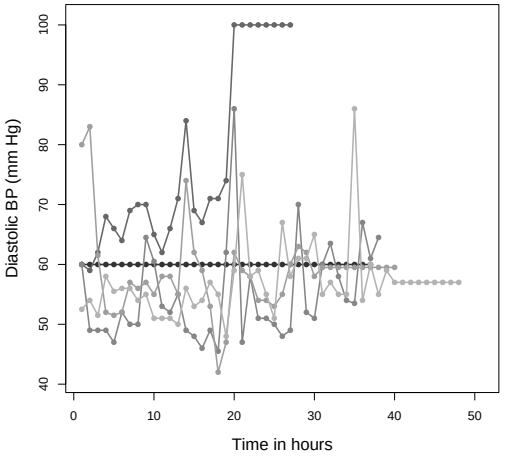


(a) Heart rate **(b)** Pulse oximetry **(c)** Temperature **(d)** Systolic blood pressure

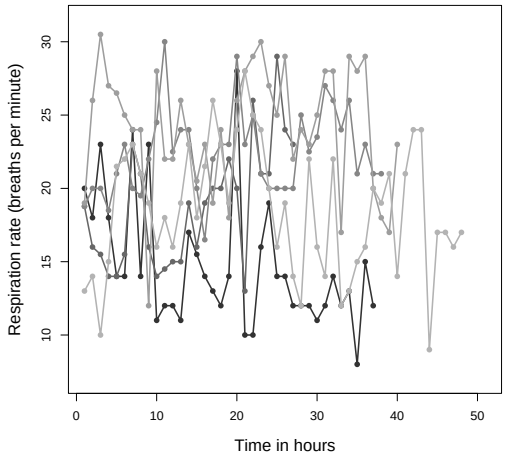
Figure 2.4: Trajectories of 4 (out of 7) vital signs of 5 randomly selected individuals.



(a) Mean arterial pressure

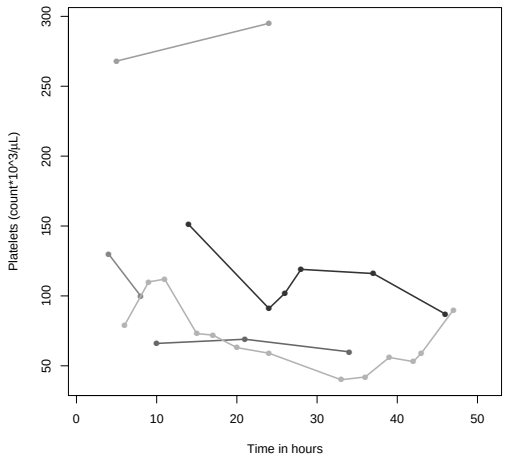


(b) Diastolic blood pressure

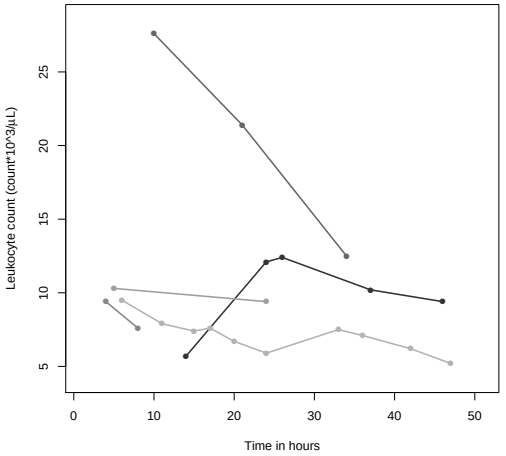


(c) Respiration rate

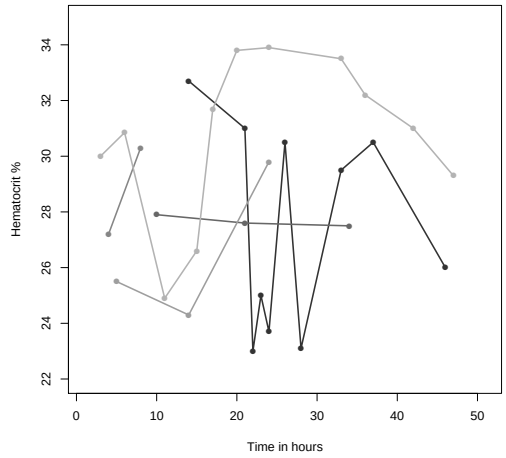
Figure 2.5: Trajectories of 3 (out of 7) vital signs of 5 randomly selected individuals.



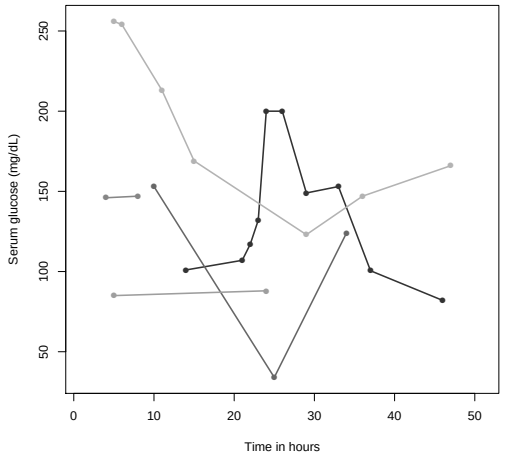
(a) Platelets



(b) Leukocyte count

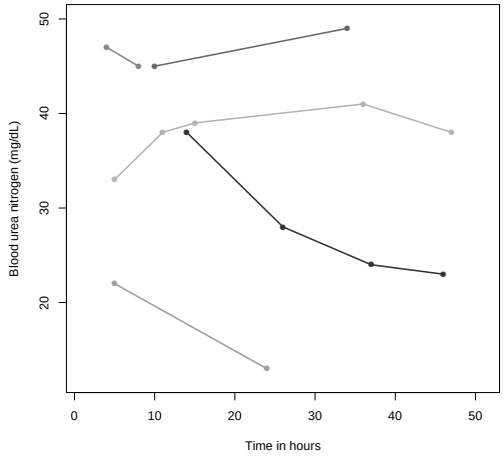


(c) Hematocrit

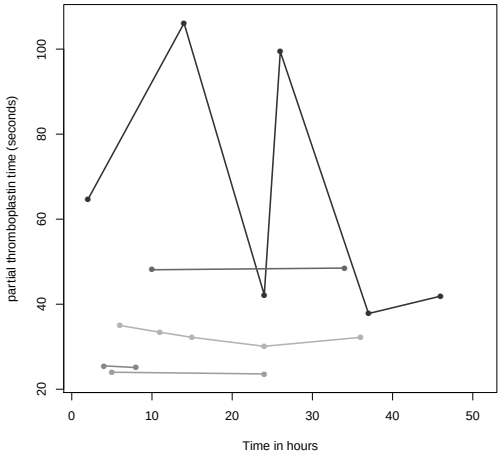


(d) Serum glucose

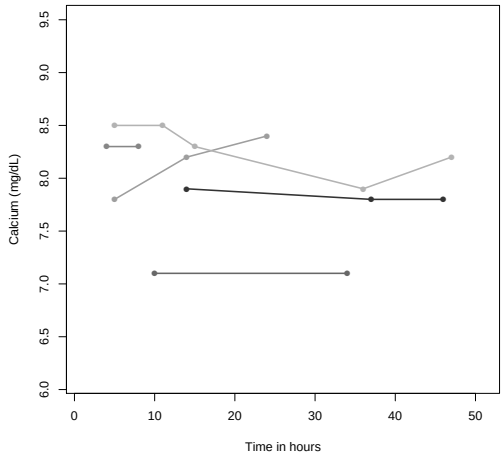
Figure 2.6: Lab sign trajectories of 5 randomly selected individuals. Lab signs include partial thromboplastin time, Leukocyte count, Hematocrit, and Glucose Serum glucose.



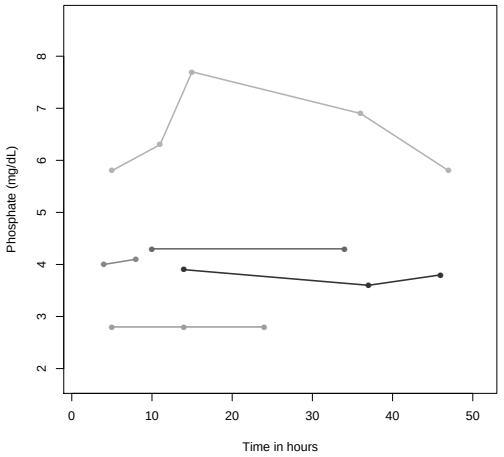
(a) Blood urea nitrogen



(b) partial thromboplastin time



(c) Calcium



(d) Phosphate

Figure 2.7: Lab sign trajectories of 5 randomly selected individuals. Lab signs include Blood urea nitrogen, partial thromboplastin time, Calcium, and Phosphate.

Figure 2.8: A pairwise scatter plot matrix of the initial recordings of 7 vital signs for 500 randomly selected individuals. Names of the vital signs are indicated on the diagonal panels. Vital signs are Heart rate (HR), Pulse oximetry (O2Sat), Temperature (Temp), Systolic Blood Pressure (SBP), Mean arterial pressure (MAP), Diastolic Blood Pressure (DBP), and Respiration rate (Resp).

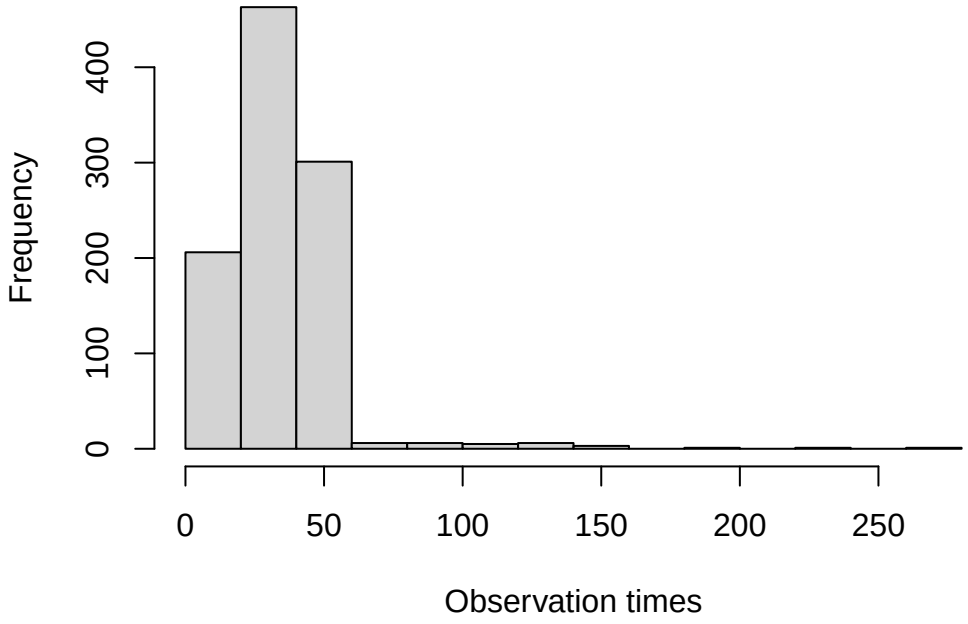


Figure 2.9: The distribution of follow-up times in the selected subsample.

2.2.2 Early Diagnosis of Sepsis

In order to assess the performance of the prediction models in terms of both the accuracy and the capacity of early diagnose, a normalised utility score was proposed in the Pysionet Challenge. Reyna et al. (2019b) mentioned the two advantages of applying this utility score. First of all, this metric is specifically designed to encourage early predictions, while penalising late or missed predictions as well as false alarms. It takes into account the clinical utility of sepsis detection and treatment, providing a comprehensive measure. Second of all, conventional scoring metrics like the Area Under the Curve (AUC) do not offer explicit incentives for early detection or penalties for false alarms or overtreatment.

The normalised utility score is defined as follows:

$$U_{normalised} = \frac{U_{total} - U_{noprediction}}{U_{optimal} - U_{noprediction}}, \tag{2.1}$$

$$U_{total} = \sum_{i=1}^{N} \sum_{t=1}^{T_i} U(i,t). \tag{2.2}$$

The term $U(i,t)$ in equation (2.2) is the utility score for each prediction, i.e. the score for the prediction of individual i at time point t, and U_{total} is the total score for a diagnosis model over N individuals. $U_{noprediction}$ is the utility score for a diagnosis model that returns only label 0. $U_{optimal}$ represents the utility score that the best diagnosis model can obtain (i.e. model that can predict every observations for all individuals correctly). $U_{total} - U_{noprediction}$ represents the 'distance' in the utility score between the diagnosis method and the method only predicts 0. Therefore the closer T_{total} to $U_{optimal}$, the closer $U_{normalised}$ to 1. $U_{normalised}$ is a measure of how well a diagnosis model is compared with the best model.

It is clear that the utility score $U(i,t)$ is a variable that depends on both the individual i and the time t. It serves the purpose of quantifying the desirability or utility associated with a specific outcome or decision. In our context, a function quantifying the utility of a diagnosis made at each time points is needed. Given that individuals have different sepsis onset times, the utility score should be formulated as a function of the individuals i, reflecting their unique characteristics. Additionally, it should also consider the perception of the ideal time for diagnosis, therefore incorporating the time variable t.

When using a method that can diagnose sepsis for each individual at different time points, the utility score for each prediction can be calculated. The total utility score can serve as a metric that can capture the trade-offs between early and late diagnoses, offers insight into the overall performance of the method.

The two graphs in Figure 2.10 illustrate the utility function proposed in the PhysioNet Challenge 2019. The lower graph illustrates the utility score for an individual without sepsis. The true diagnosis outcomes, indicated by the gray dots, should consistently be negative (represented as 0). For an incorrect prediction (where the predicted diagnosis is positive, represented as 1), the utility function assigns a score of -0.05 to this prediction. Conversely, if the prediction is accurate, the utility function assigns a score of 0.

The situation becomes more complicated when considering septic individuals. The upper graph depicts the utility function for an individual who experiences sepsis onset at hour 48. According to the definition of Reyna et al. (2019b), the optimal time for diagnosis is six hours prior to sepsis onset. At this specific time point, the utility score for a correct prediction will reach its maximum value. However, any false predictions made after this time point will be penalised. The red dots on the graph represent the rewards assigned for correct predictions. On the other hand, the blue dots represent the penalisation applied for incorrect predictions. it should be noted that the upper plot of Figure 2.10 only depicts the utility score for a septic individual with $t_{sepsis} = 48$, different individuals will have different values for t_{sepsis}.

This utility score serves to encourage early detection by rewarding correct predictions made closer to the optimal diagnosis time and penalising late detection, thereby promoting hourly prediction of sepsis.

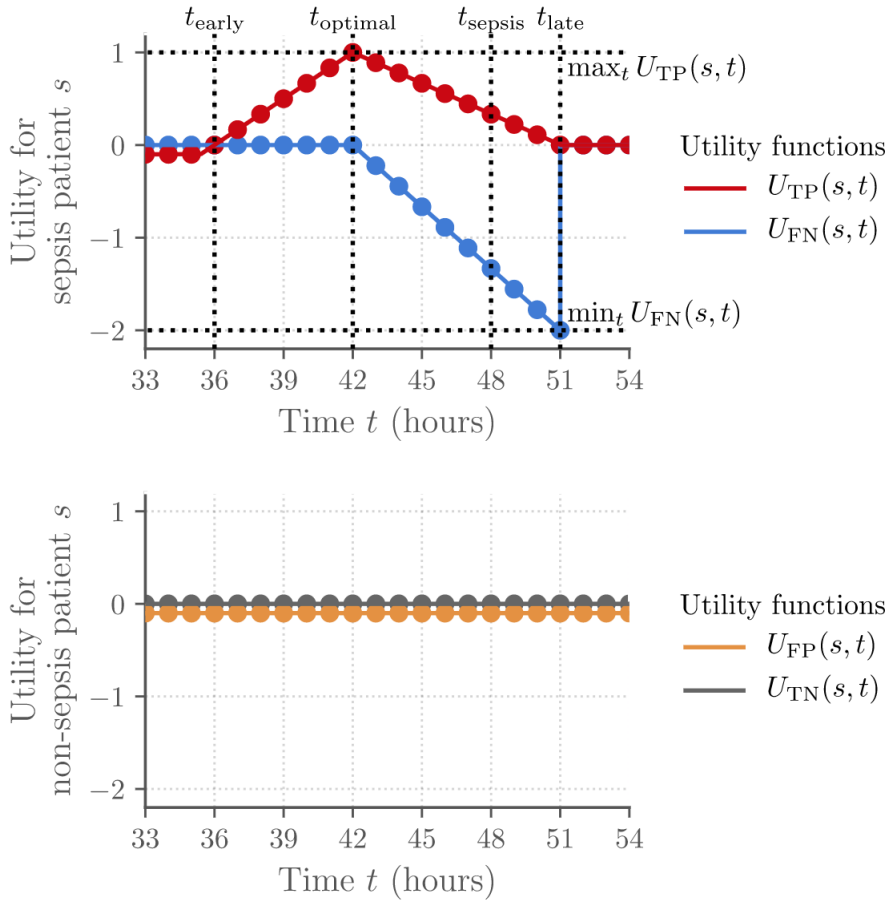


Figure 2.10: Utility function proposed by PysioNet Challenge. Taken from Reyna et al. (2019b)

Figure 2.10 illustrates that the utility function of each individual can be described using four terms: $U_{TP}(s,t)$, $U_{FN}(s,t)$, $U_{FP}(s,t)$ and $U_{TN}(s,t)$. Let $U(s,t)$ denote the utility score of the prediction for each individual s at time point t. The variable s in this function carries the same meaning as the previously defined i, both representing the individual. The utility score $U(s,t)$ equals:

$U_{TP}(s,t)$, positive prediction at time t for sepsis individual s,

$U_{FN}(s,t)$, negative prediction at time t for sepsis individual s,

$U_{FP}(s,t)$, positive prediction at time t for non-sepsis individual s,

$U_{TN}(s,t)$, negative prediction at time t for non-sepsis individual s.

The term 'positive prediction' refers to the diagnosis of sepsis, while the term 'negative prediction' refers to the diagnosis of non-sepsis. It should be noted that the utility function $U(i,t)$ (or $U(s,t)$) is different for septic and non-septic individuals.

In order to address early diagnosis, in this thesis, we propose the utilisation of a time-dependent loss function that promotes early detection and penalises delayed detection. The sepsis is diagnosed by minimising the expected loss, and the expected loss can also encourage the early diagnosis and penalise the late diagnosis. In order to accomplish this, it is necessary to establish a time-dependent loss function. Ideally, a method that attains the highest utility score should also minimise the loss. In this thesis, we construct the time-dependent loss function based on the utility score provided by the Physionet Challenge 2019.

By defining the loss function based on the utility score, we establish a direct link to the decision-making process. Through the optimisation of the loss function, our objective is to generate predictions that align with the desired outcomes and preferences outlined by the utility function.

Based on the information provided in Figure 2.10, we write out the function of the utility score explicitly.

First of all, we write out the utility score for non-septic individuals. According the bottom graph in Figure 2.10 the utility score is:

$$U(i,t) = \begin{cases} U_{FP}(i,t) = -0.05 & \text{for false positive prediction,} \\ U_{TN}(i,t) = 0 & \text{for true negative prediction,} \end{cases} \tag{2.3}$$

where $U_{FP}(i,t)$ is the utility score when a non-septic individual is misclassified to a septic individual, and $U_{TN}(i,t)$ is the utility score for a correct non-septic diagnosis for this individual. The utility scores for non-septic individuals are defined to be constants. From Equation (2.3), it is clear that the misclassication of a non-septic individual to be septic will be penalised. While on the contrary, correctly classify a non-septic individual will neither be rewarded or penalised.

Now let us rewrite the utility score of septic individuals. The top graph in Figure 2.10 present the utility function of an individual who has $t_{sepsis} = 48$. However,

different individual will have different t_{sepsis}. Therefore, the utility score of septic individuals is a function defined with respect to the time of prediction and time of diagnosis by the clinicians. Let t_{sepsis} denote the time that the septic individual i is diagnosed by clinicians (the index i is omitted for simplicity). As seen in Figure 2.10, it is considered to be ideal to diagnose the individual 6 hours before the clinicians, denoted by $t_{optimal} = t_{sepsis} - 6$. For the septic individuals, non-septic prediction after $t_{optimal}$ is increasingly penalised, and the reward for correct diagnosis is peaked at $t_{optimal}$. The utility score for septic individuals consists of

$$U(i,t) = \begin{cases} U_{TP}(i,t) & \text{for false positive prediction,} \\ U_{FN}(i,t) & \text{for true negative prediction,} \end{cases}$$

where $U_{FN}(i,t)$ is the utility score for the prediction misclassified a septic individual to be non-septic at time t, and and $U_{TP}(i,t)$ is the utility score when the individual is correctly diagnosed. Therefore, we rewrite the functions of $U_{TP}(i,t)$ and $U_{FN}(i,t)$ are follows:

$$U_{TP}(i,t) = \begin{cases} -0.05 & (t \leq t_{sepsis} - 12) \\ 1/6 & (t = t_{sepsis} - 11) \\ 2/6 & (t = t_{sepsis} - 10) \\ \ldots & \ldots \\ 1 & (t = t_{optimal}) \\ 8/9 & (t = t_{optimal+1}) \\ 7/9 & (t = t_{optimal+2}) \\ \ldots & \ldots \\ 0 & (t \geq t_{sepis} + 3). \end{cases} \tag{2.4}$$

$$U_{FN}(i,t) = \begin{cases} 0 & (t \leq t_{optimal}) \\ -2/9 & (t = t_{optimal+1}) \\ -4/9 & (t = t_{optimal+2}) \\ \ldots & \ldots \\ -2 & (t \geq t_{sepis} + 3). \end{cases} \tag{2.5}$$

It should be noted that we directly adopted the settings in the Physionet Challenge that 6 hours before the clinicians is the optimal time of diagnosis. As an acute disease, sepsis can progress rapidly, by diagnosing sepsis within 6 hours, patients who are at a higher risk of developing severe complications can receive immediate and appropriate care. As shown in Figure 2.9, the follow-up time for most of the individuals is around 25 hours. Therefore, selecting a 6-hour optimal diagnosis time is a reasonable choice considering the time-sensitive nature of sepsis and the need for early intervention.

Additionally, we propose the use of a loss function as a framework for disease diagnosis. This method can be further extended to incorporate utility scores of different shapes tailored to specific situations as specified by clinicians and domain experts.

In Chapter 5, we will propose a framework for sepsis early diagnosis. The framework consists of two-steps. The first step is fitting a joint model, and the second step is calculating an expected loss. The loss function we proposed is defined based on Equations (2.3), (2.4) and (2.5).

2.2.3 Notation

The j^{th} historical observations of each individual i till time j are denoted by $\mathcal{H}_{ij} := [\mathbf{z}_i, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{ij}], j = 1, 2, \ldots, J_i$, where $\mathbf{z}_i \in \mathcal{R}^{6\times 1}$ represents the demographic variables the individual i, and $\mathbf{y}_{ij} \in \mathcal{R}^{34\times 1}$ is the time-dependent covariates observed at j^{th} time.

In ideal case, the time-dependent covariates are measured hourly, from time $t = 1$ to $t = T_i$, the maximum observation time of individual i. Let u_j denote the time where the j^{th} observation occurs. In addition, for each individual, a sequence of label D_{ij} provided by the clinicians indicates whether the individual is diagnosed to be septic or not at the corresponding time. $D_{ij} = 1$ indicates the individual i is diagnosed to have sepsis at the j^{th} observation. On the contrary, $D_{ij} = 0$ indicates the individual does not have sepsis.

The aim of this early diagnosis of sepsis is to develop a method such that for

a individual i at any time $t, 0 < t < T_i$, the method is able to predict the label D_{ij} based on the historical data $\mathcal{H}_{ij}$, where $u_j \leq t$.

Chapter 3

Competing Risks and General Mixed Models

3.1 Competing Risks

In survival analysis, we often encounter situations where multiple events can occur, and when one event happens, it prevents the occurrence of other events. These competing events are common in fields like medical research, where individuals may die from different causes. Recognizing and accounting for competing events is important to fully grasp the complexity of the situation and accurately interpret the data.

It is important to explore models that consider the relationship between competing risks. The analysis of competing risks using multi-state models is discussed in studies such as Putter et al. (2007) and Andersen et al. (2002).

Suppose that the total number of events is K, then the model of $K+1$ states can be applied, where state 0 represents the initial state of each individual, and state 1 to state K represents the occurrence of each event. It is a special case of multi-state models since the transition between states are simply governed by the transition intensity matrix $\mathbf{Q}$, where $\mathbf{Q}$ is equivalent to an square matrix with dimension $K+1$ with its first row equals the transition intensities $q_{01},\ldots,q_{0K}$ and all the rest elements equals to 0.

3.2 Multi-State Model

3.2.1 Introduction

Multi-state models are commonly used in the analysis of a process in which individuals transit between various states. The information about the process usually consists of the records of the change of state and other outcomes of the individuals during the process. For instance, when modelling the disease progression, the disease stages can be considered as different states, and then the course of the disease can be described by the multi-state model. A continuous-time multi-state model is defined by the transition intensity functions, which describes the change of state over time.

Suppose there are K different states $r = 1,\ldots,K$. Let $\zeta(t)$ denote the state occupancy of an individual at time t, and let $\mathcal{H}(t)$ represents the history of state occupancy over time interval $[0,t)$. A multistate-state model in continuous-time with K different states can be specified using transition intensity functions, which is defined as:

$$q_{rs}(t|\mathcal{H}(t)) = \lim_{\Delta t \to 0} \frac{P\Big(\zeta(t+\Delta t) = s | \zeta(t) = r, \mathcal{H}(t)\Big)}{\Delta t}, \tag{3.1}$$

for transitions between state r and state s, where $r \neq s$. The transition intensity (3.1) is the instantaneous rate at which the individual will move from state r to state s at time t. One important property of transition intensities is as follows:

$$q_{ss}(t) = -\sum_{r \neq s} q_{rs}(t). \tag{3.2}$$

For convenience, the transition intensities are often presented in the form of a matrix. Let $\mathbf{Q}$ denote the transition intensity matrix, the element of the matrix is the corresponding transition intensity, i.e. $[\mathbf{Q}(t)]_{r,s} = q_{rs}(t)$. And according to the property (3.2), all rows of $\mathbf{Q}(t)$ sum to zero.

For a multi-state process, it is often of interest to investigate how the transition intensities are related to covariates. This can be achieved by specifying the transi-

tion intensities as functions of covariates.

3.2.2 Modelling Effects of Covariates

Let $\mathcal{H}_i(t) := [\mathbf{z}_i, \mathbf{y}_i(1), \ldots, \mathbf{y}_i(t)]$ represents the covariates observed of an individual over time period $[0,t]$, where $\mathbf{z}_i$ are the constant covariates and $\mathbf{y}_i(1), \ldots, \mathbf{y}_i(t)$ are the time-dependent covariates. For simplicity we omit the index i. In order to take into account the covariates, it is common to define the transition intensities as follows:

$$q_{rs}(t|\mathcal{H}(t)) = q_{0.rs}(t) \exp\left(\boldsymbol{\alpha}_{rs}^{\top}\mathbf{y}(t) + \boldsymbol{\lambda}_{rs}^{\top}\mathbf{z}\right). \tag{3.3}$$

The term $q_{0.rs}(t)$ in equation (3.3) is the baseline intensity, which is the transition intensity at time t for individual with covariates $\mathbf{y}(t) = \mathbf{0}, \mathbf{z} = \mathbf{0}$. The second term is a log-linear model of the covariates, where $\boldsymbol{\alpha}_{rs}$ and $\boldsymbol{\lambda}_{rs}$ are the regression coefficients of the corresponding covariates.

If the model follows the continuous-time assumption, i.e. time $t \in \mathcal{R}_{\geq 0}$, consequently the covariates change with respect to time should be defined on all non-negative real numbers. However, in practice, the value of the covariates are only available at the observation times, which makes it impossible to directly calculate the transition intensities at the time apart from the observation time points.

To solve this problem, sometimes the piecewise-constant assumption is applied. Take the transition intensity (3.3) for example. Suppose covariates $\mathbf{y}(t)$ are only available at time points $t = u_j, j = 1, \ldots, J$. By applying the piecewise-constant assumption, it is assumed that $\mathbf{y}$ remains constant within intervals $[u_j, u_{j+1})$. Then the transition intensity at any time point t are:

$$q_{rs}(t|\mathcal{H}(t)) := q_{0.rs}(u_j) \exp\left(\boldsymbol{\alpha}_{rs}^{\top}\mathbf{y}(u_j) + \boldsymbol{\lambda}_{rs}^{\top}\mathbf{z}\right),\ u_a < t < u_b.$$

However, this assumption is less accurate if the trajectory of the time-dependent covariates $\mathbf{y}$ is less smooth and the observation time points are sparse.

As shown in the data description in Chapter 2, the clinical data of sepsis consists of time-dependent covariates which is highly fluctuated. The piecewise

constant approximation does not apply to our situation. In this thesis, we utilise the continuous-time multi-state model within the joint model framework. The time-dependent covariates are modelled with mixed model, and their mean can be estimated at any time point. Therefore we did not use the piecewise constant assumption for the time-dependent covariates. However, as the covariates change continuously with time, the transition intensities defined on it will change with time as well. This introduces problems to the computing of the transition probabilities as there is no established method to directly derive transition probabilities for a changing transition intensity. A piecewise constant approximation for the transition probabilities is needed. The details will be explained in Section 3.2.3.

3.2.3 Transition Probability

Apart from transition intensities, it is also of interest to investigate other characteristics related to the multi-state process, for example, the transition probabilities. The transition probabilities are also of importance because the likelihood function of the multi-state model is based on them.

The transition probability of moving from state $\zeta(t)$ to state $\zeta(t+\Delta t)$ over time-interval $(t, t+\Delta t)$ is defined as:

$$p_{rs}(t, t+\Delta t|\mathcal{H}(t)) = P\Big(\zeta(t+\Delta t) = s|\zeta(t) = r, \mathcal{H}(t)\Big). \tag{3.4}$$

Similar to the transition intensities, a transition probability matrix is constructed with its elements represents corresponding transition probabilities between different states. Let $\mathbf{P}(t_1, t_2)$ denote the transition probability matrix over time interval $(t, t+\Delta t)$, where $[\mathbf{P}(t_1, t_2)]_{r,s} = p_{rs}(t_1, t_2)$. Then according to the properties of the transition probabilities, all elements of matrix $\mathbf{P}$ are non-negative and all rows of $\mathbf{P}$ sum to 1. The transition probabilities are calculated from the transition intensities.

Suppose the following assumptions hold for a multi-state model:

1. It is a continuous-time model, transition can take place at any $t > 0$.

2. The Markovian property holds. Therefore the transition probabilities satisfies:

$$P\Big(\zeta(t+\Delta t)=s|\zeta(t)=r,\mathcal{H}(t)\Big)=P\Big(\zeta(t+\Delta t)=s|\zeta(t)=r\Big).$$

3. It is time-homogeneous, i.e. the transition probability $P\Big(\zeta(t+\Delta t)=s|\zeta(t)=r\Big)$ does not depend on the time t.

Then this multi-state model is a time-homogeneous continuous-time Markov model (Ross (2014)), and the transition intensities can be simplified to:

$$q_{rs}(t|\mathcal{H}(t))=q_{rs}(t)=\lim_{\Delta t\to 0}\frac{P\Big(\zeta(t+\Delta t)=s|\zeta(t)=r\Big)}{\Delta t}.$$

Solving the matrix differential equations, the transition probability can be written as a matrix exponential of the transition intensity matrix (see Ross (2014) and Van den Hout (2017)):

$$\begin{aligned}\mathbf{P}(t,t+\Delta t)&=\exp\Big(\Delta t\mathbf{Q}\Big)\\&=\mathbf{I}+\Delta t\mathbf{Q}+\frac{1}{2!}\Big(\Delta t\mathbf{Q}\Big)^2+\ldots,+\frac{1}{n!}\Big(\Delta t\mathbf{Q}\Big)^n.\end{aligned}\tag{3.5}$$

In practice when the size of the state space is large, it is computationally infeasible to evaluate equation (3.5) directly. Different techniques have been proposed to approximate the transition probabilities and the matrix exponential. See, for instance Sidje and Stewart (1999) and Reibman and Trivedi (1988) for more details.

Now consider a continuous-time multi-state model with transition intensities depend on covariates, and the covariates are only available at time points $u_j, j=1,\ldots,J$. Suppose the transition intensities are defined as parametric functions of the time-dependent covariates using equation (3.3). Since the Markovian property and the time-homogeneous assumption are violated, the transition probability can no longer be calculated with equation (3.5). Similar to Section 3.2.2, we can apply the piecewise-constant approximation to the transition probabilities and assume it remain constant within the time intervals $[u_j,u_{j+1})$. Therefore within the intervals

$[u_j, u_{j+1})$ the equation (3.5) still holds:

$$\mathbf{P}(u_j, u_{j+1}) = \exp\Big((u_{j+1} - u_j)\mathbf{Q}(u_j)\Big)$$

The transition matrix over any time interval $[t_1, t_2)$ is then approximated by the multiplication of a sequence of transition matrices:

$$\mathbf{P}(t_1, t_2) = \mathbf{P}(t_1, u_{i+1}|Q(u_i))\mathbf{P}(u_{i+1}, u_{i+2}|Q(u_{i+1}))\ldots\mathbf{P}(u_{j-1}, t_2|Q(u_{j-1})),$$

where $u_i \leq t_1 < u_{i+1}$ and $u_{j-1} \leq t_2 < u_j$.

3.2.4 Parameter Estimation

The parameters of the model can be estimated by maximizing the likelihood function, which can be derived using the transition probabilities and transition intensities.

Let $\boldsymbol{\theta}$ denote the set of all the parameters used to define the multi-state model. For a dataset consists of N individuals, the total likelihood is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} L_i(\boldsymbol{\theta}). \tag{3.6}$$

Therefore according to the maximum likelihood estimation, the estimated parameters are obtained by maximizing the joint likelihood of all individuals:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \prod_{i=1}^{N} L_i(\boldsymbol{\theta}), \tag{3.7}$$

where $L_i(\boldsymbol{\theta})$ is the likelihood of individual i.

Assume that the the individual i transits from state r to state s exactly at time points t_{ik}. The likelihood contribution of individual i is given by

$$L_i(\boldsymbol{\theta}) = \prod_{r\neq s}\Big[\prod_{\forall t_{ik}} q_{rs}(t_{ik}|\mathcal{H}_i(t_{ik}))\Big]\exp\Big(\int_0^{\infty} \mathbb{1}_{\{s(u)=r\}} q_{rr}(u|\mathcal{H}_i(u))du\Big). \tag{3.8}$$

See more details about the deriving of the likelihood in, for instance, Cook and

Lawless (2018).

The numerical methods used to estimate parameters $\boldsymbol{\theta}$ are introduced in Section 3.5. In addition, general multi-state models can be fitted using R packages such as `msm` (Jackson et al. (2011)).

3.3 Mixed Models for Time-Dependent Covariates

As discussed in Section 3.2.2, the transition intensity of an individual is a function of the covariates. Therefore to predict the transition intensities in the future, the value of these covariates need to be predicted first. Therefore we need a model to describe the changes in the covariates with respect to time. To achieve this the linear mixed model is used in this study. In the following sections, the linear mixed model will be introduced, and parameter estimation will be discussed.

3.3.1 Mixed Model

Regression models are used to predict the quantitative outcome y_i given a set of observations $\mathbf{w}_i \in \mathcal{R}^{1\times p}$. For example the linear regression model with a single covariate w_i:

$$y_i = \beta_0 + \beta_1 w_i + \varepsilon_i, \tag{3.9}$$

where β_0 is the intercept, β_1 is the slope, and y_i's are assumed to be conditionally independent given w_i. In this model, ε_i is called the error term, and follows a normal distribution $\varepsilon_i \overset{i.i.d}{\sim} N(0,\sigma^2)$, where σ is a constant.

In our study, the time-dependent covariates were measured repeatedly for each individual, therefore there exists correlation between observations from the same individual. The assumption of independence is violated. The mixed model is then applied. Let y_{ij} represents the j-th repeated observation of individual i. For exam-

ple, the linear mixed model is a simple extension of the linear model above:

$$\begin{aligned} y_{ij} &= b^*_{i0} + b^*_{i1} w_{ij} + \varepsilon_{ij}, \\ b^*_{i0} &= \beta_0 + b_{i0}, \\ b^*_{i1} &= \beta_1 + b_{i1} \\ \mathbf{b}_i^\top &= (b_{i0}, b_{i1})^\top \sim N(0, \Sigma), \qquad \varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma_e^2), \end{aligned} \tag{3.10}$$

where σ_e is a constant. Similar to the linear regression model, β_0 and β_1 are the population mean of the intercept and the slope, and ε_{ij} are the random noises. Terms b_{i0} and b_{i1} are random effects that follow a bivariate normal distribution with mean equals $\mathbf{0}$ and the variance-covariance matrix Σ.

By introducing the random effects $\mathbf{b}_i$, The linear mixed model allows for modelling the between- and within-individual variations. From the equations (3.10) we have:

$$\begin{aligned} E[Y_{ij} | b_{i0}, b_{i1}] &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) w_{ij}, \\ E[Y_{ij}] &= \beta_0 + \beta_1 w_{ij}. \end{aligned} \tag{3.11}$$

It is clear that b_{i0} is the difference between the intercept of individual i and the mean intercept, and b_{i1} is the difference between the slope of individual i and the mean slope.

The general form of the mixed model can be obtained by extending the linear mixed model above. For response variables Y_{ij} violates the normality assumptions, for instance, it follows a distribution from the exponential family. One can specify a link function $g(\cdot)$ and assumes the relationship between the response variable and the covariates as

$$E[Y_{ij}] = g(\beta_0 + \beta_1 w_{ij}).$$

We can further extend this model by assuming the distribution of the response variable does not follow any distributions from the exponential family, e.g. the

skew normal, skew t distributions (Azzalini (1985)).

3.3.2 Mixed Model for Location, Scale, and Shape

The common linear model assumes the residuals follows a conditional normal distribution with a constant. However, in practice, the residuals always violate this assumption because the response variables always have much more complicated structures and are difficult to be described by the family of the most popular linear mixed models. For instance, in the ELSA data, we observe a decline in the mean number of animal names recalled as individuals age, accompanied by a decrease in variance. This suggests that the variability in the outcome may not remain constant over time and could be influenced by the predictor variables. Accounting for changes in variance over time improves the robustness and accuracy of the model. In addition, we want to further extend the generalisability of our joint model by relaxing the normal assumption of the residuals.

It is very common that in the study of clinical data that the observed records usually have special meanings and are often positive, for example the number of words recalls, the heart rate, the blood pressure, etc. And they can have complicated distributions such as positive skewed, zero-inflated, multi-modal, etc. Mihaylova et al. (2011) summarised different methods to tackle these difficulties. In our study, we adopted the framework from Rigby et al. (2019) to model the time-dependent covariates.

GAMLSS is proposed for the univariate regression. In this framework the response variables and the explanatory variables can be linked not only by linear functions, but also can be non-linear or even non-parametric. The GAMLSS approach improved the common GLM in two ways. First of all, the response variable is allowed to follow any distribution, rather than distributions from the exponential family. Secondly, not only the location parameter, but any parameter controls the distribution, e.g. the parameter for the scale and shape can be linked to the explanatory variables with any kind of links. This framework provides much more flexibility compared with the common LMs, GLMs, and GAMs.

In our study, for simplicity we only consider the parametric models. The extensions to the non-parametric models could be a valuable improvement in the future.

Based on the framework of GAMLSS, we assume the error term follows a distribution controlled by three parameters: the location, scale, and shape parameters. For example, a skew t distribution is controlled by the parameters for the location, sclae, and the skewness. The parameters can be modelled by a constant, for a function of the linear predictor. For example, in the study of ELSA data, we assume the error term of the number of animal names people can recall follows a skew normal distribution, and the variation of it decrease as age grows:

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim SN(0, \nu, \sigma_{ij}) \\ \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) age_{ij}, \\ \log(\sigma_{ij}) &= \gamma_0 + \gamma_1 age_{ij}, \end{aligned} \tag{3.12}$$

In particular we considered to use the skew t distribution and the skew normal distribution to model the response variables. For comparison, we also fitted our model with the normal distribution.

In this study we use the Type 1 skew normal distribution from the GAMLSS package. This distribution is controlled by three parameters: μ, σ and ν. Parameter μ is the parameter for the location, it describes the center of the distribution. The parameter σ is for the scale parameter, describes the the spreadness or the variability for the distribution. The parameter ν describes the shape of the parameter. Notice that when $\nu = 0$, we re-obtain the normal distribution. The density function for the Type 1 skew normal distribution is as follows:

$$f(y|\mu, \sigma, \nu) = \frac{2}{\sigma}\phi(\frac{y-\mu}{\sigma})\Phi(\nu\frac{y-\mu}{\sigma}), \tag{3.13}$$

where the range of the parameters are$(-\infty < \mu < +\infty)$, $\sigma > 0$, and $(-\infty < \nu < +\infty)$. The range of y is $(-\infty < y < +\infty)$. Notations $\phi(\cdot)$ and $\Phi(\cdot)$ represents the density function and the cumulative density function for the standard normal distribution.

The Type 1 skew t distribution is controlled by 4 parameters. Apart from the

parameters μ, σ and ν for location, scale, and shape, there is also a parameter τ controls the degree of freedom for the t-distribution. The distribution is given by:

$$f(y|\mu,\sigma,\nu) = \frac{2}{\sigma} f_{Z_1}(\frac{y-\mu}{\sigma}) F_{Z_2}(\nu \frac{y-\mu}{\sigma}), \tag{3.14}$$

where Z_1 and Z_2 have Student t-distribution with degrees of freedom greater than 0. More details about the distributions can be found in Rigby et al. (2019).

3.3.3 Parameter Estimation

Various methods can be applied for parameter estimation. The maximum likelihood estimation using the EM algorithm and restricted maximum likelihood estimation (RMLE). See, for instance Fitzmaurice et al. (2012). In this section, we will introduce the parameter estimation via maximizing the marginal likelihood.

Consider the model given by equation (3.10). Since $y_{ij}, j = 1,\ldots,n_i$ are independent condition on w_{ij} and $B_{i0} = b_{i0}, B_{i1} = b_{i1}$. The likelihood function of individual i is given by:

$$\prod_{j=1}^{n_i} f(y_{ij}|w_{ij}, B_{i0} = b_{i0}, B_{i1} = b_{i1}). \tag{3.15}$$

The marginal likelihood is given by marginalizing over the random variables B_{i0}, B_{i1}, given by:

$$L_i(\boldsymbol{\beta},\boldsymbol{\theta}) = \int \prod_{j=1}^{n_i} f(y_{ij}|w_{ij}, B_{i0} = b_{i0}, B_{i1} = b_{i1}) dG(b_{i0}, b_{i1}), \tag{3.16}$$

where $\boldsymbol{\beta}$ is the set of fixed coefficients for the linear mixed model, and $\boldsymbol{\theta}$ is the set of parameters govern the distribution $G(b_{i0}, b_{i1})$. Then both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated by maximizing the total marginal likelihood $\prod_{i=1}^{N} L_i(\boldsymbol{\beta},\boldsymbol{\theta})$.

Apart from parameter estimation, it is also of interest to estimate the individual-specific random effects b_{i0}, b_{i1} for individual i after the linear mixed model is fitted.

The conditional probability of the random effects given the data is given by:

$$f(B_{i0} = b_{i0}, B_{i1} = b_{i1} | y_{ij}, w_{ij}) = \frac{f(B_{i0} = b_{i0}, B_{i1} = b_{i1}, y_{ij} | w_{ij})}{f(y_{ij} | w_{ij})} \tag{3.17}$$

$$= \frac{L_i(\hat{\beta}, \hat{\theta}) g(b_{i0}, b_{i1} | \hat{\theta})}{\int L_i(\hat{\beta}, \hat{\theta}) dG(b_{i0}, b_{i1} | \hat{\theta})} \tag{3.18}$$

Then the individual-specific random effects b_{i0}, b_{i1} can be estimated by calculating the posterior mean or via the maximum a posteriori, i.e. the mode of the posterior distribution.

3.4 Joint Modelling of Competing Risks and Time-Dependent Covariates

In studies with both time-to-event data and repeatedly measured covariates, it is of interest to study the associations between them. The study of the joint modelling of longitudinal data and time-to-event data was greatly developed by Wulfsohn and Tsiatis (1997) and Faucett and Thomas (1996). In Rizopoulos (2012) and Tsiatis and Davidian (2004) the joint model for the time-to-event data and the longitudinal data was discussed in detail. Shared random effect models are one of the most commonly used joint models. It was mentioned that the advantage of applying the joint model is two-fold: (i) it provides an approach to take into account the time-dependent markers with noise when the primary interest is modelling the survival time; (ii) it is able to model is longitudinal data more accurately since it can take into account the effect of the associated events. In Ferrer et al. (2016) a shared random effects joint model for the longitudinal outcomes and multi-state process was discussed.

In this section, we present an example of the joint model consists of a the multi-state sub-model for the time-to-event data and a linear mixed model for the time-dependent covariate.

Suppose y_{ij} is the value of a time-dependent covariate observed at time j of

individual i. The linear mixed model for y_{ij} is given by:

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij} \\ \mu_{ij} &= b^*_{i0} + b^*_{i1} w_{ij} + \varepsilon_{ij} \\ b^*_{i0} &= \beta_{i0} + b_{i0}, \\ b^*_{i1} &= \beta_{i1} + b_{i1}, \\ (b_{i0}, b_{i1})^\top &\sim G(b_{i0}, b_{i1}),\ \varepsilon_{ij} \sim N(0, \sigma_e^2), \end{aligned} \tag{3.19}$$

where b_{i0} and b_{i1} are random effects jointly follows distribution $G(\cdot)$. For simplicity we assume the distribution $G(\cdot)$ is a bivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ. Notice that according to the model above, the term μ_{ij} is considered as the unobservable true value of the covariates. Suppose that $j = 1, \ldots, T_i$, the likelihood contribution from the sub-model for the time-dependent covariates is then given by:

$$\prod_{j=1}^{T_i} f(y_{ij} | w_{ij}, B_{i0} = b_{i0}, B_{i1} = b_{i1}). \tag{3.20}$$

For the modelling of the time-to-event data, instead of defining the transition intensity as a function of the value of the covariates, μ_{ij} is used. We build this model based on the established methods discussed in Ferrer et al. (2016). The two processes are therefore linked by the shared random effects. Recall the multi-state model described in Section 3.2.2, here for simplicity we assumed that time-dependent covariate y_{ij} is univariate, then the intensities of individual i at time t is therefore given by:

$$q_{rs}(t|\mu_{ij}) = q_{0.rs}(t) \exp\left(\alpha_{rs}\mu_{ij} + \boldsymbol{\lambda}_{rs}^\top \mathbf{z}_i\right), \tag{3.21}$$

where $j \leq t < j+1$. The likelihood contribution from the submodel for competing risks is

$$\prod_{r \neq s} \Big[\prod_{\forall t_{ik}} q_{rs}(t_{ik} | \mu_i(t_{ik})) \Big] \exp\Big(\int_0^\infty \mathbb{1}_{\{s(u)=r\}} q_{rr}(u | \mu_i(u)) du \Big). \tag{3.22}$$

Similar to Section 3.3, the marginal likelihood of individual i is given by:

$$L_i(\boldsymbol{\theta}) = \int \prod_{r \neq s} \Big[\prod_{\forall t_{ik}} q_{rs}(t_{ik}|\mu_i(t_{ik})) \Big] \exp\Big(\int_0^{\infty} \mathbb{1}_{\{s(u)=r\}} q_{rr}(u|\mu_i(u)) du \Big). \\ \times \prod_{j=1}^{T_i} P(y_{ij}|w_{ij}, B_{i0} = b_{i0}, B_{i1} = b_{i1}) dG(b_{i0}, b_{i1}). \tag{3.23}$$

All parameters $\boldsymbol{\theta}$ of the joint model are estimated by maximizing the total marginal likelihood function.

3.5 Numerical Methods

The maximum likelihood estimates of the parameters can be obtained by maximizing the likelihood function numerically. In R software (R Core Team (2016)), this can be achieved using the general-purpose optimisation function like `optim` and `nlminb`.

The standard errors can be estimated using the results obtained from the optimisation. The Hessian matrix can be at the estimated parameters, and the standard errors can be obtained by taking the square root of the diagonal elements of the inverse of the Hessian matrix.

For likelihood functions which involve integral, there are many numerical integration functions in R software can be used. For example, function `integrate` is for the one-dimensional integration, functions `int2` (Swihart and Lindsey (2016)) and `integral2` (Borchers (2019)) can perform two-dimensional integration. In addition, the Monte Carlo integration can be used to approximate the integration.

For marginal likelihood functions like equation (3.16) and equation (3.23), if we assume that the random effects (B_{i0}, B_{i1}) follows a multivariate normal distribution, then the integral can be approximated using the Gaussian-Hermite quadrature method.

The performance of Gaussian-Hermite quadrature approximation depends on the choice of the number of nodes. The value of the integrand are calculated on the grids. The approximation will be more accurate if more nodes are used. Through

out the data analysis in this thesis, we use 30 nodes for the Gaussian-Hermite quadrature. In Chapter 8 we present an example code for calculating the marginal likelihood using two-dimensional Gaussian-Hermite quadrature approximation.

In the current model we assume there are two random effects. However when the number of random effects increases, the time cost to calculate the integration by Gaussian-Hermite quadrature approximation will grow exponentially. For the parameter estimation of joint models, there exist different frameworks. In Ferrer et al. (2016) and Murray and Philipson (2022), the EM algorithm was used to alleviate the numerical problems. In Chapter 6 we will discuss this problem in detail. We use Bayesian inference instead to alleviate this problem. The Bayesian inference of joint model is presented in Chapter 6 as well.

Chapter 4

Joint Model for Cognitive Function Data

In this chapter we present a joint model of competing risks and longitudinal outcomes. The health status are considered as competing risks and modelled with a multi-state sub-model. The longitudinal outcome is modelled by a generalised linear mixed model, where the error term is modelled by a skew normal distribution with time-dependent variance. Two sub-models are linked by shared random effects. Parameters of the joint model are estimated by maximising the marginal likelihood. The proposed method extends the Gaussian assumption for errors, and provides a flexible method which could be easily applied to disease progression modelling problem with complex longitudinal outcome structures. We apply the joint model we proposed to the English Longitudinal Study of Ageing data.

4.1 Joint Model

The joint model consists of two sub-models: a multi-state model for competing risks and a regression model for the longitudinal data. The two sub-models are linked by shared random effects.

Suppose there are N individuals in the study. For each individual i, let $[\mathbf{z}_i, y_{ij}, s_{ij}]$ denote all the observations recorded at time t_{ij}, where $j = 1, 2, \ldots, n_i$. $\mathbf{z}_i \in \mathcal{R}^{p\times 1}$ represents the demographic variables of the individual i, y_{ij} is value of the longitudinal outcome measured at each time point, and $s_{ij} \in \{1, 2, 3\}$ denote the

state observed at time t_{ij}.

4.1.1 Longitudinal Sub-model

We assume that the longitudinal outcome y_{ij} at time t_{ij} consists of the unobserved true value μ_{ij} and the error term ε_{ij}. We assume both the unobserved true longitudinal outcome and the error term changes with time, and the error term follows a skew normal distribution: $\varepsilon_{ij} \sim SN(0, \nu, \sigma_{ij}^2)$, where ν is the parameter of skewness, σ_{ij} is the standard deviation at time t_{ij}. The generalised linear mixed model is given by

$$\begin{aligned} \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \omega^\top \mathbf{z}_i, \\ \log(\sigma_{ij}) &= \gamma_0 + \gamma_1 t_{ij}, \end{aligned} \tag{4.1}$$

where the random effects are assumed to follow a bivariate normal distribution: $\mathbf{b}_i^\top = (b_{0i}, b_{1i})^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where the mean is $\mathbf{0} = (0,0)^\top$ and Σ is the variance-covariance matrix.

We assume that the longitudinal outcome is continuous, and modelled with skew normal distribution with time-dependent mean and variance. In the applications in Section 4.3, the longitudinal outcome is the number of animals individuals can name in the interviews, which is discrete, but we show that our model is applicable to it as well.

4.1.2 Competing-Risk Sub-model

Assume there are three states: $1,2,3$, and let $q_{rs}(t_{ij}), t_{ij} > 0$ be the transition intensity of individual i from state r to state s at time t_{ij}. We use the transition-specific hazard regression models to describe the dependency of the transition intensities on the covariates:

$$q_{rs}(t_{ij}|\mu_{ij}) = q_{0.rs} \exp\left(\eta_{rs} t_{ij} + \alpha_{rs}\mu_{ij} + \boldsymbol{\lambda}_{rs}^\top \mathbf{z}_i\right), \tag{4.2}$$

where $(r,s) \in \{(1,2),(1,3)\}$; η_{rs}, α_{rs}, and $\boldsymbol{\lambda}_{rs}^{\top}$ are regression coefficients corresponding to the time, the unobserved true value of the longitudinal outcome, and the demographic covariates. The term $q_{0.rs}$ in equation (4.2) is interpreted as the baseline transition intensity at time $t = 0$ for individual with covariates $\mu_{ij} = 0, \mathbf{z} = \mathbf{0}$.

It is clear from the Equation (4.2) that the random effects were included in the competing-risk sub-model through the term μ_{ij} and the coefficient η_{rs}. The sub-models for the longitudinal outcome and the competing risks are independent conditionally on the random effects. The marginal likelihood is obtained by integrating out the random effects, the details are in Section 4.2.2.

4.2 Estimation

4.2.1 Exact and Interval Censored Event Times

In Andersen and Keiding (2002) different patterns of the observation of states were discussed. In the application of the English Longitudinal Study of Ageing in Section 4.3, the three states are: $s = 1$ (healthy), $s = 2$ (death) and $s = 3$ (severe cognitive impairment), and there is a mixture of states observation patterns. For the death state, the observation is exact, the age of death of the individual is recorded. On the contrary, the transition to the severe cognitive impairment state is interval censored, because it happens between the follow-ups. This feature can be taken into account by defining different likelihood contributions, the details is in Section 4.2.

4.2.2 Maximum Likelihood

The parameters of the joint model are estimated by maximising the likelihood function, which are derived using the transition probabilities and transition intensities. For individual i, the likelihood contribution condition on the random effect of the longitudinal sub-model is:

$$\prod_{j=1}^{T_i} f(y_{ij}|\mathbf{b}_i), \tag{4.3}$$

where f is a generic notation for the density of y_{ij} defined by the model, e.g. the model defined by Equation (4.1). For the multi-state sub-model, we assume that the transition from state $s = 1$ to state $s = 2$ is exact, and the transition from state $s = 1$ to state $s = 3$ is interval censored. For the time-interval formed by a pair of successive time points $(t_{ij}, t_{i(j+1)}]$, the conditional likelihood contribution from this time-interval is of the form:

$$\begin{aligned} L(t_{ij}|\mu_{ij}) &= p_{s_{ij},s_{i(j+1)}}(t_{ij},t_{i(j+1)})q_{s_{ij},s_{i(j+1)}}(t_{1(j+1)}), && \text{if } s_{i(j+1)} = 2 \\ L(t_{ij}|\mu_{ij}) &= p_{s_{ij},s_{i(j+1)}}(t_{ij},t_{i(j+1)}), && \text{if } s_{i(j+1)} = 3 \end{aligned} \tag{4.4}$$

We assume the time-dependent covariates and the transition intensities are independent conditional on the random effects, the complete likelihood of individual i given by marginalising over the random variables is:

$$L_i(\boldsymbol{\theta}) = \int \prod_{j=1}^{T_i} f(y_{ij}|\mathbf{b}_i) \prod_{j=1}^{T_i-1} L(t_{ij}|\mu_{ij}) dG(\mathbf{b}_i), \tag{4.5}$$

where $\boldsymbol{\theta}$ is the collection of all the parameters define in Equations (4.1) and (4.2). Individuals who died before the start of the study are not observed and the data is therefore left-truncated. By using the Markovian assumption, the likelihood is defined conditioned on the first observed state, and in this way the left truncation was taken into account. In Van den Hout (2017) more details about modeling left-truncation is discussed.

The parameters $\boldsymbol{\theta}$ of the joint model in Equation (4.5) are estimated by maximising the logarithm of the total marginal likelihood $\ell = \log\Big(\prod_{i=1}^{N} L_i(\boldsymbol{\theta})\Big)$.

The maximisation of the likelihood is implemented using the R software (R Core Team (2016)). The general purpose optimiser `optim` is used. To have a more robust result the Nelder-Mead method was chosen for the `optim` function. The parameters with constraints were transformed so that the likelihood function is maximised over the unbounded parameter space. For example the standard deviations τ_0, τ_1 are estimated over $\log(\tau_0), \log(\tau_1)$. We present an example code for calculating the marginal likelihood using ELSA data in Chapter 8.

To assess the convergence of the optimisation results, we examined the impact of different initial values. By fitting initial values with small variations to the same model, we evaluated the convergence. Additionally, we employed various optimisation methods, such as Nelder-Mead and BFGS, starting from the same initial value, to determine if the results converged.

4.3 Data Analysis

4.3.1 Joint Models for ELSA data

Our joint model starts with defining the competing-risk sub-model:

$$q_{rs}(t_{ij}|\mu_{ij}) = q_{0.rs} \exp\left(\eta_{rs} t_{ij} + \alpha_{rs}\mu_{ij} + \lambda_{rs} gender_i\right), \tag{4.6}$$

and the sub-model for competing risks:

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) age_{ij}, \\ \log(\sigma) &= \gamma_0, \end{aligned} \tag{4.7}$$

where the error term of the longitudinal outcome is modelled by normal distribution with mean 0 and a constant variance (Model 4). For simplicity we also assume the random effects b_{0i} and b_{1i} are independently normal distributed with standard deviations τ_0 and τ_1 respectively.

Then we extended the distribution of the error terms to a normal distribution with time-dependent variance (Model 5):

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2) \\ \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) age_{ij}, \\ \log\left(\sigma_{ij}\right) &= \gamma_0 + \gamma_1 age_{ij}, \end{aligned} \tag{4.8}$$

We extended above model further by extending the distribution of the error

terms to a skew normal distribution with skewness parameter ν (Model 6):

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim SN(0, \nu, \sigma_{ij}^2) \\ \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) age_{ij}, \\ \log(\sigma_{ij}) &= \gamma_0 + \gamma_1 age_{ij}, \end{aligned} \tag{4.9}$$

We compared these models and their counterparts: Model 1, 2, and 3, where only one random effect was used when modelling the longitudinal outcome, i.e. $b_{1i} = 0$. The results are presented in Table 4.1. It shows that Model 6 performs better than the rest with $AIC = 11853.89$.

Model	p	-log-likelihood	AIC
(1) 1 random effect, Normal	12	11889.46	23790.92
(2) 1 random effect, Normal, sigma	13	11884.50	23782.00
(3) 1 random effect, SN, sigma	14	11861.71	23737.42
(4) 2 random effects, Normal	13	11884.52	23782.04
(5) 2 random effects, Normal, sigma	14	11878.35	23770.70
(6) 2 random effects, SN, sigma	15	11853.89	23722.78
(7) 2 random effects, gender, educ, SN, sigma	17	11797.89	23612.78
(8) 2 correlated random effects, gender, educ, SN, sigma	18	11792.92	23603.84
(1') 1 random effect, gender , educ, Normal	14	11846.49	23706.98
(2') 1 random effect, , gender , educ, Normal, sigma	15	11833.55	23682.10
(4') 2 random effects, gender, educ, Normal	15	11823.71	23662.42
(5') 2 random effects, gender, educ, Normal, sigma	16	11818.51	23653.02
(8') 2 correlated random effects, gender, educ, Normal	16	11821.97	23659.94
(8") 2 correlated random effects, gender, educ, Normal, sigma	17	11817.55	23652.10

Table 4.1: Comparison between joint models for ELSA data (N=919). SN is the abbreviation of skew normal. p is the number of parameters of the model.

Then based on this model, we include demographic variables *gender* and *educ* to the linear mixed model of the longitudinal outcome (Model 7):

$$\begin{aligned} \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) age_{ij} + \omega_1 gender_i + \omega_2 educ_i, \\ \log(\sigma_{ij}) &= \gamma_0 + \gamma_1 age_{ij}, \end{aligned} \tag{4.10}$$

where $gender = 0$ for women (1 for men) and $educ = 0$ for education fewer than ten years (1 for more than ten years).

Table 4.1 shows that Model (7) perform much better compared to Model (6), as the AIC greatly decreased from 23722.78 to 23612.78. To make a more comprehensive comparison and observe the effect of modeling the error term using a skew normal distribution, we conducted a comparison between Model (7) and a series of alternative models. The alternative models, namely Model (1'), (2'), (4'), and (5'), were constructed by introducing two demographic coefficients, gender and education level, to the existing Model (1), (2), (4), and (5). The performance of the alternative models is summarized in Table 4.1 as well. The results demonstrate that Model (7) outperformed the other models.

The parameter estimates and 95% confidence interval of Model (7) are shown in Table 4.2. The estimated value of baseline transition intensities are $\hat{q}_{0.12} = 0.010$ and $\hat{q}_{0.13} = 0.196$, indicating transition intensities to state 3 is higher than the transition intensities to state 2 at baseline age ($age = 50$). The estimated value $\exp(\hat{\eta}_{12}) = 1.080$, $\exp(\hat{\eta}_{13}) = 1.050$, both are greater than 1, and have confidence intervals that exclude 1. This indicates that as age increases, there will be higher hazards for transitioning out of state 1. Their confidence intervals are overlapping, suggests that the effect of time on the intensities of transition $1 \to 2$ and $1 \to 3$ are relatively similar.

The parameters α_{12} and α_{13} establish the connection between the time-dependent covariate 'number of animal names' and the multi-state model. The estimated values $\exp(\hat{\alpha}_{12}) = 0.940$ and $\exp(\hat{\alpha}_{13}) = 0.842$ are both smaller than 1, and both have confidence intervals that exclude 1. This indicates that for individual who obtain higher scores in the test for remembering animal names, the individual

will have lower hazard to transit into state 2 or state 3. In addition, the confidence intervals of $\exp(\alpha_{12})$ and $\exp(\alpha_{13})$ are not overlapping, therefore $\exp(\hat{\alpha}_{12})$ is significantly greater than $\exp(\hat{\alpha}_{13})$. Hence for individual who can remember more animal names, the hazard of transitioning to state 3 (severe cognitive impairment) is smaller than state 2 (death).

Furthermore, $\exp(\hat{\lambda}_{12}) = 1.658$ is greater than $\exp(\hat{\lambda}_{13}) = 1.374$, and their confidence intervals are non-overlapping, This implies that men have a higher hazard for transitioning out of state 1 than women.

Coefficients β_0 and β_1 are in the mixed model. The estimated value of the slope for age $\hat{\beta}_1 < 0$ shows that when individual getting older, the cognitive function will decline. In addition, $\hat{\gamma}_1 < 0$ shows that the scale parameter of the skew normal distribution decreases as age increases. This might show the effect that when individual getting old, they are likely to memorize fewer animal names therefore the variation of the error terms decreases with time as well.

The estimated value of the skewness parameter for the skew normal distribution is $\hat{\nu} = -1.447$, the confidence interval is $(-1.773, -1.247)$. In Rubio and Genton (2016), the Bayesian linear regression models incorporating skew-symmetric error distributions are investigated, and a detailed discussion on the shape parameter can be found. In this study they mentioned that when the absolute value of the shape parameter is less than or equal to 1, the skew normal distribution can be well approximated by a normal distribution. In Table 4.2, the point estimate $\hat{\nu}$ is close to the range $(-1, 1)$. Therefore this skew normal distribution is close to symmetric and a normal distribution can be used to approximate it. However, the confidence interval of $\hat{\nu}$ is outside the range $(-1, 1)$, therefore it suggests that the estimate of the shape parameter can be smaller than estimates, making it a more pronounced skewness than expected. In addition, Table 4.1 shows that the skew normal distribution provides a better fit to the data compared to the normal alternative distributions, this implies that the skew normal distribution has the ability to capture the unique characteristics in the ELSA data more effectively in comparison with the normal error model.

For Model (7), a fit to 16 randomly selected individuals is shown in Figure 4.1. The horizontal axis is the age of the individual corresponding to each time of the observation. It is obvious that the trajectory of the number of animal names recalled is fitted well. The parameter estimates and the confidence intervals of the model are shown in Table 4.2. The estimates of the slope for age $\hat{\beta}_1 < 0$ shows that when individual getting older, the cognitive function will decline. It also shows the effect that when people getting old, they are likely to memorise fewer animal names therefore the variance of the noise decreases with time as well, i.e. $\hat{\gamma}_2 < 0$. (women=0, men=1).

Longitudinal sub-model			Multi-state sub-model		
	Value	95% CI		Value	95% CI
τ_0	3.972	(4.017; 4.802)	$q_{0.12}$	0.010	(0.003; 0.036)
τ_1	0.100	(0.070; 0.144)	$q_{0.13}$	0.196	(0.048; 0.797)
β_0	25.225	(25.800; 27.414)	$\exp(\eta_{12})$	1.080	(1.057; 1.104)
β_1	-0.148	(-0.183; -0.114)	$\exp(\eta_{13})$	1.050	(1.025; 1.076)
ω_1	0.290	(-0.364; 0.944)	$\exp(\alpha_{12})$	0.940	(0.898; 0.984)
ω_2	3.808	(3.147; 4.469)	$\exp(\alpha_{13})$	0.842	(0.797; 0.890)
γ_1	1.833	(1.721; 1.874)	$\exp(\lambda_{12})$	1.658	(1.172; 2.344)
γ_2	-0.005	(-0.007; -0.001)	$\exp(\lambda_{13})$	1.374	(0.941; 2.007)
			ν	-1.447	(-1.703; -1.190)

Table 4.2: Parameter estimates and 95% confidence intervals in the joint model on the ELSA data (N=919). Results of Model (7).

To relax the assumption that the random effects are independent. We fitted Model (8) by assuming the random effects b_{0i} and b_{1i} follow a bivariate normal distribution. The parameter estimates and 95% confidence interval of Model (8) are shown in Table 4.3.

According to Table 4.1, it gives the smallest AIC among all models we fitted. Additionally, we have conducted a comparison between Model (8) and its counterparts with a normal distribution, namely Model (8') and Model (8"). The performance of the alternative models is summarized in Table 4.1 as well. The results demonstrate that Model (8) outperformed the other models.

Longitudinal sub-model			Multi-state sub-model		
	Value	95% CI		Value	95% CI
τ_0	4.262	(3.443; 5.275)	q_{120}	0.015	(0.004; 0.055)
τ_1	0.130	(0.077; 0.220)	q_{130}	0.253	(0.060; 1.069)
β_0	24.414	(23.497; 25.331)	$\exp(\eta_{12})$	1.068	(1.045; 1.092)
β_1	-0.151	(-0.186; -0.116)	$\exp(\eta_{13})$	1.057	(1.033; 1.083)
ω_1	0.387	(-0.260; 1.034)	$\exp(\alpha_{12})$	0.931	(0.889; 0.974)
ω_2	3.800	(3.149; 4.452)	$\exp(\alpha_{13})$	0.851	(0.807; 0.897)
γ_1	1.806	(1.730; 1.882)	$\exp(\lambda_{12})$	1.716	(1.213; 2.429)
γ_2	-0.004	(-0.007; -0.001)	$\exp(\lambda_{13})$	1.253	(0.863; 1.819)
ρ	-0.228	(-0.627;0.267)	ν	-1.510	(-1.773; -1.247)

Table 4.3: Parameter estimates and 95% confidence intervals in the joint model on the ELSA data (N=919). Results of Model (8)

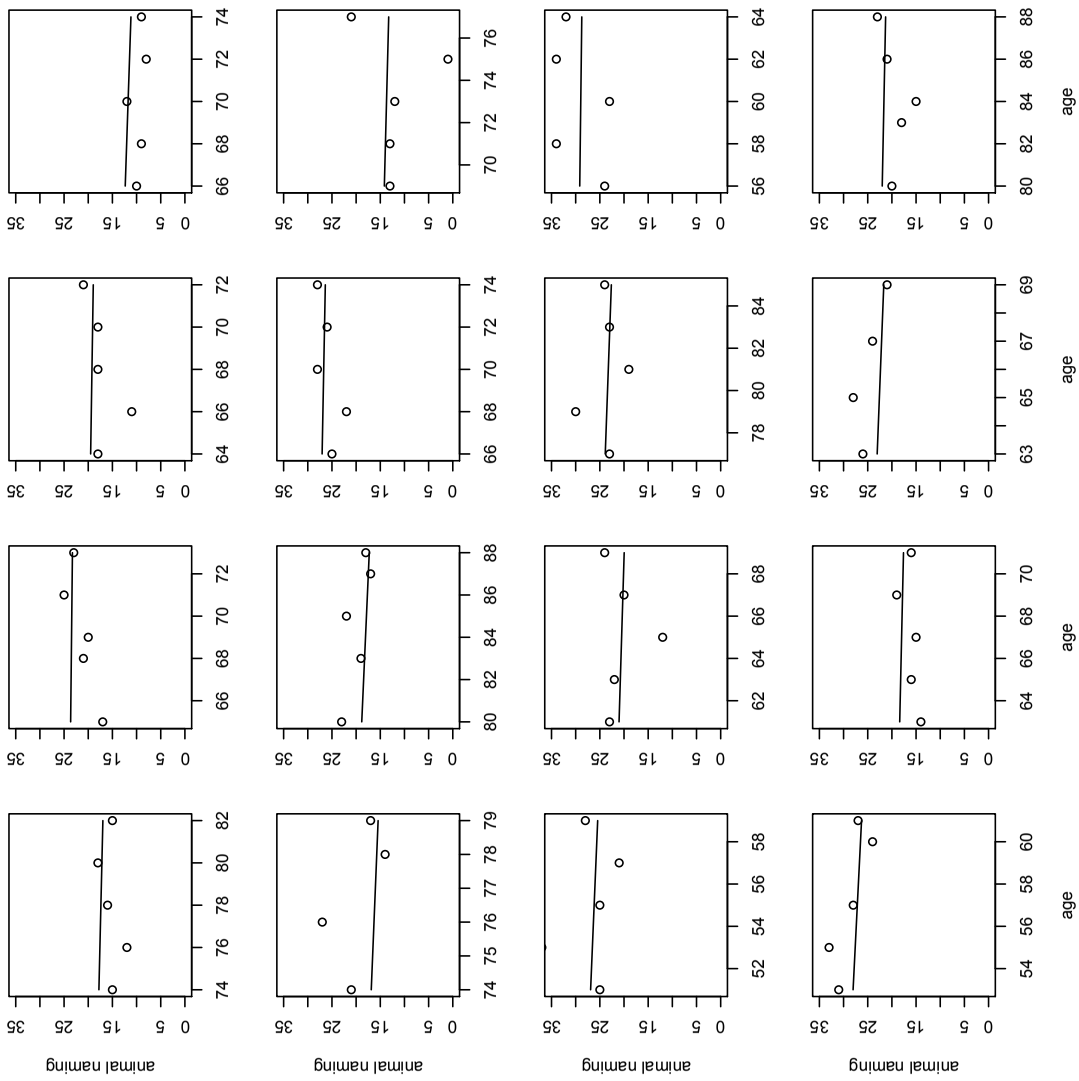


Figure 4.1: The number of animal names recalled (denoted by the dots:"o") and the fitted trajectories (presented by the lines) for 16 randomly selected individuals.

4.3.2 Prediction

In practice, there is need to predict the risk of server cognitive impairment for individuals with different follow-up times. Compared with the parameter estimates and the transition intensities, the estimated transition probabilities are sometimes a more straightforward depiction of the disease progression of interest.

In this section we show how the transition probability matrix can be estimated given the historical observations of a random individual. To illustrate this process we consider the following example generated randomly: a man with more than 10 years education. Using his 2 records of number of animal names recalled: 30 at age 65, and 25 at age 67, the joint model can be used to predict the transition probabilities with the 95% confidence intervals till age 89. First of all, using the two records of the longitudinal outcome, we can estimate the posterior expected value of the random variables $(\hat{b}_{0i}, \hat{b}_{1i}) = (3.35, 0.14)$. Now assuming that the transition probabilities are constant within a year, the transition probability estimated at age 68 is

$$\begin{aligned} &\hat{P}(t_1 = 18, t_2 = 19 | gender = 1, educ = 1, age = 68) \\ &= \begin{pmatrix} 0.981 & 0.016 & 0.003 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \end{aligned} \tag{4.11}$$

where t represents age shifted by -49 years. It is clear that at early age, the probabilities of moving to both state 2 and state 3 are very small. We can estimate 1-year transition probabilities at any time between age 65 and 89. Then we can estimate the transition probability matrix $P(t_1, t_n)$ by multiplying together the transition probability matrices over each years:

$$\hat{P}(t_1, t_n) = \hat{P}(t_1, t_2)\hat{P}(t_2, t_3) \cdots \hat{P}(t_{n-1}, t_n). \tag{4.12}$$

The estimated transition probabilities with 95% confidence intervals were illustrated in Figure 4.2. The confidence intervals were constructed from simulation with 1000 replicas. The estimated transition probabilities shown in Figure 4.2 agree with the

expectations. For male with higher educations, as age grows, the probability of having severe cognitive impairment is less, but the probability of death is growing quickly with respect to age.

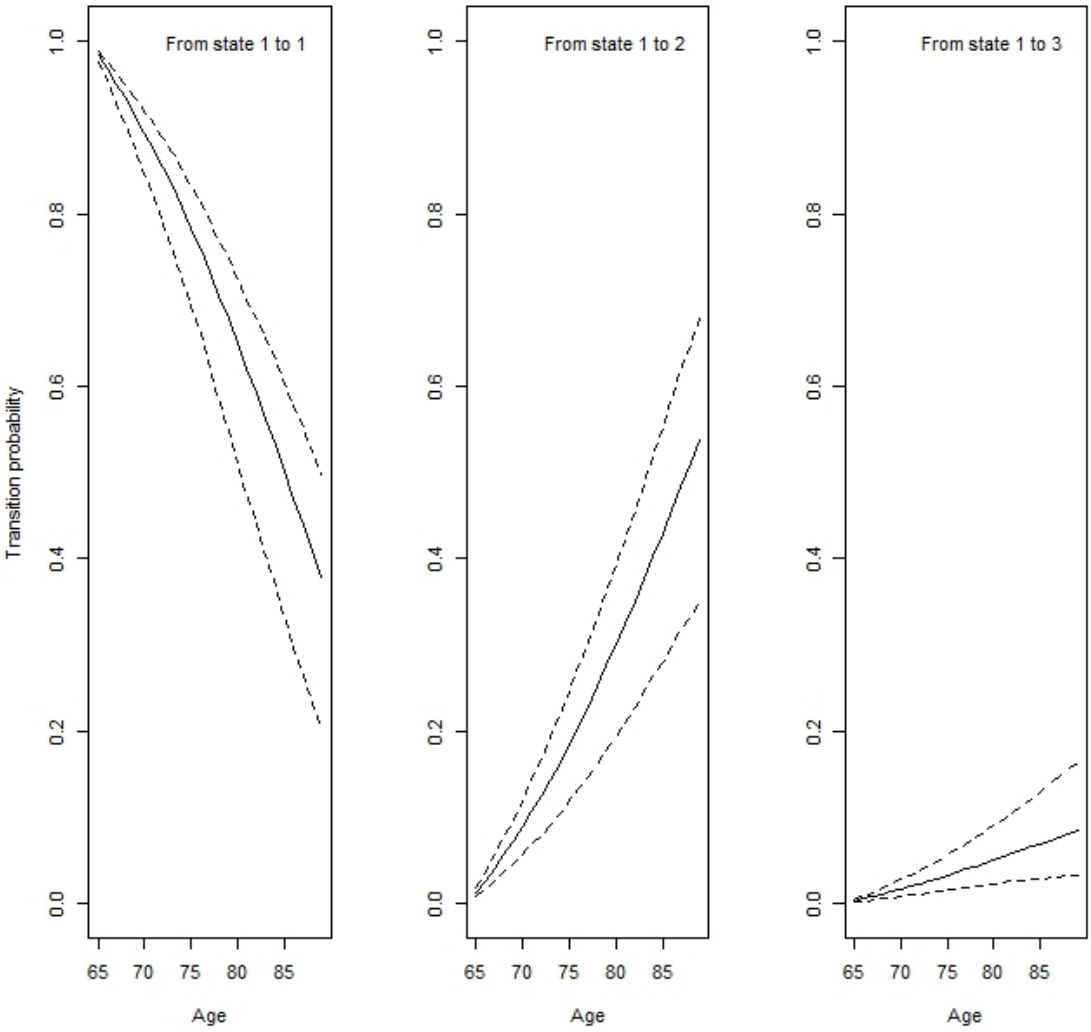


Figure 4.2: The transition probabilities of transitions from state 1 to state 1, 2, 3 are estimated for a woman aged 65 and has fewer than 10 years education. The dashed lines shows the 95% confidence intervals.

4.4 Simulation Study

For the simulation study, we employed the ADEMP structure as a framework. The ADEMP structure comprises the following components: *Aims, Data-*

generating mechanism, Methods, Estimands, and *Performance measures*.

1 The *Aims* of our simulation studies are:

(i) Verify the code in order to ensure the absence of any significant errors.

(ii) Check the correctness of the formulas and the estimation methods.

(iii) Investigate the performance of the joint model in various sample sizes.

(iv) Investigate the performance of the joint model when the correlation between the random effects are taken into account.

2 The *Data-generating mechanism* is introduced below:

Two sets of simulation studies were performed. For simulation study A, longitudinal data and multi-state data for each individual were generated based on Equations (4.6) and (4.9), assuming that the error term follows a skew normal distribution. The random effects for the intercept and the slope of the time-dependent covariates were assumed to be independent.

For simulation study B, data were generated using equations (4.6) and (4.9) as well, but with the assumption that the random effects follow a bivariate normal distribution. Let ρ denote the correlation between the random effects.

The parameter values used for data generation were chosen based on the estimates of Model (6), Model (7), and Model (8). For example, for bothe simulation studies A and B we set the parameter of skewness $\nu = -1.447$, which matches the estimate of ν in Model (7). For Simulation study B, we want to investigate the effect of modelling the correlation between the random effects. And we found that introducing the correlation between the random effects is making it much difficult to estimate the variance terms. Therefore we selected a relatively large correlation of $\rho = -0.462$ to examine the method's performance when the correlations between random intercepts are substantial. Additionally, to assess the impact of the value of τ_0 on the model, we chose two values: $\tau_0 = 1$ and $\tau_0 = 4.392$ respectively.

In both simulation studies A and B, the time scale used is the age, the longitudinal outcome is the number of animal names remembered in a recall, and the competing events are death and severe cognitive impairment. We assume the baseline age is 65 for all individuals. The follow-up time is 2 years. For each individual, we first simulate the two random effects b_{0i}, b_{1i}, then generate the longitudinal trajectories from the skew normal distribution with time-dependent variance. To make sure the longitudinal outcomes are representative of the number of animal words recalled in the ELSA data, the generated value out of the bounds $(0, 60]$ are rounded to 0 and 60. Then the survivor functions to state 2 and state 3 at each time point are computed, and the transition times T_{12}, T_{13} can be generated using the inversion method.

It is typical to vary the sample size when examining the data generation mechanism. As such, we incorporated diverse sample sizes into the design of our simulation study. For simulation studies A and B, we set the sample sizes as $N = 100, 200, 400$ respectively.

In particular for simulation study B, based on the specified parameters and follow-up intervals, we simulate the corresponding random effects for each individual by generating data from bivariate normal distribution. We use the package `mvtnorm` in `R` software to generate bivariate normal samples.

3 *Estimands:*

In the two simulation studies, the parameters that is intended to be estimated are the model parameters.

4 *Methods:*

In the parameter estimation process, The maximum likelihood estimates of the parameters can be obtained by maximising the likelihood function numerically. optimization function `optim` in `R` software was used.

The calculation of the marginal likelihood involved the use of the Gauss-

Hermite quadrature method to approximate the integration. We set the number of nodes to 30 in this method. The simulation study was conducted on a parallel High-Performance Computing (HPC) system. The choice of the number of nodes was made to restrict the optimisation process and ensure its completion within 72-hour usage limit.

The results of simulation A and B are based on 1000 replicates and 500 replicates, respectively. The reason for this choice is also influenced by the usage limit of the HPC.

5 *Performance measures:* The results of the simulation studies are assessed by bias, relative bias (%) and root-mean-square-deviation (rMSE).

Table 4.4 shows the results of the simulations results obtained through 1000 replicates of sample sizes $N = 100, 200, 400$ respectively. In this table we present the value of the parameter used for simulation in the first column. For each sample size, we present the bias, the percentage of the bias and the root-mean-square respectively. Comparing the percentage of bias for $N = 100$ and $N = 400$, all of them decrease as N increases. In addition, the root-mean-square error consistently reduces when the sample size N increases. The results of the simulation study show that sample size $N = 400$ is large enough for the implemented method to have estimates with the percentage of bias is 5% or less.

Table 4.5 and 4.6 present the estimates of the parameters obtained from the two datasets. Table 4.5 shows that when the standard error of the random intercept is small, i.e. $\tau_0 = 1$, the bias and the percentage of the bias decreases when the sample size N increases. However, when the random effects are highly correlated, larger sample sizes might be needed to obtain precise estimates. Furthermore, compared with $\tau_0 = 1$, Table4.5 shows that when the random intercept has a large variance, i.e. $\tau_0 = 4.392$, it is relatively difficult to obtain precise estimates as small sample sizes. The bias and the percentage of the bias decrease much slower compared with the two previous experiments. To conclude, time-dependent covariates outcomes that huge population-level variances might increase the difficulties in estimating the model parameters. The modelling of the correlation between random effects also

increases the need for the sample sizes for precise estimation.

Parameter		$N = 100$			$N = 200$			$N = 400$		
	Value	Bias	*%bias*	rMSE	Bias	*%bias*	rMSE	Bias	*%bias*	rMSE
$\log(\tau_0)$	1.480	0.034	0.023	0.087	0.014	0.009	0.061	0.010	0.007	0.042
$\log(\tau_1)$	-2.303	-0.117	0.051	0.216	-0.091	0.040	0.124	-0.096	0.042	0.073
β_0	26.607	-0.216	0.008	0.810	-0.124	0.005	0.536	-0.151	0.006	0.331
β_1	-0.165	0.009	0.055	0.025	0.004	0.024	0.018	0.004	0.024	0.012
γ_1	1.798	-0.024	0.013	0.076	-0.012	0.007	0.055	-0.012	0.007	0.039
γ_2	-0.004	0.000	0.000	0.003	0.000	0.000	0.002	0.000	0.000	0.001
$\log(q_{0.12})$	-4.200	-0.028	0.006	0.864	-0.010	0.002	0.470	0.037	0.009	0.288
$\log(q_{0.13})$	-1.374	0.415	0.302	0.616	0.058	0.042	0.527	0.029	0.021	0.371
α_{12}	-0.062	-0.004	0.066	0.024	0.001	0.016	0.010	0.001	0.016	0.007
α_{13}	-0.172	-0.024	0.140	0.018	-0.003	0.017	0.022	-0.001	0.006	0.015
η_{12}	0.077	0.006	0.078	0.025	0.001	0.013	0.017	0.000	0.000	0.010
η_{13}	0.049	-0.003	0.061	0.013	0.002	0.041	0.013	0.002	0.041	0.010
ν	-1.447	0.077	0.053	0.217	0.033	0.023	0.144	0.033	0.023	0.093

Table 4.4: Simulation study for $N = 100, 200, 400$. 1000 replicates. Column Value displays the true value of the parameters. Bias, percentage of bias is measured with respect to mean. rMSE represents the root mean square error.

Parameter		$N = 100$			$N = 200$			$N = 400$		
	Value	Bias	*%bias*	rMSE	Bias	*%bias*	rMSE	Bias	*%bias*	rMSE
τ_0	1.000	-0.008	0.008	0.231	0.007	0.007	0.095	0.006	0.006	0.079
τ_1	0.100	0.001	0.010	0.011	0.000	0.000	0.007	-0.001	0.010	0.005
β_0	26.607	0.003	0.000	0.250	0.005	0.000	0.147	0.009	0.000	0.106
β_1	-0.165	0.000	0.000	0.011	0.000	0.000	0.009	0.000	0.000	0.007
γ_1	0.500	0.024	0.048	0.083	0.018	0.036	0.060	0.021	0.042	0.041
γ_2	-0.004	0.000	0.000	0.003	0.000	0.000	0.002	0.000	0.000	0.001
$q_{0.12}$	0.015	0.003	0.200	0.009	0.003	0.200	0.007	0.002	0.133	0.005
$q_{0.13}$	0.253	0.058	0.229	0.164	0.030	0.119	0.103	0.037	0.146	0.084
$\exp(\alpha_{12})$	0.940	0.000	0.000	0.011	0.000	0.000	0.007	0.001	0.001	0.006
$\exp(\alpha_{13})$	0.842	0.001	0.001	0.010	0.000	0.000	0.006	0.000	0.000	0.004
$\exp(\eta_{12})$	1.080	-0.001	0.001	0.021	-0.002	0.002	0.015	-0.002	0.002	0.011
$\exp(\eta_{13})$	1.050	0.000	0.000	0.013	-0.001	0.001	0.009	-0.001	0.001	0.007
ν	-1.447	0.056	0.039	0.276	0.060	0.041	0.153	0.053	0.037	0.085
ρ	-0.462	0.020	0.043	0.151	0.014	0.030	0.091	-0.001	0.002	0.063

Table 4.5: Setting $\tau_0 = 1$. Simulation study for $N = 100, 200, 400$. 500 replicates. Column Value displays the true value of the parameters. Bias, percentage of bias is measured with respect to mean. rMSE represents the root mean square error.

Parameter		$N = 100$			$N = 200$			$N = 400$		
	Value	Bias	*%bias*	rMSE	Bias	*%bias*	rMSE	Bias	*%bias*	rMSE
τ_0	4.392	-0.335	0.076	0.367	-0.324	0.074	0.367	-0.324	0.073	0.322
τ_1	0.100	-0.010	0.100	0.009	-0.011	0.110	0.009	-0.014	0.140	0.007
β_0	26.607	0.366	0.014	0.402	0.413	0.016	0.402	0.375	0.014	0.358
β_1	-0.165	0.001	0.006	0.012	0.001	0.006	0.012	0.001	0.006	0.007
γ_1	0.500	0.016	0.032	0.066	0.018	0.036	0.066	0.015	0.030	0.044
γ_2	-0.004	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.001
$q_{0.12}$	0.015	0.002	0.133	0.005	0.002	0.133	0.005	0.002	0.133	0.003
$q_{0.13}$	0.253	0.050	0.197	0.108	0.042	0.166	0.108	0.028	0.111	0.062
$\exp(\alpha_{12})$	0.940	0.001	0.001	0.009	0.002	0.002	0.009	0.000	0.000	0.004
$\exp(\alpha_{13})$	0.842	0.000	0.000	0.009	-0.003	0.004	0.009	0.000	0.000	0.008
$\exp(\eta_{12})$	1.080	-0.001	0.001	0.009	-0.002	0.002	0.009	-0.002	0.002	0.005
$\exp(\eta_{13})$	1.050	-0.001	0.001	0.008	0.000	0.000	0.008	-0.001	0.001	0.005
ν	-1.447	0.136	0.094	0.235	0.134	0.093	0.235	0.133	0.092	0.112
ρ	-0.462	0.414	0.896	0.189	0.409	0.885	0.189	0.394	0.853	0.165

Table 4.6: Setting $\tau_0 = 4.392$. Simulation study for $N = 100, 200, 400$. 500 replicates. Column Value displays the true value of the parameters. Bias, percentage of bias is measured with respect to mean. rMSE represents the root mean square error.

Chapter 5

Early Sepsis Diagnosis Using Joint modelling of Competing Risks and Covariates

The task of early diagnosis of disease is to make predictions on the disease progression based on the health records of individuals. For the sepsis early diagnosis task introduced in Chapter 2, the aim is to find out sepsis patients as early as possible, in order to gain a higher utility score.

It is straightforward to solve the sepsis early detection problem as a classification problem. With the development of devices, the speed of computing is getting faster, therefore it is more convenient to apply the machine learning and deep-learning algorithms to real-world problems. Since classification problem is an important topic in supervised learning, lots of algorithms to can be chosen.

However directly applying the methods for classification, some important characteristics of this dataset were neglected: (i) the labels are not consistent. Both non-septic patients and septic patients have labels ”0”, but the changes of their biomarkers are different, and these labels should not be treated equally. (ii) There are uncertainties in the label of sepsis. Since labels are given by the clinicians, the diagnosis time is always after the sepsis onset time. Therefore directly using the training set with these labels will lead to the methods failed to diagnose sepsis at the optimal time. In previous studies the label was preprocessed, the diagnosis time

was moved forward for 6 hours. However, this method lacks justification and will introduce extra noise.

Ripley and Ripley (2001) discussed the relationship between the classification problem and survival analysis. In our approach, we show that the two events diagnosed to be 'septic' and 'nonseptic' can be modelled as two competing events. To penalize the late prediction, we design a time-dependent loss function and propose to make diagnoses by minimizing the expected loss.

The framework we proposed for the early diagnosis consists of two steps. First of all, the individual-specific time-to-event distribution is estimated by joint modelling the survival data and the time-dependent covariates. In the second step, the label minimizing the expected loss is chosen. The competing risks and the covariates are modelled using a joint model based on multi-state model and a mixed model.

5.1 Statistical Decision Theory

Decision theory involves the rational process of decision-making, providing a framework to make choices in the face of uncertainty and various outcomes. It aims to optimize decisions by evaluating the associated risks, rewards, and probabilities of different actions. Peterson (2017) offers a comprehensive introduction to decision theory. The book introduces the concept of utility, which represents the value of an outcome as perceived by decision-makers. One widely used decision rule for making choices under risk is to maximize expected utility. In the case of the sepsis early diagnosis problem, the utility score for each diagnosis result of every patient at each time point is defined by the utility functions 2.4 and 2.5. In the framework for disease early diagnosis we proposed, we define a loss function based on the utility function, and the minimizing of the expected loss is equivalent to the maximizing of the expected utility.

More detailed introduction can be found in, for instance Berger (2013), which focuses on statistical decision theory and its connection to Bayesian analysis, and

Parmigiani and Inoue (2009) which offers an introduction to various decision-making frameworks.

In this section, we present an introduction to statistical decision theory, following the approach outlined in Bishop and Nasrabadi (2006). We show that the classification problem can be solved by minimizing the loss function. We will define a time-dependent loss function specific to the sepsis early diagnosis problem.

Suppose that $D \in \mathcal{R}$ and $\mathbf{Y} \in \mathcal{R}^{p\times 1}$ are random variables with joint distribution $f_{\mathbf{Y},D}(\mathbf{y},d)$. In many applications, the goal is to predict the corresponding values of D given values of $\mathbf{Y}$. The prediction is achieved by finding a function f where the difference between $f(\mathbf{Y})$ and D is minimized. To measure the loss generated, a loss function $L(D,f(\mathbf{Y}))$ is introduced. Given $\mathbf{Y}=\mathbf{y}$, it is ideal to choose the $f(\mathbf{y})$ which minimizes the loss function. However, the loss function $L(v,f(\mathbf{y}))$ depends on d, which value is unknown and we aim to predict. Therefore, instead of minimizing the loss function directly, we can minimize the conditional expectation given $\mathbf{y}$:

$$E_D[L(D,f(\mathbf{Y}))|\mathbf{y}] = \int_v L(v,f(\mathbf{y}))f_{D|\mathbf{Y}}(v|\mathbf{y})dv. \tag{5.1}$$

Consider a classification problem, where D is a random variable which can only take value 0 or 1. Therefore two types of loss could arise in the decision-making procedure: L_{01} loss stands for the loss incurred when misclassified an observation $\mathbf{x}$ from class 0 to class 1, and vise versa for L_{10}. Then using the method above, the classification of $\mathbf{x}$ is given by minimizing the equation below:

$$L_{01}P(D=0|\mathbf{Y}=\mathbf{y}) + L_{10}P(D=1|\mathbf{Y}=\mathbf{y}). \tag{5.2}$$

See more detailed discussion on statistical decision theory in, for instance Bishop and Nasrabadi (2006) and Friedman et al. (2001).

5.2 Prediction and Evaluation

The diagnosis of sepsis is usually considered as a classification problem. According to Section 5.1, it can be solved by minimizing the expected loss. In practice, the clinicians can only make diagnosis many hours later than the sepsis onset time. Therefore applying the classification methods directly with the labels provided by the clinicians are less likely to provide early diagnosis. To tackle this challenge, we propose to use a time-dependent loss function that penalizes late predictions. In this section, we first define this loss function then derive the expected loss.

Recall the utility score function introduced in Chapter 2. For patients that eventually have sepsis (i.e., with at least one *SepsisLabel* entry of 1), the classifiers is rewarded if it predict sepsis onset between 12 hours before and 3 hours after t_{sepsis}, where the maximum reward is 1. We penalize classifiers that do not predict sepsis or predict sepsis more than 12 hours before t_{sepsis}, where the maximum penalty for very early detection is a -0.05 and the maximum penalty for late detection is -2.0.

For patients that do not eventually have sepsis (i.e., all *SepsisLabel* entries equals 0), we penalize classifiers that predict to be sepsis onset, where the maximum penalty for false alarms is a parameter (0.05; equal to the very early detection penalty). We neither reward nor penalize those that do not predict sepsis.

For the prediction, according to the definition of the utility score, it is clear that the reward of the prediction depends on whether a patient is a septic patient or not, furthermore, it is related to the onset time of the sepsis. Therefore to make predictions at time points $t = 1, 2, \ldots, T_i$ given the historical data updated hourly, it is natural to introduce a time-dependent loss. Similar to the utility score, we define the loss as a function of t and t_{sepsis} as well. The value is chosen based on the utility score in equations (2.3), (2.4) and (2.5).

In line with equation (5.2), term L_{01} is the false positive loss, i.e. the loss incurred if a non-septic patient is misclassified to be septic. We define L_{01} to be a constant: $L_{01} = 0.05$.

Similarly, the term L_{10} is the false negative loss generated while the septic patient is classified to be nonseptic. Since we want the algorithm to classify the

septic patient as early as possible, we assume the loss incurred varies with respect to the differences between t and t_{sepsis}. To be more specific, when $t \leq t_{sepsis} - 12$, there will be no penalty if the algorithm is failed to correctly diagnose septic patients 12 hours before the clinicians and L_{10} will be zero. For $t_{sepsis} - 12 < t$, there will be a slight penalty if the algorithm cannot diagnose the septic patient and the loss will increase as time t increases. $L_{10}^{(t)}(t_{sepsis})$ is given by:

$$L_{10}^{(t)}(t_{sepsis}) = \begin{cases} 0 & (t \leq t_{sepsis} - 12) \\ 1/6 & (t = t_{sepsis} - 11) \\ \dots & \dots \\ 1 & (t = t_{sepsis}) \\ \dots & \dots \\ 2 & (t \geq t_{sepsis} + 9). \end{cases} \tag{5.3}$$

Notice that the loss function L_{10} is different for each individual since the t_{sepis} of each septic patient is different. Therefore it is natural to consider the time of diagnosis as a random variable T_{sepis}.

Therefore to predict the label of patient i at time t, provided the historical observations $\mathcal{H}(t)$ (for simplicity the index i is omitted), the expected loss is given by:

$$\begin{aligned} & E_L\left[L^{(t)}(D, T_{sepsis})|t, \mathcal{H}(t)\right] \\ & = E_{T_{sepsis}}\left[E_{L|t_{sepsis}}\left[L^{(t)}(D, t_{sepsis})|\, T_{sepsis} = t_{sepsis}, t, \mathcal{H}(t)\right]\right] \\ & = \sum_v \int_{t_{sepsis}=1}^{\infty} L^{(t)}(D = v, t_{sepsis}) P(D = v | T_{sepsis} = t_{sepsis}, t, \mathcal{H}(t)) f(t_{sepsis}|t, \mathcal{H}(t)). \end{aligned} \tag{5.4}$$

In our case, the label D can only take value 0 or 1. Now consider the conditional probability $P(D = v|\, T_{sepsis} = t_{sepsis}, t, \mathcal{H}(t))$ in equation (5.4). It is natural to de-

fined that:

$$P(D=0|T_{sepsis}=t_{sepsis},t,\mathcal{H}(t))=\begin{cases}1 & (t\leq t_{sepsis}-12)\\ 0 & (\text{otherwise}),\end{cases}$$

$$P(D=1|T_{sepsis}=t_{sepsis},t,\mathcal{H}(t))=\begin{cases}1 & (t>t_{sepsis}-12)\\ 0 & (\text{otherwise}),\end{cases}$$

since we will only classify an individual to be septic when $t > t_{sepsis} - 12$. Therefore equation (5.4) is simplified to:

$$\underbrace{\int_{t_{sepsis}=1}^{t+12} L_{10}^{(t)}(t_{sepsis}) f(t_{sepsis}|\mathcal{H}(t))}_{\text{(a) false negative loss}} + \underbrace{\int_{t_{sepsis}=t+12}^{\infty} L_{01}^{(t)}(t_{sepsis}) f(t_{sepsis}|\mathcal{H}(t))}_{\text{(b) false positive loss}}. \quad (5.5)$$

When making a diagnosis, we classify the individual to class 0 (non-septic) if the value of term (a) is smaller than term (b), and vise versa.

To calculate the two terms in equation (5.5), we need to estimate the conditional distribution of t_{sepsis}. We propose to achieve this by joint modelling the competing risks and the time-dependent covariates. Then the transition probability $P_{12}(t_j, t_{j+1}|\mathcal{H}(t_{j+1}))$ can be used to approximate the distribution of t_{sepsis}. The details will be discussed in the next section.

5.3 Joint Model of Competing Risks and Covariates

In the context of early sepsis diagnosis, individuals were excluded from the study once they received a non-septic diagnosis. Furthermore, the study involved hourly measurements due to the acute nature of sepsis. As a result, the diagnoses of "sepsis" and "non-sepsis" can be regarded as competing events, given that individuals diagnosed as non-septic will not develop sepsis during the study's timeframe.

In our multi-state model we defined three states (displayed in Figure 2.3): State 1 - under risk; State 2 - onset of sepsis; State 3 - diagnosed to be non-septic/dropout. Transitions are permitted are from State 1 to State 2, and State 1 to State 3. $\mathbf{y}(t)$

are the values of the time-dependent covariates at time t, $\mathbf{z}$ are the demographic covariates. This multi-state model is governed by two transition hazards q_{12} and q_{13}, defined as:

$$q_{rs} = q_{0.rs} \exp\left(\boldsymbol{\alpha}_{rs}^{\top}\mathbf{y}(t) + \boldsymbol{\lambda}_{rs}^{\top}\mathbf{z} + \eta_{rs}t\right).$$

The parameters of the joint modelling of competing risks and time-dependent covariates are estimated by maximising the marginal likelihood function or using Bayesian inference (see Chapter 6).

After obtaining the parameters for the joint model, the transition probability $P_{1,2}(t_1,t_2)$ of an individual over any time-intervals can be estimated. Then the equation (5.5) can be approximated by:

$$\begin{aligned} E_L[L^{(t)}(D,t_{sepsis})|t,\mathcal{H}(t)] = &\sum_{t_{sepsis}=t+1}^{\tau} L_{01}[P_{1,2}(0,t_{sepsis}) - P_{1,2}(0,t_{sepsis}-1)] \\ &+ \sum_{t_{sepsis}=1}^{t} L_{10}^{(t)}(t_{sepsis})[P_{1,2}(0,t_{sepsis}) - P_{1,2}(0,t_{sepsis}-1)], \end{aligned} \tag{5.6}$$

where τ is a large number, and $\tau > \max_i(T_i)$ for this method to valid. In the data analysis, we let $\tau = 300$ hours.

5.4 Data Analysis

In this section, we present the results of the joint models using the clinical data of sepsis. We fitted four different joint models to the time-dependent covariate respiratory rate. We start defining the joint models with the competing-risk sub-model:

$$q_{rs}(t_{ij}|\mu_{ij}) = q_{0.rs} \exp\left(\eta_{rs}t_{ij} + \alpha_{rs}\mu_{ij} + \lambda_{rs}^{(1)}age_i + \lambda_{rs}^{(2)}gender_i\right),$$

where age_i and $gender_i$ are the demographic variables for the *gender* and *age*. The variable t_{ij} is the corresponding time the covariates are recorded. We assume the transition intensity only depend on one time-dependent covariate, and the term μ_{ij}

is the mean of the covariate. From Model (i) to Model (iv), we kept the structure of the competing-risk sub-model unchanged, and apply 4 different structures to model the time-dependent covariates. Now we define the sub-model for time-dependent covariate.

Model (i). General mixed model with normal distribution.

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) t_{ij}, \qquad b_i = (b_{0i}, b_{1i}) \sim \mathcal{N}(0, \Sigma). \end{aligned} \tag{5.7}$$

In this mixed model we assume the error terms e_{ij} follows a normal distribution with mean 0 and variance σ^2. The term y_{ij} represents the observed value of the time-dependent covariates. For the true value of the time-dependent covariates μ_{ij}, we assume it changes linearly with respect to the time t_{ij}. To describe the individual-level intercept and slope we introduced two random variables b_{0i} and b_{1i}, and assume they jointly follows a bivariate normal distribution with mean $(0,0)$ and the variance-covariance matrix Σ.

Model (ii). General mixed model with normal distribution, variances of error terms changes over time.

$$\begin{aligned} y_{ij} &= \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2) \\ \log(\sigma_{ij}) &= \gamma_0 + \gamma_1 t_{ij}, \\ \mu_{ij} &= \beta_0 + b_{0i} + (\beta_1 + b_{1i}) t_{ij}, \qquad b_i = (b_{0i}, b_{1i}) \sim \mathcal{N}(0, \Sigma). \end{aligned} \tag{5.8}$$

Model (ii) is an extension of Model (i). We assume the variance of the normal error term is no longer a constant but changes over time t_{ij}.

Model (iii). General mixed model with skew normal distribution, variances of error terms changes over time.

$$
\begin{aligned}
&y_{ij} = \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim SN(0, \nu, \sigma_{ij}) \\
&\log(\sigma_{ij}) = \gamma_0 + \gamma_1 t_{ij}, \\
&\mu_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) t_{ij}, \qquad b_i = (b_{0i}, b_{1i}) \sim \mathcal{N}(0, \Sigma).
\end{aligned} \tag{5.9}
$$

We extend Model (ii) by extending the distribution of the error terms from the normal distribution to a skew normal distribution. The parameter controls the skewness is denoted by ν.

Model (iv). General mixed model with skew t distribution, variances of error terms changes over time.

$$
\begin{aligned}
&y_{ij} = \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim ST(0, \nu, \tau, \sigma_{ij}) \\
&\log(\sigma_{ij}) = \gamma_0 + \gamma_1 t_{ij}, \\
&\mu_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) t_{ij}, \qquad b_i = (b_{0i}, b_{1i}) \sim \mathcal{N}(0, \Sigma).
\end{aligned} \tag{5.10}
$$

We extend Model (ii) by extending the distribution of the error terms from the normal distribution to a skew t distribution. The parameter controls the skewness is denoted by ν, and the parameter of the degree of freedom is τ.

Before we fit the models above, the records of the respiratory rate were pre-processed. The missing values of each patient were imputed with the nearest available observation before the missing value. To avoid numerical problems, the mean of the observations was shifted to 0. The models were fitted to a subset of $N = 1000$ patients, which were randomly sampled from the whole dataset from one hospital. The number of repeated observations of each patient is between 20 to 300.

Since in the four models above we assumed the random effects follow the normal distribution, the Gauss-Hermite quadrature method was applied to approximate the integrals in the marginal likelihood function. Then the function `optim` in R

software was used to obtain the maximum likelihood estimates of parameters.

The -log-likelihood and AIC of the above models are shown in table 5.1. It shows that the model (iv) has the smallest AIC value.

5.5 Numerical Problem

It is important to notice that the estimation process is sensitive to the approximation of the marginal likelihood. In this study the Gaussian-Hermite approximation method was applied, which accuracy depends on the number of nodes. It is straightforward that the more number of nodes we use, the more accurate marginal likelihood we can obtain. However more number of nodes requires longer computing time. For example, when we model the covariate respiratory rate using Model (i), we can fix a set parameters and compare the differences in the approximated marginal likelihood using different number of nodes respectively. Figure 5.1 present the changes in the approximated value of log-likelihoods using different number of nodes. It is clear that the approximation of the log-likelihood have not converge yet when number of nodes equals 60. However, to limit the computing time we cannot choose as many nodes as possible. And to balance the time computing time and the accuracy, we set the number of nodes to be 30.

Furthermore, In Table 5.2 we compare the estimates of a subset of parameters in Model (iv) obtained using maximum marginal likelihood and Bayesian inference (this will be discussed in Chapter 6). It is clear that marginal likelihood tends to underestimate the variance terms, which is expected based on the features of maximum marginal likelihood estimation. Therefore, despite increasing the sample size, the underestimation of variance terms still exists. The estimates of σ_e, σ_{b1} and σ_{b2}

Model	Covariate	p	-log-likelihood	AIC
(i)	Resp	16	22810.84	45637.68
(ii)		17	22223.03	44462.06
(iii)		18	22184.97	44385.94
(iv)		19	21270.06	42556.12

Table 5.1: Comparison between joint models for Sepsis data (N=1000). The time-dependent covariate is the respiratory rate. p is the number of parameters for the corresponding model.

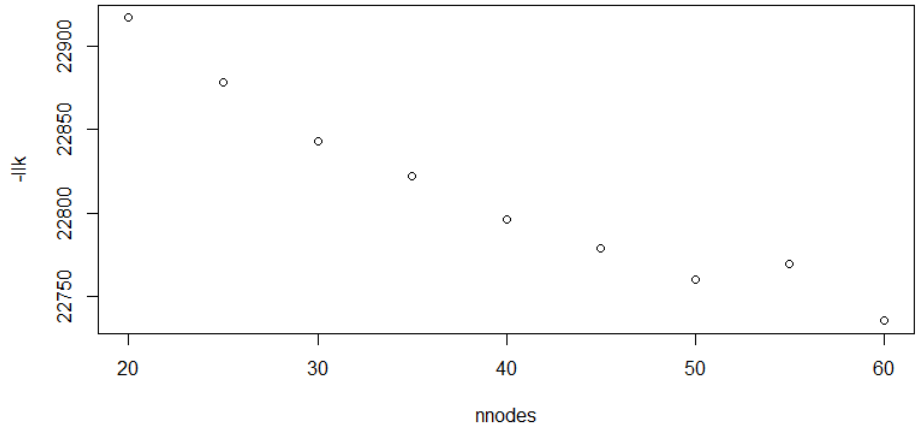


Figure 5.1: Value of minus log-likelihood computed using different number of nodes.

	Marginal Likelihood			Bayesian Inference		
	Value	95% CI		Value	95% CI	
σ_e	0.362	0.359	0.364	1.182	1.176	1.188
σ_{b1}	0.348	0.337	0.360	0.392	0.373	0.410
σ_{b2}	0.111	0.107	0.115	0.125	0.117	0.134
β_1	1.796	1.780	1.813	1.812	1.796	1.834
β_2	0.050	0.047	0.052	0.035	0.031	0.039
ρ	-0.427	-0.473	-0.378	-0.458	-0.612	-0.343

Table 5.2: Comparing the estimates of parameters of Model (iv) using Maximum marginal likelihood and Bayesian inference. L and U are the boundaries of the 95% confidence interval.

using maximum marginal likelihood are smaller than the estimates obtained using Bayesian inference.

To solve above problems different approximation method can be used. In Chapter 6 we will show that by using Bayesian inference this numerical problem can be alleviated.

Chapter 6

Bayesian Inference for Joint Models

In Chapter 5 we illustrated the joint model using the ELSA data, where only one time-dependent covariate was modelled with the general mixed sub-model for the longitudinal outcome. In that general mixed model, there were two random effects, and the Gauss-Hermite quadrature approximation was applied to numerically estimate the value of the double integral. This joint model can be extended to take into account multiple time-dependent covariates simultaneously. However, as the number of random effects increases, integration in higher dimensions is needed. In this case, using the Gauss-Hermite quadrature becomes less efficient, since the number of evaluation points will grow exponentially. This is known as the ”curse of dimensionality”, a term first used by Bellman (1966).

There are various ways to overcome the computational complexity of integration in high-dimensional space, and among them, the Monte Carlo method discussed by Jerrum and Sinclair (1996) is often the primary choice. However, as a method based on sampling, it introduces randomness to the approximation of the integrals. therefore it is problematic to apply the general purpose optimiser as we did in the previous chapter. The Bayesian inference can help us to alleviate the difficulties. Within the Bayesian framework, we estimate the parameters using the posterior expectations.

This chapter consists of two sections. First of all we will briefly explain the methods for Bayesian inference, including the concepts of the Bayesian inference, and the Bayesian model comparison criteria. Then we will introduce the Bayesian

inference for joint models. Finally, we will apply the joint models to the clinical data of sepsis, and then present the early prediction of sepsis based on the joint models.

6.1 Bayesian Inference

Both the Bayesian and frequentist approaches are important in statistical inference, and over the years people have been debating which approach is better. Briefly speaking, the frequentist approach is purely data-driven. On the contrary, the Bayesian inference considers probability as a measure of the belief, and incorporates the prior beliefs into the data analysis. In most of the statistical books introductions to both approaches can be found. Books written by Bolstad and Curran (2016); Gelman et al. (1995) and Hoff (2009) focus on Bayesian inference only, provide information from theory to practice. Bayesian inference is also a very popular technique in areas such as artificial intelligence, engineering, etc. Hastie et al. (2009) systematically introduces the statistical learning methods and discusses the model inference methods including both maximum likelihood and Bayesian inference. Bishop and Nasrabadi (2006) focuses on the Bayesian viewpoint of the pattern recognition and machine learning methods, presenting a profound understanding of the Bayesian inference. In particular, the Bayesian inferential procedure for the multi-state model can be found in Welton and Ades (2005) and Van den Hout (2017). In Van den Hout (2017), the Bayesian inference of a frailty model was discussed and illustrated with the ELSA data.

As summarised in Bayarri and Berger (2004), the debate over the superiority of one approach to the other should continue, as the philosophies underlying the frequentist and Bayesian inference are different. Let us denote the data and the parameter of interest by $\mathbf{x}$ and $\boldsymbol{\theta}$ respectively. The probability mass function and the probability density function are denoted by $p(\cdot)$. For simplicity we use notation $\mathbf{x}$ to denote the probability distribution $p(\mathbf{X} = \mathbf{x}) = p(\mathbf{x})$. In the world of the frequentist, the unknown parameter $\boldsymbol{\theta}$ is considered to be fixed, and the inference of them only

depends on the data $\mathbf{x}$, i.e. this approach only focuses on the distribution $p(\mathbf{x}|\boldsymbol{\theta})$. On the contrary, in Bayesian statistics, $\boldsymbol{\theta}$ is considered as a random variable, and a prior distribution is given to it based on the knowledge or the belief. The Bayesian inference is based on the Bayes Rule:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}, \tag{6.1}$$

where the term $p(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood function. Notice that the marginal distribution $p(\mathbf{x})$ is a constant with respect to $\boldsymbol{\theta}$, therefore can be considered as a normalising constant of the posterior distribution. The Equation (6.1) expresses that after observing data, the prior distribution $p(\boldsymbol{\theta})$ can be updated by the likelihood and we obtain the posterior of the parameters of interest, i.e.

$$\text{posterior} \propto \text{likelihood} * \text{prior}. \tag{6.2}$$

When there is little prior knowledge available for the parameters, weakly informative priors can be used, then the posterior is dominated by the likelihood. And it was mentioned in Ghosh et al. (2006); Lemoine (2019) that assigning the uniform prior will lead to inference similar to the frequentist approach. In this thesis, we will not discuss whether the Bayesian or frequentist approach is more appropriate to be applied to our problem. Instead, we only want to use the Bayesian inference as a technique to alleviate the high-computational complexity when calculating the marginal likelihood. Therefore, we utilise the property that using the weakly informative prior will produce results that coincide with the maximum likelihood estimation, assign vague prior distributions, and finally make inference of the parameters which control the joint model from their posterior distributions. In the analysis of the clinical data of sepsis, we use the Normal distribution $N(0, 100^2)$ as the weakly informative prior for the fixed effect coefficients and uniform distributions such as $U(-10, 10)$ as priors for the standard deviation of the normally distributed residual error.

6.2 High-dimensional integration

In this section we will briefly discuss the high-dimensional integration and introduce the Markov Chain Monte Carlo methods. Let $f(\boldsymbol{\theta})$ denote any transformation of the parameter $\boldsymbol{\theta}$. Now consider the integral:

$$E[f(\boldsymbol{\theta})|\mathbf{x}] = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}. \tag{6.3}$$

The computation of this integral is usually a complex problem, especially in high-dimensional space. For example, in Chapter 5, we applied the general mixed model to the number of animal names the individual can recall from a list. Two random effects were specified to account for the individual level intercept and the slope with respect to age. The two-dimensional Gaussian-Hermite quadrature approximation was applied to numerically approximate the marginal likelihood. If we set the number of nodes to 20, then for a two-dimensional Gaussian-Hermite quadrature approximation, the integrand function will be calculated 400 times. If we want to extend the general mixed model, allow it to model two time-dependent outcomes simultaneously and specify two random effects for each time-dependent outcome, the marginal likelihood will be a four-dimensional integration. If we still set the number of nodes to 20, then the integrand function will be calculated for 20^4 times. This will make the maximisation of the marginal likelihood very inefficient. Therefore we need a different method to approximate the marginal likelihood.

The Monte Carlo integration can be used to approximate the integration:

$$E[f(\boldsymbol{\theta})|\mathbf{x}] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)}), \tag{6.4}$$

where $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$ are samples drawn from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. The naive Monte Carlo integration performs poorly in high dimensions as well. As the dimension grows, the range of the integral will be too large to be sampled from. However, it is not necessary for the samples to be independent. The Markov Chain Monte Carlo (MCMC) methods draw samples from a Markov chain having equilibrium distribution $p(\boldsymbol{\theta}|\mathbf{x})$. This provides an efficient method to calculate the

high-dimensional integration efficiently.

However, since the MCMC methods approximate the integral by sampling, it introduces randomness when approximating the marginal likelihood. Using the general purpose optimiser as we did in Chapter 5 might lead to problematic results.

To alleviate the problems above, we specify weakly informative priors to the parameters of the joint model. We then make inferences based on the posterior distributions, which ideally converge towards the maximum likelihood estimates (MLE). The data analysis is performed directly using the program WinBUGS 1.4. WinBUGS is a statistical software developed in the 1990s (Lunn et al. (2000)). WinBUGS implements MCMC algorithms, such as the Gibbs sampler and the Metropolis-Hastings algorithm, to obtain posterior samples from Bayesian models. Although WinBUGS may be outdated and less efficient for more complicated models, it is powerful enough for the joint model with multi-state model and general mixed model and is robust in estimation.

By incorporating Bayesian inference, our objective is to leverage simulation-based estimation for approximating the marginal likelihood. This thesis stays dedicated to the frequentist approach and does not involve a transition to the Bayesian paradigm.

6.3 Model Comparison Criteria

The deviance information criterion (DIC) is a Bayesian version generalisation of the model comparison criteria Akaike information criterion (AIC). It measures the adequacy of a model and also penalise if the model complexity increases. The deviance is defined based on the posterior distribution of the likelihood:

$$D(\mathbf{x}, \boldsymbol{\theta}) = -2 \log p(\mathbf{x}|\boldsymbol{\theta}). \tag{6.5}$$

The DIC is defined as

$$DIC = D(\mathbf{x}, E[\boldsymbol{\theta}]) + 2p_D, \tag{6.6}$$

where $\boldsymbol{\theta}$ is the set of all model parameters. The term $D(\mathbf{x}, E[\boldsymbol{\theta}])$ is known as the plug-in deviance, which is evaluated using the posterior means of the model parameters. The model complexity is described using the effective number of parameters p_D, which is defined as

$$p_D = E[D(\mathbf{x}, \boldsymbol{\theta})] - D(\mathbf{x}, E[\boldsymbol{\theta}]), \tag{6.7}$$

where $E[D(\mathbf{x}, \boldsymbol{\theta})]$ is the expected deviance. The model with a smaller DIC is preferred. However, there exists a limitation for using the DIC, it is only valid when the posterior distribution is asymptotically normal.

6.4 Bayesian Inference for Joint Models

In this section we will introduce the Bayesian inference for the joint model, then we will illustrate our method using the clinical data for sepsis from the PysioNet challenge Reyna et al. (2019b). The joint model consists of a sub-model for the longitudinal data and a sub-model for the competing risks. We extend the general mixed model so that it can model multiple time-dependent covariates simultaneously. For computational simplicity, in this study we only take into account two time-dependent covariates. Due to the low observing frequencies, there are no enough information to fit mixed models to the lab signs. In addition, due to the high missingness rate in the vital signs, we apply the general mixed model only to the three vital signs *heart rate* (HR), *mean arterial pressure* (MAP), and *respiration rate* (Resp)

To model the disease progression of sepsis, the competing-risk sub-model is a three-state multi-state model. As we discussed in Chapter 2, the three states are: s = 1 (under risks of sepsis), s = 2 (sepsis onset) and s = 3 (non-sepsis/dropout). The transitions are allowed from state 1 to state 2, and from state 1 to state 3. Though we have labels provided by clinicians which indicate whether an individual has sepsis or not, because we consider the labels as imprecise and want to generate labels that allow for early diagnosis, we assume both the transitions to the state 2 and state 3 are

interval censored, because it can happen between the follow-ups. On the other hand, since the follow-up time is 1 hour, modelling the transition as interval-censored or exact censored should be very similar.

The dependencies of the transition intensities on the covariates are modelled by the transition-specific hazard regression models, given by:

$$q_{rs}(t_{ij}|\mu_{ij}^{(1)},\mu_{ij}^{(2)}) = q_{0.rs}\exp\left(\eta_{rs}t_{ij} + \alpha_{rs}^{(1)}\mu_{ij}^{(1)} + \alpha_{rs}^{(2)}\mu_{ij}^{(2)} + \lambda_{rs}^{(1)}gender_i + \lambda_{rs}^{(2)}age_i\right), \tag{6.8}$$

where the time-scale t is time in hours, re-scaled by $1/10$, and $gender = 0/1$ for women/men. The term *age* represents the age of each individual. Since the clinical data of sepsis is observed on the scale of hours, the covariate *age* is a constant for each individual. $\mu_{ij}^{(1)}$ and $\mu_{ij}^{(2)}$ are expected values of two different time-dependent covariates for individual i at time point j.

Now we construct the sub-model for the time-dependent covariates. We start with the sub-model for time-dependent covariates with no random effects (Model A):

$$\begin{aligned} y_{ij}^{(k)} &= \mu_{ij}^{(k)} + \varepsilon_{ij}^{(k)}, \qquad \varepsilon_{ij}^{(k)} \sim \mathcal{N}(0,{\sigma^{(k)}}^2) \\ \mu_{ij}^{(k)} &= \beta_0^{(k)} + \beta_1^{(k)}t_{ij}, \end{aligned} \tag{6.9}$$

where $k = 1,2$. The error terms of the longitudinal outcomes are modelled by normal distribution with mean 0 and a constant variance.

Based on the Equation (6.8), the transition probabilities moving from state 1 to state 2 or state 3 over the 1 hour time interval $(t_{ip}, t_{ip}+1)$ are given by:

$$\begin{aligned} P_{1,2}(t_{ip},t_{ip}+1) &= \Big(1-\exp(-q_{1,1}(t_{ip}))\Big)\frac{q_{1,2}(t_{ip})}{q_{1,2}(t_{ip})+q_{1,3}(t_{ip})}, \\ P_{1,3}(t_{ip},t_{ip}+1) &= \Big(1-\exp(-q_{1,1}(t_{ip}))\Big)\frac{q_{1,3}(t_{ip})}{q_{1,2}(t_{ip})+q_{1,3}(t_{ip})}. \end{aligned} \tag{6.10}$$

To make inference on the posterior distributions using WinBUGS, the prior distributions needed to be specified. For each of the 14 fixed-effects parameters β's, η's and λ's, weakly informative prior densities are specified by $p(\cdot) \sim N(0, 100^2)$.

For the standard deviations σ_k's, we specify vague priors $U(-10,10)$. Van den Hout (2017) presents the implementation of the Bayesian multi-state continuous-time survival model in BUGS. Since we have a three-state sub-model for the competing risks, we adopt a similar model implementation. It is proposed that the survival data observed in each time-intervals $(t_{i1},t_{i2}],(t_{i2},t_{i3}],\ldots,(t_{i,N-1},t_{iN}]$ should be modelled independently. Let $S_{i,j}$ denote the state of individual i at time j, and the transitions to state s_{ij} given state $s_{i,j-1}$ should be modelled with multinomial distribution, where the corresponding probabilities equal to the transition probabilities at time point j. For convenience, we code the state transitions in each hourly interval as $(1,0,0),(0,1,0)$ and $(0,0,1)$, and denote the observed code at each time point as $\bar{S}_{i,j}$. Then we specify the likelihoods as

$$S_{i,j}|s_{i,j-1},\boldsymbol{\theta} \sim Multinomial(\mathbf{p},1), \tag{6.11}$$

where $\boldsymbol{\theta}$ is the vector with the parameters in equations (6.8),(6.9) and (6.10), and $\mathbf{p}=(P_{1,1}(t_{ij}),P_{1,2}(t_{ij}),P_{1,3}(t_{ij}))$.

Now we extend the sub-model for the time-dependent covariates based on Equation (6.9) in Model A. For each of the time-dependent covariates, we introduce two random effects to take into account the individual-level intercept and slope (Model B). The equations are as follows:

$$\begin{aligned} y_{ij}^{(k)} &= \mu_{ij}^{(k)} + \varepsilon_{ij}^{(k)}, \qquad \varepsilon_{ij}^{(k)} \sim \mathcal{N}(0,{\sigma^{(k)}}^2) \\ \mu_{ij}^{(k)} &= \beta_0^{(k)} + b_0^{(k)} + (\beta_1^{(k)} + b_1^{(k)})t_{ij}. \end{aligned} \tag{6.12}$$

We assume that all the random effects $\mathbf{B}=(B_0^1,B_2^1,B_1^2,B_2^2)$ follows a multivariate normal distribution with mean $\mathbf{0}$ and the variance-covariance matrix Σ. The correlations between two time-dependent covariates are also taken into account by the variance-covariance matrix Σ.

The implementation of this model is similar to the Model A. In addition, we need to specify priors for the four-dimensional random effects $\mathbf{B}$:

Model	Vital Signs	p	DIC
Model A 1-3	HR and MAP	24	615386.7
	HR and Resp	24	565218.0
	MAP and Resp	24	554004.5
Model B 1-3	HR and MAP	28	218099.6
	HR and Resp	28	140700.5
	MAP and Resp	28	169099.4

Table 6.1: Comparison between joint models for sepsis data (N=1000). p is the number of parameters used in the model.

$$\begin{aligned} \mathbf{B} &\sim N(\mathbf{0}, \Sigma), \\ \Sigma^{-1} &\sim W_4(\mathbf{R}, p), \end{aligned} \tag{6.13}$$

where W_4 represent a four-dimensional Wishart distribution, $\mathbf{R} > 0$ is the scale matrix and p is the degree of freedom. In Bayesian statistics, the Wishart distribution is used as the conjugate prior for the precision matrix of the multi-variate normal distribution.

We fit both Model A and Model B to pairwise combinations of three vital signs. The values of the observed vital signs and the time t were rescaled to $1/10$ to avoid numerical problems. We generate the initial values using the package `lme4`. For all the models, the MCMC consists of two chains with each 10,000 sample iteration. Among them, the first 4000 iterations are in the burn-in procedure and therefore discarded. The DIC of the models are displayed in Table 6.1.

According to the Table 6.1, joint model fitted time-dependent covariates *heart rate* and *respiration rate* (Model B (2)) gives the smallest DIC among all models we fitted.

The posterior mean and the 95% credible intervals are shown in Table 6.2. The estimates $\hat{q}_{0.12} = 0.0003$ and $\hat{q}_{0.13} = 0.053$ indicating transition intensity to state 3 is much higher than the transition intensities to state 2 at baseline ($t = 0$). The order of magnitude of $\hat{q}_{0.12}$ is smaller than $\hat{q}_{0.13}$, and this in line with our observation in Table 2.4 that it is 10 times more likely to move to state 3 (non-sepsis) than state 2 (onset of sepsis).

The parameters $\alpha_{12}^{(1)}$, $\alpha_{13}^{(1)}$ and $\alpha_{12}^{(2)}$, $\alpha_{13}^{(2)}$ establish the connection between the multi-state model and the time-dependent covariates *heart rate* and *respiratory rate* respectively. The credible interval of $\hat{\alpha}_{12}^{(1)}$ includes 0, indicating that $\hat{\alpha}_{12}^{(1)}$ is not significantly different from 0. The estimated value $\hat{\alpha}_{13}^{(1)} = -0.005$ has a credible interval excluding 1, indicating that $\hat{\alpha}_{13}^{(1)}$ is significantly smaller than 0. Hence for individuals who have higher heart rate will have lower hazard to transit to state 3. Similarly, the credible interval of $\hat{\alpha}_{13}^{(2)}$ includes 0, and the credible interval of $\hat{\alpha}_{13}^{(2)} = -0.205$ excludes 0, indicating that individuals who have a higher respiratory rate will have a lower hazard to transit to state 3 (non-sepsis). This is in line with the syndrome of sepsis.

Furthermore, the estimate $\hat{\lambda}_{13}^{(1)} = -0.166$ is smaller than 0, implying that men have a lower hazard for transitioning from state 1 to state 3 compared to women. The estimate $\hat{\lambda}_{13}^{(2)} = 0.403$ is greater than 0, implying that older individuals have a higher hazard for transitioning from state 1 to state 3. The estimate $\hat{\eta}_{12} = 0.161$ and $\hat{\eta}_{13} = 1.051$ are greater than 0, indicating that as time increases, individuals will have a higher hazard for transitioning out of state 1. The credible intervals of $\hat{\eta}_{12}$ and $\hat{\eta}_{13}$ both exclude 0, and are non-overlapping, indicating that the difference between $\hat{\eta}_{12}$ and $\hat{\eta}_{13}$ is significant, as time increases, individuals have higher hazards to transit in to state 3.

It should be noted that for parameters describing transitions from state 1 to state 2, the credible intervals of their estimates are relatively large, especially the credible intervals for $\hat{\alpha}_{12}^{(2)}$, $\hat{\lambda}_{12}^{(1)}$ and $\hat{\lambda}_{12}^{(2)}$, this might due to there are relatively fewer individuals transit to state 2 than state 3.

Coefficients $\beta_0^{(1)}, \beta_1^{(1)}$ and $\beta_0^{(1)}, \beta_0^{(2)}$ are in the mixed model. The estimated values of the slopes for time $\hat{\beta}_1^{(1)} < 0$ and $\hat{\beta}_1^{(2)} < 0$ show that when time increases, the heart rate and respiratory rate will decline. The parameters $\sigma^{(1)}$ and $\sigma^{(2)}$ are standard deviations of the residual errors of covariates *heart rate* and *respiratory rate*. The estimates $\hat{\sigma}^{(1)} = 1.972$ indicating that the variations in the heart rate is quite large.

The variance-covariance matrix Σ describes the correlations between the ran-

dom effects of the two time-dependent covariates as specified in Equation (6.13). $\hat{\Sigma}_{11} = 2.499$, $\hat{\Sigma}_{22} = 0.196$, $\hat{\Sigma}_{33} = 0.154$ and $\hat{\Sigma}_{44} = 0.016$ are the estimated values of the variances of the random effects $b_0^{(1)}, b_1^{(1)}, b_0^{(2)}$ and $b_1^{(2)}$. The estimated covariance between the random intercept and slope of the covariate *heart rate* is $\hat{\Sigma}_{34} = -0.022 < 0$, implying that the two random effects are negatively correlated. The estimated covariances $\hat{\Sigma}_{13} = 0.183$, $\hat{\Sigma}_{23} = -0.036$, $\hat{\Sigma}_{14} = -0.019$ and $\hat{\Sigma}_{24} = 0.020$, captures the dependencies between the two time-dependent covariates. $\hat{\Sigma}_{13} > 0$ implying that $b_0^{(1)}$ and $b_0^{(2)}$ are positively correlated. This implies that individuals with higher baseline heart rates are likely to have higher baseline respiratory rates. Similarly, $\hat{\Sigma}_{24}$ implies the slope that the heart rate and respiratory rate changes with respect to time are positively correlated.

Longitudinal sub-model			Multi-state sub-model		
	Value	95% CI		Value	95% CI
$\beta_0^{(1)}$	8.532	(8.486; 8.569)	$q_{0.12}$	0.0003	(0.0001; 0.0009)
$\beta_1^{(1)}$	-0.741	(-0.085; -0.061)	$q_{0.13}$	0.053	(0.035; 0.087)
$\beta_0^{(2)}$	1.812	(1.796; 1.834)	η_{12}	0.161	(0.120; 0.200)
$\beta_1^{(2)}$	-0.151	(-0.186; -0.116)	η_{13}	1.057	(1.033; 1.083)
$\sigma^{(1)}$	1.972	(1.920; 2.027)	$\alpha_{12}^{(1)}$	0.001	(-0.010; 0.012)
$\sigma^{(2)}$	0.700	(0.697; 0.702)	$\alpha_{13}^{(1)}$	-0.005	(-0.008; -0.002)
Σ_{11}	2.499	(2.294; 2.732)	$\alpha_{12}^{(2)}$	-0.079	(-0.531; 0.401)
Σ_{12}	-0.293	(-0.348; -0.243)	$\alpha_{12}^{(2)}$	-0.205	(-0.338; -0.083)
Σ_{13}	0.183	(0.140; 0.226)	$\lambda_{12}^{(1)}$	0.078	(-0.054; 0.1951)
Σ_{14}	-0.019	(-0.034; -0.005)	$\lambda_{13}^{(1)}$	-0.166	(-0.196; -0.123)
Σ_{22}	0.196	(0.174; 0.218)	$\lambda_{12}^{(2)}$	0.357	(-0.380; 0.891)
Σ_{23}	-0.036	(-0.049; -0.024)	$\lambda_{13}^{(2)}$	0.403	(0.218; 0.579)
Σ_{24}	0.020	(0.016; 0.026)			
Σ_{33}	0.154	(0.140; 0.168)			
Σ_{34}	-0.022	(-0.027; -0.019)			
Σ_{44}	0.016	(0.014; 0.018)			

Table 6.2: Model B (2) for the sepsis data (N=1000), fitted with time-dependent covariates heart rate and respiratory rate simultaneously. This table illustrates posterior means and 95% credible intervals for the parameters in the joint model.

For Model B (2), a fit to 9 randomly selected individuals is shown in Figure 6.1 and 6.2. The individual-specific random effects are estimated by the posterior expectation of the random effects. Let $\boldsymbol{\theta}^*$ denote the posterior mean of the parameters obtained from the Model B (2). The historical observations of the individuals for prediction are represented by $\mathbf{y}_i$. Then the posterior expectation of the random effects is given by Equation (6.14).

$$E[\mathbf{B}_i|\mathbf{y}_i] = \int \mathbf{b}_i p(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^*) d\mathbf{b}_i = \int \mathbf{b}_i \frac{p(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\theta}^*)}{p(\mathbf{y}_i)} d\mathbf{b}_i. \tag{6.14}$$

Due to the high rate of missing data, we did not present the fitting of joint models for other time-dependent covariates in this study. However, it should be noticed that the time-dependent covairates systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP) are all blood pressure measurements, which means they are highly correlated. When the covariates in a prediction model exhibit high correlation, it can significantly influence the model's performance. This correlation often leads to collinearity, make it difficult to assess the individual effects of each covariate. Moreover, it can lead to unstable coefficient estimates and inflated standard errors, affecting the reliability of the model. To avoid these problems, dimension reduction methods and be applied.

However, the primary focus of this study is on the joint model and disease diagnosis, which rely on the utilisation of a joint model and a time-dependent loss function. Hence we conduct our analysis on pairs of covariates to demonstrate the effectiveness of the joint model for multiple time-dependent covariates and competing risks. Additionally, when using the proposed joint model for prediction purposes, it becomes essential to predict the mean of the time-dependent covariates at each time point. The inclusion of multiple covariates, even when they are correlated, contributes to improving covariate prediction.

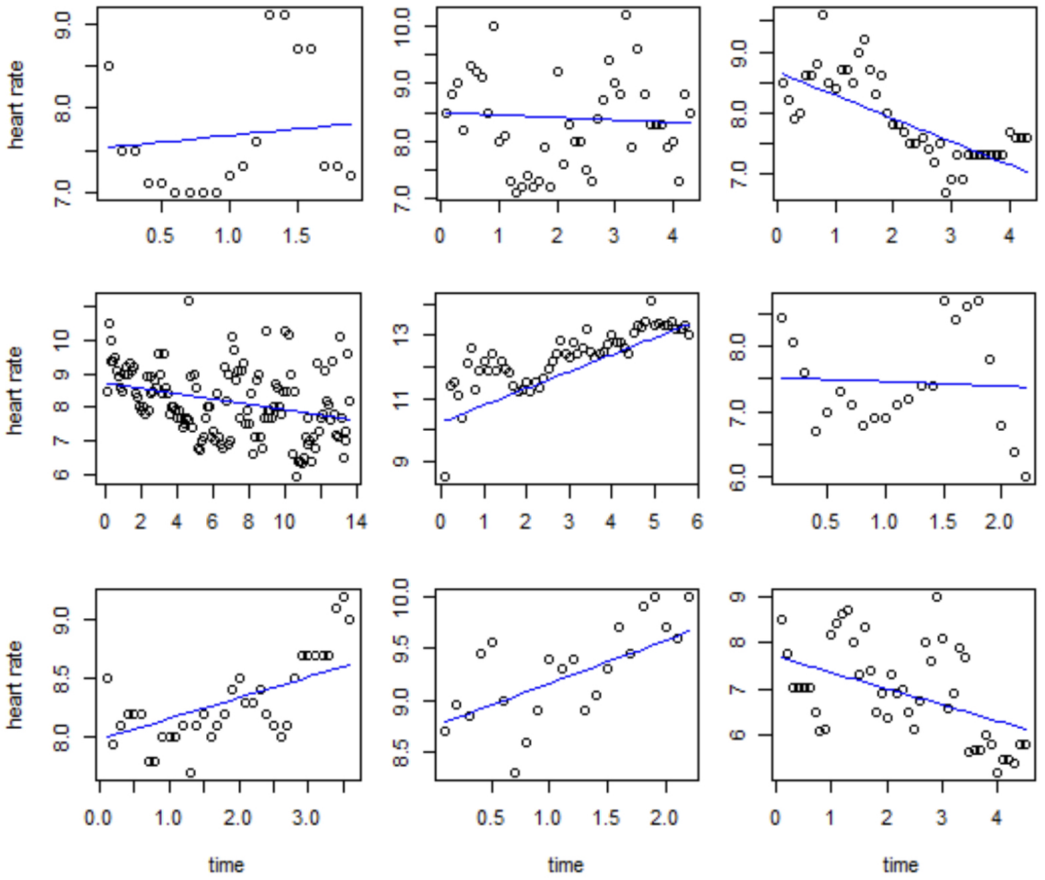


Figure 6.1: The observed heart rate (denoted by the dots:"o") and the fitted trajectories (presented by the blue lines) for 9 randomly selected individuals. The scales of the x and y axis in the plot is $1/10$ of the original data.

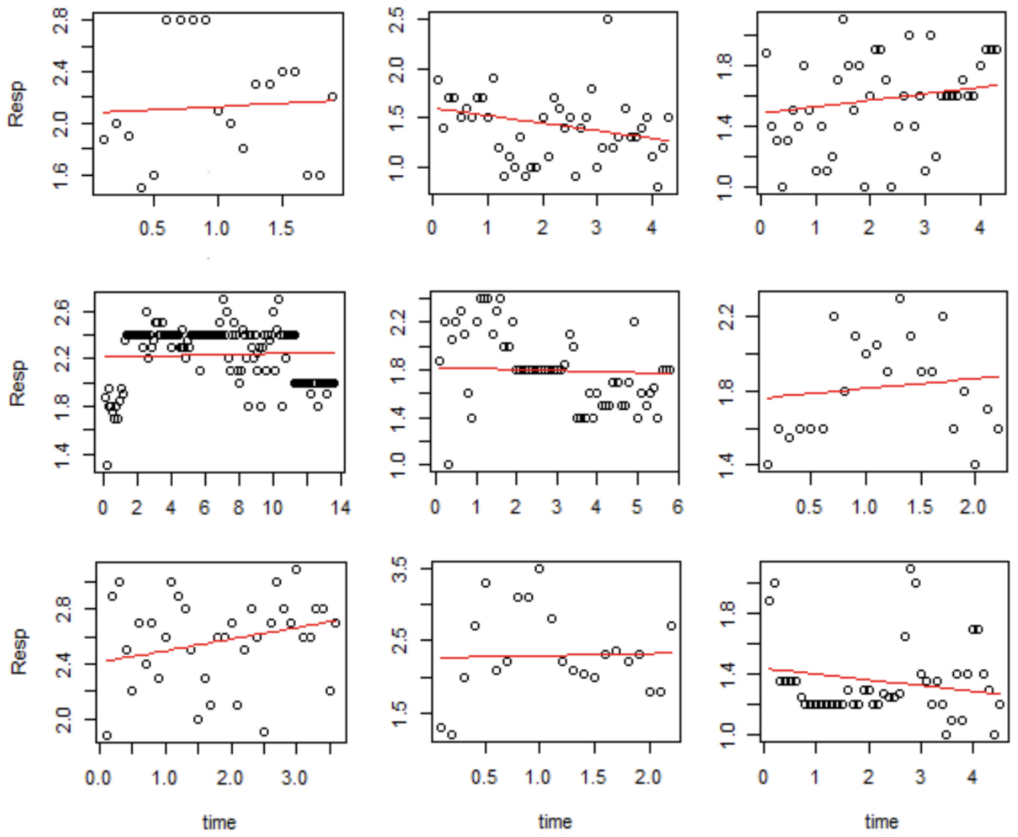


Figure 6.2: The observed respiratory rate (denoted by the dots:"o") and the fitted trajectories (presented by the red lines) for 9 randomly selected individuals. The scales of the x and y axis in the plot is $1/10$ of the original data.

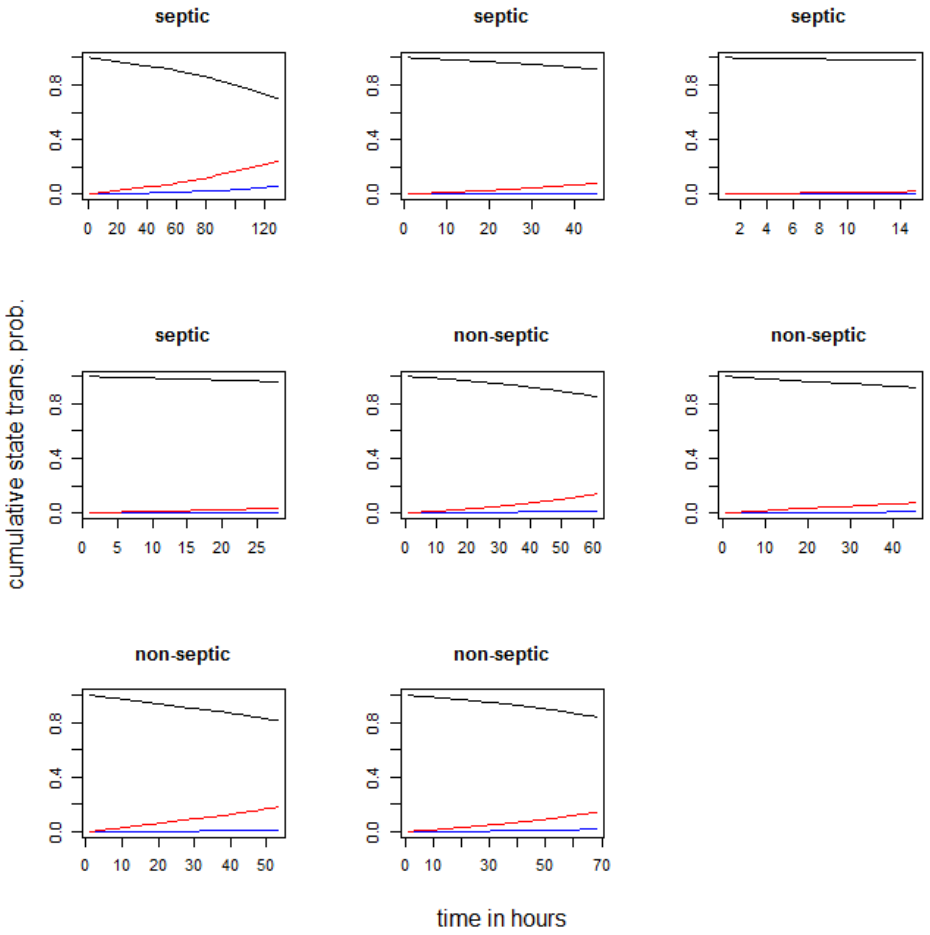


Figure 6.3: The estimated cumulative state transition probabilities for 8 randomly selected individuals. The colour black represents the transition probability from state $1 \rightarrow 1$, red represents the transition probability from state $1 \rightarrow 3$, and blue represents the transition probability from state $1 \rightarrow 2$. Subtitles septic/non-septic of each plot indicate whether the individual has sepsis or not. The time scale is in hours. The cumulative state transition probabilities are estimated till the end of the follow-up.

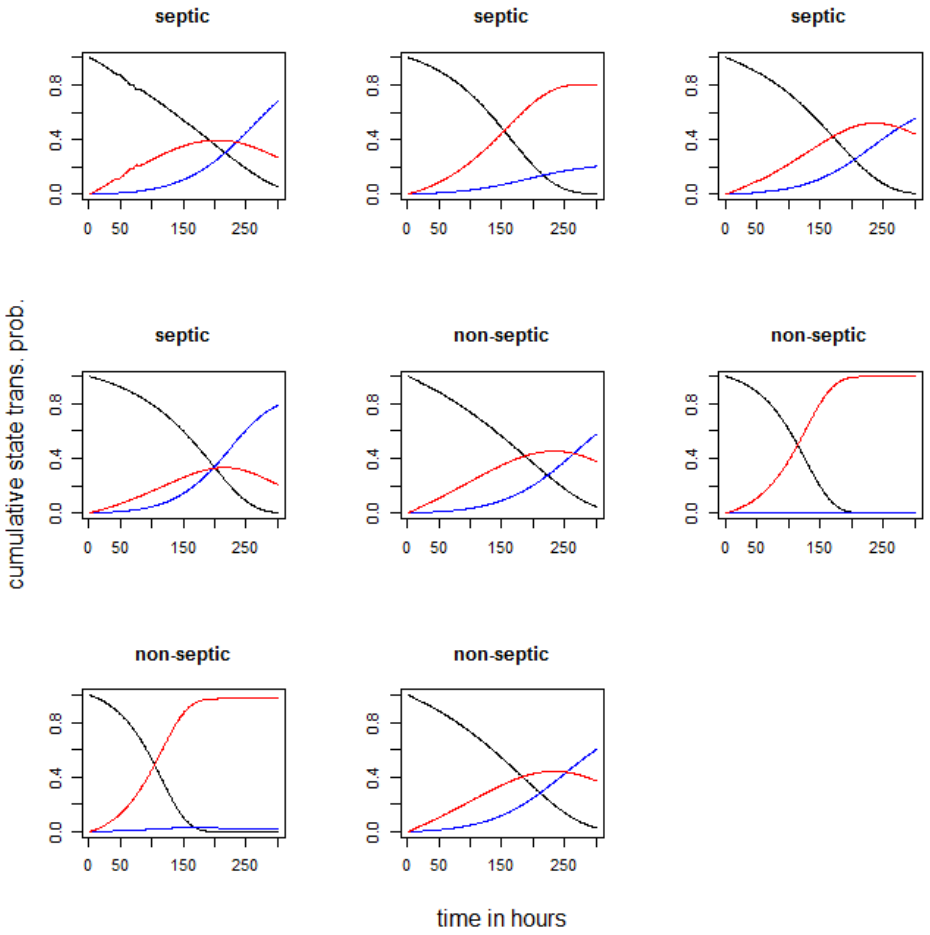


Figure 6.4: The estimated cumulative state transition probabilities for 8 randomly selected individuals. The colour black represents the transition probability from state $1 \rightarrow 1$, red represents the transition probability from state $1 \rightarrow 3$, and blue represents the transition probability from state $1 \rightarrow 2$. Subtitles septic/non-septic of each plot indicate whether the individual has sepsis or not. The time scale is in hours. The cumulative state transition probabilities are estimated for 300 hours.

6.5 Prediction

In this section, we illustrate the sepsis early diagnosis framework we proposed, applied to the clinical data of sepsis based on Model B (2) mentioned above. Additionally, we compare our method with a logistic regression model. The aim of this analysis is to explain and clarify the concept of the sepsis diagnosis framework we developed. By comparing it with a model based on logistic regression, we can better understand the unique aspects and benefits of our framework.

We estimated the parameters of Model B (2) and logistic regression using the same subsample of 1000 randomly selected individuals from the clinical data of sepsis. The details about this subsample can be found in Section 2.2.1. The parameter estimates for Model B (2) were presented in Table 6.2, and we will directly utilise the results in this section.

To evaluate the performance, both models were applied to a validation set with sample size 1000, which consists of a different subsample of the clinical data of sepsis. This evaluation allows us to assess how well the statistical models perform on previously unseen data. In this subsample, there are 64 sepsis individuals and 936 non-sepsis individuals. To better evaluate the performance of each model, we examine the prediction results at each time point. For convenience, if, at a certain time point, an individual is predicted to have sepsis based on the corresponding observation (and historical observations), we label the corresponding observation as positive. This subsample consists of a total of 39269 observations, according to the information provided by clinicians in the dataset, among them there are 934 observations have positive diagnosis results and 38335 observations have negative diagnosis results.

Now we explain how the sepsis early diagnosis framework works. Recall Equation (5.6), to predict the value of the label D for each individual we need to estimate the transition probabilities at each time interval then calculate the weighted summation. To be more specific, given the historical data of a random individual, we need to first estimate the individual-specific random effects given the historical observations. Based on that, we can further estimate the mean of the time-dependent

covariates, calculate transition intensities and transition probabilities, and finally use these values to make the diagnosis.

Figure 6.1 and 6.2 illustrate that for each individual, given the historical data, our method can provide reasonable estimates of the individual level random effects for the two longitudinal outcomes heart rate and respiratory rate included in Model B (2). The mean of the time-dependent covariates estimated reflects the trend of the change in the longitudinal outcomes.

Figure 6.3 illustrates the estimated transition probabilities of 8 randomly selected individuals. 4 of the individuals are septic indiviudals and 4 of them are non-septic indiviudals. For any individual at any time point, the estimates of the transition probabilities from state 1 to state 2 are small compared to the transition probabilities to state 3. This result is expected, since the proportion of septic individuals in the whole data set is relatively small. However, we can still observe that for indiviudals finally end up in state 3 (i.e. nonseptic) the transition probabilities to state 3 are significantly larger compared with septic individuals.

Figure 6.4 also illustrates the estimated transition probabilities for 300 hours for 4 septic individuals and 4 non-septic individuals. In general, transition probabilities to state 2 increases faster for the septic indiviudals compared with non-septic indiviudals.

At each time point, given the observed historical data, we are able to estimate the individual-level intercepts and slopes for the two time-dependent covariates, then we are able to estimate the averages of the two covariates from time point $t = 0$ to $t = 300$. Based on this we can estimate the hourly transition probabilities P_{12}. We can finally apply Equation (5.6) to predict the labels, and use the utility function described in Section 5.2 to compute the utility Score U_{total}.

The choice of the loss function has a significant impact on the predictions. We perform predictions by rescaling L_{01} using a rescale parameter. For each rescaled L_{01} value, we compute the true positive rate (TPR) and false positive rate (FPR). Note that the TPR is the proportion of actual positive will test positive, while the FPR is the proportion of negative events wrongly categorised as positive. We use

different values of L_{01} from 5×10^{-4} to 10, to obtain the corresponding TPR and FPR respectively. The resulting receiver operating characteristic curve (ROC) is illustrated in Figure 6.5.

The reason we only rescale L_{01} is that, unlike L_{10}, it is a scalar and therefore easier to manipulate. In addition, by only modifying L_{01} we can still maintain the shape of the function L_{10} which could encourage early diagnosis of sepsis.

It should be noted that the ROC curve visually represents the trade-off between the true positive rate (TPR) and false positive rate (FPR) at different threshold values. In our method, the loss functions L_{01} and L_{10} serve as thresholds. Generally, a ROC curve that is closer to the point (0,1) indicates better overall performance. However, selecting the optimal threshold depends on the specific problem, the desired trade-off between false positives and false negatives, and the associated costs or consequences of misclassification. In our case, if we use the utility score introduced in Section 2.2.3 as an assessment of performance, it is crucial to note that the penalty for falsely classifying a septic individual as non-septic (false negative) is much more severe than falsely classifying a non-septic individual as septic.

We found that the loss functions defined in Chapter 5 tend to penalise false negative predictions too much, resulting in a high number of false positive predictions. When we used $L_{01} = 0.05$ in the prediction procedure, out of 39269 observations, only 348 were false negatives, but 13721 observations were false positives. The true positive rate was 0.627, and the false positive rate was 0.358. This indicates a trade-off where a higher true positive rate is achieved at the cost of a higher false positive rate.

To address this imbalance, we tuned the value of the loss function from $L_{01} = 0.05$ to $L_{01}^* = 0.5$. With this adjustment, 544 observations were false negatives, while only 1,579 observations were false positives. The number of false positive observations decreased significantly compared to the previous case. This demonstrates that the choice of the loss function value, such as $L_{01} = 0.05$, has a substantial impact on the prediction performance. By increasing the penalty for false positives, the false positive rate will be effectively reduced.

We compare our method with a logistic regression model. Logistic regression is a widely used model for predicting binary outcomes based on predictor variables. In addition, performing logistic regression is straightforward using R programming. For simplicity, when fitting the logistic regression, the correlation within individuals were ignored. Each observation are assumed to be independent. The binary response variable in this context is defined as whether or not an individual is diagnosed with sepsis based on the observations recorded at the corresponding time points. Similar to our method, we include covariates *heart rate*, *respiratory rate*. In addition, we take into account the individual's gender and age, as well as the corresponding observation time t_{ij}. The relationship between the covariates and the response D_{ij} can be specified as follows:

$$p(D_{ij}=1)=\frac{1}{1+\exp^{-(\beta_0+\beta_1 y_{ij}^{(1)}+\beta_2 y_{ij}^{(2)}+\gamma_1 gender_i+\gamma_2 age_i+\eta t_{ij})}}, \tag{6.15}$$

where $y^{(1)}$ and $y^{(2)}$ are longitudinal outcomes of heart rate and respiratory rate.

After obtaining the predicted probabilities for each individual in the validation set, we can predict whether the individual has sepsis or non-sepsis by selecting a threshold value. Individuals with predicted probabilities above this threshold will be classified as septic individuals, while those with probabilities below it will be classified as non-septic.

For example, if we set the threshold to be 0.05, meaning we classify D_{ij} as 1 when $p(D_{ij}=1)>0.05$, the resulting TPR and FPR will be 0.169 and 0.013, respectively. In this case, there are 776 observations classified as false negatives, while only 489 observations are classified as false positives.

On the other hand, if we lower the threshold to 0.025, the number of false negative observations decreases to 634, but the number of false positive observations increases to 1,089.

For different threshold values between 0 and 1, we calculate the TPR and FPR respectively, which are illustrated in Figure 6.5.

As discussed earlier, our proposed method demonstrates that the choice of loss

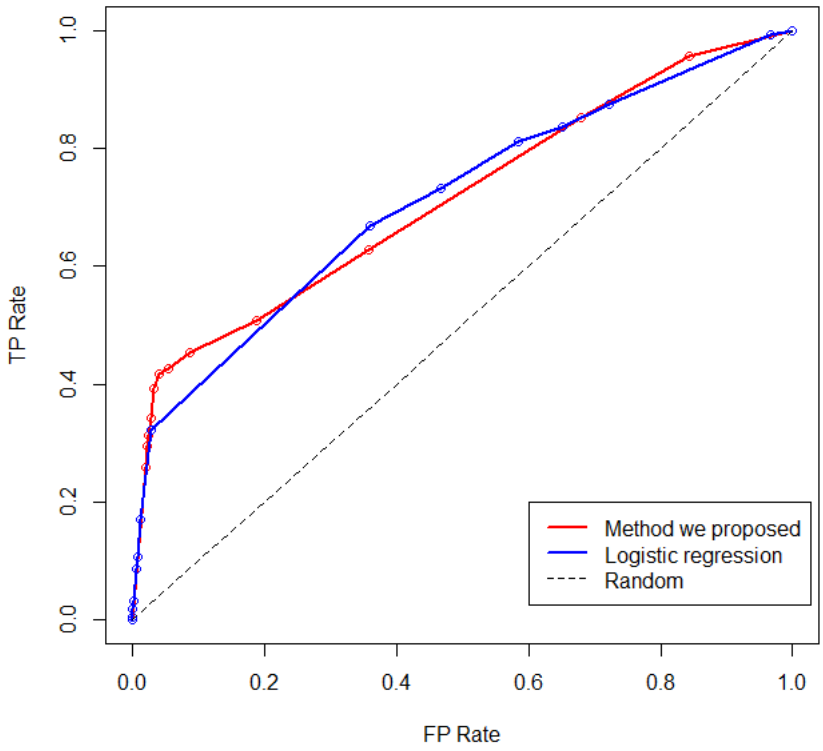


Figure 6.5: ROC of prediction based on method we proposed and the logistic regression.

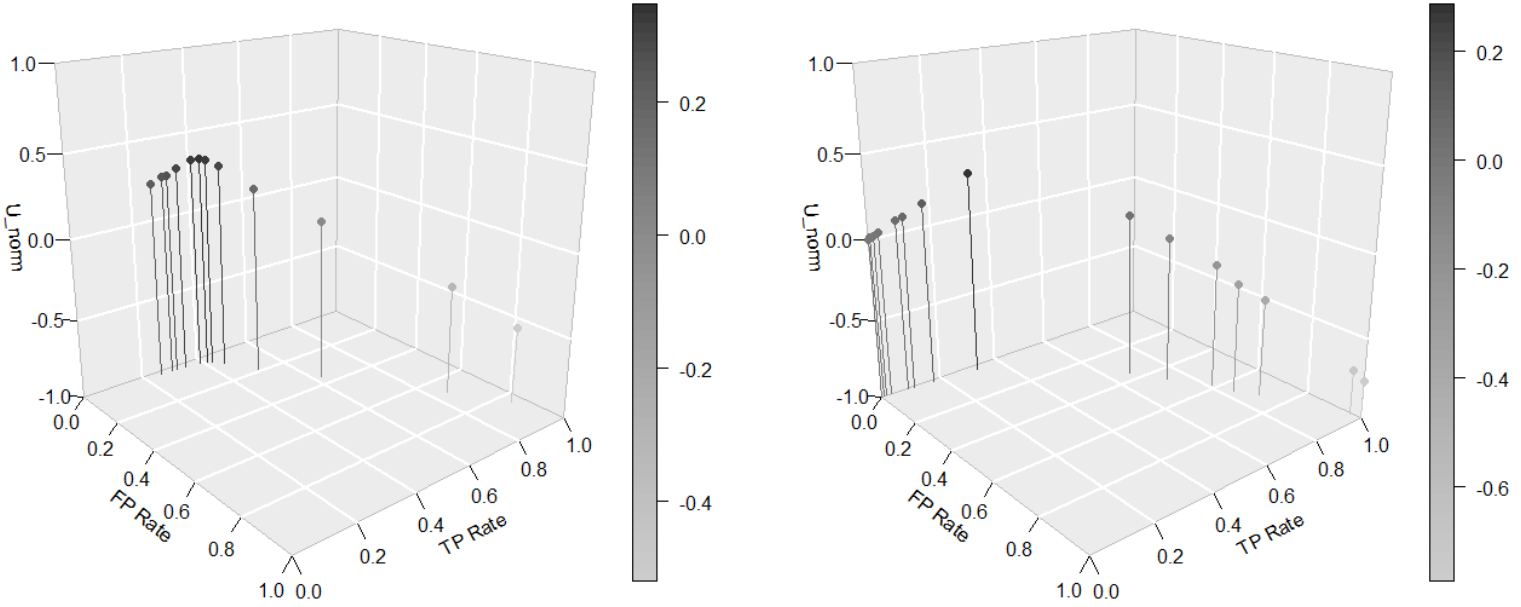
function affects the prediction results. Similarly, when employing a logistic regression model, the selection of different thresholds leads to variations in the prediction outcomes. Therefore to better evaluate the performance of both methods using the normalised utility score defined in Equation (2.1), for both methods, we computed the normalised utility scores using different loss functions/thresholds. We illustrate the normalised utility score with the corresponding TPR and FPR. The results are illustrated in Figure 6.6.

It is clear from the Figure 6.6 that, as both TPR and FPR increase from 0 to 1, the normalised utility score initially rises and then declines. This trend suggests that initially increasing the number of correctly identified positive instances (higher TPR) and minimising false positive predictions (lower FPR) positively impacts the normalised utility score. When both TPR and FPR are 0, it indicates that no individuals are classified as sepsis. In this case, the total utility score is equivalent to

the utility score when all predictions are 0, resulting in a normalised utility score of zero. When both TPR and FPR approach 1, it implies that most individuals are predicted as sepsis. Although the utility score penalises false positive predictions less in this scenario, the cumulative penalty on the entire dataset remains substantial since the majority of individuals are non-sepsis. Consequently, the normalised utility score decreases as a result.

By comparing the two plots in Figure 6.6, it can be observed that our proposed method consistently achieves higher normalised utility scores compared to the method based on logistic regression. Starting from an initial value of $U_{\text{norm}} = 0.237$, our method reaches its peak at $U_{\text{norm}} = 0.348$ with a true positive rate (TPR) of 0.418 and a false positive rate (FPR) of 0.041. On the other hand, the logistic regression method starts at $U_{\text{norm}} = 0.008$ and reaches its highest point at $U_{\text{norm}} = 0.286$ with a TPR of 0.321 and an FPR of 0.028. Overall, our method consistently outperforms the logistic regression method in terms of normalised utility scores.

This result suggests that our method captures and utilises the available information more effectively, leading to better predictions and decision-making. The higher normalised utility scores achieved by our proposed method implies that our method has the potential to be used in sepsis-early diagnosis problem. With further extensions, it can be applied to other disease diagnosis problems.



(a) Method we proposed

(b) Logistic regression

Figure 6.6: Relationship Between Normalised Utility, True Positive Rate (TPR), and False Positive Rate (FPR) for both methods.

Chapter 7

Conclusion

In many medical research, it is common that both longitudinal and time-to-event data are available. Simultaneously modelling the longitudinal data and the time-to-event data is important. This can be achieved by applying a joint model that consists of a sub-model for the longitudinal data and a sub-model for the survival data, where the two sub-models are linked by random effects.

In this study, we proposed to use a general mixed model for the modelling of the longitudinal outcomes. It allows for a more flexible assumption of the shapes of the longitudinal outcomes. In addition, we proposed an extension of the sub-model for the longitudinal outcomes, which allows us to model the correlated longitudinal outcomes simultaneously. Furthermore, we proposed a new framework for the early diagnosis of sepsis based on the joint model. The joint model combining a three-state sub-model and a general mixed sub-model was used to model the survival data and the longitudinal data simultaneously. For the prediction procedure, to achieve early diagnosis, we proposed a time-depending loss function to penalise late prediction. The diagnosis of each individual at each time point was made based on the expected loss.

However, there remain more challenges. For instance, longitudinal data usually have high-dimensional and complicated structures in practice, which imposes difficulties on the model choice and variable selections. It also increases the time needed to fit the model. Additionally, in this study, when fitting the joint model, we directly use the diagnosis outcomes provided by the clinicians, which might be

imprecise. We created an ad hoc time-dependent loss function to achieve early diagnosis and made predictions based on it. However, this might introduce undesired artifacts. In this section, we will discuss the potential improvements of our proposed method.

Extending joint models to accommodate complex data structures has been the subject of extensive research. As we introduced in Section 1.2, many examples of such extensions can be found in the literature. For instance, incorporating association structures between time-to-event data and longitudinal data, as well as extending time-to-event data to handle competing risks. In particular, there has been a lot of research on modelling time-dependent covariates.

In this study, we extended the mixed model for the longitudinal data by relaxing the normal assumption of the error term and proposed to use the skew normal distribution for the error term. Additionally, we took into account the variability in the variance of the error term over time. We applied our model to the ELSA data and observed consistently lower AIC values compared to the alternatives using normal error assumptions. This result highlights that our model fits the data better and performs better overall. It demonstrated that the proposed joint model effectively captures the underlying patterns in the data compared to existing approaches. As a result, it can provide reliable coefficient estimates and accurate predictions for disease progression. To conclude, the joint model we proposed demonstrates its ability to handle longitudinal data with complex error distributions, showing its generalizability and practicality.

In our study, our primary focus was on extending the distribution of the error term while specifying the linear effects of covariates. We restricted our methods to parametric linear prediction functions due to limitations in computation resources and time. Section 7.1 will provide a more detailed discussion of potential future research regarding the modelling of time-dependent covariates within joint models.

In addition, we have proposed a joint model that considers competing risks and multiple time-dependent covariates. However, as the number of random effects increases within the joint model, integration in a high-dimensional space becomes

necessary, leading to numerical challenges. To tackle this issue, we have proposed to use the simulation-based integration to approximate marginal likelihoods. However, this approach makes the optimisation process unstable, which makes it difficult to obtain reliable estimates. To overcome this problem, we introduced Bayes inference and specified weakly informative priors, therefore the posteriors converge towards the maximum likelihood estimates (MLE). The application of this method to clinical data on sepsis has demonstrated its effectiveness in accurately estimating the model parameters.

For the sepsis early diagnosis problem, the approach we proposed involves using a joint model to simultaneously model the time-dependent covariates and time-to-event data. This enables us to predict the value of the time-dependent covariates at any given time point and estimate the transition probabilities over any given time interval. Based on the transition probabilities, we can compute the expected loss and make predictions. In contrast to machine learning algorithms introduced in Section 1.5, our method offers good interpretability. Within this framework, the loss function we used for minimising the expected loss is specifically tailored for the sepsis early diagnosis problem. However, it is worth noting that the method can be generalised by adapting the loss function to other diseases, based on the knowledge of clinicians. We will discuss the potential improvement of the framework we proposed for the disease diagnosis problem in Section 7.2.

7.1 Modelling Time-Dependent Covariates

The modelling of longitudinal outcomes is an important part of the joint model. Based on the types of the longitudinal outcomes, there are various types of regression models can be chosen from. The most commonly used sub-models are linear regression models and generalized linear regression models. However, applying these always requires rigorous assumptions. For example, the error terms of longitudinal outcomes should follow a normal distribution, or distributions from the exponential family. However, this assumption can be easily violated, since in many

research the longitudinal outcomes usually have complicated structures. In this study, we proposed a general mixed sub-model for the longitudinal data based on the GAMLSS method. In the data analysis sections, for simplicity, we only fitted parametric models to the ELSA data and the Sepsis data. However, compared to the parametric approach, non-parametric models have more flexibility. The regression curves can be estimated without making strong assumptions about the shape of the true regression function. Detailed discussions about the non-parametric regression can be found in Härdle (1990). Since the GAMLSS are semi-parametric regression models, based on it we can extend the current model to the semi-parametric and even the non-parametric models.

Another challenge in using joint modelling for the diagnosis of sepsis is imposed by the large number of outcomes with different characteristics. In the study of the early diagnosis for sepsis, the dataset consists of 6 time-invariant demographic variables and 34 time-varying markers, where 26 of them are sparse longitudinal data. The complex structure makes it difficult to correctly specify the dependence of the transition intensities on the markers and the trajectory of the markers. In addition, a heavy computational burden will be imposed by incorporating all markers into the joint model. Furthermore, it was discussed that the high dimensionality has an adverse impact on both classification and regression problems, for instance in Fan and Fan (2008) and Fan and Lv (2010).

One way to solve this problem is to incorporate variable selection into the multi-state model. In Fan and Lv (2010) a detailed literature review on variable selection methods is provided. In particular, the penalty-based variable selection schemes, for instance, the LASSO (Tibshirani (1996)) can be applied. By imposing constraints on the coefficients, this parameter estimation of the regression model and the variable selection can be done simultaneously.

Apart from variable selection, there are other ways to incorporate high dimensional covariates. In Ferrer et al. (2019) and Suresh et al. (2017), two approaches for dynamic predictions: joint modelling and landmarking were compared and discussed. It was suggested that if the model is correctly specified, the joint model has

better performance. But when it is difficult to estimate the stochastic process of the time-varying markers, for example when the longitudinal data are sparse and the dimension of the longitudinal data is high, the landmark approaches can provide a good approximation.

7.2 Modelling the Uncertainty in Labels

In this thesis, a multi-state model is used for the modelling of the time-to-event data. Currently, in the multi-state model, we assumed that the transitions between states are exactly observed. However, for septic patients, the time of diagnosis by clinicians is always later than the real onset time of sepsis. Therefore there exists uncertainty in the label provided in the dataset, and the assumption that the transitions are exactly observed is violated.

In order to take into account this uncertainty, we introduced a time-depending loss function, therefore in the prediction step, the diagnosis after the true sepsis onset time will be penalized. However, this will produce undesired artefacts, since the choice of the value of the loss function only depends on the utility score. To alleviate the problem above, methods that can take into account the uncertainties in the labels should be investigated in the future.

Since it is already known that true sepsis onset time is always earlier than the diagnosis time, it is reasonable to model the event time of ”onset of sepsis” as left censored. And by modelling the left-censoring in the data, we can take into account the uncertainty of the labels provided by clinicians and achieve higher classification accuracy.

To sum up, in this thesis, we extended the commonly used sub-models for the longitudinal outcomes. Using general mixed model for location, scale, and shape. It allows for the longitudinal outcomes that have much more complicated structures. We further extend it by imposing random effects that follow multivariate normal distribution, therefore it allows us to fit multiple correlated longitudinal outcomes simultaneously. We fitted the sepsis data using sub-models for two correlated lon-

gitudinal outcomes. It can be extended to higher dimensions if needed. In addition, we proposed a two-stage framework for sepsis diagnosis, and it can be extended to any disease diagnosis.

Chapter 8

Appendix

8.1 Code for Joint Modelling of Competing Risks and Covariates

In this section, we present an example code for the marginal likelihood of a joint model based on the three-state multi-state model and the general mixed model for the ELSA data. The transition intensities are:

$$q_{rs}(t_{ij}|\mu_{ij}) = q_{0.rs}\exp\left(\eta_{rs}t_{ij} + \alpha_{rs}\mu_{ij} + \lambda_{rs}gender_i\right),$$

and the general mixed model for the time-depending covariate number of animal names individuals can recall:

$$\begin{aligned}
&y_{ij} = \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2),\\
&\mu_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})age_{ij} + \omega_1 gender_i + \omega_2 educ_i,\\
&\log\left(\sigma_{ij}\right) = \gamma_0 + \gamma_1 age_{ij}.
\end{aligned}$$

We assume the random effects b_{0i} and b_{1i} follows a bivariate normal distribution. The extension to models with more covariates and more random effects is straightforward. The integral is approximated using the Gauss-Hermite quadrature method. In addition, to avoid underflow problems, the log-sum-exp trick is applied when computing the marginal likelihood.

The coding of the marginal likelihood is:

```
# generate gauss.hermite
library (statmod)
nnodes <- 25
quad <- gauss.quad(nnodes, "hermite")
print("quad")
print(quad)
nodes <- quad$nodes
print("nodes")
print(nodes)
wi <- quad$weights

# redefine weights
weights <- expand.grid(b1=wi, b2=wi)
weights <- weights$b1 * weights$b2
weights <- log(weights)

#====== margllk.R =======
marg.all <- function(Mydata_all, params){

    # read in parameters
    sigma.b1 <- exp( params[1] )
    sigma.b2 <- exp( params[2] )
    beta1 <- params[3]
    beta2 <- params[4]
    beta3 <- params[5] # longi: gender
    gamma1 <- params[6] # for sigma.e
    gamma2 <- params[7]
    q120 <- exp(params[8])
    q130 <- exp(params[9])
    alpha11 <- params[10] # competing: # of animal name
    alpha12 <- params[11]
    alpha21 <- params[12] # competing: age
    alpha22 <- params[13]
    alpha31 <- params[14] # competing: gender
    alpha32 <- params[15]
    rho <- 2*exp(params[16])/( 1+ exp(params[16])) - 1 # correlation

    # define integration grids (order important !)
    # b2
    b2 <- sqrt(2) * sigma.b2 * nodes # dim = nnodes
    # b1|b2
    # mu.1c2 <- rho*( sigma.b1/sigma.b2) * b2 # dim = nnodes * nnodes
    # sigma.1c2 <- sigma.b1^2*(1-rho^2) # constant
    mu.b1 <- 0
    mu.b2 <- 0
    b.grids <- expand.grid(b1=nodes, b2=b2)
    # b1 is actually b1|b2
    b.grids$b1 <- sqrt(2) * sigma.b1 * sqrt(1-rho^2) * b.grids$b1 +
              mu.b1 + rho* sigma.b1/sigma.b2 * (b.grids$b2-mu.b2)

    # function-marg.single
    marg.single <- function( id.k ){
        singleData <- subset(Mydata_all, id==id.k)
        n.row <- nrow(singleData)
```

```
        # define function integrand ===
        integrand <- function(bi1, bi2){

        #---- llk.lm -----
        mu.cfani <- beta1 + bi1 + (beta2 + bi2) * singleData$age +
                    beta3 * singleData$sex
        sigma.e <- exp(gamma1 + gamma2 * singleData$age)
        llk.lm <- dnorm( x=singleData$cfani, mean = mu.cfani,
                    sd = sigma.e, log=TRUE)

        #---- llk.comp ----
        llk.comp <- 0
        q12 <- q120 * exp(  alpha11 * mu.cfani + alpha21 * singleData$age +
        alpha31 * singleData$sex )
        q13 <- q130 * exp(  alpha12 * mu.cfani + alpha22 * singleData$age +
        alpha32 * singleData$sex )
        q11 <- -q12-q13

        time.interval <- singleData$age[2: n.row] -
                         singleData$age[1: (n.row-1)]
        p11 <- q11[-length(q11)] * time.interval

        l = max(singleData$state)
        # state 3 = dementia = interval
        if (l==3){
        # P13 = [1-P11 ]  * q13/ q12 + q13
        llk.comp <- sum( p11[-length(p11)] ) #log of P11(t1, t_n-2)
        p13 <- log( 1-exp(p11[length(p11)]) ) + log(q13[length(q13)-1])
                -log( q12[length(q12)-1] + q13[length(q13)-1]  )
        llk.comp <- llk.comp + p13
        }
        # state 2 = death = exact
        if (l==2){ llk.comp <- sum( p11) + log( q12[n.row-1]  ) }
        if (l==1){ llk.comp <- sum( p11)  }
        # ---------------------------

    return(  llk.comp + llk.lm )
    }
    # ============================

  approx <- mapply(integrand, bi1=b.grids$b1, bi2=b.grids$b2 )
  approx <- weights +  approx
  approx <-  max(approx) + log( sum(  exp(approx-max(approx)) ) )
  approx <- log(1/pi) + approx
  return(approx)
}

all.likelihood <- lapply(unique(Mydata_all$id),  marg.single  )
all.likelihood <- unlist( all.likelihood )
all.likelihood <- sum( all.likelihood)
print(-all.likelihood)
return( -all.likelihood)
}
```

At the end of this code, the function returns -logarithm marginal likelihood. In some extreme cases, it is possible that some joint probabilities are too small that the arithmetic underflow occurs. To avoid these cases, instead of computing the logarithm after integrating the marginal likelihood, we calculate the logarithm of the likelihood functions before the integration. Furthermore, we can consider the integration as a summation problem in programming. As a matter of fact, the Gaussian-Hermite quadrature approximation is exactly a summation problem. Therefore we adopted the log-sum-exp trick to the Gaussian-Hermite quadrature approximation and derived a new method to approximate the 2-dimensional logarithm marginal likelihood.

8.2 Code for Bayesian Inference of Joint Model

In this section we present an example code for the Bayesian inference of the joint model in the WinBUGS. The model was introduced in Section 6.4.

```
# BUGS code used for CR and LME
# Model specification:
model
{

#  Three-state model:
#::#rows of data, for sepsis N=1000 is 37848
for(i in 1:37848) {


 #:  x1 and x2 (two time varying covariates)
 x1[i] ~  dnorm(mu1[i], prec1)
 mu1[i] <- alpha[1]  + ( alpha[3]  )*t[i]
 x2[i] ~  dnorm(mu2[i], prec2)
 mu2[i] <- alpha[2]  + ( alpha[4]  )*t[i]
 # Transition intensities:
 log(q[i,1,2]) <- beta[1] + beta[3]*t[i] + beta[5]*x3[i] + beta[7]*x4[i] +
                            beta[9]*mu1[i] + beta[11]*mu2[i]
 log(q[i,1,3]) <- beta[2] + beta[4]*t[i] + beta[6]*x3[i] + beta[8]*x4[i] +
                            beta[10]*mu1[i] + beta[12]*mu2[i]
 # Transition probabilities for observed interval:
 P[i,1,1] <- exp( -(q[i,1,2]+q[i,1,3])*time[i] )
 P[i,1,2] <- (q[i,1,2]/(q[i,1,2]+q[i,1,3]))*
                             ( 1-exp( -(q[i,1,2]+q[i,1,3])*time[i] ) )
 P[i,1,3] <- (q[i,1,3]/(q[i,1,2]+q[i,1,3]))*
                             ( 1-exp( -(q[i,1,2]+q[i,1,3])*time[i] ) )
 P[i,2,1] <- 0 ;  P[i,2,2] <- 1 ; P[i,2,3] <- 0
 P[i,3,1] <- 0 ;  P[i,3,2] <- 0 ; P[i,3,3] <- 1

 # Move out of state 1 or not:
 r[i,1:3] ~ dmulti(P[i,current[i],1:3],1)
```

```
 }

# In Use:
# beta[1] + beta[3]*t[i] + beta[5]*x4[i] + beta[7]*x1[i]
# Not using:
for(i in 1:37848){
   id[i] ~ dnorm(0.0,0.001)
   x5[i] ~ dnorm(0.0,0.001)
   x6[i] ~ dnorm(0.0,0.001)
   x7[i] ~ dnorm(0.0,0.001)
}

# Residual precision and sd:
log.prec1 ~ dunif(-10,10)
log(prec1) <- log.prec1;
sd1 <- 1/sqrt(prec1)
# Residual precision and sd:
log.prec2 ~ dunif(-10,10)
log(prec2) <- log.prec2;
sd2 <- 1/sqrt(prec2)

# Priors for intensities parameters:
for(f in 1:12){ beta[f]~dnorm(0.0,0.001) }

# Priors for LME:
for(f in 1:4){ alpha[f] ~ dnorm(0.0,0.001)}

}
```

Bibliography

Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018.

Hanan Al-Hadeethi, Shahab Abdulla, Mohammed Diykh, Ravinesh C Deo, and Jonathan H Green. Adaptive boost ls-svm classification approach for time-series signal classification in epileptic seizure diagnosis applications. *Expert Systems with Applications*, 161:113676, 2020.

Per Kragh Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2):91–115, 2002.

Per Kragh Andersen, Steen Z Abildstrom, and Susanne Rosthøj. Competing risks as a multi-state model. *Statistical methods in medical research*, 11(2):203–215, 2002.

Eleni-Rosalina Andrinopoulou, Dimitris Rizopoulos, Johanna JM Takkenberg, and Emmanuel Lesaffre. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in medicine*, 33(18):3167–3178, 2014.

Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

Adelchi Azzalini. A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, pages 171–178, 1985.

William G Baxt. Application of artificial neural networks to clinical medicine. *The lancet*, 346(8983):1135–1138, 1995.

M Jésus Bayarri and James O Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

William M Bolstad and James M Curran. *Introduction to Bayesian Statistics*. John Wiley & Sons, 2016.

Hans W. Borchers. *pracma: Practical Numerical Math Functions*, 2019. URL `https://CRAN.R-project.org/package=pracma`. R package version 2.2.9.

Yueh-Yun Chi and Joseph G Ibrahim. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2):432–445, 2006.

David Collett and Alan Kimber. *Modelling survival data in medical research*. CRC press, 2014.

Richard J Cook and Jerald F Lawless. *Multistate Models for the Analysis of Life History Data*. CRC Press, 2018.

Victor De Gruttola and Xin Ming Tu. Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, pages 1003–1014, 1994.

Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, et al. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

Robert M Elashoff, Gang Li, and Ning Li. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64 (3):762–771, 2008.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605, 2008.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Cheryl L Faucett and Duncan C Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, 15(15):1663–1685, 1996.

Loïc Ferrer, Virginie Rondeau, James Dignam, Tom Pickles, Hélène Jacqmin-Gadda, and Cécile Proust-Lima. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in medicine*, 35(22):3933–3948, 2016.

Loïc Ferrer, Hein Putter, and Cécile Proust-Lima. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Statistical methods in medical research*, 28(12):3649–3666, 2019.

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons, 2012.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York, 2001.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

Jayanta K Ghosh, Mohan Delampady, and Tapas Samanta. *An Introduction to Bayesian Analysis: Theory and Methods*, volume 725. Springer, 2006.

Robert D Gibbons, Donald Hedeker, and Stephen DuToit. Advances in analysis of longitudinal data. *Annual review of clinical psychology*, 6:79, 2010.

Wolfgang Härdle. *Applied Nonparametric Regression*. Number 19. Cambridge university press, 1990.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

Peter D Hoff. *A First Course in Bayesian Statistical Methods*, volume 580. Springer, 2009.

Xin Huang, Gang Li, Robert M Elashoff, and Jianxin Pan. A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis*, 17(1):80, 2011.

Joseph G Ibrahim, Haitao Chu, and Liddy M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.

Christopher H Jackson et al. Multi-state models for panel data: the msm package for r. *Journal of Statistical Software*, 38(8):1–29, 2011.

Mark Jerrum and Alistair Sinclair. The markov chain monte carlo method: an approach to approximate counting and integration. *Approximation Algorithms for NP-hard problems, PWS Publishing*, 1996.

Bodin Khwannimit, Rungsun Bhurayanontachai, and Veerapong Vattanavanit. Comparison of the performance of sofa, qsofa and sirs for predicting mortality and organ failure among sepsis patients admitted to the intensive care unit in a middle-income country. *Journal of critical care*, 44:156–160, 2018.

Christopher Kok, V Jahmunah, Shu Lih Oh, Xujuan Zhou, Raj Gururajan, Xiaohui Tao, Kang Hao Cheong, Rashmi Gururajan, Filippo Molinari, and U Rajendra Acharya. Automated prediction of sepsis using temporal convolutional network. *Computers in Biology and Medicine*, 127:103957, 2020.

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

Simon Lambden, Pierre Francois Laterre, Mitchell M Levy, and Bruno Francois. The sofa score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23(1):1–9, 2019.

A Lawrence Gould, Mark Ernest Boye, Michael J Crowther, Joseph G Ibrahim, George Quartey, Sandrine Micallef, and Frederic Y Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, 34(14): 2181–2195, 2015.

Nathan P Lemoine. Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *Oikos*, 128(7):912–928, 2019.

Xin Li, G André Ng, and Fernando S Schlindwein. Convolutional and recurrent neural networks for early detection of sepsis using hourly physiological data from patients in intensive care unit. In *2019 Computing in Cardiology (CinC)*. IEEE, 2019.

Haiqun Lin, Charles E McCulloch, and Susan T Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in medicine*, 21(16):2369–2382, 2002.

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

Romany Fouad Mansour, Adnen El Amraoui, Issam Nouaouri, Vicente García Díaz, Deepak Gupta, and Sachin Kumar. Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems. *IEEE Access*, 9: 45137–45146, 2021.

Katya Mauff, Ewout W Steyerberg, Giel Nijpels, Amber AWA van der Heijden, and Dimitris Rizopoulos. Extension of the association structure in joint models to include weighted cumulative effects. *Statistics in medicine*, 36(23):3746–3759, 2017.

Borislava Mihaylova, Andrew Briggs, Anthony O'Hagan, and Simon G Thompson. Review of statistical methods for analysing healthcare resources and costs. *Health economics*, 20(8):897–916, 2011.

Karla Monterrubio-Gómez, Nathan Constantine-Cooke, and Catalina A Vallejos. A review on competing risks methods for survival analysis. *arXiv preprint arXiv:2212.05157*, 2022.

James Murray and Pete Philipson. A fast approximate em algorithm for joint models of survival and multivariate longitudinal data. *Computational Statistics & Data Analysis*, 170:107438, 2022.

Nosheen Nasir, Bushra Jamil, Shahla Siddiqui, Najeeha Talat, Fauzia A Khan, and Rabia Hussain. Mortality in sepsis and its relationship with gender. *Pakistan journal of medical sciences*, 31(5):1201, 2015.

World Health Organization et al. Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions. 2020.

Graziella Orru, William Pettersson-Yeo, Andre F Marquand, Giuseppe Sartori, and Andrea Mechelli. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, 36(4):1140–1152, 2012.

Grigorios Papageorgiou, Mostafa M Mokhles, Johanna JM Takkenberg, and Dimitris Rizopoulos. Individualized dynamic prediction of survival under time-varying treatment strategies. *arXiv preprint arXiv:1804.02334*, 2018.

Grigorios Papageorgiou, Katya Mauff, Anirudh Tomer, and Dimitris Rizopoulos.

An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application*, 6:223–240, 2019.

Giovanni Parmigiani and Lurdes Inoue. *Decision theory: Principles and approaches*. John Wiley & Sons, 2009.

Martin Peterson. *An introduction to decision theory*. Cambridge University Press, 2017.

Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `http://www.R-project.org/`.

Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, 56(9):455, 2020.

Andrew Reibman and Kishor Trivedi. Numerical transient analysis of markov models. *Computers & Operations Research*, 15(1):19–36, 1988.

Matthew A Reyna, Chris Josef, Russell Jeter, Supreeth P Shashikumar, Benjamin Moody, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early Prediction of Sepsis from Clinical Data – the PhysioNet Computing in Cardiology Challenge 2019 (version 1.0.0). `https://doi.org/10.13026/v64v-d857`, 2019a.

Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)*. IEEE, 2019b.

Robert A Rigby, Mikis D Stasinopoulos, Gillian Z Heller, and Fernanda De Bastiani. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press, 2019.

Brian D Ripley and Ruth M Ripley. Neural networks as statistical methods in survival analysis. *Clinical applications of artificial neural networks*, 237:255, 2001.

Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

Dimitris Rizopoulos. *Joint Models for Longitudinal and Time-to-event data: With applications in R*. CRC Press, 2012.

Sheldon M Ross. *Introduction to Probability Models*. Academic Press, 2014.

Francisco J Rubio and Marc G Genton. Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis. *Statistics in medicine*, 35(14):2441–2454, 2016.

Javad Salimi Sartakhti, Mohammad Hossein Zangooei, and Kourosh Mozafari. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (svm-sa). *Computer methods and programs in biomedicine*, 108(2):570–579, 2012.

Mark D Schluchter. Methods for the analysis of informatively censored longitudinal data. *Statistics in medicine*, 11(14-15):1861–1870, 1992.

Roger B Sidje and William J Stewart. A numerical study of large sparse matrix exponentials arising in markov chains. *Computational Statistics & Data Analysis*, 29(3):345–368, 1999.

Rodney Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric competing risks analysis using bayesian additive regression trees. *Statistical methods in medical research*, 29(1):57–77, 2020.

D Mikis Stasinopoulos and Robert A Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46, 2008.

Andrew Steptoe, Elizabeth Breeze, James Banks, and James Nazroo. Cohort profile: the english longitudinal study of ageing. *International journal of epidemiology*, 42(6):1640–1648, 2013.

Krithika Suresh, Jeremy MG Taylor, Daniel E Spratt, Stephanie Daignault, and Alexander Tsodikov. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59(6):1277–1300, 2017.

Bruce Swihart and Jim Lindsey. *rmutil: Utilities for Nonlinear Regression and Repeated Measurements Models*, 2016. URL `https://CRAN.R-project.org/package=rmutil`. R package version 1.1.0.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

Ardo Van den Hout. *Multi-State Survival Models for Interval-Censored Data*. CRC Press, 2017.

Ardo Van den Hout and Graciela Muniz-Terrera. Joint models for discrete longitudinal outcomes in aging research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):167–186, 2016.

Hans C Van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

Hans C Van Houwelingen and Hein Putter. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime data analysis*, 14:447–463, 2008.

J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, 1996.

Nicky J Welton and AE Ades. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, 25(6):633–645, 2005.

Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

Xubo Yue and Raed Al Kontar. Joint models for event prediction from time series and survival data. *Technometrics*, 63(4):477–486, 2021.

Ming Zheng and John P Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.