LANCASTER UNIVERSITY

Longitudinal and survival statistical methods with applications in renal medicine

by

Özgür Asar

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Health and Medicine
Medical School

August 2015

Declaration of Authorship

I, Özgür Asar, declare that this thesis titled, 'Longitudinal and survival statistical methods with applications in renal medicine' and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at Lancaster University.
- This work was not submitted to a research degree at another institute.
- Where any part of this thesis has previously been submitted for publication or published in a scholarly journal, this has been clearly stated.
- Where there are multiple authors in the published/submitted works that constitute this thesis, my roles in these works are clearly indicated and these roles are approved by my supervisor and co-authors.
- Where I have consulted the published work of others, this is always clearly cited.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

Signed: _____

Date: _____

"Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."

Jules Verne, A Journey to the Center of the Earth

LANCASTER UNIVERSITY

Abstract

Faculty of Health and Medicine
Medical School

Doctor of Philosophy

by Özgür Asar

In this thesis, we develop statistical methodology to find solutions to contemporary problems in renal research. These problems include 1) assessing the association of the underlying kidney function and the risk of survival events, 2) early detection of progression towards renal failure amongst primary care patients, and 3) long-term influences of acute kidney injury occurrences on the subsequent kidney function. Joint modelling of longitudinal and time-to-event outcome and Cox model with time-varying covariate are considered to answer the first problem. Whilst parameters are estimated by maximum likelihood (ML) using an expectation-maximisation (EM) algorithm for the former model, by partial likelihood for the latter. Results show that Cox model underestimates the association parameter between the longitudinal and survival processes, and joint models correct this. A longitudinal model with a non-stationary stochastic process is developed for the second problem. Parameters are estimated by ML using a Fisher-Scoring algorithm. Based on the results of this model, we obtain the predictive distribution of meeting the clinical guideline for detecting progression. Results show that there are patients with very high probability and emerging behaviour of progression. By these probabilities, we aim to inform clinical decision-making. Another longitudinal model with a class of stationary stochastic processes and heavy tailed response distribution is developed for the third problem. Parameters are estimated by ML using an EM algorithm, and random effects are predicted using the conditional distribution of random effects given data. Results show that AKI might have serious impacts on kidney function such that on average the loss of kidney function doubles after having an AKI. Nonetheless, there are substantial between patient heterogeneity in terms of this influence. The `R` package `lmenssp` which enables inference for a range of mixed models with non-stationary stochastic processes is developed and its core features are presented.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Peter J. Diggle, for providing me an excellent environment to pursue my PhD studies. Not only his deep knowledge in statistics, but also his open-mindedness and optimism taught me a lot. I wish I became such a supervisor and mentor to my students.

I would like to thank my external examiner Prof. John Matthews and my internal examiner Dr. Andrew Titman for their time, careful reading and constructive criticisms. I would also like to thank Dr. Chris Jewell for being the independent chair of my viva and the celebration afterwards.

I am grateful to Prof. Philip Kalra, Dr. James Ritchie and Dr. Inés Sousa for the collaboration during my PhD studies.

I consider myself very lucky to being a part of CHICAS. I especially thank them for the after-viva celebration. I would also like to thank Mrs. Catherine Thompson for her helps with the logistics. I gratefully acknowledge helpful discussions with Dr. Benjamin Taylor.

I gratefully acknowledge the financial support for my PhD studies from Health e-Research Centre and Lancaster University.

Last but not least, I am grateful to my mother, Atike, and my sister, Keziban, for their devotion and sacrifices.

Contents

List of Figures

List of Tables

Dedicated to my mother and sister

Chapter 1

General introduction

1.1 Longitudinal data analysis

Longitudinal data comprise repeatedly collected data on each study subject (e.g. patients). At each of the data collection times, information on a number of variables is collected, e.g. date of the follow-up, serum creatinine, blood pressure etc., which we call time-varying variables. However, some of the variables are collected only at study entry, e.g. gender, called time-independent variables. The data collection times might be pre-specified, e.g. in clinical trials, or might not be pre-specified, e.g. in observational studies. Here, the latter might impose further aspects of data analysis as the observation times might depend on the phenomenon of interest. Repeated measurements on the same subjects are typically dependent, whereas measurements on different subjects are typically independent. The dependency amongst repeated measures makes longitudinal data analysis a special statistical research area and requires development of novel statistical methods, both exploratory and confirmatory, and new statistical software.

The random effects modelling framework (also known as mixed effects models) gives a very useful and widely used data-analytic class for longitudinal data analysis. In these models, whilst the relationships between the explanatory variables are captured by regression parameters (also known as fixed effects) as in classical multiple linear regression, the dependencies amongst the repeated measurements are captured by subject-specific random parameters. Here, the subject-specific parameters are typically assumed to be drawn from a statistical distribution. The working assumption with random effects models is that given the random effects the repeated measures are independent. Early development of these models is attributed to the seminal paper of Laird and Ware (1982). A useful special case of this framework is the well-known random intercept and random slope model, especially for longitudinal data sets consisting of short series. Diggle (1988) proposed a flexible approach to modelling the dependency amongst the repeated measurements by including stationary Gaussian processes as serially correlated random effects. This proposal is especially useful for longitudinal data sets with long series. The author also proposed the use of the variogram as a diagnostic tool for the shape of the correlation function of the repeated measurements. Taylor, Cumberland and Sy (1994) considered the use of non-stationary stochastic processes. Jennrich and Schluchter (1986) and Lindstrom and Bates (1988) presented parameter estimation by maximum likelihood (ML) using numerical algorithms, e.g. Newton-Raphson methods and the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). Robinson (1991) discusses the prediction of the random effects by best linear unbiased prediction. Book length materials of longitudinal data analysis include Verbeke and Molenberghs (2000), Diggle *et al.* (2002), Fitzmaurice *et al.* (2009) and Fitzmaurice, Laird and Ware (2011) amongst others. For the details of variogram, see Diggle *et al.*

(2002) from a perspective of longitudinal data analysis and Diggle and Ribeiro (2007) for a geostatistical data analysis perspective. For stochastic processes, the interested reader is referred to Ross (1996). In continuous longitudinal data analysis, the distribution of the repeated measurements is typically assumed to be multivariate normal. However, in some real life examples, this assumption might not be appropriate, e.g. the underlying distribution might have heavier tails than the multivariate normal. Pinheiro, Liu and Wu (2001) proposed random effects models for multivariate t distributed repeated measures and algorithms for ML estimation via EM algorithms. Rosa, Padovani and Gianola (2003) consider mixed models with a class of heavy tailed distributions from a Bayesian perspective. Software for longitudinal data analysis is now widely available. We specifically consider the R programming language (R Development Core Team, 2014) which is an open-software platform. The `nlme` (Pinheiro *et al.*, 2013) and `lme4` (Bates *et al.*, 2013) packages facilitate fitting random effects models. Whilst both packages are able to fit the Laird-Ware model, random effects models with stationary Gaussian process can also be fitted by the former package. The procedures in both packages assume multivariate normal distributions for the continuous repeated measures.

Whilst more details on longitudinal data analysis and stochastic process modelling can be found in Chapters 3 and 4, heavy tailed extensions can be found in Chapter 4. Software on mixed models with non-stationary stochastic processes can be found in Chapter 5.

1.2 Survival data analysis

Survival data comprise time to a certain event from a well-defined time origin. The event might be a single (and possibly terminating) event, e.g. death, or recurrent events, e.g. asthma attacks. A number of explanatory variables (both time dependent and independent) are collected together with the time-to-event data. Time-to-event data are greater than zero and typically follow a right-skewed statistical distribution. Scientific interest is mostly on hazard and survival functions. However, what makes analysis of survival data a special statistical research area is censoring, defined as not knowing the exact time of the survival event. There are three types of censoring mechanisms: 1) left, 2) interval, and 3) right-censoring. Interval censoring covers the other two and is the subject-matter when the time to event is known to be between two time points. If we set the lower time point to 0, what we obtain is left censoring. Similarly, if we set the upper time point to infinity, we obtain right censoring. Amongst these censoring mechanisms, the typical and most extensively studied censoring mechanism, also the one we typically experience in renal medicine, is right-censoring. Right-censoring can be better understood by the following explanation. Not all the study subjects experience

the survival event and we do not know the time-to-event for the subjects who do not experience the survival event, only that the time-to-event is greater than or equal to the latest observed time. Analysis of survival data requires development of novel statistical methods and software, as for analysis of longitudinal data.

Methods for survival data analysis started by comparing survival curves amongst groups, with Kaplan-Meier survival curves and the log-rank test (Kaplan and Meier, 1958; Kalbfleisch and Prentice, 2002). Incorporation of explanatory variables became available with the Cox proportional hazards model (Cox, 1972). This is a semiparametric regression method, since the baseline hazard function, i.e. hazard for the patients with all covariates taking 0, is left unspecified. Parameters are estimated by the partial likelihood method, which was shown to have good properties, e.g. efficiency and consistency (Cox, 1975). The Cox model accommodates only time-independent covariates and assumes that the hazard ratio between covariate sub-groups through time is constant, i.e. proportional hazards. The model was then extended to time-varying covariates based on the counting process approach (Andersen and Gill, 1982). This allows models with non-proportional hazards to be built. Random effects models for survival data (also known as frailty models) are also available, e.g. see Vaida and Xu (2000). An interesting recent review of methods for survival data analysis can be found in Oakes (2013). Whilst the beginners of survival data analysis might refer to Kleinbaum and Klein (2012) as a self learning text, for more advanced methods the interested reader is referred to Kalbfleisch and Prentice (2002), Lawless (2003) and Lee and Wang (2003). For survival data analysis with a stochastic processes point of view, one might refer to Andersen *et al.* (1993) and Aalen, Borgan and Gjessing (2008). The `survival` package (Therneau, 2013) in `R` provides survival curve comparisons and fitting of a wide range of survival models, including the Cox model, time-varying survival models and frailty models.

More details on survival data methods can be found in Chapter 2.

1.3 Simultaneous analysis of longitudinal and survival data: joint modelling

In prospective studies, longitudinal and survival data are typically collected at the same time. Simultaneous analysis of these two types of data is of scientific interest and drives many studies. With such analysis, whilst the questions that are related to longitudinal and survival processes can be answered, questions on the association between these processes can also be answered, potentially making better use of the available data. For example, in renal research, the medical area that motivates this thesis, the following

three questions might be of scientific interest and can be simultaneously answered by joint analysis: 1) how does the kidney function evolve through time and differ in terms of explanatory variables, 2) how does the hazard for renal replacement therapy evolve through time and differ in terms of explanatory variables, and 3) how is the hazard for renal replacement therapy influenced by the underlying kidney function? Historically, analyses of longitudinal and survival data evolved separately as two major areas of medical statistics. The intersection of interests from these two areas has yielded a fresh, popular and rapidly growing research area, called joint analysis of longitudinal and survival data.

Both longitudinal and survival data processes are continuous time stochastic processes, with the values of a variable (e.g. kidney function) in continuous time for the former, the hazard for the survival event in continuous time for the latter. These are the underlying data generating mechanisms, but are unobserved in real life. What we observe in real life are the imperfect measurements of the longitudinal process at intermittent and possibly irregularly spaced time points, and time to a survival event which is possibly subject to censoring. These make the joint analysis of longitudinal and survival data challenging. Initial attempts to combine these areas started with accommodating longitudinal data as a time-varying covariate into the survival models (Andersen and Gill, 1982). However, this approach assumes that the covariate is measured perfectly and is available at each event time. But it is now known that measurement error (Carroll *et al.*, 2006) in the longitudinal covariate biases the association parameter between these data towards zero (Prentice, 1982). A one-step improvement to this approach is the two-stage analysis of longitudinal and survival data (Tsiatis, DeGruttola and Wulfsohn, 1995). This includes modelling the longitudinal data first separately and obtaining the error-removed predictions at the event times in the first step. Then, in the second step these predicted values are considered as the time-varying covariates. Although undoubtedly this approach is an improvement, it is still not optimal and introduces bias, especially on the association parameter for the relationship between the longitudinal and survival processes (Sweeting and Thompson, 2011). Simultaneous analysis of these data, via joint models, made it possible to overcome the drawbacks of the aforementioned methods. Early works that facilitated simultaneous modelling of longitudinal and survival data are Berzuini and Larizza (1996), Faucett and Thomas (1996) with a Bayesian paradigm, and Wulfsohn and Tsiatis (1997) with ML via EM algorithms. These works combine the random effects for longitudinal data and proportional hazards model for survival data; the latter may also be viewed as a time-varying frailty model. Henderson, Diggle and Dobson (2000) proposed flexible parametrisation of the association structure between the two processes and introduced the inclusion of the stationary Gaussian processes as the serially correlated random effects. Parameters are estimated

by ML estimation via EM algorithms. Guo and Carlin (2004) inspected the flexible parametrisation of the association structure with a Bayesian perspective. Wang and Taylor (2001) considered inclusion of non-stationary stochastic processes and estimated the parameters with Bayesian methods. There are now many papers published on joint modelling which have extended the aforementioned works by modifying either the longitudinal or survival sub-models. A parametric model for the survival part of the joint model can be found in Diggle, Sousa and Chetwynd (2007). Up to date reviews and the aforementioned extensions can be found in Sousa (2011), McCrink, Marshall and Cairns (2013), Gould *e*t al. (2015). A book length material solely devoted to joint modelling is the book of Rizopoulos (2012). Individual chapters on joint modelling are Chapter 7 of Ibrahim, Chen and Sinha (2001) and Chapter 8 of Wu (2009). Software for joint models include the `R` packages `JM` (Rizopoulos, 2010) and `joineR` (Philipson *et al.*, 2012) for ML estimation, and `JMbayes` (Rizopoulos, 2014) for Bayesian methods.

More details on the methodology and software of joint models can be found in Chapter 2.

1.4 Renal function and two kidney diseases: a brief overview

1.4.1 Renal function

Kidneys are bean-shaped organs that constitute the upper part of the urinary system. Their roles include filtering blood, i.e. removing waste products, and regulating blood volume and pressure, amongst others (Field, Pollock and Harris, 2010). The glomerulus is the kidney's filtration unit. A single kidney includes about one million glomeruli. It is accepted that glomerular filtration rate (GFR) is the best overall measure of kidney function/health (Stevens *et al.*, 2006). Normal GFR values are expected to be approximately 130 ml/min/1.73m^2 of body-surface area for a young man, and 120 ml/min/1.73m^2 for a young woman (depending on age and body size). GFR less than 60 ml/min/1.73m^2 indicates chronic kidney disease (CKD), and GFR less than 15 ml/min/1.73m^2 indicates end-stage renal disease and preparation for renal replacement therapy (RRT), i.e. dialysis or transplantation. Direct measurement of GFR is expensive and difficult in routine clinical practice. Alternatively, estimated versions, called eGFR, are widely used. There are many formulae to obtain eGFR which combine kidney function biomarkers, e.g. serum creatinine (SCr), and demographic factors, e.g. gender, age and ethniticity, in a deterministic manner. Two popular ones are the Modification of Diet in Renal Disease (MDRD, Levey *and others* (1999)) and Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI, Levey *and others* (2009)). Here, kidney function biomarkers are easy

to measure in a routine blood test, but are known to be subject to substantial measurement error, which is also inherited into eGFR. These noisy variables are biochemicals that are expected to be removed from the body by the kidneys, but their levels in blood might also be associated with other factors. For example, SCr is a muscle breakdown product, and heavy drinkers are known to lose muscle. Such a person might have a higher level of SCr than expected, yet have a pair of well-functioning kidneys. Another example is that some drugs are known to inhibit creatinine clearance, and might lead to measuring higher levels of SCr in the tests (Stevens *et al.*, 2006).

1.4.2 Two kidney diseases

Chronic Kidney Disease (CKD; El Nahas and Levin (2009); Levey and Coresh (2012); Jha *et al.* (2013); Arici (2014)) is defined based on either the existence of kidney damage, e.g. increased protein level in urine, or decreased kidney function, e.g. GFR < 60 mL/min/1.73m^2, that continues for at least three months. The term CKD covers a wide range of kidney diseases that affect the function and structure of the kidneys. It might be detected during routine blood tests, and some treatments can slow the progression towards renal failure. However, it is usually detected during the assessment of co-morbid conditions at possibly advanced stages, because kidney disease can be asymptomatic for many years. Most common causes are diabetes mellitus, cardiovascular diseases and hypertension. The risk factors include old age, obesity and family history. Many CKD patients lose their kidney function gradually and do not experience kidney failure in their lifetime. However, for some patients the disease is aggressive and might result kidney failure in a short term. CKD is now accepted to be a major public health problem with approximate worldwide prevalence of 10%.

Acute Kidney Injury (AKI; Bellomo, Kellum and Ronco (2012); Lameire *et al.* (2013); KDIGO (2012)) is defined as a sudden fall in the kidneys' excretory function, typically within a defined period of time, e.g. 48 hours. For example, if the SCr levels increase between 1.5 and 2 fold within 48 hours, it is called a stage 1 AKI. It is mostly a reflection of a disease that might also affect the kidneys, e.g. chest infection or heart attack. AKI is common and potentially catastrophic amongst seriously ill patients, e.g. patients in intensive care units. Impacts of AKI range between slight loss of kidney function without actual damage, to complete renal failure and even death. Approximately 2 million deaths worldwide are attributed to AKI (Murugan and Kellum, 2011). Typically, there is no specific treatment for AKI; treatment is only supportive. The focus is mostly on identification and treatment of the primary disease that drives AKI. However, some patients, e.g. with AKI due to vasculitis, are known to response to some specific therapeutic treatment options. Temporary dialysis is another option for treating AKI.

CKD and AKI are now accepted to be highly associated syndromes, such that it is occasionally argued that the distinction between the two is artificial (Chawla and Kimmel, 2012; Chawla *et al.*, 2014). History of AKI is a risk factor for developing CKD. AKI might worsen the severity of an existing CKD, and CKD patients are known to be at high risk of experiencing an AKI.

More details on CKD and AKI can be found in Chapter 4.

1.5 Motivating cohort studies

1.5.1 Salford primary care cohort

This is a cohort study in Salford, Greater Manchester, UK, on patients who do not have any kidney diseases, but are at high risk of developing such a disease. These patients are identified as having predisposing conditions for renal diseases, e.g. having diabetes mellitus, hypertension, cardiovascular diseases etc. The main aim of this study is to detect any progression towards kidney disease at an early stage. Any patient who meets the above criteria can enter the study whilst it is running. There is no rigid follow-up schedule for the blood sample collections. Whilst on some occasions the timings of the follow-ups are based on physicians' decisions, in others the patients decide when to go to the hospital, e.g. when they feel sick. Therefore, the follow-up times are typically irregularly spaced and the maximum number of repeated measures per patient differs between patients. Recorded variables include demographic characteristics of the patients, dates of follow-ups, bio-marker measurements, treatment and co-morbidity history. Amongst these variables, whilst demographic variables are collected only once at baseline, the others are collected at the subsequent follow-ups. In addition to these variables, date of death is recorded if a patient dies, and the follow-up terminates. Date of death is typically confirmed from UK Office of National Statistics. The minimum lag between two successive measurements is a day in most cases. When there are multiple blood measurements on the same day, we set the arithmetic average of the multiple data as the data of that day. The data set we have covers the cohort between March 7, 1997 and March 22, 2007 with 22,910 patients and a total of 392,780 repeated measurements. A refined version of the data set for a single patient is given in Table 1.1. This patient was a male and 74.2 years old when he entered the study. He provided data at 7 subsequent follow-ups, and died at the age of 75. More details of the data and the detailed statistical analysis can be found in Chapter 3. A longitudinal data version of the Salford primary care cohort study can be found at

http://biostatistics.oxfordjournals.org/content/early/2014/12/16/biostatistics.kxu053/suppl/DC1

TABLE 1.1: A portion of the data from Salford primary care cohort for a single patient with identification number (ID) 48700. Gender takes 1 for females, 0 for males; Age0 is age at baseline (in years); Age is age at measurement; Stime is age at death or data lock (March 22, 2007); Death is an indicator variable: 0 = alive, 1 = died.

ID	SCr	eGFR	Gender	Age0	Age	Stime	Death
48700	106	59.204	0	74.196	74.196	75.039	1
48700	105	59.737	0	74.196	74.921	75.039	1
48700	101	62.475	0	74.196	74.927	75.039	1
48700	93	68.716	0	74.196	74.932	75.039	1
48700	96	66.243	0	74.196	74.938	75.039	1
48700	95	67.036	0	74.196	75.006	75.039	1
48700	109	57.197	0	74.196	75.036	75.039	1
48700	88	73.220	0	74.196	75.039	75.039	1

1.5.2 CRISIS cohort

Chronic Renal Insufficiency Standards Implementation Study (CRISIS, Eddington *et al.*, 2010; Hoefield *et al.*, 2010) is an on-going cohort study on all-cause CKD. It is run by the Salford Royal NHS Foundation Trust. Recruitment of the patients was started on October 1, 2002. Patients with eGFR less than 60 ml/min/1.73m^2 without RRT are invited to attend the study, and those who have signed the consent form are recruited. The main aims of the study are to understand the associated factors of CKD and the time-course of the disease. The set of variables recorded is similar to that of the Salford primary care cohort. Since CRISIS is a study on CKD patients, RRT is another possibility as a survival event. Data are censored at RRT or death (whichever comes first). Data have been collected annually by protocol. Our data set covers CRISIS cohort from October 1, 2002 to July 30, 2012. There are 1,611 patients with a total of 3,154 follow-up measurements. We aim to investigate with this data set the time-course of CKD and the association between kidney function and risk for survival events. More details can be found in Chapter 2. We are also able to obtain the records between the planned (annual) follow-ups for the patients in CRISIS through the electronic records of Salford Royal Hospital. In this data set, we have records between November 15, 2000, and February 28, 2013. There are 2,289 patients with a total of 48,382 repeated measurements. We aim to investigate with this data set the influences of AKI events on long-term kidney function amongst patients with existing CKD. More details regarding this data set can be found in Chapter 4.

1.6 Aims, contributions and organisation

This thesis is motivated by real-life medical problems that arise in renal research. We aim to develop novel statistical methods to provide solutions to such problems. The statistical methods considered in the thesis currently cover methods for longitudinal and survival data analysis. The thesis provides the following contributions:

- Chapter 2: we investigate the association between kidney function and risk of having RRT in a case-study. We prepared an accessible tutorial on joint modelling of longitudinal and survival data with detailed applications. `R` codes are also provided.
- Chapter 3: our primary aim in this study is to flag primary care patients in terms of referral to secondary care. This is an important problem, since progression towards renal failure might be asymptomatic for many years. A linear mixed effects model with a non-stationary stochastic process is developed.
- Chapter 4: we investigate the influence of AKI occurrences on long term kidney health by statistical modelling. Such a study is important in terms of understanding the natural history of kidney function after acute kidney injury occurrences. To the best of our knowledge, this study is the first one that uses advanced statistical modelling to inspect the aforementioned phenomenon. A linear mixed model with stationary stochastic processes and a heavy tailed response distribution is developed.
- Chapter 5: an R package, `lmenssp`, is developed for mixed effects models with non-stationary stochastic processes. Details of the usage and applications are provided.

Each chapter of this thesis can be read separately since they include their own introduction, conclusion and discussion sections. Therefore we decided to keep the amount of information and references short in this chapter. Some guidance to the readers is as follows. Chapter 2 can be seen as a very detailed introduction to the thesis. `R` codes regarding Chapters 2, 3 and 4 are provided at the end of each chapter. In Chapter 6, we provide general discussions and mention future work.

We focus on analysis of continuous univariate longitudinal data and time to a non-recurrent event from a confirmatory data analysis perspective, and mostly use available exploratory data analysis methods. Therefore, in this chapter we have reviewed the

literature on longitudinal continuous data and non-recurrent survival data with an emphasis on statistical modelling. The interested reader might refer to the books we cited for the other aspects.

1.7 Role in the published/submitted works

The role of Özgür Asar in the submitted/published papers presented in the Chapters 2, 3, 4 and 5 is as follows:

- Chapter 2: statistical analysis, preparation of the manuscript, and final version of the paper.
- Chapter 3: statistical analysis and manuscript preparation revising and extending an unpublished technical report by Peter J. Diggle and Ines Sousa, and final version of the paper.
- Chapter 4: statistical analysis and preparation of the manuscript.
- Chapter 5: development of the `R` package and preparation of the manuscript.

Bibliography

Arici M. (2014). *Management of chronic kidney disease: a clinican's guide*, New York: Springer - Verlag.

Aalen O. O., Borgan Ø. and Gjessing H. K. (2008). *Survival and event history analysis: a process point of view*, New York: Springer - Verlag.

Andersen P. K. and Gill R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.

Andersen P. K., Borgan O., Gill R. D. and Keiding N. (1993). *Statistical models based on counting processes*, New York: Springer - Verlag.

Bates D., Maechler M., Bolker B. and Walker S. (2014). lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1-7, http://CRAN.R-project.org/package=lme4.

Bellomo R., Kellum J. A. and Ronco C. (2012). Acute kidney injury. *Lancet* **380**, 756–766.

Berzuini C. and Larizza C. (1996). A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 1100–1120.

Carroll R. J., Ruppert D., Stefanski L. A. and Crainiceanu C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. Boca Raton: Chapman & Hall/CRC.

Chawla L. S. and Kimmel P. L. (2012). Acute kidney injury and chronic kidney disease: an integrated clinical syndrome. *Kidney International* **82**, 516–524.

Finlay S., Bray B., Lewington A. J., Hunter-Rowe C. T., Banerjee A., Atkinson J. M. and Jones M. C. (2014). Acute kidney injury and chronic kidney disease as interconnected syndromes. *New England Journal of Medicine* **371**, 58–66.

Cox D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society - Statical Methodology* **34**, 187–220.

Cox D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Diggle P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

Diggle P. J., Heagerty P., Liang K.Y. and Zeger S. L. (2002). *Analysis of longitudinal data*, 2nd edition. Oxford: Oxford University Press.

Diggle P. J. and Ribeiro, P. J. Jr. (2007). *Model-based geostatistics*. New York: Springer - Verlag.

Diggle P. J., Sousa I. and Chetwynd A. (2007). Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. *Statistics in Medicine* **26**, 2981–2998.

Dempster A. P., Laird N. M. and Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society - Statistical Methodology* **39**, 1–38.

Eddington H., Hoefield R., Sinha S., Chrysochou C., Lane B., Foley R. N., Hegarty J., New J., O'Donoghue D. J., Middleton R. J. and Kalra P. A. (2010). Serum phosphate and mortality in patients with chronic kidney disease. *Clinical Journal of the American Society of Nephrology* **5**, 2251–2257.

El Nahas M. and Levin A. (2009). *Chronic Kidney Disease: A practical Guide to Understanding and Management*. Oxford: Oxford University Press.

Faucett C. L. and Thomas D. C. (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* **15**, 1663-1685.

Field M., Pollock C. and Harris D. (2010). *The Renal System*, 2nd edition. Elsevier.

Fitzmaurice G. M., Davidian M., Verbeke G. and Molenberghs G. (2011). *Longitudinal data analysis*, Boca Raton: Chapman & Hall/CRC.

Fitzmaurice G. M., Laird N. M. and Ware J. H. (2011). *Applied longitudinal analysis*, 2nd edition. New Jersey: John Wiley & Sons.

Gould *et* al.. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* 34, 2181–2895.

Guo X. and Carlin B. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistician* **58**, 16–24.

Henderson R., Diggle P. J. and Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Hoefield R. A., Kalra P. A., Baker P., Lane B., New J. P., O'Donoghue D. J., Foley R. N. and Middleton R. J. (2010). Factors associated with kidney disease progression and mortality in a referred CKD population. *American Journal of Kidney Diseases* **56**, 1072–1081.

Ibrahim J., Chen M. and Sinha D. (2001). *Bayesian Survival Analysis*, New York: Springer-Verlag.

Jennrich R. I. and Schluchter M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

Jha V., Garcia-Garcia G., Iseki K., Li Z., Naicker S., Plattner B., Saran R., Wang A. Y. and Yang C. W. (2013). Chronic kidney disease: global dimension and perspectives. *Lancet* **382**, 260–272.

Kalbfleisch J. D. and Prentice R. L. (2002). *The statistical analysis of failure time data*, 2nd edition. New York: John Wiley & Sons.

Kaplan E. L. and Meier P. (1958). Nonparametric estimation for incomplete observations. *Journal of the American Statistical Association* **93**, 457–481.

Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group (2012). KDIGO Clinical Practice Guideline for Acute Kidney Injury. *Kidney International, Supplementary* **2**, 1–138.

Kleinbaum D. G. and Klein M. (2012). *Survival analysis: a self learning text*, 3rd edition. New York: Springer-Verlag.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lameire N. H., Bagga A., Cruz D., De Maeseneer J., Endre Z., Kellum J. A., Liu K. D., Mehta R. L., Pannu N., Van Biesen W. and Vanholder R. (2013). Acute kidney injury: an increasing global concern. *Lancet* **382**, 170–179.

Lawless J. F. (2003). *Statistical models and methods for lifetime data*, 2nd edition. New Jersey: John Wiley & Sons.

Lee E. T. and Wang J. W. (2003). *Statistical methods for survival data analysis*. New Jersey: John Wiley & Sons.

Levey A. S., Bosch J. P., Lewis J. B., Greene T., Rogers N. and Roth D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of Internal Medicine* **130**, 461–470.

Levey A. S., Stevens L. A., Schmid C. H., Zhang Y. L., Castro A. F., Feldman H. I., Kusek J. W., Eggers P., Lente F. V., Greene T. and Coresh J. for the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* **150**, 604–612.

Levey A. S. and Coresh J. (2012). Chronic kidney disease. *Lancet* **379**, 165–180.

Lindstrom M. J. and Bates D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**, 1014–1022.

McCrink L. M., Marshall A. H. and Cairns K. J. (2013). Advances in joint modelling: a review of recent developments with application to the survival of end stage renal disease patients. *International Statistical Review* **81**, 249–269.

Murugan R. and Kellum J. A. (2011). Acute kidney injury: whats the prognosis? *Nature Reviews Nephrology* **7**, 209–217.

Oakes D. (2013). An introduction to survival models: in honor of Ross Prentice. *Lifetime Data Analysis* **19**, 442–462.

Philipson P., Sousa I., Diggle P. J., Williamson P., Kolamunnage-Dona R. and Henderson R. (2012). joineR: joint modelling of repeated measurements and time-to-event data, R package version 1.0-3, http://CRAN.R-project.org/package=joineR.

Pinheiro J. C., Liu C. and Wu Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10**, 249–276.

Pinheiro J., Bates D., DebRoy S., Sarkar D. and the R Development Core Team (2013). nlme: linear and nonlinear mixed effects models, R package version 3.1-109, http://CRAN.R-project.org/package=nlme.

Prentice R. L. (1982). Covariate measurement error and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-342.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, URL http://www.R-project.org/.

Rizopoulos D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**, 1-33.

Rizopoulos D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*, Boca Raton: Chapman & Hall/CRC.

Rizopoulos D. (2014). JMbayes: joint modeling longitudinal and time-to-event data under a Bayesian approach, R package version 0.6-1, URL: http://CRAN.R-project.org/package=JMbayes.

Robinson G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–32.

Rosa G. J. M., Padovani C. R. and Gianola D. (2003). Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* **45**, 573–590.

Ross S. M. (1996). *Stochastic Processes*, 2nd edition. New Jersey: John Wiley & Sons.

Sousa I. (2011). A review of joint modelling of longitudinal measurements and time-to-event. *REVSTAT* **9**, 57–81.

Stevens L. A., Coresh J., Greene T. and Levey A. S. (2006). Assessing kidney function - measured and estimated glomerular filtration rate. *New England Journal of Medicine* **354**, 2473–2483.

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

Taylor J. M. G., Cumberland W. G. and Sy J. P. (1994). A stochastic process model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**, 727–736.

Therneau T. (2013). A package for survival analysis in S, R package version 2.37-4, 2013, URL: http://CRAN.R-project.org/package=survival.

Tsiatis A. A., DeGruttola V. and Wulfsohn M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

Vaida F. and Xu R. (2000). Proportional hazards model with random effects. *Statistics in Medicine* **19**, 3309–3324.

Verbeke G. and Molenberghs G. (2000). *Linear mixed models for longitudinal data*, New York: Springer - Verlag.

Wang Y. and Taylor J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905.

Wu L. (2009). *Mixed effects models for complex data*, Boca Raton: Chapman & Hall/CRC.

Wulfsohn M. S. and Tsiatis A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Chapter 2

Joint modelling of repeated measurement and time-to-event data: an introductory tutorial

This chapter is based on the following paper:

Asar Ö, Ritchie R, Kalra PA and Diggle PJ (2015). Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. *International Journal of Epidemiology*, **44(1)**: 334–344.

Abstract

Backgound: The term 'joint modelling' is used in the statistical literature to refer to methods for simultaneously analysing longitudinal measurement outcomes, also called *repeated measurement* data, and time-to-event outcomes, also called *survival* data. A typical example from nephrology is a study in which the data from each participant consist of repeated estimated glomerular filtration rate (eGFR) measurements and time to initiation of renal replacement therapy (RRT). Joint models typically combine linear mixed effects models for repeated measurements and Cox models for censored survival outcomes. Our aim in this paper is to present an introductory tutorial on joint modelling methods, with a case study in nephrology.

Methods: We describe the development of the joint modelling framework and compare the results with those obtained by the more widely used approaches of conducting separate analyses of the repeated measurements and survival times based on a linear mixed effects model and a Cox model, respectively. Our case study concerns a data set from the Chronic Renal Insufficiency Standards Implementation Study (CRISIS). We also provide details of our open-source software implementation to allow others to replicate and/or modify our analysis.

Results: The results for the conventional linear mixed effects model and the longitudinal component of the joint models were found to be similar. However, there were considerable differences between the results for the Cox model with time-varying covariate and the time-to-event component of the joint model. For example, the relationship between kidney function as measured by eGFR and the hazard for initiation of RRT was significantly underestimated by the Cox model that treats eGFR as a time-varying covariate, because the Cox model does not take measurement error in eGFR into account.

Conclusions: Joint models should be preferred for simultaneous analyses of repeated measurement and survival data, especially when the former is measured with error and the association between the underlying error-free measurement process and the hazard for survival is of scientific interest.

Key words: Chronic kidney disease, cohort study, epidemiology, joint modelling of longitudinal and survival data, measurement error, medical statistics, statistical software

Key Messages

- Longitudinal studies often include both repeated measurement and survival outcomes. Common practice is to analyse these data separately. This is mostly due to the lack of awareness of the available tools for simultaneous analysis.
- Measurement error in a time-varying covariate biases the estimate of the underlying association with the hazard for survival towards zero. Joint modelling corrects this.
- Joint modelling of longitudinal and survival data is preferable to separate analyses, both to make optimal use of the available information and to obtain unbiased estimates of the model parameters.
- The availability of publicly available software to fit joint models and examples on their use will encourage wider use of joint models.

2.1 Introduction

Prospective medical studies typically record a variety of covariates on each subject, some fixed (e.g. gender), others time-varying (e.g. age), together with two fundamentally different kinds of outcome: longitudinal data at a regular or irregular sequence of time-points, also called *repeated measurements* (e.g. estimated glomerular filtration rate, eGFR); and time-to-event outcomes, also called *survival data* [e.g. time to initiation of renal replacement therapy (RRT)].

Repeated measurement and survival data require different statistical methods, and are traditionally analysed separately. Typical properties of these data are: (i) repeated measurement sequences are intermittently collected and subject to measurement error; (ii) occurrence of the survival event terminates the underlying measurement process, potentially in an informative manner; and (iii) the underlying measurement process affects the hazard for survival. Together, these properties imply that separate analysis of repeated measurement and survival outcomes is potentially inefficient, because it does not fully exploit the dependence between the repeated measurement process and the hazard for survival, and leads to biased estimation of the association between the two, because it ignores measurement error.

For example, in nephrology it is important to understand the relationship between changes in a patients renal function over time and the corresponding changes in their survival prognosis; but neither the changes in renal function nor the hazard for RRT are directly observable at all times. For this reason, we need to build a statistical model that

relates these unobservable quantities to each other and to the observable data. These data consist of the intermittently measured, error-prone and possibly informatively censored eGFR measurements and the observed, possibly censored, times to RRT for each patient in the study.

Statistical methods for repeated measurement and survival data have generated extensive, but largely separate, literatures. For book-length reviews, see for example Diggle *et al.* (2002) or Fitzmaurice, Laird and Ware (2011) for the former, and Kalbfleisch and Prentice (2002) or Kleinbaum and Klein (2012) for the latter.

Recently, simultaneous analysis of these two types of data has become possible through the development of the so-called *joint modelling* methods: see, for example, Wulfsohn and Tsiatis (1997), Henderson, Diggle and Dobson (2000), Diggle, Sousa and Chetwynd (2007) and Rizopoulos (2012). Much of the early methodological work was stimulated by problems arising in AIDS research (Tsiatis, DeGruttola and Wulfsohn, 1995; Wulfsohn and Tsiatis, 1997). More recently, joint modelling methods have been adopted in other areas of clinical research, including cancer (Ibrahim, Chu and Chen, 2010), cardiovascular disease (Andrinopoulou *et al.*, 2012) and kidney transplantation studies (Garre *et al.*, 2008; Daher Abdi *et al.*, 2013). However, joint modelling methods remain under-used, and the absence of an accessible introduction in the epidemiological literature inhibits their wider adoption. The aim of this paper is to provide an introductory tutorial on joint modelling embedded in a specific application in nephrology and including an illustration of open-source software for joint modelling that is available within the R (R Development Core Team, 2013) computing environment.

The paper is organised as follows. We first provide details of the data set that we use throughout the paper, and define the required statistical terminology. We then formulate repeated measurement, survival and joint models for these data and discuss their basic properties. We then use the models to investigate the effect of changes in kidney function on the hazard for RRT, and discuss our findings. R scripts to reproduce our analyses are provided in the online supplementary material (available as Supplementary data at *IJE* online).

2.2 Materials and Methods

2.2.1 Patient population

Patients were selected from the Chronic Renal Insufficiency Standards Implementation Study (CRISIS; Hoefield *et al.* (2010); Eddington *et al.* (2010)) run by Salford Royal

NHS Foundation Trust (SRFT). CRISIS is an ongoing prospective observational study of outcome in all-cause chronic kidney disease (CKD) that has continued to recruit patients since 1 October 2002. Patient records are updated at annual nephrology follow-ups by trained research nurses. Renal function is estimated using the four-variable MDRD equation (Levey *et al.*, 1999):

$$\text{eGFR} = 175 \times \left(\frac{\text{SCr}}{88.4}\right)^{-1.154} \times \text{age}^{-0.203} \times 0.742^{\text{I(female)}} \times 1.21^{\text{I(black)}} \qquad (2.1)$$

where SCr denotes serum creatinine. In our analysis, we ignored the ethnicity term in Equation 2.1, because most of the patients in the data set we analysed were Caucasian (96.3%). Predefined study end-points are death (confirmed by the Office for National Statistics) and initiation of RRT, defined as chronic haemodialysis, peritoneal dialysis or transplantation. In this paper, we consider data collected until 30 July 2012 and initiation of RRT as the survival outcome. There are 1611 patients with a total number of 3154 follow-up measurements.

2.2.2 Explanation of statistical terms

Repeated measures are eGFR measurements belonging to the same patient but performed at different times, here corresponding to hospital visits. *Measurement times* are the follow-up times at each hospital visit, defined as the years elapsed between study entry and hospital visits. *Measurement error* is the difference between the computed value of eGFR and the true (isotopic) GFR. A *time-constant* or *baseline* covariate is one whose value does not change over time, e.g. gender, whereas a *time-varying* covariate is one whose value does change over time and is available at all times, e.g. age. Covariates are to be regarded as *inputs* to a biomedical system, whereas outcome variables, here repeated measurements of eGFR and time to initiation of RRT, are to be regarded as *outputs*.

A *survival* outcome is the time, from a defined origin, at which an event of clinical interest (e.g. initiation of RRT) occurs. Typically, survival times T, can be either *observed* or *censored*, the latter meaning that observation of the subject in question is terminated before the event of clinical interest occurs; hence the data tell us that T is at least T_0, but we do not know the exact value of T. In our example, patients who had either died or were still alive but had not begun RRT by 30 July 2012 are censored for initiation of RRT, and we know only that initiation of RRT happened, if at all, after 30 July 2012. Censoring is *non-informative* if it is statistically independent of the outcome of interest, *informative* otherwise. In our example, censoring due to death could be

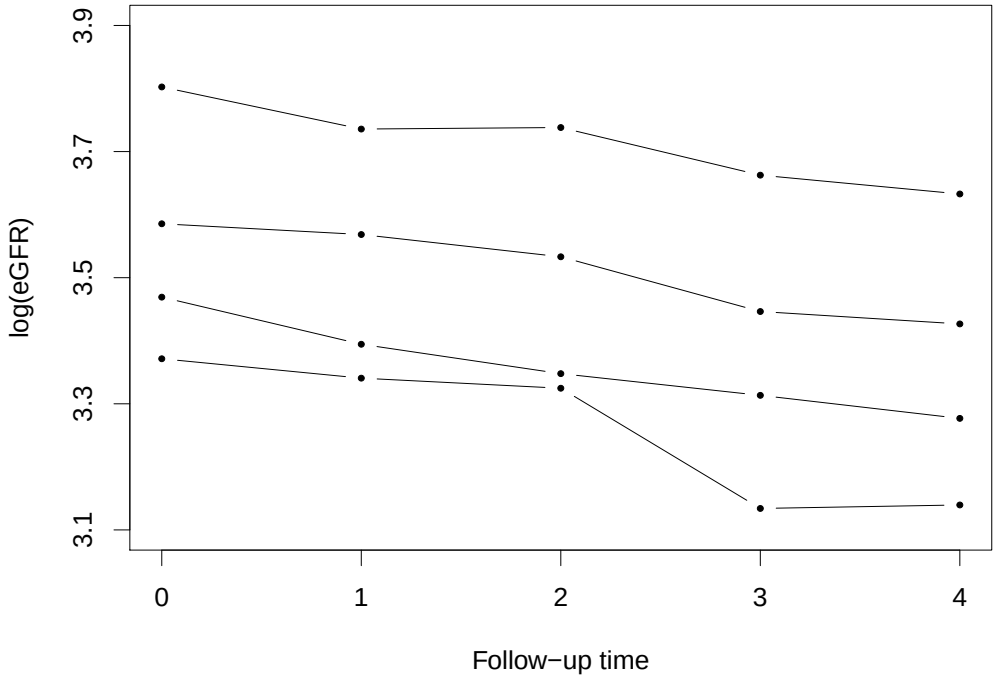

FIGURE 2.1: Hypothetical longitudinal data for four patients with five follow-ups.

informative or non-informative, depending on the cause of death, whereas censoring at the study end-date, 30 July 2012, is unambiguously non-informative.

Finally, a *random effect* is a patient-specific coefficient that represents between-patient heterogeneity in an outcome variable that cannot be explained by measured covariates. This is best understood through an example. Figure 2.1 shows hypothetical data on eGFR measured annually over 5 years for four different patients. All four patients show an approximately linear decrease in eGFR over time, but from clearly different initial values. A simple mathematical representation of this might be:

$$Y_{ij} = A_i + \beta * t_{ij} + Z_{ij}, \tag{2.2}$$

where Y_{ij} is the j^{th} $(j = 1, \ldots, 5)$ measurement for subject i $(i = 1, \ldots, 4)$ at time t_{ij} $(t_{ij} = 1, \ldots, 4)$ and Z_{ij} is the corresponding (random) measurement error. The slope, β is the same for all patients whereas the values of A_i differ among patients. In treating A_i as a random effect, we are assuming that its values are drawn from a statistical distribution. If patients also differed in their rate of decrease in eGFR, we would replace the *fixed effect* β by a random effect, Bi, again assumed to be drawn from a statistical distribution. A useful way to think about random effects is as proxies for unmeasured

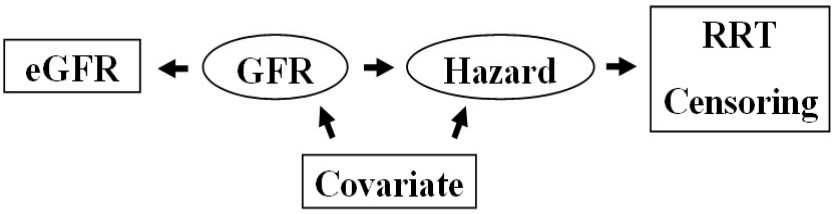

FIGURE 2.2: The underlying mechanism of the longitudinal and survival processes. Rectangles denote observed outcomes, ellipses unobserved quantities and arrows directed statistical dependencies. The causal chain of interest runs from GFR to hazard for RRT, whereas eGFR does not 'cause' RRT but is statistically related to RRT through its dependence on the unobserved GFR.

covariates.

2.2.3 Rationale for joint modelling

The distinction between covariates and outcome variables is an operational one, in the sense that the same biological construct may be regarded as an input or an output in different studies. For example, in renal research we could envisage a study whose primary objective was to investigate the relationship between blood pressure and the hazard for end-stage renal failure. In that context, blood pressure would be an input, and progression to end-stage renal failure an output. However, we could equally envisage a study of the efficacy of an antihypertensive medication in which dose would be an input and blood pressure an output. A second, more technical distinction is that in order to obtain an unbiased estimate of the effect of a covariate on an outcome variable using standard survival analysis methods, it is necessary that the covariate can be measured at all times and without error (Carroll *et al.*, 2006). Hence in the present context, if our only objective was to understand the effect of renal function on the hazard for initiation of RRT *and* we were able to monitor error-free GFR continuously over time, we would treat GFR as a time-varying covariate and formulate a statistical model for a patient's hazard given their GFR history. But this is infeasible. We therefore need to model eGFR measurements and time to initiation of RRT jointly in order to understand the relationship between the underlying error-free GFR and the hazard for initiation of RRT. This is shown schematically in Figure 2.2.

The two components of the resulting joint model, which we explain in detail in the next section of the paper, are:

i. a linear mixed model for the time-course of eGFR;

ii. and a proportional hazards model for the time to initiation of RRT with time-varying random effects.

Both components of the joint model will include terms for measured covariates and unmeasured, error-free GFR, which we treat as a time-varying, patient-specific random effect, GFR(t). This framework, and in particular the linkage of the two components of the joint model through a shared random effect, allow us to answer a range of questions simultaneously according to the goals of each specific application. For example:

i. what is the typical pattern of progression in GFR, and how is this affected by baseline or time-varying covariates;

ii. and how do changes in level of GFR affect survival prognosis?

2.2.4 Links between missing data mechanisms and joint modelling

The term 'missing data' refers to data that are intended to be observed, but are unobserved for some reason (Little and Rubin, 2002; Molenberghs and Kenward, 2007; Ibrahim and Molenberghs, 2009). This is a common occurrence in both longitudinal and cross-sectional studies, either in the explanatory variables or the outcome variables or both. Missing data in longitudinal studies can be either *intermittent*, i.e. a patient might miss some of their hospital visits and return to the study later, or *drop-out*, i.e. a patient might leave the study prematurely. Missing data can be: (i) missing completely at random; (ii) missing at random; or (iii) missing not at random (MNAR). Details of these mechanisms can be found in any material on missing data; see for example, Little and Rubin (2002), Molenberghs and Kenward (2007) and Ibrahim and Molenberghs (2009). There is a close link between drop-out and joint modelling in that drop-out time can be considered as a survival outcome. An operational distinction is that, in the missing data literature, drop-out is typically inferred from a patient's failure to present at a scheduled follow-up time and treated as a discrete-time outcome whereas, in the joint modelling literature, event-time of interest is either recorded exactly or right-censored at the study end-time. The MNAR case is of particular interest in the present context, because it implies that the drop-out time conveys information about the unobserved, error-free longitudinal measurement process over and above that provided by the longitudinal measurements that are observed before drop-out. An example of MNAR in the context of nephrology would be a randomised clinical trial in which patients are likely to drop out when they perceive a lack of benefit from the diet.

2.3 Explanation of statistical methods

2.3.1 Repeated measurements

The most widely used class of models for repeated measurement data is the linear mixed effects model (Laird and Ware, 1982). This is defined by:

$$Y_{ij} = Y_i^*(t_{ij}) + Z_{ij} = X_{ij}\beta + W_{ij}B_i + Z_{ij}. \tag{2.3}$$

Here, Y_{ij} denotes the j^{th} eGFR measurement for the i^{th} patient, a value which is typically measured with error, $Y_i^*(t)$ denotes the true GFR level at time t, and Z_{ij} denotes measurement error. A patient's true GFR level can be decomposed into two components: fixed effects $X_{ij}\beta$; and random effects, $W_{ij}Bi$. The fixed effects represent the expected behaviour of kidney function, averaged over all patients who share the same covariate information; hence, X_{ij} is a vector containing the values of covariates that relate to the i^{th} patient at the time of their j^{th} eGFR measurement. The effects of changing the values of the covariates are represented by the corresponding elements of the regression parameter vector β. The random effects describe how patient-specific true GFR levels deviate from their expected behaviour. Each W_{ij} is a vector containing the values of covariates that relate to the i^{th} patient at the time of their j^{th} eGFR measurement, whereas B_i is analogous to β but, rather than taking a fixed unknown value, varies randomly among patients. Typically the B_i are assumed to follow a zero mean multivariate Normal distribution. There is no requirement for the same covariates to be included in the fixed and random effect components of the model, but typically the latter is a subset of the former. Both X_{ij} and W_{ij} can include time-constant and time-varying covariates.

Parameters are typically estimated by maximum or restricted maximum likelihood (ML and REML, respectively), and the random effects B_i are predicted by their conditional expectations given the data, so as to minimize mean square prediction error. ML is a general method for estimating parameters in complex statistical models that is known to have desirable theoretical properties (see for example, Pawitan (2001)). Many familiar elementary statistical methods can be derived as special cases of ML or REML estimations including, for example, t-tests, linear regression and generalized linear modelling. Note that no survival information is considered in Equation 2.3.

2.3.2 Survival times

The most widely used model for analysing survival data is the Cox proportional hazards model (Cox, 1972). This is given by:

$$\lambda_i(t|K_i) = \lambda_0(t)\exp(K_i\alpha). \tag{2.4}$$

Here, $\lambda_i(t|K_i)$ is the hazard for the i^{th} patient to experience the event of interest, e.g. initiation of RRT, at time t. It depends on time-constant covariates represented by the elements of a vector of covariates K_i with associated regression parameters α, and a baseline hazard function $\lambda_0(t)$ that represents the hazard for (possibly hypothetical) patients all of whose covariates take the value zero. In most applications, the main interest is in estimating α, which describes how the covariates affect the relative, rather than absolute, hazard. An attractive feature of the Cox proportional hazards methodology is then that it allows the baseline hazard to be left unspecified. If the baseline hazard is of interest, it can be estimated non-parametrically, or modelled parametrically with a specified class of lifetime distributions (Kalbfleisch and Prentice, 2002). A common choice for a parametric specification is the Weibull hazard, which follows a power law, $\lambda_0(t) = \lambda_0 k t^{k-1}$ where $\lambda_0 > 0$ and $k > 0$. Piecewise constant or regression spline function are popular choices for $\lambda_0(t)$ if more flexible parametric modelling is required (Rizopoulos, 2012). Estimates of α are typically obtained by maximizing the partial likelihood (Cox, 1972, 1975) in the Cox model, or by ML estimation in parametric models.

In principle, time-varying covariates can be added to the model given in Equation 2.4 by making the elements of K_i functions of time, hence $K_i(t)$; the resulting model is known as the Cox model with time-varying covariate. However, this requires all time-varying covariates to be measured continuously and without measurement error, which is only feasible in special cases, for example where the elements of $K_i(t)$ are functions of time itself. Note however, that in the Cox model any function of time alone is absorbed into the baseline hazard, $\lambda_0(t)$, and therefore cannot be estimated using the partial likelihood. In some applications, continuous measurement of time-varying covariates is induced by interpolating between actual measurements, but this is both an artificial device and also takes no account of measurement error. A key advantage of joint modelling is its ability to handle irregularly and imperfectly measured time-varying covariates correctly.

Inclusion of random effects in the model given in Equation 2.4 is comparatively straightforward. The simplest example takes the form:

$$\lambda_i(t|K_i) = \lambda_0(t)\exp(K_i\alpha + A_i) \tag{2.5}$$

where the A_i are drawn from a statistical distribution. In the survival literature, the quantity $\exp(A_i)$ is called the frailty of the i^{th} patient, and the distribution of the A_i is scaled so that the average frailty is 1.

2.3.3 Joint modelling

A joint model for data on eGFR and time to initiation of RRT can now be defined by the following two equations:

$$Y_{ij} = Y_i^*(t_{ij}) + Z_{ij} = X_{ij}\beta + W_{ij}B_i + Z_{ij}, \tag{2.6}$$

$$\lambda_i(t|K_i, Y_i^*(t)) = \lambda_0(t)\exp(K_i\gamma_1 + Y_i^*(t)\gamma_2). \tag{2.7}$$

Here, γ_2 measures the relationship between the unmeasured, error-free GFR process, $Y_i^*(t)$, and the time to initiation of RRT. The fundamental feature of joint modelling is that repeated measurement and survival data are modelled simultaneously. The algorithm for estimating the parameters of the joint model given in Equations 2.6 and 2.7 also exploits the model assumptions to predict the values of $Y_i^*(t)$ at all times t, and thereby to estimate the associated regression parameter γ_2 while making proper allowance for the measurement error in the observed eGFR values. The model can be extended to include particular features of the error-free GFR processes $Y_i^*(t)$. For example, rate of change in GFR can be added to Equation 2.7 as an additional term with its own regression parameter, i.e. $Y_i^{*'}(t)\gamma_3$. Other possible extensions include interactions of kidney function with a set of baseline covariates; lagged or cumulative effect of kidney function on the hazard for survival. Alternatively, only the random effect component of the longitudinal sub-model might be included in the survival sub-model instead of the current kidney function level, $Y_i^*(t)$. For further details, see Chapter 5 of Rizopoulos (2012) and Henderson, Diggle and Dobson (2000).

Parameters of joint models are typically estimated by maximizing the likelihood, and random effects are predicted by their conditional expectations given all of the data. The interpretations of the parameters of a joint model are the same as for their linear mixed effects and Cox components.

The benefits of joint modelling are not cost free. The main disadvantages of joint modelling are the increase in computational effort required to fit the models and the relative scarcity of software to enable their routine use. The former is only a significant problem when dealing with very large data sets, in particular data sets with large numbers of

repeated measurements on each subject. The latter is being addressed by the development of packages such as JM (Rizopoulos, 2010) and joineR (Philipson *et al.*, 2012) that run within the open-source R computing environment.

2.4 Framework for statistical analysis

Since some patients missed their annual data updates, there are intermittent missing values in the data set, which we treated as missing at random. Each patient contributed both repeated measurement and survival outcomes; the former are the repeated measurements of kidney function as determined by eGFR, the latter are (possibly censored) times to initiation of RRT. In our analysis, we treated death before initiation of RRT as a right-censored event-time. Our analysis comprised three main steps: (i) separate longitudinal analysis of repeated eGFR measurements; (ii) separate survival analysis of time to initiation of RRT using the Cox model with eGFR treated as a time-varying covariate by carrying forward each observed value of eGFR at a constant level until the next observed value on the same patient; and (iii) joint analysis of the repeated eGFR measurements and time to initiation of RRT. In the first step, we built a linear mixed effects model with repeated eGFR data as the response variable, ignoring the potentially informative nature of the censoring of each eGFR sequence by the occurrence of the survival event. In the second step, we analysed RRT with *observed* repeated eGFR measurements treated as a time-varying covariate. In the third step, we considered joint analysis of repeated eGFR measurements and time to initiation of RRT with the current (unobserved) value of GFR included in the survival submodel in addition to the baseline covariates. We analysed log-transformed eGFR (Y = log(eGFR)) data throughout for the following three reasons. First, this transformation leads to an approximately linear relationship with age and approximately symmetrical scatter about the long-term trend. Second, analyses of log(eGFR) and log-transformed creatinine would be equivalent when the former is re-adjusted with age and gender (see Equation 2.1). Finally, using a log transformation leads to an interpretation of the fitted model in terms of relative, rather than absolute, change in eGFR, which relates more directly clinical guidelines for monitoring changes in kidney function.

2.5 Results

2.5.1 Study population

All of the log(eGFR) measurements are shown in Figure 2.3. Individual trajectories for 10 patients randomly selected among those with at least three observations are highlighted. Median age at recruitment was 67.2 years (IQR 55.6-74.9); 603 (37.4%) of the patients were female; 1551 (96.3%) of the patients were Caucasian; 509 (31.6%) had SRFT as the base hospital. Mean log(eGFR) at the first hospital visit was 3.4 (standard deviation 0.5); and mean log(eGFR) at the last hospital visit was 3.3 (standard deviation 0.6). The log(eGFR) measurements at first and last visits are displayed in the upper and lower panels of Figure 2.4, respectively. In total, 516 (32%) of the patients had diabetes mellitus, 1086 (67.4%) were current or former smokers and 342 (21.2%) had a history of coronary artery disease (defined as previous myocardial infarction and/or coronary revascularization procedure). Median follow-up period was 3.9 years (IQR 2.2-5.6); 304 (19%) of the patients experienced initiation of RRT. The Kaplan-Meier survival plot for RRT as the survival outcome is displayed in Figure 2.5.

2.5.2 Separate analysis

Details of the variables considered in the longitudinal model for log(eGFR) and in the survival model are presented in Table 2.1. Note that we decompose age at measurement into age at recruitment and time since recruitment (in years) in order to differentiate between cross-sectional and longitudinal effects of age. To explain this distinction, consider the following simple regression model, in which age_0 denotes age on entry to the study:

$$log(eGFR) = \alpha + \beta_1 * age_0 + \beta_2 * (age - age_0) + \text{noise}.$$

In this model, the parameters β_1 and β_2 represent the cross-sectional and longitudinal effects of age, respectively. If $\beta_1 = \beta_2$ the model reduces to:

$$log(eGFR) = \alpha + \beta_1 * age + \text{noise}.$$

but there is no reason in general why this should be so. For example, if the stages of the disease for different patients are approximately the same at each follow-up, the cross-sectional effect β_1 would be approximately zero; but as patients typically lose renal function over time, the longitudinal effect β_2 would be negative.

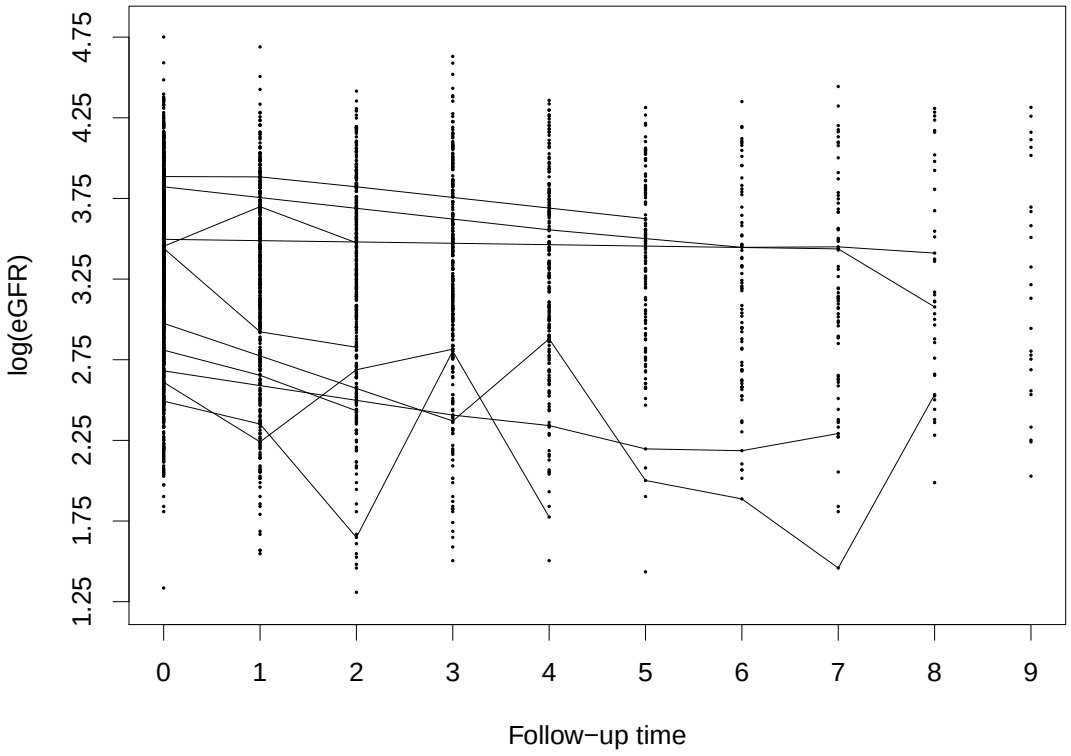

FIGURE 2.3: All log(eGFR) measurements. Trajectories for 10 randomly selected patients are shown as connected line-segments.

2.5.3 Longitudinal model

We fit the so-called random-intercept-and-random-slope version of the model given in Equation 2.3, namely:

$$Y_{ij} = Y_i^*(t_{ij}) + Z_{ij} = X_{ij}\beta + A_i + B_i t_{ij} + Z_{ij}, \tag{2.8}$$

where t_{ij} denotes time since recruitment. The fixed effects estimates from the separate longitudinal model for change in log(eGFR) are presented in Table 2.2. Kidney function was found to decrease with increasing age at study start [Estimate = -0.005, 95% confidence interval (CI) -0.007, -0.003] and with increasing time under observation (Estimate = -0.064, 95% CI -0.073, -0.056). Recall that parameter estimates represent relative rather than absolute changes. Hence, a 1-year increase in age at study start was associated with a relative decrease of 0.5% (=(exp(-0.005)-1)*100) in expected eGFR. Similarly, a 1-year increase in time under observation was associated with a relative decrease of 6.2% in expected eGFR. Patients living in the catchment area of SRFT were found to have 15.1% higher expected eGFR at recruitment compared with the patients

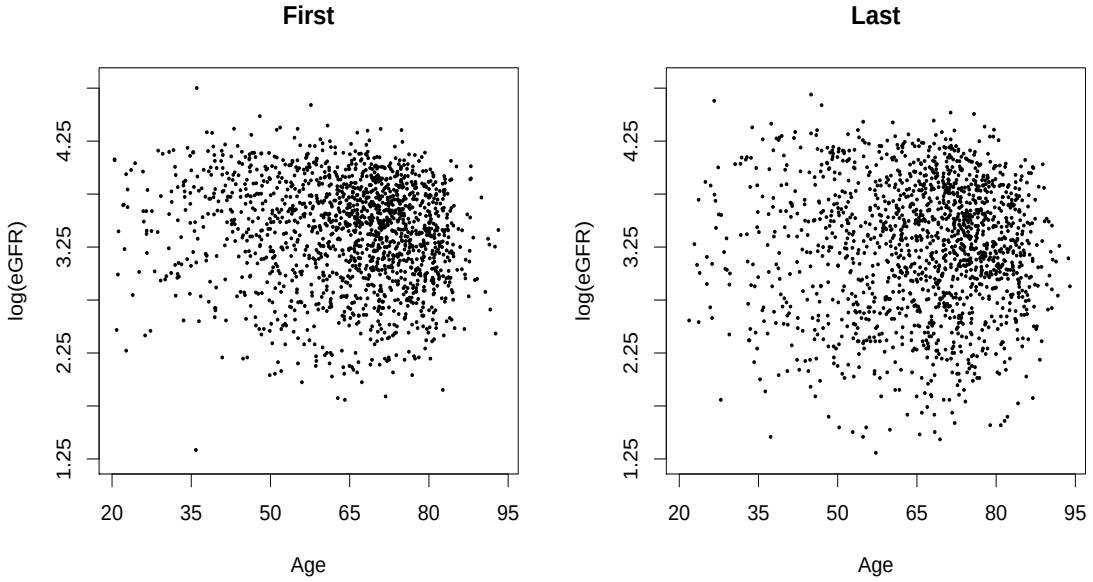

FIGURE 2.4: log(eGFR) measurements at the first (left panel) and the last (right panel) follow-ups.

TABLE 2.1: Covariates used in analyses of the CRISIS data.

Variable	Explanation
Baseline age	(Date at study start date at birth)/365.25
Follow-up	Age at measurement age at baseline
Hospital base	1 if base hospital is SRFT, 0 otherwise
Gender	1 if male, 0 if female
Smoking	1 if ex or current smoker, 0 if never smoked
Alcohol	1 if alcohol consumer, 0 if abstinent from alcohol
Diabetes	1 if type I or type II diabetes, 0 if no diabetes
Co-morbidity	1 if having at least one of myocardial infarction, coronary artery bypass surgery or stenting, 0 otherwise

initially managed at satellite units. Males were found to have 8.3% higher expected eGFR than females. Patients who had type 1 or type 2 diabetes were found to have 9.2% lower expected eGFR than non-diabetic patients. No differences in expected eGFR were found between ex or current smokers and non-smokers, between alcohol consumers and abstainers from alcohol or between patients who did or did not have at least one comorbidity event; respective p-values were 0.601, 0.098 and 0.468.

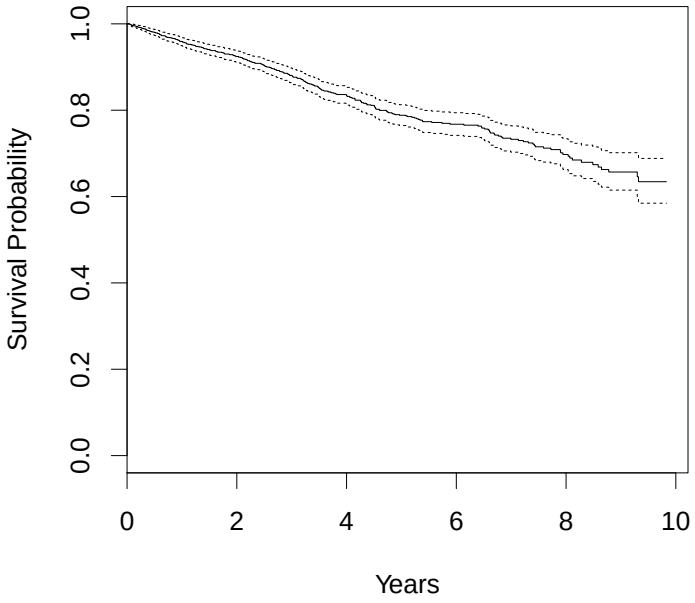

FIGURE 2.5: Kaplan-Meier survival plot for RRT as the survival event.

TABLE 2.2: Estimated regression parameters, 95% confidence intervals (95% CI), standard errors (SE), p-values (p) and percentage relative effects (RE %) in separate longitudinal analysis of the CRISIS data set. The response variable is log transformed eGFR. RE % corresponding to an estimate $\hat{\beta}$, expressed as expected percentage change in eGFR, calculated as $(exp(\hat{\beta}) - 1) * 100$.

Variable	Estimate (95% CI)	SE	p	RE %
Intercept	3.585 (3.463, 3.707)	0.062	<0.001	NA
Baseline age (years)	-0.005 (-0.007, -0.003)	0.001	<0.001	-0.5
Follow-up (years)	-0.064 (-0.073, -0.056)	0.004	<0.001	-6.2
Hospital base	0.141 (0.088, 0.194)	0.027	<0.001	+15.1
Gender	0.080 (0.028, 0.133)	0.027	0.003	+8.3
Smoking	0.014 (-0.040, 0.068)	0.028	0.601	+1.4
Alcohol	0.043 (-0.008, 0.093)	0.026	0.098	+4.4
Diabetes	-0.097 (-0.151, -0.044)	0.027	<0.001	-9.2
Comorbidity	-0.024 (-0.087, 0.040)	0.032	0.468	-2.4

2.5.4 Survival model

Results obtained from a Cox model for the survival outcome are displayed in Table 2.3. Time-varying eGFR was associated with the hazard for RRT. The corresponding parameter estimate was -2.510 (95% CI -2.691, -2.330); hence, a 1% reduction in eGFR was associated with a 2.5% (=exp(2.510/100)) increased risk of RRT, with lower and upper

TABLE 2.3: Estimated regression parameters, 95% confidence intervals (95% CI), standard errors (SE), p-values (p), hazard ratios (HR) and related 95% confidence intervals, in analysis of CRISIS data set with Cox model with time-varying covariate for RRT as the event

Variable	Estimate (95% CI)	SE	p	HR (95% CI)
log(eGFR)	-2.510 (-2.691, -2.330)	0.092	<0.001	0.081 (0.068, 0.097)
Baseline age (years)	-0.030 (-0.038, -0.021)	0.004	<0.001	0.971 (0.963, 0.979)
Hospital base	-0.368 (-0.645, -0.091)	0.141	0.009	0.692 (0.525, 0.913)
Gender	0.139 (-0.106, 0.385)	0.125	0.267	1.149 (0.899, 1.469)
Smoking	0.362 (0.112, 0.612)	0.128	0.005	1.436 (1.119, 1.844)
Alcohol	0.055 (-0.180, 0.290)	0.120	0.647	1.056 (0.835, 1.336)
Diabetes	-0.109 (-0.366, 0.149)	0.131	0.409	0.897 (0.693, 1.161)
Comorbidity	-0.043 (-0.403, 0.317)	0.184	0.815	0.958 (0.669, 1.373)

95% confidence interval limits of 2.4% (=exp(2.330/100)) and 2.7% (=exp(2.691/100)), respectively. Baseline age was also associated with the hazard for RRT, but unexpectedly in the negative direction; the corresponding estimated hazard ratio per year of age was 0.971 (95% CI 0.963, 0.979). Patients whose base hospital was not SRFT were estimated to have 44.5% higher hazard for RRT than patients whose base hospital was SRFT. Patients who were ex or current smokers were estimated to have a 43.6% higher hazard for RRT than patients who never smoked. On the other hand, there were no differences between males and females, between alcohol consumers and abstainers from alcohol, between diabetic and non-diabetic patients or between patients who did or did not have at least one comorbidity event; related p-values were 0.267, 0.647, 0.409 and 0.815, respectively.

2.5.5 Joint model

The longitudinal sub-model of the joint model was a random-intercept-and-random-slope model, as given in Equation 2.8. Results are shown in Table 2.4. The results for the longitudinal sub-model were consistent with the results from the separate longitudinal analysis. This might be explained by the fact that the modelling assumptions are the same and the censoring of the eGFR measurements due to RRT is not severe. The differences in magnitudes of the parameter estimates were negligible and there was no material difference in terms of statistical significance. In contrast, material differences were found in the results for the survival processes. This shows the importance of recognizing the measurement error in the observed values of eGFR. Hence, in what follows we discuss only the results for the survival sub-models.

TABLE 2.4: Results for joint modelling analysis of CRISIS data set. For the longitudinal sub-model estimated parameters and related 95% confidence intervals (95% CI), standard errors (SE), p-values (p) and precentage relative effects (RE %) are reported. For the survival sub-model, estimated parameters, related 95% confidence intervals, standard errors, p-values, hazard ratios (HR) and related 95% confidence intervals are reported.

Longitudinal sub-model				
Variable	Estimate (95% CI)	SE	p	RE %
Intercept	3.614 (3.495, 3.732)	0.060	<0.001	NA
Baseline age (years)	-0.005 (-0.007, -0.003)	0.001	<0.001	-0.5
Follow-up (years)	-0.073 (-0.081, -0.064)	0.004	<0.001	-7.0
Hospital base	0.133 (0.081, 0.185)	0.027	<0.001	+14.2
Gender	0.077 (0.026, 0.129)	0.026	0.003	+8.0
Smoking	0.021 (-0.032, 0.074)	0.027	0.430	+2.2
Alcohol	0.038 (-0.011, 0.088)	0.025	0.130	+3.9
Diabetes	-0.096 (-0.149, -0.044)	0.027	<0.001	-9.2
Comorbidity	-0.029 (-0.093, 0.035)	0.032	0.373	-2.8
Survival sub-model				
Variable	Estimate (95% CI)	SE	p	HR (95% CI)
log(GFR)	-3.656 (-4.042, -3.270)	0.197	<0.001	0.026 (0.018, 0.038)
Baseline age (years)	-0.036 (-0.046, -0.026)	0.005	<0.001	0.964 (0.955, 0.974)
Hospital base	-0.190 (-0.509, 0.128)	0.162	0.241	0.827 (0.601, 1.136)
Gender	0.204 (-0.085, 0.493)	0.148	0.167	1.226 (0.918, 1.637)
Smoking	0.480 (0.186, 0.774)	0.150	0.001	1.616 (1.204, 2.169)
Alcohol	0.036 (-0.239, 0.311)	0.140	0.796	1.037 (0.788, 1.365)
Diabetes	0.085 (-0.210, 0.380)	0.151	0.572	1.089 (0.811, 1.463)
Comorbidity	-0.112 (-0.534, 0.310)	0.215	0.603	0.894 (0.586, 1.363)

Most importantly, we found that a 1% reduction in GFR level was associated with a 3.7% (95% CI 3.3%, 4.1%) increased hazard for RRT, compared with 2.7% (95% CI 2.4%, 2.9%) previously obtained in the separate analysis of the survival outcome. There was no difference between patients whose base hospital was or was not SRFT (p-value = 0.241), whereas in the separate analysis of survival outcome, base hospital was associated with the hazard for RRT (p-value = 0.009).

2.6 Discussion

Our aim in this paper has been to present an introductory tutorial on simultaneous analysis of repeated measurement and time-to-event outcome data, using the *joint modelling* approach that features most prominently in the biostatistical literature. This approach incorporates the most widely used methods that have been developed for separate analysis of the two types of outcome, namely linear mixed effects modelling and Cox proportional hazards modelling with frailty, and combines the two by linking their respective

random effects. The principal advantage of this approach over separate analyses of each outcome is the correct treatment of noisy and incompletely observed time-varying covariate information, which enables unbiased estimation of the relationship between the two.

We have reported an analysis of a data set from the CRISIS cohort on CKD patients, where the repeated measurements are serial eGFR measurements and the survival outcome is time to initiation of RRT. The results demonstrate the usefulness of kidney function as a predictor for the hazard for initiation of RRT. This was substantially underestimated in a separate analysis of time to initiation of RRT treating eGFR as a time-varying covariate, because the separate analysis fails to take account of the measurement error in eGFR. In general, measurement error in a covariate biases the estimate of the associated regression parameter towards zero (Wulfsohn and Tsiatis, 1997; Ibrahim, Chu and Chen, 2010; Prentice, 1982; Sweeting and Thompson, 2011).

We found an unexpected result that increased baseline age was associated with a decreased hazard for initiation of RRT. This is most likely explained by the fact that regression associations do not generally equate to causal relationships. Another explanation might be that younger patients with poor kidney function were given priority for RRT.

Our analysis of time to initiation of RRT as the survival outcome alone, treating death as a right-censored event time, could be criticised for failing to take account of the status of RRT and death as asymmetrical competing risks, in the sense that RRT necessarily precedes death, whereas death automatically censors initiation of RRT. Joint models for such competing risks have been considered, for example in the work of Williamson (2008), but are beyond the scope of this paper; see also Chapter 5.5 of Rizopoulos (2012). Similarly, we did not include proteinurea or blood pressure data in our analyses because both are subject to measurement error and are collected intermittently at irregular times. We would argue that analysis of proteinurea and blood pressure together with eGFR strictly requires *multivariate* joint modelling. This is an area of current methodological research and cannot be implemented routinely in freely available software packages.

The joint modelling framework we considered in this paper is also called the shared random effects model. An alternative approach to this formulation is the latent class model. For a recent review, see Proust-Lima *et al.* (2014). Latent class models can be fitted by the R package lcmm (Proust-Lima *et al.*, 2014).

We have presented joint modelling results obtained by ML estimation. An alternative approach for parameter estimation is Bayesian inference (Ibrahim, Chen and Sinha, 2001; Guo and Carlin, 2004). This can be applied to a limited class of models with

the R package JMbayes (Rizopoulos, 2014). Model fit can be assessed by Akaike or Bayesian Information Criteria (Zhang *et al.*, 2014). The sensitivity of the fixed effects parameters in the longitudinal sub-model might be investigated by, for example, index of local sensitivity to nonignorability (Viviani, Rizopoulos and Alfó, 2014). Diagnostic tools can be applied to multiple-imputation based empirical residuals for the longitudinal sub-model of the joint models (Rizopoulos, Verbeke and Molenberghs, 2010). Methods to inspect the association between longitudinal and time-to-event data, to investigate empirical residuals of the longitudinal sub-model after fitting a joint model and to detect influential observations within the joint modelling framework can be found in Dobson and Henderson (2003).

Software for the separate analysis of longitudinal and survival data is widely available. For the results reported in this paper, we used the R packages nlme (Pinheiro *et al.*, 2013) and survival (Therneau, 2013) for longitudinal and survival data analysis, respectively. Software for joint modelling is becoming increasingly available in statistical packages, for example in the R packages JM (Rizopoulos, 2010), joineR (Philipson *et al.*, 2012) and JMbayes (Rizopoulos, 2014). The R scripts for the analyses reported in this paper are provided as Supplementary data, available at *IJE* online.

2.7 Online supplementary material: R codes

We have two data sets, one for the longitudinal data and one for the survival data, named as `longitudinal.data` and `survival.data` in `R`, respectively. The explanations of the variable names of these data sets are given below:

- mdrd: eGFR measurements
- age.0: baseline age
- fu: follow-up
- salford: base hospital
- male: gender
- smoking: smoking status
- alcohol: alcohol consumption
- diabetes: diabetic conditions
- mcp: co-morbidity

- stime: survival time
- rrt.event.ind: event indicator variable for RRT
- start: the follow-up time at which the measurements are taken
- stop: the sub-sequent follow-up time of start, but the last value is the survival time
- rrt.event: time varying event indicator for RRT for each start stop intervals

The models were fitted by using the following script:

```
##### Separate analyses of longitudinal and survival data

### Longitudinal data analysis

R> install.packages(nlme)
R> library(nlme)

R> lme.fit <- lme(log(mdrd) ~ age.0 + fu + salford + male + smoking +
                    alcohol + diabetes + mcp,
                    random = ~ fu | ID, method = "ML",
                    data = longitudinal.data)
R> summary(lme.fit)


### Survival data analysis

## Cox model with time-varying covariate for RRT
R> install.packages(survival)
R> library(survival)

R> tdep.rrt <- coxph(Surv(start, stop, rrt.event) ~ log(mdrd) + age.0 +
                      salford + male + smoking + alcohol + diabetes +
                      mcp, data = longitudinal.data)
R> summary(tdep.rrt)


##### Joint model
```

```
R> install.pacakges(JM)
R> library(JM)

## JointModel function in the JM package fits joint models. It requires
## the fitted objects of separate longitudinal and survival (Cox
## proportional hazards model with baseline covariate) models.

# RRT

R> lme.fit <- lme(log(mdrd) ~ age.0 + fu + salford + male + smoking +
                    alcohol + diabetes + mcp, random = ~ fu | ID,
                    method = "ML", data = longitudinal.data)
R> cox.rrt <- coxph(Surv(stime, rrt.event.ind) ~ age.0 + salford +
                    male + smoking + alcohol + diabetes + mcp,
                    data = survival.data, x = TRUE)
R> jm.rrt <- jointModel(lme1, cox.rrt, timeVar = "fu",
                    method = "piecewise-PH-aGH", iter.EM = 150,
                    verbose = T)
R> summary(jm.rrt)
```

Bibliography

Al-Aly Z., Zeringue A., Fu J., Rauchman M. I., McDonald J. R., El-Achkar T. M., Balasubramanian S., Nurutdinova D., Xian H., Stroupe K., Abbott K. C. and Eisen S. (2010). Rate of kidney function decline associates with mortality. *Journal of the American Society of Nephrology* **21**, 1961–1969.

Andrinopoulou E. R., Rizopoulos D., Jin R., Bogers A. J., Lesaffre E. and Takkenberg J. J. (2012). An introduction to mixed models and joint modeling: analysis of valve function over time. *The Annals of Thoracic Surgery* **93**, 1765–1772.

Carroll R. J., Ruppert D., Stefanski L. A. and Crainiceanu C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. Boca Raton: Chapman & Hall/CRC.

Cox D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society - Statistical Methodology* **34**, 187–220.

Cox D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Daher Abdi Z., Essig M., Rizopoulos D., Le Meur Y., Prómaud A., Woillard J. B., Rórolle J. P., Marquet P. and Rousseau A. (2013). Impact of longitudinal exposure to mycophenolic acid on acute rejection in renal-transplant recipients using a joint modeling approach. *Pharmacological Research* **72**, 52-60.

Diggle P. J., Heagerty P., Liang K.Y. and Zeger S. L. (2002). *Analysis of longitudinal data*, 2nd edition. Oxford: Oxford University Press.

Diggle P. J., Sousa I. and Chetwynd A. (2007). Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. *Statistics in Medicine* **26**, 2981–2998.

Dobson A. and Henderson R. (2003). Diagnostics for joint longitudinal and dropout time modelling. *Biometrics* **59**, 741–751.

Eddington H., Hoefield R., Sinha S., Chrysochou C., Lane B., Foley R. N., Hegarty J., New J., O'Donoghue D. J., Middleton R. J. and Kalra P. A. (2010). Serum phosphate

and mortality in patients with chronic kidney disease. *Clinical Journal of the American Society of Nephrology* **5**, 2251–2257.

Fitzmaurice G. M., Laird N. M. and Ware J. H. (2011). *Applied longitudinal analysis*, 2nd edition. New Jersey: John Wiley & Sons.

Garre F. G., Zwinderman A. H., Geskus R. B. and Sijpkens Y. W. J. (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society - Statistics in Society* **171**, 299-308.

Guo X. and Carlin B. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistician* **58**, 16–24.

Henderson R., Diggle P. J. and Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Hoefield R. A., Kalra P. A., Baker P., Lane B., New J. P., O'Donoghue D. J., Foley R. N. and Middleton R. J. (2010). Factors associated with kidney disease progression and mortality in a referred CKD population. *American Journal of Kidney Diseases* **56**, 1072–1081.

Ibrahim J. G. and Molenberghs G. (2009). Missing data methods in longitudinal studies: a review. *Test* **18**, 1–43.

Ibrahim J., Chen M. and Sinha D. (2001). *Bayesian Survival Analysis*, New York: Springer-Verlag.

Ibrahim J. G., Chu H. and Chen L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**, 2796–2801.

Kalbfeisch J. D. and Prentice R. L. (2002). *The statistical analysis of failure time data*, 2nd edition. New York: John Wiley & Sons.

Kleinbaum D. G. and Klein M. (2012). *Survival analysis: a self learning text*, 3rd edition. New York: Springer-Verlag.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Levey A. S., Bosch J. P., Lewis J. B., Greene T., Rogers N. and Roth D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of Internal Medicine* **130**, 461–470.

Little R. J. A. and Rubin D. B. (2002). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons.

Molenberghs G. and Kenward M. G. (2007). *Missing Data in Clinical Studies*, New York: John Wiley & Sons.

Pawitan Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford: Oxford University Press.

Philipson P., Sousa I., Diggle P. J., Williamson P., Kolamunnage-Dona R. and Henderson R. (2012). joineR: joint modelling of repeated measurements and time-to-event data, R package version 1.0-3, http://CRAN.R-project.org/package=joineR.

Pinheiro J., Bates D., DebRoy S., Sarkar D. and the R Development Core Team (2013). nlme: linear and nonlinear mixed effects models, R package version 3.1-109, http://CRAN.R-project.org/package=nlme.

Prentice R. L. (1982). Covariate measurement error and parameter estimation in a failure time regression model. *Biometrika* **69**, 331–342.

Prous-Lima C., Sóne M., Taylor J. M. and Jacqmin-Gadda H. (2014). Joint latent class models for longitudinal and time-to-event data: a review. *Statistical Methods in Medical Research* **23**, 74–90.

Proust-Lima C., Philipps V., Diakite A. and Liquet B. (2014). lcmm: extended mixed models using latent classes and latent processes, R package version 1.6.4, http://CRAN.R-project.org/package=lcmm (26 September 2014, date last accessed).

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, URL http://www.R-project.org/.

Rizopoulos D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**, 1-33.

Rizopoulos D., Verbeke G. and Molenberghs G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics* **66**, 20–29.

Rizopoulos D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*, Boca Raton: Chapman & Hall/CRC.

Rizopoulos D. (2014). JMbayes: joint modeling longitudinal and time-to-event data under a Bayesian approach, R package version 0.6-1, URL: http://CRAN.R-project.org/package=JMbayes (5 May 2014, date last accessed).

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

Therneau T. (2013). A package for survival analysis in S, R package version 2.37-4, URL: http://CRAN.R-project.org/package=survival.

Tsiatis A. A., DeGruttola V. and Wulfsohn M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

Viviani S., Rizopoulos D. and Alfó M. (2014). Local sensitivity to non-ignorability in joint models. *Statistical Modelling* **14**, 205–228.

Williamson P., Kolamunnage-Dona R., Philipson P. and Marson A. G. (2008). Joint modelling of longitudinal and competing risks data. *Statistics in Medicine* **27**, 6426–6438.

Wulfsohn M. S. and Tsiatis A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Zhang D., Chen M. H., Ibrahim J. G., Boye M. E., Wang P. and Shen W. (2014). Assessing model fit in joint models of longitudinal and survival data with applications to cancer trials. *Statistics in Medicine* **33**, 4715–4733.

Chapter 3

Real-time monitoring of progression towards renal failure in primary care patients

This chapter is based on the following paper:

Diggle P. J., Sousa I. and Asar Ö. (2015). Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*, **16(3)**: 522–536.

Abstract

Chronic renal failure is a progressive condition that, typically, is asymptomatic for many years. Early detection of incipient kidney failure enables ameliorative treatment that can slow the rate of progression to end-stage renal failure, at which point expensive and invasive renal replacement therapy (dialysis or transplantation) is required. We use routinely collected clinical data from a large sample of primary care patients to develop a system for real-time monitoring of the progression of undiagnosed incipient renal failure. Progression is characterised as the rate of change in a person's kidney function as measured by the estimated glomerular filtration rate (eGFR), an adjusted version of serum creatinine level in a blood sample. Clinical guidelines in the UK suggest that a person who is losing kidney function at a relative rate of at least 5% per year should be referred to specialist secondary care. We model the time-course of a person's underlying kidney function through a combination of explanatory variables, a random intercept and a continuous-time, non-stationary stochastic process. We then use the model to calculate for each person the predictive probability that they meet the clinical guideline for referral to secondary care. We suggest that probabilistic predictive inference linked to clinical criteria can be a useful component of a real-time surveillance system to guide, but not dictate, clinical decision-making.

Key words: Dynamic modelling; Kidney failure; Longitudinal data analysis; Non-stationarity; Real-time prediction; Renal medicine; Stochastic processes.

3.1 Introduction

In this paper, we consider the problem of using routinely collected data from a large sample of people in primary care to monitor the progression of undiagnosed incipient renal failure. The problem is important because chronic renal failure can be asymptomatic for many years. Early detection followed by initiation of ameliorative treatment can slow the rate of progression to end-stage renal failure and so postpone the need for expensive, invasive and often scarce renal replacement therapy (dialysis or transplantation). Progression towards renal failure is characterised by a sustained fall in a person's glomerular filtration rate (GFR). However, direct measurement of GFR is expensive. For this reason, many specialist renal treatment centres now use as a clinical indicator

of a person's renal function an *estimated* glomerular filtration rate (eGFR). This is calculated from a person's age, sex, ethnicity and their level of serum creatinine (SCr) as determined by a blood sample. A widely used formula is the Modification of Diet in Renal Disease equation (MDRD, Levey *and others*, 1999),

$$\text{eGFR} = 175 \times \left(\frac{\text{SCr}}{88.4}\right)^{-1.154} \times \text{age}^{-0.203} \times 0.742^{\text{I(female)}} \times 1.21^{\text{I(black)}}. \qquad (3.1)$$

In common with other published formulae, (3.1) expresses a multiplicative relationship between eGFR (in ml/min per $1.73m^2$ of body surface area) and SCr (in μmol/L), adjusted by age, sex and ethnicity. In our study, information on ethnicity is not available but the population is mostly Caucasian, and we have ignored the ethnicity component of (3.1).

The data that we analyse consist of repeated measurements of eGFR taken at irregular, person-specific follow-up times for 22,910 primary care patients who have been diagnosed with pre-disposing conditions for renal failure; these co-morbidities, and other relevant baseline information, are included in the data as a set of explanatory variables attached to each person's record. The data were collected as part of a longitudinal cohort study run by the Salford Royal Hospital Foundation Trust (SRFT), Greater Manchester, UK.

Our proposed strategy, anticipating to some extent results from the preliminary analysis of the data as reported in Section 3.5, is to build a dynamic regression model in which a subject's rate of change in log-transformed GFR, relative to the expected profile for all people with the same values of the explanatory variables, is modelled as a stochastic process $B(t)$, which is realised independently for each person. Using the fitted model, we then evaluate the predictive distribution of $B(t)$ for each person, conditional on their data up to and including time t. Under this approach, a person would be flagged as a candidate for referral to a specialist treatment unit if and when the predictive probability that their current rate of decrease in GFR exceeds 5% per year is at least p_c, where p_c is a threshold value to be specified by the clinician. The choice of p_c will determine the balance between sensitivity and specificity.

The remainder of the paper is organized as follows. In Section 3.2 we give a more detailed description of the SRFT data. Section 3.3 describes our proposed model. Section 3.4 describes associated methods for parameter estimation and for prediction of $B(t)$. Section 3.5 describes the application to the SRFT data, including model diagnostics and individual predictions. Section 3.6 presents two simulation studies conducted to investigate the properties of the estimators of the model parameters and the influence of distributional and variance structure misspecifications on the predicted probabilities. Section 3.8 is a concluding discussion.

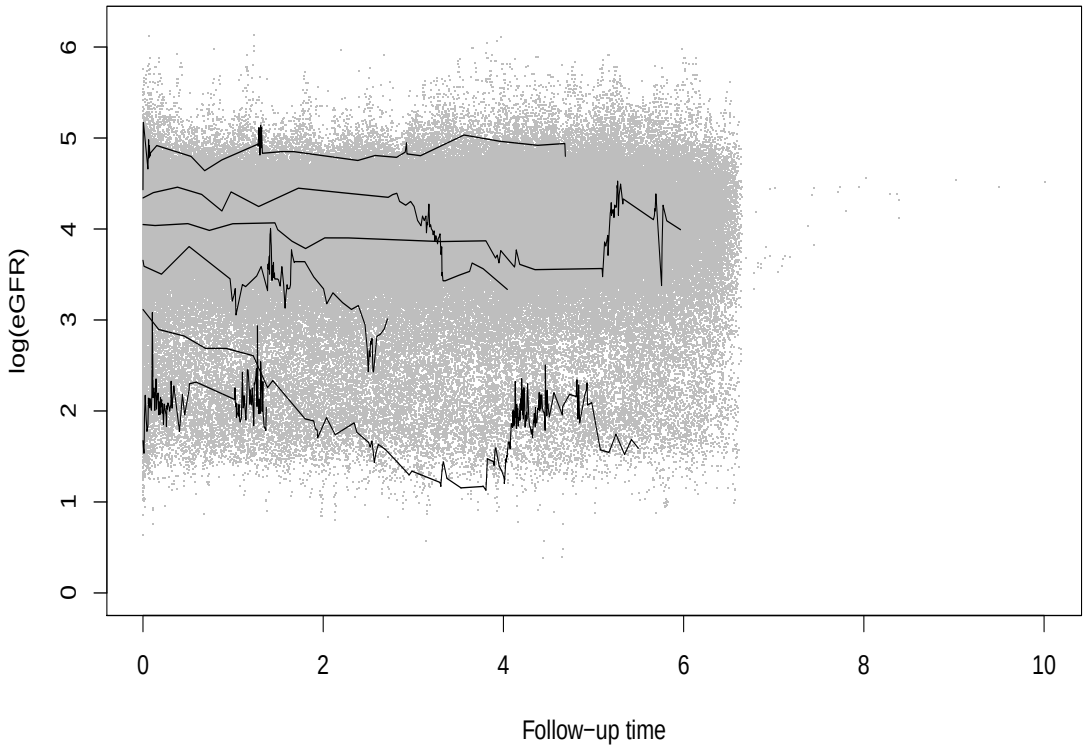

FIGURE 3.1: Log-transformed eGFR measurements against follow-up time (in years). Data from a representative sample of 6 patients are highlighted as black lines.

3.2 Data

The SRFT data set contains information on 22,910 patients who entered the study between March 7, 1997 and March 22, 2007 and met the criterion of being at risk for renal failure. The patients provided a total of 392, 870 values of eGFR. Of the 22,910 patients, 11, 833 (51.65%) were male. Their baseline age ranged between 13.74 and 102.10 years with a median of 67.19 years. The number of eGFR values per patient ranged between 1 and 305 with a median of 12. Total follow-up time ranged from 0 (i.e. only one eGFR value) to 10.02 years with median of 4.46 years. Figure 3.1 shows the complete set of log-transformed eGFR values as a grey scatterplot, with longitudinal trajectories for a representative sample of 6 patients highlighted as black lines. The data exhibit substantial variation in eGFR, both between patients and over time within patients.

3.3 Model formulation

We consider a general model for the longitudinal eGFR trajectories, of the form

$$Y_{ij} = \mu_i(t_{ij}) + U_i + W_i(t_{ij}) + Z_{ij}. \tag{3.2}$$

In (3.2), Y_{ij} denotes the log-transformed eGFR response for subject i ($i = 1, \ldots, M$) at time t_{ij} ($j = 1, \ldots, n_i$). The function $\mu_i(t)$ is the expected value of the response, which we represent as a multiple linear regression, hence $\mu_i(t_{ij}) = \boldsymbol{X}_i(t_{ij})\boldsymbol{\alpha}$, where $\boldsymbol{X}_i(t_{ij})$ denotes a set of explanatory variables and $\boldsymbol{\alpha}$ denotes the corresponding set of fixed effects regression parameters to be estimated. The U_i are independent $\mathrm{N}(0, \omega^2)$ random variables that represent time-constant differences amongst patients that cannot be explained by the linear regressions. The $W_i(t)$ are independent copies of a zero-mean, continuous-time stochastic process representing change in a patient's GFR over time that cannot be explained by the linear regressions. We assume that this continuous-time process is Gaussian, and therefore specified by its covariance function, $\gamma(s,t) = \mathrm{Cov}(W_i(s), W_i(t))$. Finally, the Z_{ij} are mutually independent $\mathrm{N}(0, \tau^2)$ random variables representing measurement error in the determination of Y_{ij}. Note that the model expresses eGFR as a noisy version of GFR.

The scalar-valued random effects U_i in (3.2) could be replaced by a second multiple linear regression, $\boldsymbol{X}_i^*(t_{ij})\boldsymbol{U}_i$, where now the $\boldsymbol{U}_i$ are mutually independent, multivariate Normal random variables. The term $W_i(t_{ij})$ in (3.2) might then be omitted. A widely used example of such specification is the random-intercept-and-slope model in which $\boldsymbol{X}_i(t_{ij}) = (1, t_{ij})$ (Laird and Ware, 1982). However, this is seldom realistic for long series (Henderson, Diggle and Dobson, 2000). Also, both the clinical context and visual inspection of longitudinal trajectories for individual patients, typical examples of which are shown in Figure 3.1, suggest that a random slope is too inflexible. Instead, we model $W_i(t)$ as the integral of a continuous-time random walk,

$$W_i(t) = \int_0^t B_i(v)dv, \tag{3.3}$$

where $B_i(v)$, the rate of change at time v, is Brownian motion. We set $B_i(0) = 0$ for every patient i, so that the random effects U_i represent each patient's deviation from their expectation at the time of their first eGFR measurement. The conditional distribution of $B_i(t)$ given $B_i(s)$ for some $s < t$ is Normal, with mean $B_i(s)$ and variance $\sigma^2(t-s)$. It follows that unconditionally, and using $[\cdot]$ to mean "the distribution of", $[B_i(t)] = \mathrm{N}(0, \sigma^2 t)$ and $\mathrm{Cov}(B_i(s), B_i(t)) = \sigma^2 \mathrm{min}(s,t)$. It then follows in turn that

$[W_i(t)] = \mathrm{N}(0, \sigma^2 t^3/3)$ and

$$\mathrm{Cov}(W_i(s), W_i(t)) = \sigma^2 \frac{\min(s,t)^2}{2}\left(\max(s,t) - \frac{\min(s,t)}{3}\right).$$

The process $B(t)$ is Markov, but $W_i(t)$ is not, i.e. in general $[W_i(t)|W_i(s), W_i(q)] \neq [W_i(t)|W_i(s)]$ for $q \leq s \leq t$. The bivariate process $(B_i(s), W_i(t))$ is bivariate Gaussian with zero means and cross-covariance structure

$$\mathrm{Cov}(B_i(s), W_i(t)) = \sigma^2 \frac{\min(s,t)^2}{2}, \tag{3.4}$$

and *is* Markov. For details, see Chapter 8 of Ross (1996) and Robinson (2010).

In what follows, we write the model equation (3.2) in the following condensed form,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{U}_i + \boldsymbol{W}_i + \boldsymbol{Z}_i. \tag{3.5}$$

Here, $\boldsymbol{Y}_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))^T$, $\boldsymbol{X}_i = (\boldsymbol{X}_i(t_{i1})^T, \ldots, \boldsymbol{X}_i(t_{in_i})^T)^T$ with $\boldsymbol{X}_i(t_{ij}) = (1, X_{i1}(t_{ij}), \ldots, X_{ip}(t_{ij}))$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_p)^T$, $\boldsymbol{U}_i = U_i\mathbf{K}_i$ where $\mathbf{K}_i$ denotes an $n_i \times 1$ matrix of ones, $\boldsymbol{W}_i = (W_i(t_{i1}), \ldots, W_i(t_{in_i}))^T$ and $\boldsymbol{Z}_i = (Z_i(t_{i1}), \ldots, Z_i(t_{in_i}))^T$.

3.4 Inference

3.4.1 Estimation

The distributional properties of $\boldsymbol{U}_i$, $\boldsymbol{B}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$ as defined in Section 3.3 induce a multivariate Normal distribution for $\boldsymbol{Y}_i$,

$$[\boldsymbol{Y}_i] = \mathrm{MVN}\left(\boldsymbol{X}_i\boldsymbol{\alpha}, \boldsymbol{V}_i(\boldsymbol{\phi})\right),$$

where $\boldsymbol{X}_i$ and $\boldsymbol{\alpha}$ are as before. Also,

$$\boldsymbol{V}_i(\boldsymbol{\phi}) = \omega^2\boldsymbol{J}_i + \sigma^2\boldsymbol{R}_i + \tau^2\boldsymbol{I}_i, \tag{3.6}$$

where $\boldsymbol{\phi} = (\omega^2, \sigma^2, \tau^2)^T$, $\boldsymbol{J}_i$ is an $n_i \times n_i$ matrix of ones, $\boldsymbol{R}_i$ is an $n_i \times n_i$ matrix with (j,k)th element

$$\frac{\min(t_{ij}, t_{ik})^2}{2}\left(\max(t_{ij}, t_{ik}) - \frac{\min(t_{ij}, t_{ik})}{3}\right),$$

and $\boldsymbol{I}_i$ is an $n_i \times n_i$ identity matrix.

We assume that repeated observations belonging to different patients are independent, i.e. $\text{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_i') = 0$ for $i \neq i'$. Then, the log-likelihood function can be written as

$$L(\boldsymbol{\theta}) = \text{Constant} - \frac{1}{2}\sum_{i=1}^{M} \log\left(\det\left(\boldsymbol{V}_i(\boldsymbol{\phi})\right)\right) - \frac{1}{2}\sum_{i=1}^{M} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}\right)^T \boldsymbol{V}_i(\boldsymbol{\phi})^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}\right), \tag{3.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\phi}^T)^T$ and "det" denotes determinant of a square matrix. We obtain the maximum likelihood estimates (MLE) of the parameters, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\phi}}$, by a Fisher-Scoring algorithm as described in Jennrich and Schluchter (1986). Let $\hat{\boldsymbol{\phi}}^m$ and $\hat{\boldsymbol{\alpha}}^m$ denote the values of $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ at the mth scoring step. Given $\hat{\boldsymbol{\phi}}^m$, set

$$\hat{\boldsymbol{\alpha}}^{m+1} = \left(\sum_{i=1}^{M} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\phi}}^m)\boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{M} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\phi}}^m)\boldsymbol{Y}_i\right). \tag{3.8}$$

and update $\boldsymbol{\phi}$ to

$$\hat{\boldsymbol{\phi}}^{m+1} = \hat{\boldsymbol{\phi}}^m + \boldsymbol{I}^{-1}_{\hat{\boldsymbol{\phi}}^m \hat{\boldsymbol{\phi}}^m} \boldsymbol{S}_{\hat{\boldsymbol{\phi}}^m}. \tag{3.9}$$

In (3.9), the (r, s)th element of $\boldsymbol{I}_{\hat{\boldsymbol{\phi}}^m \hat{\boldsymbol{\phi}}^m}$ is

$$\left\{\boldsymbol{I}_{\hat{\boldsymbol{\phi}}^m \hat{\boldsymbol{\phi}}^m}\right\}_{rs} = \frac{1}{2}\sum_{i=1}^{M} \text{trace}\left(\boldsymbol{V}_i(\hat{\boldsymbol{\phi}}^m)^{-1} \frac{\partial \boldsymbol{V}_i(\boldsymbol{\phi})}{\partial \phi_r} \boldsymbol{V}_i(\hat{\boldsymbol{\phi}}^m)^{-1} \frac{\partial \boldsymbol{V}_i(\boldsymbol{\phi})}{\partial \phi_s}\right), \quad r, s = 1, 2, 3, \tag{3.10}$$

and the rth element of $\boldsymbol{S}_{\hat{\boldsymbol{\phi}}^m}$ is

$$\left\{\boldsymbol{S}_{\hat{\boldsymbol{\phi}}^m}\right\}_r = \frac{1}{2}\sum_{i=1}^{M} \text{trace}\left(\boldsymbol{V}_i(\hat{\boldsymbol{\phi}}^m)^{-1} \left(\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}^{m+1}\right)\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}^{m+1}\right)^T - \boldsymbol{V}_i(\hat{\boldsymbol{\phi}}^m)\right) \boldsymbol{V}_i(\hat{\boldsymbol{\phi}}^m)^{-1} \frac{\partial \boldsymbol{V}_i(\boldsymbol{\phi})}{\partial \phi_r}\right), \quad r = 1, 2, 3. \tag{3.11}$$

In (3.10) and (3.11), the first partial derivatives of $\boldsymbol{V}_i(\boldsymbol{\phi})$ are calculated by $\frac{\partial \boldsymbol{V}_i(\boldsymbol{\phi})}{\partial \omega^2} = \boldsymbol{J}_i$, $\frac{\partial \boldsymbol{V}_i(\boldsymbol{\phi})}{\partial \sigma^2} = \boldsymbol{R}_i$ and $\frac{\partial \boldsymbol{V}_i(\boldsymbol{\phi})}{\partial \tau^2} = \boldsymbol{I}_i$. In our data analysis and simulations, we assess convergence by the criterion, $\sqrt{(\hat{\boldsymbol{\phi}}^m - \hat{\boldsymbol{\phi}}^{m+1})^T(\hat{\boldsymbol{\phi}}^m - \hat{\boldsymbol{\phi}}^{m+1})} < 10^{-10}$.

At convergence, the large sample variance-covariance matrices of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\phi}}$ can be obtained by

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{M} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\phi}})\boldsymbol{X}_i\right)^{-1} \tag{3.12}$$

and $\text{cov}(\hat{\boldsymbol{\phi}}) = \boldsymbol{I}^{-1}_{\hat{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}}$, respectively.

3.4.2 Prediction

The conditional distributions $[U_i|\boldsymbol{Y}_i, \boldsymbol{\theta}]$, $[W_i(t_{ik})|\boldsymbol{Y}_i^k, \boldsymbol{\theta}]$ and, especially, $[B_i(t_{ik})|\boldsymbol{Y}_i^k, \boldsymbol{\theta}]$, where $\boldsymbol{Y}_i^k = (Y_{i1}, \ldots, Y_{ik})^T$, are of scientific interest. The first two are relevant to an individual's prognosis, whilst the third is the predictive distribution of the underlying rate of change, which is the primary target in our application. The explicit forms of these distributions can be obtained using the properties of multivariate Normal distribution (Anderson, 1984) as

$$[U_i|\boldsymbol{Y}_i; \boldsymbol{\theta}] = \mathrm{N}\left(\omega^2 \boldsymbol{K}_i{}^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - X_i\boldsymbol{\alpha}\right), \omega^2\left(1 - \omega^2 \boldsymbol{K}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{K}_i\right)\right), \tag{3.13}$$

$$[W_i(t_{ik})|\boldsymbol{Y}_i^k; \boldsymbol{\theta}] = \mathrm{N}\left(\frac{\sigma^2}{2}\boldsymbol{F_i^k}^T \left(\boldsymbol{V}_i^k\right)^{-1}\left(\boldsymbol{Y}_i^k - \boldsymbol{X}_i^k\boldsymbol{\alpha}\right), \sigma^2\left\{\frac{t_{ik}^3}{3} - \frac{\sigma^2}{4}\boldsymbol{F_i^k}^T\left(\boldsymbol{V}_i^k\right)^{-1}\boldsymbol{F}_i^k\right\}\right), \tag{3.14}$$

$$[B_i(t_{ik})|\boldsymbol{Y}_i^k; \boldsymbol{\theta}] = \mathrm{N}\left(\frac{\sigma^2}{2}\boldsymbol{L_i^k}^T \left(\boldsymbol{V}_i^k\right)^{-1}\left(\boldsymbol{Y}_i^k - \boldsymbol{X}_i^k\boldsymbol{\alpha}\right), \sigma^2\left\{t_{ik} - \frac{\sigma^2}{4}\boldsymbol{L_i^k}^T\left(\boldsymbol{V}_i^k\right)^{-1}\boldsymbol{L}_i^k\right\}\right), \tag{3.15}$$

where $\boldsymbol{V}_i^k$ is the variance-covariance matrix of $\boldsymbol{Y}_i^k$, $\boldsymbol{K}_i$ is as before, $\boldsymbol{F}_i^k = \left(t_{i1}^2\left(t_{ik} - \frac{t_{i1}}{3}\right), \ldots, t_{ik}^2\left(t_{ik} - \frac{t_{ik}}{3}\right)\right)^T$ and $\boldsymbol{L}_i^k = (t_{i1}^2, \ldots, t_{ik}^2)^T$. Here, we suppress the dependence of $\boldsymbol{V}_i$ on $\boldsymbol{\phi}$. In our application, we substitute the maximum likelihood estimates of $\boldsymbol{\theta}$ into (3.13), (3.14) and (3.15), because estimation errors are negligible compared with prediction errors. In smaller samples, Bayesian prediction with diffuse priors would provide a convenient, albeit pragmatic, means of accommodating estimation error.

Similarly, we use properties of the multivariate Normal distribution to obtain the predictive distributions needed for forecasting W_i and B_i at time t_{ik} with lead-time u as

$$[W_i(t_{ik}+u)|\boldsymbol{Y}_i^k; \boldsymbol{\theta}] = \mathrm{N}\left(\frac{\sigma^2}{2}\boldsymbol{F}_i^{k,u^T}\left(\boldsymbol{V}_i^k\right)^{-1}\left(\boldsymbol{Y}_i^k - \boldsymbol{X}_i^k\boldsymbol{\alpha}\right), \sigma^2\left\{\frac{(t_{ik}+u)^3}{3} - \frac{\sigma^2}{4}\boldsymbol{F}_i^{k,u^T}\left(\boldsymbol{V}_i^k\right)^{-1}\boldsymbol{F}_i^{k,u}\right\}\right), \tag{3.16}$$

$$[B_i(t_{ik}+u)|\boldsymbol{Y}_i^k; \boldsymbol{\theta}] = \mathrm{N}\left(\frac{\sigma^2}{2}\boldsymbol{L}_i^{k,u^T}\left(\boldsymbol{V}_i^k\right)^{-1}\left(\boldsymbol{Y}_i^k - \boldsymbol{X}_i^k\boldsymbol{\alpha}\right), \sigma^2\left\{(t_{ik}+u) - \frac{\sigma^2}{4}\boldsymbol{L}_i^{k,u^T}\left(\boldsymbol{V}_i^k\right)^{-1}\boldsymbol{L}_i^{k,u}\right\}\right), \tag{3.17}$$

where $\boldsymbol{F}_i^{k,u} = \left(t_{i1}^2\left(t_{ik}+u-t_{i1}/3\right), \ldots, t_{ik}^2\left(t_{ik}+u-t_{ik}/3\right)\right)^T$. Note that the expectation of $[B_i(t_{ik}+u)|\boldsymbol{Y}_i^k; \boldsymbol{\theta}]$ is the same for all u, but, its variance increases with u. Whenever a new response $Y_{i,k+1}$ becomes available, at time $t_{ik}+u$ say, we update the predictions accordingly.

3.5 Application: SRFT data set

Using log(eGFR) rather than eGFR as our response variable is better-matched to the clinical criterion for referral to specialist secondary care and improves the empirical fit of our data to the linear model. On the log-transformed scale eGFR is equivalent to SCr adjusted for sex and age. Nevertheless, we include sex and age as explanatory variables because standard formulae such as (3.1) are not optimal for prediction in particular sub-populations; see, for example, page 606 of Levey *and others* (2009). It follows that our predictive inferences are unaffected by whether we use log(eGFR) or log(SCr) as the response variable. Our main reason for working with eGFR rather than directly with SCr is that this is more easily interpretable by renal physicians.

We decompose age into age at entry and follow-up time in order to separate cross-sectional and longitudinal effects of age. This decomposition is strongly supported by a likelihood ratio criterion: for our final model, the maximised log-likelihood is $312,251.81$ when age at measurement-time is included as a single explanatory variable, $312,425.83$ when the cross-sectional and longitudinal effects of age are separated. Based on our preliminary analyses we use a piece-wise linear model for the longitudinal age effect with a change of slope at age 56.5 years. We code sex as 0 for males, 1 for females. The explicit form of our linear model for a single patient at the jth measurement is therefore

$$\log(\text{eGFR})_{ij} = \alpha_0 + \alpha_1 \text{ Gender}_i + \alpha_2 \text{ Baseline age}_i + \alpha_3 \text{ t}_{ij} + \alpha_4 \text{ max(0, age - 56.5)}_{ij} + U_i + W_i(t_{ij}) + Z(t_{ij}), \tag{3.18}$$

and the corresponding rate of change in kidney function is

$$\alpha_3 + \alpha_4 \text{I(age} > 56.5) + B_i(t_{ij}). \tag{3.19}$$

3.5.1 Estimation

Table 3.1 shows the maximum likelihood estimates and standard errors for the regression parameters; all are (unequivocally) significantly different from zero. The estimate $\hat{\alpha}_0 = 4.6006$ establishes the average level of kidney function on entry, whilst $\hat{\alpha}_1 = -0.08768$ indicates that at study entry, females had 8.4% ($= 100*(\exp(-0.0877)-1)$) worse kidney function on average than males. The estimate $\hat{\alpha}_2 = -0.0048$ indicates that the cross-sectional effect of age is a loss of kidney function at a rate of approximately 0.5% per year of age at entry; this is more or less in line with the general population, as would be expected. The estimate $\hat{\alpha}_3 = -0.0232$, corresponds to an average loss of kidney function

TABLE 3.1: Maximum likelihood estimates of the model parameters and the corresponding standard errors (SE).

Parameter	Estimate	SE
α_0	4.6006	0.0203
α_1	-0.0877	0.0048
α_2	-0.0048	0.0004
α_3	-0.0232	0.0011
α_4	-0.0075	0.0006
ω^2	0.1111	0.0012
σ^2	0.0141	0.0002
τ^2	0.0469	0.0001

at a rate of 2.3% per year until and including age 56.5. This value, together with the estimate $\hat{\alpha}_4 = -0.0075$ indicates that the loss of kidney function accelerates to 2.9% after age 56.5. These estimates of the loss of kidney function reflect the fact that all of the patients have been diagnosed with pre-disposing conditions for renal failure. Maximum likelihood estimates of the covariance parameters are $\hat{\omega}^2 = 0.1111$, $\hat{\sigma}^2 = 0.0141$ and $\hat{\tau}^2 = 0.0469$.

3.5.2 Diagnostics

We apply diagnostic checks on our fitted model, using empirical residuals calculated as follows. Firstly, for each subject i let $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}$. Now, denote by $\hat{\boldsymbol{V}}_i$ the fitted variance matrix of $\boldsymbol{Y}_i$, obtained by substituting $\hat{\boldsymbol{\phi}}$ into (3.6), and write $\hat{\boldsymbol{V}}_i = \boldsymbol{S}_i\boldsymbol{S}_i^T$, where $\boldsymbol{S}_i$ is lower triangular. Finally, define the transformed (or standardised) empirical residual vector for patient i as $\boldsymbol{r}_i^* = \boldsymbol{S}_i^{-1}\boldsymbol{r}_i$.

To check the assumed form of the regression model, we inspect scatterplots of the residuals against the fitted values $\hat{Y}_{ij} = \boldsymbol{X}_{ij}\hat{\boldsymbol{\alpha}}$, and against the follow-up times, t_{ij}. Figure 3.2 shows the two scatterplots with superimposed LOWESS curves (Cleveland, 1979) obtained using the `R` function `lowess` with the default value for the smoothing parameter. There is no discernible systematic pattern in either of the scatterplots and the fitted smooth curves are close to zero, suggesting a reasonable fit.

To check the assumed form of the covariance structure, we use the variogram of the empirical residuals (Diggle *and others*, 2002). The theoretical variogram is the function $\gamma(u)$, where $\gamma(u_{ijk}) = \frac{1}{2}E(r_{ij}^* - r_{ik}^*)^2$ and $u_{ijk} = |t_{ij} - t_{ik}|$. The empirical variogram is obtained by calculating empirical variogram ordinates, $g_{ijk} = \frac{1}{2}(r_{ij}^* - r_{ik}^*)^2$ and averaging all g_{ijk} corresponding to each unique value of u_{ijk} or, if follow-up times are completely

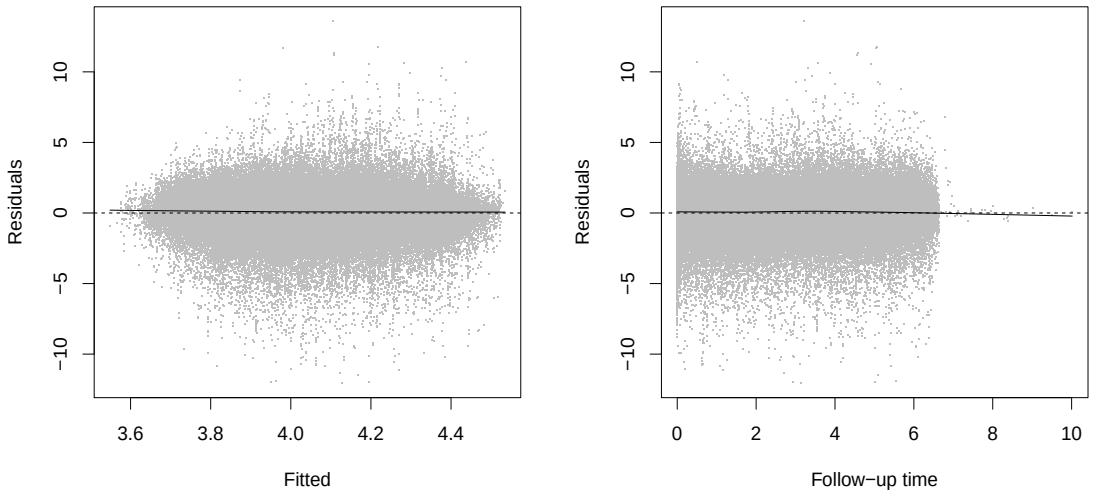

FIGURE 3.2: Left panel: Scatterplot of fitted values versus standardised residuals. Right panel: Scatterplot of follow-up time (in years) versus standardised residuals. The dashed line is the zero line, the solid line a LOWESS smooth.

irregular, by averaging all g_{ijk} corresponding to values of u_{ijk} within a pre-specified set of intervals. In a well-fitting model, the empirical variogram of the standardised residuals should fluctuate randomly around 1 when drawn against the lag, $u^*_{ijk} = |t^*_{ij} - t^*_{ik}|$, where $t^*_{i.}$ are elements of $\boldsymbol{t}^*_i = \boldsymbol{S}_i^{-1}\boldsymbol{t}_i$ (Fitzmaurice, Laird and Ware, 2011). Figure 3.3 shows the empirical variogram for the fitted model. The empirical variogram ordinates show a decreasing trend from approximately 1.5 to 1 over the range 0 to 2 of differences of transformed times, and thereafter fluctuate around 1.

We further assess the appropriateness of the assumed variance structure by comparing the variances of the (unstandardised) empirical residuals r_{ij} and the theoretical variance implied by our model, $\text{var}(Y_{ij}) = \omega^2 + \sigma^2 t_{ij}^3/3 + \tau^2$, plugging-in the estimates $\hat{\omega}^2$, $\hat{\sigma}^2$ and $\hat{\tau}^2$. The empirical and theoretical variances are drawn in Figure 3.4. The empirical variances were calculated from binned residuals through time, with bin size of one week, but with baseline data treated separately, i.e. variances are calculated separately at baseline and over follow-up measurements between 0+ and 7 days after baseline. The empirical variance of the baseline residuals is substantially smaller than the variances of the rest of the residuals, whose variance increases with time but at a substantially slower rate than the theoretical variance of the fitted model.

We also inspect the distributional assumptions of our model by examining the standardised residuals which are expected to follow standard Normal distribution. Figure 3.5 shows four diagnostic plots. Collectively, these indicate that the residual distribution

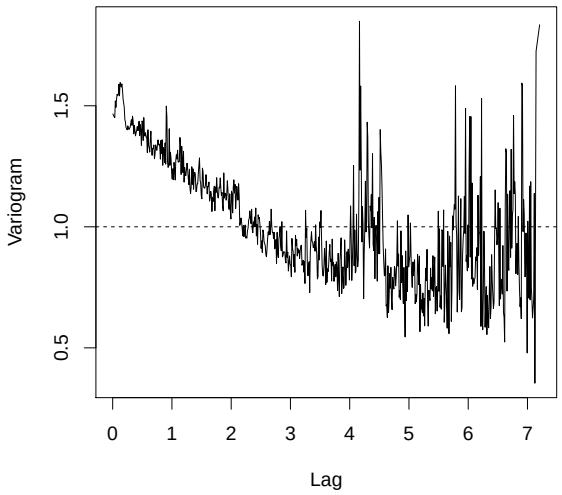

FIGURE 3.3: The empirical variogram based on the transformed residuals against the lag based on the transformed time-scale. The variogram ordinates are averaged over bins with width 0.01. Bins with fewer than 30 residuals are omitted.

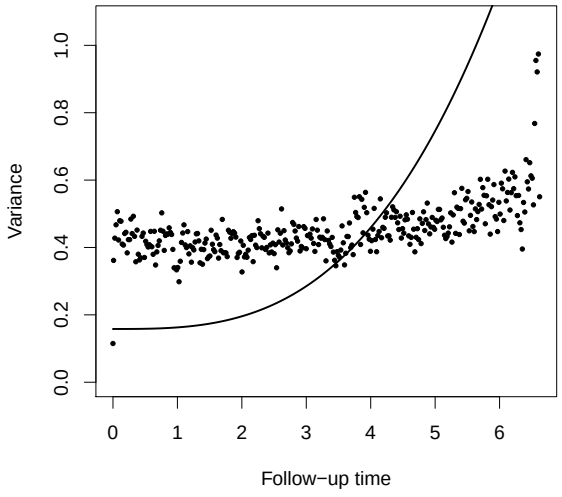

FIGURE 3.4: The variances of the raw residuals over follow-up time, in years, (dots) and the theoretical variance function of the fitted model (solid line). Residuals are binned through time with bin size of one week and bins with less than 30 elements are omitted. Baseline data are treated separately.

has the expected properties in the main body of the distribution, but heavier tails than the standard Normal distribution.

Overall, the diagnostic plots show discrepancies between the data and fitted model, whose influence on predictive performance we investigate in Section 3.6.

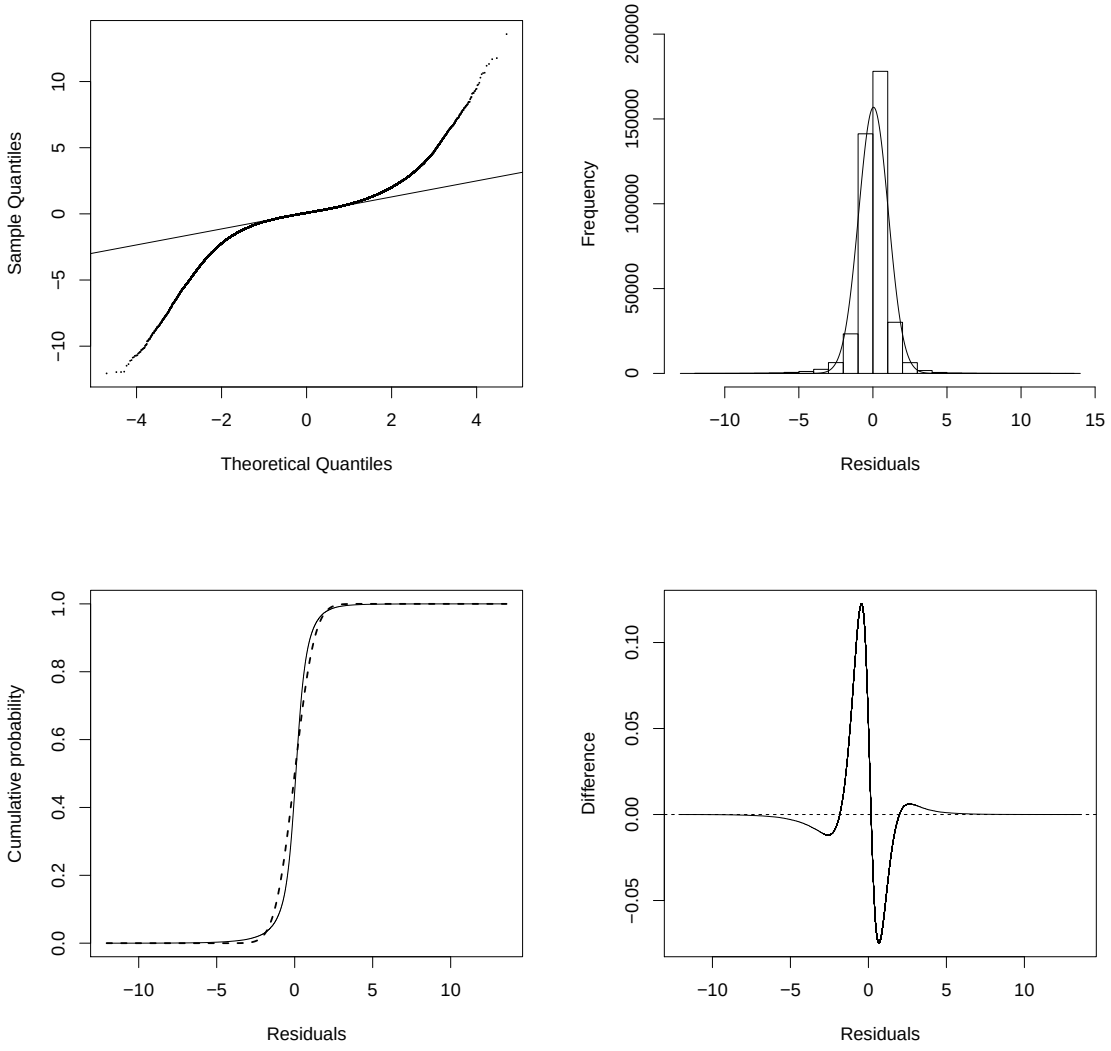

FIGURE 3.5: Diagnostics plots on distributional assumptions based on standardised residuals. Upper left panel: quantile-quantile plot. Upper right panel: histogram with Normal density superimposed. Lower left panel: empirical (solid line) and theoretical (dashed line) cumulative distribution functions. Lower right panel: the difference between the empirical and theoretical distribution functions.

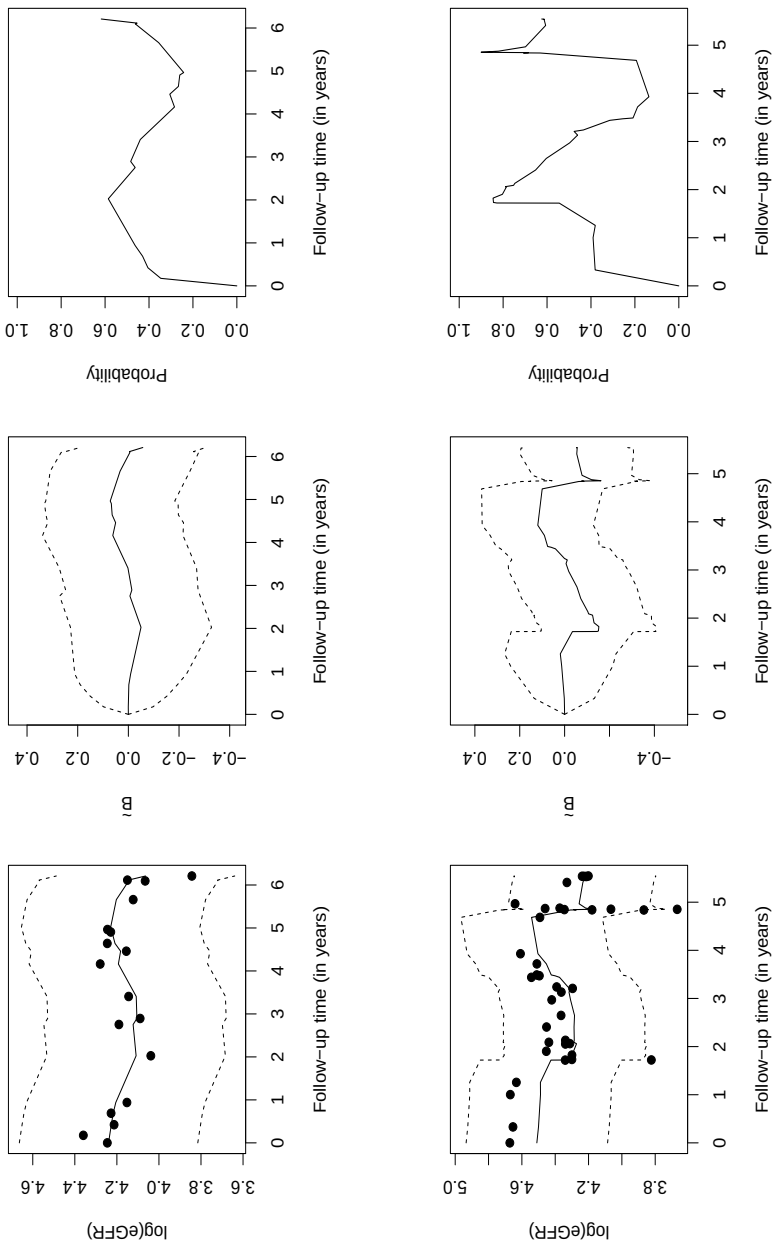

FIGURE 3.6: Plots of the predictions for two selected patients. Rows 1 and 2 correspond to patients $i = 100$ and $i = 9000$, respectively. Column 1 shows observed values of log(eGFR) (solid dots), predictive means (solid lines) and predictive 2.5% and 97.5% predictive quantiles (dashed lines). Column 2 shows predictive means (solid lines) and 2.5% and 97.5% predictive quantiles (dashed lines) of the underlying rate of change in log(eGFR). Column 3 shows the predictive probabilities, $p_i^*(t)$ that the underlying rate of change is less than -0.05.

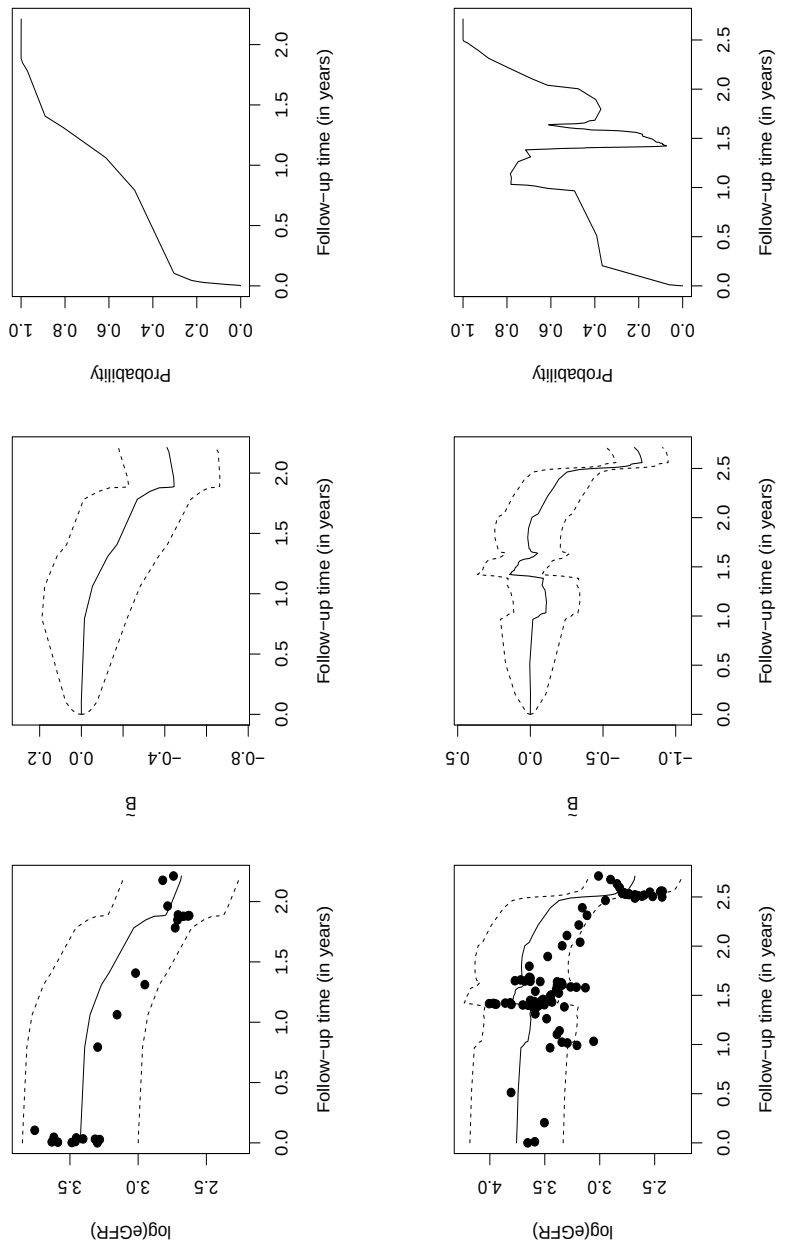

FIGURE 3.7: Plots of the predictions for two more selected patients, $i = 9600$ (row 1) and $i = 1278800$ (row 2). Details as for Figure 3.6.

3.5.3 Prediction

Figures 3.6 and 3.7 show predictions for four selected patients. Point predictions and prediction intervals for log(eGFR) were calculated from the mean, 2.5% and 97.5% quantiles of the conditional distribution, $[Y_{ij}|\boldsymbol{X}_{ij}, \hat{\boldsymbol{\alpha}}, \tilde{U}_i, \tilde{W}_i(t_{ij}), \hat{\tau}^2] = \mathrm{N}(\boldsymbol{X}_{ij}\hat{\boldsymbol{\alpha}} + \tilde{U}_i + \tilde{W}_i(t_{ij}), \hat{\tau}^2)$, where $\tilde{U}_i$ and $\tilde{W}_i(t_{ij})$ correspond to the mean of $[U_i|\boldsymbol{Y}_i; \boldsymbol{\theta}]$ and $[W_i(t_{ik})|\boldsymbol{Y}_i^k; \boldsymbol{\theta}]$, given in (3.13) and (3.14), respectively. Point and interval predictions for the rate of change were similarly calculated from the conditional distribution $[B_i(t_{ik})|\boldsymbol{Y}_i^k; \boldsymbol{\theta}]$ given in (3.15). The predictive probability, at each follow-up time, for the underlying rate of change being less than -0.05 was calculated as

$$p_i^*(t_{ik}) = P(B_i(t_{ik}) < -0.05 - \hat{\alpha}_3 - \hat{\alpha}_4 \mathrm{I}(\mathrm{age}_{ik} > 56.5)|\boldsymbol{Y}_i^k; \boldsymbol{\theta}). \tag{3.20}$$

The individual predictions for log(eGFR) seem very reasonable. The predicted means smooth out the erratic fluctuations in measured log(eGFR) and almost all the observed log(eGFR) measurements are covered by the 95% confidence intervals.

The predictive probability graphs of $p_i^*(t)$ as defined by (3.20) are shown in the third column of Figures 3.6 and 3.7. Our view is that these should guide, rather than determine, clinical decision-making. From this point of view, the four selected patients show interestingly different patterns.

1. For patient $i = 100$, the predictive probability $p_i^*(t)$ rises to a value slightly greater than 0.5 after two years of follow-up, and thereafter fluctuates between about 0.3 and 0.6. The appropriate clinical response would likely depend on factors other than those that can be encoded in a statistical model, for example the patient's general frailty and any co-morbidities.

2. For patient $i = 9000$, $p_i^*(t)$ rises sharply to approximately 0.8 after two years, drops equally sharply between two and four years, then rises again. This is not atypical of patients experiencing progression towards renal failure. Indeed, *this patient may well have received treatment within their primary care setting to reverse an acute loss of kidney function.* This pattern is one example of something that the model cannot be expected to capture, but which does not necessarily negate its value as a predictive tool.

3. For patient $i = 9600$, $p_i^*(t)$ rises inexorably from 0 to 1 during the first two years of follow-up. For this patient, referral to secondary care is clearly indicated.

TABLE 3.2: Results of the simulation study 1. Columns give the parameter name (Parameter), the mean (Mean), percentage bias (Bias (%)) and standard deviation (SD) of the parameter estimates, the mean of the nominal standard errors according to standard likelihood asymptotic theory (meSE), and the percentage coverage of the corresponding approximate 95% confidence intervals (CP%).

Parameter	Value	Mean	Bias (%)	SD	meSE	CP(%)
α_0	4.6006	4.5874	-0.2869	0.1458	0.1379	94.0
α_1	-0.0877	-0.0842	-4.0437	0.0323	0.0323	94.0
α_2	-0.0048	-0.0046	-4.5885	0.0029	0.0027	94.0
α_3	-0.0232	-0.0227	-1.9979	0.0071	0.0072	95.6
α_4	-0.0075	-0.0078	3.3830	0.0041	0.0038	93.4
ω^2	0.1111	0.1103	-0.6764	0.0078	0.0079	95.4
σ^2	0.0141	0.0141	-0.0806	0.0010	0.0010	94.2
τ^2	0.0469	0.0469	-0.0120	0.0008	0.0008	95.8

4. For patient $i = 1278800$, the progress of $p_i^*(t)$ is qualitatively similar to that for patient $i = 113$, but on a shorter time-scale and with an unequivocal indication of referral by 2.5 years.

3.6 Simulations

3.6.1 Simulation study I

We first conducted a simulation study to investigate the properties of the parameter estimates under the model given by (3.18). For each simulation, the values of the explanatory variables were those of a random sample of 500 from the 22,910 patients in the SRFT data. The random effects, U_i, $W_i(t_{ij})$ and Z_{ij} were then simulated from their assumed distributions as described in Section 3.3. The true parameter values for the simulations were the estimated values from our analysis of the SRFT data, as reported in Table 3.1. The simulation was replicated 500 times. The mean and standard deviation of the total number of repeated measurements in the simulated data-sets were 8,599 and 410, the variation being a consequence of the variation in the number of repeated measurements per patient in the SRFT data. Table 3.2 summarises the results. The parameter estimates are approximately unbiased. The empirical standard deviations of the parameter estimates are close to the mean of the nominal asymptotic standard errors. Coverage is close to the nominal rate of 95%.

TABLE 3.3: Results of the simulation study 2. Columns give the mean (Mean) and standard deviation (SD) of the area under the ROC curve, calculated form 500 replicate simulations for each cases 1, 2 and 3.

	Mean	SD
Case 1	0.7574	0.0186
Case 2 (variance structure misspecification)	0.7373	0.0278
Case 3 (heavy-tailed residuals)	0.7558	0.0186

3.6.2 Simulation study II

We conducted a second simulation study to examine the robustness of the predictive probabilities $p_i^*(t)$ to the covariance structure misspecification as depicted in Figures 3.3 and 3.4, and to the heavy-tailed residual distribution as depicted in Figure 3.5.

For each simulation, the values of the explanatory variables were those of a random sample of 500 from the 22,910 patients in the SRFT data. We considered three data-generating mechanisms as follows. In the first case, we generated data according to (3.18) as for simulation study I, again setting the parameter values as the estimates reported in Table 3.1. In the second case, we added a random effect, $U_i^* \mathrm{I}(t > 0)$, where $U_i^* \sim \mathrm{N}(0, 0.25)$ and is independent of U_i. The value 0.25 for the variance of U_i^* was based on the empirical variances displayed in Figure 3.4, whilst the assumed independence of U_i and U_i^* was based on a preliminary analysis of the SRFT data set using a linear mixed model with U_i and U_i^* as the random effects. In the third case, we reverted to the model given by (3.18), but with the Z_{ij} simulated from a t distribution with degrees of freedom 3, scaled by $\tau = \sqrt{0.0469}$. In all three cases, we simulated $B_i(t_{ij})$ together with $W_i(t_{ij})$ as a bivariate process whose properties are as described in Section 3.3.

Properties of the predictive probabilities, $p_i^*(t)$, for the underlying rate of change being less than -0.05 are summarised by their area under the receiver operating characteristics (ROC) curve. Area under the ROC curve is calculated as time-invariant, since the random effects are simulated from their marginal distribution. The simulations are replicated 500 times for each case. The mean numbers of repeated measurements, over 500 replicate simulations, were 8574, 8533 and 8582 in cases 1, 2 and 3, with standard deviations 448, 398 and 409, respectively. Table 3.3 shows the mean and standard deviation of the area under the ROC curve in each case. The influence of either covariance structure misspecification or heavy-tailed residuals is negligible.

3.7 Computational aspects

All computations were programmed in R and run on a PC with Windows 7 32bit, 4.00 GB RAM and 3.00GHz processor. We have written an R package, `lmenssp`, available at http://CRAN.R-project.org/package=lmenssp, that implements parameter estimation and plug-in prediction for a range of non-stationary Gaussian process models. The supplementary material, available at *Biostatistics* online includes the SRFT dataset and gives an exemplary R script for data analysis reported in Section 3.5. The computational time required for the estimation of the parameters was 60 minutes for the SRFT data set, 34 seconds for a simulated data set with 500 patients and 8,462 repeated measurements. The predictions require less computational time, since we estimate the model parameters once and plug-in these estimates into the predictive distributions for each patient separately. For example, computational times required for predictions were 0.9, 1.0, 2.8 and 8.8 seconds for patients with 10, 100, 203 and 305 observations, respectively.

3.8 Discussion

We have used a large set of longitudinal clinical data to develop a statistical model for real-time monitoring of progression towards end-stage renal failure. Our specific objective was to provide predictive probabilities for the event that the underlying rate of change in a patient's kidney function is less than -0.05, i.e. a loss of at least 5% of kidney function per year. The value -0.05 is taken from current UK guidelines for referral of a primary care patient to specialist secondary care. Our model is a linear mixed effects model in which between and within patient heterogeneities are captured by a random intercept and integrated Brownian motion, respectively.

We found discrepancies between the assumed and empirical distribution and variance structure of the residuals. However, simulations showed that the impact of these discrepancies on the predictive performance of the model is negligible. The finding regarding the influence of heavy-tailedness is in agreement with the results of Sweeting and Thompson (2012). We considered capturing the behaviour of the empirical variance depicted in Figure 3.4 by adding a random effect to (3.18) for post-baseline measurements, i.e. $U_i^* \mathrm{I}(t_{ij} > 0)$ where $U^* \sim \mathrm{N}(0, \omega^{*2})$, but this resulted in a non-identifiable model. We also considered modelling only post-baseline data, differences from baseline and differences between successive observations. However, none of these gave any improvement in diagnostic performance. Similarly, inclusion of the baseline co-morbidity variables did not improve the diagnostics. As an alternative to the integrated random walk specification for the serially correlated random effects, we considered specifying $W_i(t_{ij})$ in (3.18) as

an integrated Ornstein-Uhlenbeck process (Taylor, Cumberland and Sy, 1994). For this specification the underlying rate of change follows a stationary Gaussian process with exponential correlation function. However, the resulting fit suggested that the rate of change in kidney function behaves as white noise, which is biologically implausible.

The discussion of Figures 3.6 and 3.7 in Section 3.5.3 makes the general point that the evolution of kidney function in individual patients shows features that are unlikely to be captured by any statistical model. These features arise for a number of reasons, including the imperfection of eGFR as a measure of true kidney function (for example, the underlying serum creatinine assay is affected by changes in muscle mass) and the transient effects of minor acute illnesses that go unrecorded. For observational time series of this kind, fit to the data is less important than the ability to address the primary research question which is the probability that a patient is losing kidney function at a rate of 5% or more per year.

Our current algorithm for prediction requires inversion of matrices of dimension n_i by n_i, where n_i is the current number of eGFR measurements available for the ith patient. In principle, it would be computationally more efficient to use a Kalman filter algorithm (Kalman, 1960) to update predictive probabilities based on the most recent results, rather than re-calculating from scratch whenever new eGFR measurements are added to the data. By exploiting the result that the bivariate process $(B_i(t), W_i(t))$ is Markov, we can represent our model as a local linear trend model (Robinson, 2010; Durbin and Koopman, 2012). However, this formulation assumes that the rate of change is constant between successive measurements, which introduces an element of approximation. In our data, the maximum value of n_i is 305, most are much smaller and the associated exact computations are not burdensome. Similarly, in principle we might want to allow for parameter uncertainty when calculating the predictive probabilities, for example by assigning Bayesian priors to the model parameters, $\boldsymbol{\theta}$ say, and replacing the plug-in predictive distribution, $[B_i(t_{ik})|\boldsymbol{Y}_i; \hat{\boldsymbol{\theta}}]$, by $\int [B_i(t_{ik})|\boldsymbol{Y}_i; \boldsymbol{\theta}][\boldsymbol{\theta}|\boldsymbol{Y}_i]d\boldsymbol{\theta}$. However, in our application the difference between the two is negligible as the prediction error in $B_i(t)$ dominates the estimation error in $\boldsymbol{\theta}$.

3.9 Online supplementary material: R codes

```
## R script to apply to SRFT data set the methodology used in the paper

# loading the package
R> library(lmenssp)
```

```
# reading the data set into R
R> srft.data <- read.csv("srft.data.csv", header = T)

# obtaining the parameter estimates
R> fit.ibm <- lmenssp(log(egfr) ~ sex + bage + fu + pwl,
+                     data = srft.data, id = srft.data$id,
+                     process = "ibm", timeVar = srft.data$fu,
+                     tol = 1e-10, maxiter = 100, silent = FALSE)
R> fit.ibm

# obtaining the predictive distributions of [U_i|Y_i],
# [W_i(t_{ik})|Y_i^k] and [B_i(t_{ik})|Y_i^k]
# for patients with IDs = 100, 9000, 9600 and 1278800, respectively

R> subj.id <- c(100, 9000, 9600, 1278800)[1]

R> filter <- filtered(log(egfr) ~ sex + bage + fu + pwl,
+                     data = srft.data, id = srft.data$id,
+                     process = "ibm", timeVar = srft.data$fu,
+                     estimate = fit.ibm$estimate[, 1],
+                     subj.id = subj.id)
R> filter

# drawing Figures 3.6 & 3.7

R> Xi   <- as.matrix(cbind(1, srft.data[srft.data$id == subj.id,
+                           c("sex", "bage", "fu", "pwl")]))
R> Yi    <- log(as.matrix(srft.data[srft.data$id == subj.id, c("egfr")]))
R> timei <- srft.data[srft.data$id == subj.id, "fu"]
R> pwli  <- srft.data[srft.data$id == subj.id, "pwl"]

R> fitted.mean <- Xi %*% fit.ibm$estimate[1 : 5, 1, drop = F] +
+                 matrix(rep(filter$u[1, "mean"], nrow(Xi))) +
+                 matrix(filter$w[, "mean"])
R> fitted.var  <- fit.ibm$estimate[8, 1]

R> threshold   <- -0.05 - fit.ibm$estimate[4, 1] -
+                 fit.ibm$estimate[5, 1] * as.numeric(I(pwli > 0))

R> par(mfrow = c(1, 3))

R> pllim <- fitted.mean - 1.96 * sqrt(fitted.var)
R> pulim <- fitted.mean + 1.96 * sqrt(fitted.var)
```

```
R> plot(timei, Yi, type="p", ylab="log(eGFR)",
+       xlab = "Follow-up time (in years)",
+       ylim = c(min(min(Yi), min(pllim)) - 0.01,
+                max(max(Yi), max(pulim)) + 0.01),
+       pch = 19, col = "black")
R> points(timei, fitted.mean, col = "black", type = "l")
R> points(timei, pllim, col = "black", type = "l", lty = 2)
R> points(timei, pulim, col = "black", type = "l", lty = 2)

R> plot(timei, filter$b[, "mean"], ylab = expression(tilde(B)),
+       xlab = "Follow-up time (in years)",
+       ylim = c(min(qnorm(0.025, mean = filter$b[, "mean"],
+                          sd = sqrt(filter$b[, "variance"]),
+                          lower.tail = TRUE)) - 0.1,
+                max(qnorm(0.975, mean = filter$b[, "mean"],
+                          sd = sqrt(filter$b[, "variance"]),
+                          lower.tail = TRUE)) + 0.1),
+       type = "l")
R> points(timei, qnorm(0.025, mean = filter$b[, "mean"],
+                      sd = sqrt(filter$b[, "variance"]),
+                      lower.tail = TRUE),
+       type = "l", lty = 2, col = "black")
R> points(timei, qnorm(0.975, mean = filter$b[, "mean"],
+                      sd = sqrt(filter$b[, "variance"]),
+                      lower.tail = TRUE),
+       type = "l", lty = 2, col = "black")

R> plot(timei, pnorm(threshold, mean = filter$b[, "mean"],
+                    sd = sqrt(filter$b[, "variance"])), ylim=c(0,1),
+       ylab = "Probability", xlab = "Follow-up time (in years)",
+       pch = 20, type = "l")

# calculating the probabilities of meeting the clinical guideline
# of losing renal function
# at a rate of at least 5% per year
R> pnorm((threshold), filter$b[, "mean"], sqrt(filter$b[, "variance"]))
```

Bibliography

Anderson T. W. (1984). *An introduction to multivariate statistical analysis*, 2nd edition. New York: John Wiley & Sons.

Cleveland W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Diggle P. J., Heagerty P., Liang K.Y. and Zeger S. L. (2002). *Analysis of longitudinal data*, 2nd edition. Oxford: Oxford University Press.

Durbin J. and Koopman S. J. (2012). *Time series analysis by state space methods*, 2nd edition. Oxford: Oxford University Press.

Fitzmaurice G. M., Laird N. M. and Ware J. H. (2011). *Applied longitudinal analysis*, 2nd edition. New Jersey: John Wiley & Sons.

Henderson R., Diggle P. J. and Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Jennrich R. I. and Schluchter M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

Kalman R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Levey A. S., Bosch J. P., Lewis J. B., Greene T., Rogers N. and Roth D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of Internal Medicine* **130**, 461–470.

Levey A. S., Stevens L. A., Schmid C. H., Zhang Y. L., Castro A. F., Feldman H. I., Kusek J. W., Eggers P., Lente F. V., Greene T. and Coresh J. for the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* **150**, 604–612.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, URL http://www.R-project.org/.

Robinson G. K. (2010). Continuous time Brownian motion models for analysis of sequential data. *Journal of the Royal Statistical Society - Applied Statistics* **59**, 477–494.

Ross S. M. (1996). *Stochastic Processes*, 2nd edition. New Jersey: John Wiley & Sons.

Sweeting M. J. and Thompson S. G. (2012). Making predictions from complex longitudinal data, with application to planning monitoring intervals in a national screening programme. *Journal of the Royal Statistical Society - Statistics in Society* **175**, 569–586.

Taylor J. M. G., Cumberland W. G. and Sy J. P. (1994). A stochastic process model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**, 727–736.

Chapter 4

Acute kidney injury amongst chronic kidney disease patients: a case-study in statistical modelling

This chapter is based on the following paper:

Asar Ö., Ritchie J., Kalra P. A. and Diggle P. J. (2015). Acute kidney injury amongst chronic kidney disease patients: a case-study in statistical modelling. Submitted to *Statistics in Medicine*.

Abstract

Chronic kidney disease (CKD) and acute kidney injury (AKI) are two important kidney related health problems. The former is defined as enduring kidney damage or decreased kidney function, the latter as a sudden fall in the kidneys' excretion function. These two conditions are now recognised to be strongly associated. In this study, we investigate the influence of AKI on long-term kidney function amongst patients already diagnosed with CKD. We develop a longitudinal statistical model with Matérn correlation structure and multivariate t-distributed stochastic components. We also specify three change-points that describe the typical trajectory of an AKI event. We use maximum likelihood estimation implemented with an expectation-maximisation algorithm to estimate the model parameters, and best linear unbiased prediction to predict the random effects. Our case-study uses data from an on-going cohort study of chronic kidney patients at Salford Royal NHS Foundation Trust, Manchester, UK. A simulation study is conducted to investigate the properties of the estimators.

Key words: Dynamic modelling, longitudinal data analysis, mixed-effects models, renal medicine, robust distributions, stochastic modelling.

4.1 Introduction

Chronic kidney disease (CKD) covers a wide range of disorders that affect the function and structure of the kidneys (El Nahas and Levin, 2009; Levey and Coresh, 2012). It is defined as kidney damage and/or decreased kidney function that continues for at least three months. CKD can be asymptomatic for many years and is usually detected during the assessment of co-morbid conditions. Most cases are characterised by a gradual decline in kidney function, which might take decades or a small minority of patients might not progress at all, whilst aggressive cases can lead to kidney failure within months. Common causes of CKD are cardiovascular disease and diabetes mellitus. It has now been recognised as an important public health problem, with a worldwide prevalence of 8-10%, whilst prevalence in high-risk sub-populations can exceed 50% (Jha *et al.*, 2013; Eckardt *et al.*, 2013).

Acute kidney injury (AKI) is defined as a sudden fall in kidneys' excretion function, over a period of hours or, at most, a few days. It represents an acute response to any of a number of disorders that affect the kidneys (Bellomo, Kellum and Ronco, 2012). AKI

is common, and potentially catastrophic, in hospitalised and critically-ill patients. For example, Finlay *et al.* (2013) reports that the prevalence of AKI amongst patients who were admitted to acute medical units was 18%. Its severity ranges from mild renal impairment to complete renal failure and death. Even early-stage and short-term episodes of AKI need to be treated carefully to minimise the risk of severe complications (Chawla *et al.*, 2014). AKI is associated with high mortality and long-term complications (Kerr *et al.*, 2014). For example, Bellomo, Kellum and Ronco (2012) reports that mortality rates due to AKI amongst criticaly-ill patients were 53% and 44.7% in two different clinical trials. Although selected cases of AKI, e.g. vasculitis, have specific therapeutic treatment options, in most cases intervention is limited to identification and treatment of the underlying cause.

CKD and AKI are highly associated syndromes: an AKI episode can lead to the onset of CKD or worsen the severity of existing CKD (Chawla and Kimmel, 2012; Eckardt *et al.*, 2013; Lameire *et al.*, 2013; Chawla *et al.*, 2014). Conversely, CKD confers an approximately ten-fold increase in the risk for AKI (Chawla *et al.*, 2014). Nevertheless, the full implications of AKI for long-term kidney function remain an open research area (Lameire *et al.*, 2013; Chawla *et al.*, 2014).

In this study, we investigate the influence of AKI on the long-term evolution of kidney function amongst CKD patients. Our data are taken from the Chronic Renal Insufficiency Standards Implementation Study (CRISIS; Eddington *et al.*, 2010; Hoefield *et al.*, 2010). AKI events are declared retrospectively by inspection of repeated serum creatinine (SCr) measurements, according to the guidelines given by the Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group guideline (KDIGO, 2012). We develop a clinically-informed, formal description of an AKI episode that includes three change-points in the longitudinal evolution of a patient's kidney function. A gap of greater than three days between blood tests is used to define a new AKI episode.

We develop a linear mixed effects model for a patient's repeated measurements that combines fixed effects, a random intercept, a stationary stochastic process with Matérn covariance structure (Matérn, 1960) and mutually independent random measurement error. Diggle (1988) proposed a model of this kind under Gaussian distributional assumptions and powered correlation family. We extend this by allowing stochastic components of the model to follow a multivariate t distribution, because diagnostic tests of the Gaussian model gave clear evidence of long-tailed behaviour in the CRISIS data set, and Matérn correlation family is wider than the powered correlation family. We estimate model parameters by maximum likelihood (ML), using an expectation-maximisation (EM) algorithm (Dempster, Laird and Rubin, 1977) that exploits the normal-gamma

hierarchical representation of the t distribution. We use exploratory methods to identify change-points in the longitudinal profile during an AKI episode. All computations are carried out in R (R Development Core Team, 2014) by exploiting the `lmenssp` package (Asar and Diggle, 2014).

The paper is organised as follows. In Section 4.2, we give details of the data set, the clinical definition of an AKI episode and our algorithm for identifying each AKI episode within the CRISIS data set. In Section 4.3, we describe the formulation of our statistical model and the associated inferential methods. In Section 4.4, we present the results from our analysis of the CRISIS data set, and Section 4.5 provides diagnostic results. We conduct a simulation study in Section 4.6 and close the paper with a concluding discussion. Whilst `R` scripts are provided in Section 4.8, a sensitivity analysis on censoring longitudinal trajectories of the patients with no AKI events can be found in the appendix.

4.2 Data, definition of AKI and the change-points

4.2.1 CRISIS cohort

Our data are obtained from the Chronic Renal Insufficiency Standards Implementation Study (CRISIS) run by the Salford Royal NHS Foundation Trust (SRFT). CRISIS is an on-going observational study of all-cause CKD. In the UK, renal services are based on a hub and spoke system with a single main centre supporting care at many smaller local trusts. For patients in the CRISIS study the main centre is SRFT, with six district general hospitals. Measurement of a patient's renal function uses the four-variable MDRD equation (Levey *et al.*, 1999),

$$\text{eGFR} = 175 \times \left(\frac{\text{SCr}}{88.4}\right)^{-1.154} \times \text{age}^{-0.203} \times 0.742^{\text{I(female)}} \times 1.21^{\text{I(black)}}, \qquad (4.1)$$

where eGFR denotes estimated glomerular filtration rate (in mL/min 1.73m^2 of body surface area) and SCr serum creatinine (in μmol/L), except that the ethnicity factor in (4.1) is ignored, since most of the patients in the CRISIS data set were Caucasian.

The data set covers 2,289 patients who entered the study between 15/11/2000 and 28/02/2013. The patients provided a total of 48,382 repeated measurements of eGFR, in 5,480 (11.3%) of which the patients were hospitalised. Amongst the 2,289 patients, 1,415 (61.8%) were male, 2,202 (96.2%) were Caucasian, 682 (29.8 %) had SRFT as their base hospital, 1,540 (67.3%) were current- or ex-smokers, 1,097 (47.9%) were alcohol-consumers, 723 (31.6%) had been diagnosed with diabetes mellitus and 455 (19.9%) had

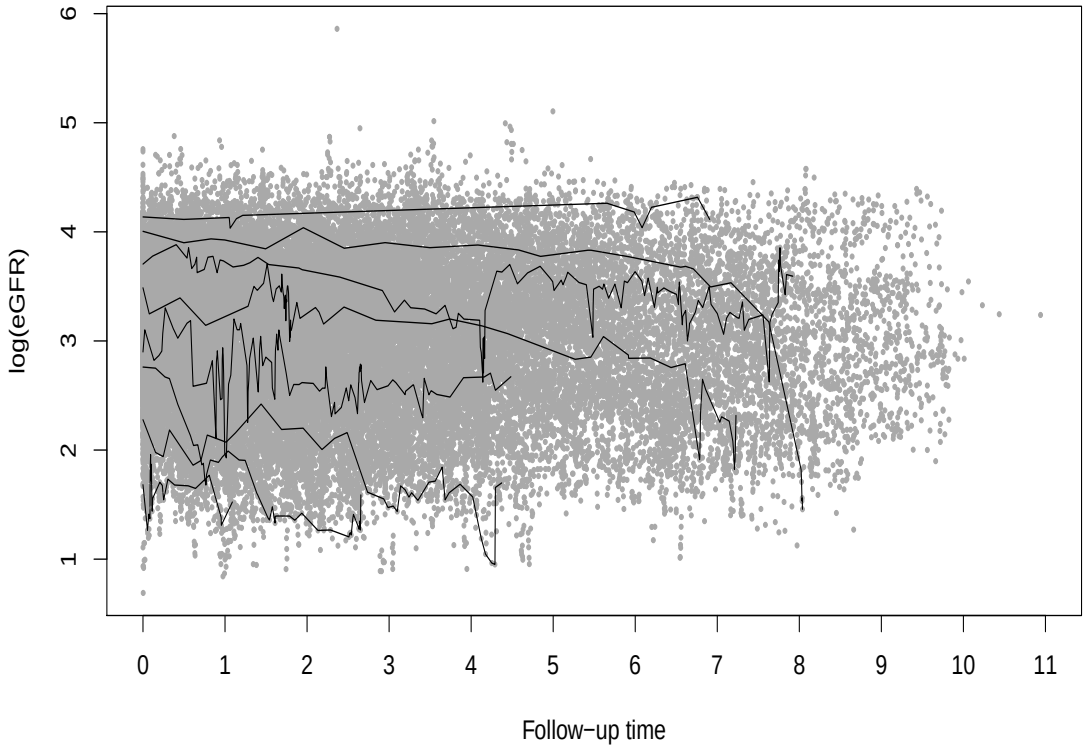

FIGURE 4.1: Log-transformed eGFR measurements against follow-up time (in years) in background as grey scatter-plot. Data from a representative sample of 8 patients are highlighted as black lines.

a history of coronary artery disease, defined as previous myocardial infarction and/or coronary re-vascularisation procedure. Baseline age ranged between 20 and 94.3 (median=66.8). The number of repeated measurements per patient ranged between 1 and 280 (median=13). Total follow-up time ranged between 0 (i.e. only one measurement) and 10.9 years (median=2.6).

The log-transformed eGFR measurements are displayed in Figure 4.1 as a gray scatter-plot, with longitudinal trajectories for a representative sample of 8 patients highlighted as black lines. These data suggest considerable heterogeneity between patients, in respect of both their level of kidney health on entering the study and the pattern of longitudinal evolution in their kidney function. Individual longitudinal trajectories of eGFR typically show non-linear features, some at least of which are associated with AKI episodes.

TABLE 4.1: AKI stages based on the KDIGO AKI guideline. RC denotes relative change in SCr, calculated as $(\mathrm{SCr}_t - \mathrm{SCr}_s)/(\mathrm{SCr}_s)$, where s and t are two time points and $s < t$.

Stage	Criterion
1	$0.5 \leq \mathrm{RC} < 1$ or increase of 26.5 μmol/L or greater
2	$1 \leq \mathrm{RC} < 2$
3	$2 \leq \mathrm{RC}$ or increase to or greater than 353.6 μmol/L or initiation of RRT

4.2.2 Clinical definiton of AKI

AKI events are decided based on inspecting repeated SCr measurements retrospectively, according to the KDIGO AKI guideline (KDIGO, 2012, pg. 8), presented in Table 4.1. In-patient episodes are considered to be the main periods where AKI might have occurred. This is inspected by comparing the following SCr measurement combinations:

- most recent out-patient measurement and first in-patient measurement,
- first in-patient measurement and other in-patient measurements, and
- successive in-patient measurements.

Out-patient AKI episodes are defined as in Table 4.1 where measurements are separated by less than two days. Where successive out-patient measurements are performed within two to seven days of each other, AKI episodes are defined only by relative change in serum creatinine. Multiple AKI events occurring within three days are considered as belonging to a single AKI episode with the most severe grade.

The distribution of the number of AKI events per patients for the CRISIS data set are displayed in Table 4.2. For example, 1,576 (68.9%) patients did not experience any AKI events, 408 (17.8%) experienced only one AKI event (any stage), 126 (5.5%) experienced two AKI events (any stage), and so on. In total, 713 (31.1%) patients experienced at least one AKI event. These patients had total amount of 1,863 AKI events, where 1,198 (64.3%) of the events were stage 1, 37 (2.0%) stage 2 and 628 (33.7%) stage 3. Whilst the patients who did not have any AKI had total amount of 21,176 (43.6%) repeated measurements, the ones who had AKI had 27,401 (56.4%) repeated measurements.

TABLE 4.2: AKI distribution of the CRISIS data set.

AKI events	Patients	AKI events	Patients
0	1,576 (68.9%)	5	24 (1.0%)
1	408 (17.8%)	6	15 (0.7%)
2	126 (5.5%)	7	11 (0.5%)
3	59 (2.6%)	≥ 8	43 (1.8%)
4	27 (1.2%)		

4.2.3 Data

In this study, we specifically consider the influence of the first AKI events on kidney function and censor the longitudinal trajectories at the second AKI. In total, 8,514 repeated measurements on 305 patients (number of patients who had more than one AKI event) are discarded because of this censoring. Data falling into the first AKI episodes are also omitted as they are abrupt observations and the interest is in comparing trajectories for pre- and post-AKI periods. 1,234 repeated measurements on 713 patients (number of patients who had at least one AKI event) are discarded, because they fall into the first AKI episode.

Amongst the first AKI events, 379 (53.2%) are stage 1, 5 (0.7%) stage 2 and 329 (46.1%) stage 3. As the number of stage 2 AKI is low, we proceed by combining stage 2 and 3 into a single stage. The patients with stage 1 AKI had total number of 11,153 repeated measurements, of which 6,954 (62.4%) belong to post-AKI period, 4,199 (37.6%) belong to pre-AKI period. The patients with stage 2 or 3 AKI had total number of 6,500 repeated measurements, of which 5,425 (83.5%) belong to pre-AKI period, 1,075 (16.5%) belong to post-AKI period.

Eight patients' data are displayed in Figures 4.2 and 4.3. Solid dots denote log-transformed eGFR measurements. Amongst these patients, whilst the ones with ID = 474 and 1220 did not experience any AKI, the ones with ID = 26 and 46 experienced stage 1 AKI, at years 1.74 and 1.73, respectively, the ones ID = 430 and 875 experienced stage 2 AKI, at years 1.86 and 1.75, respectively, and the ones with ID = 1 and 187 experienced stage 3 AKI, at years 8.03 and 4.15, respectively. Time at AKI are indicated on the $x - axis$.

4.2.4 Change-points

We assume that there are three change-points in the longitudinal trajectories of the patients who experienced AKI. Amongst these, second change-points are placed at the follow-up times at which AKI events occur, c_i. These change-points are patient-specific

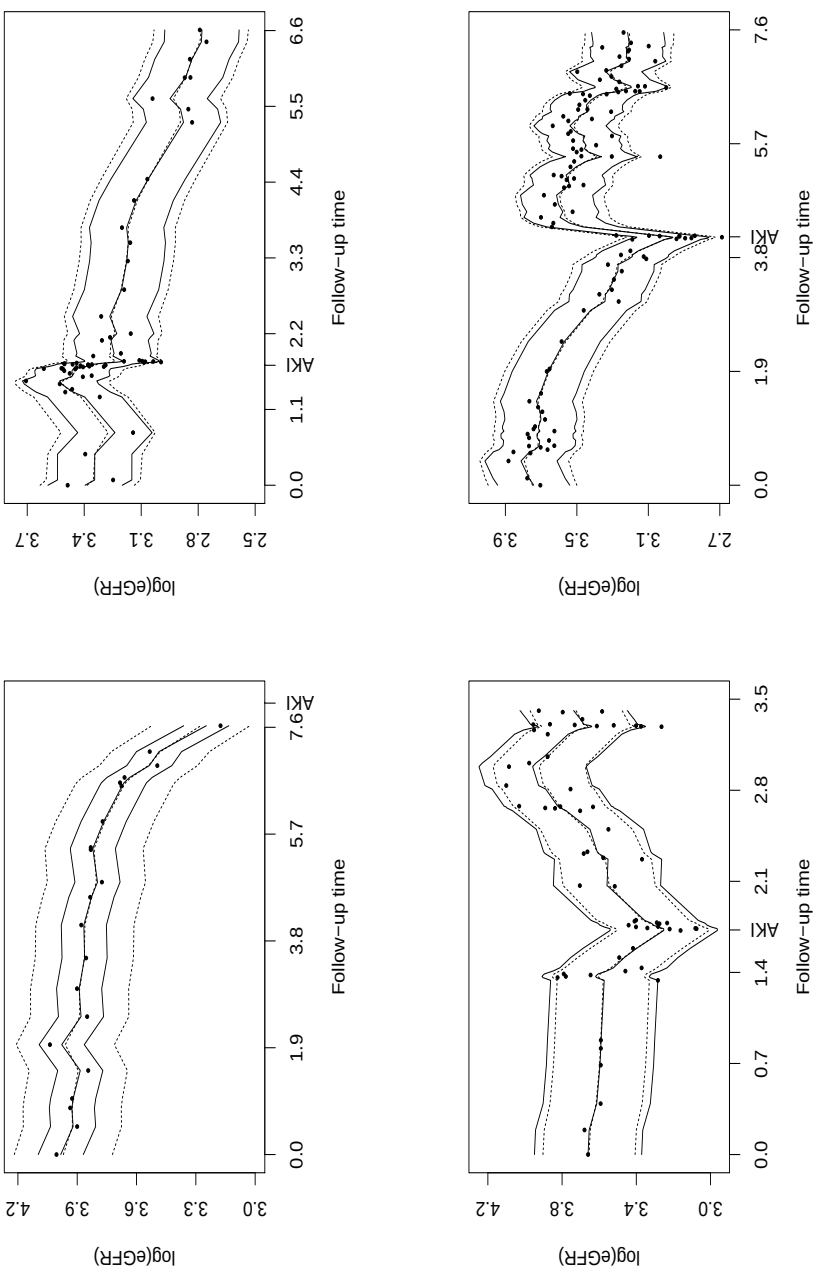

FIGURE 4.2: Observed data and predictions for four subjects. Upper left panel: patient with ID=1 (stage 3 AKI, at year 8.03), upper right panel: ID=26 (stage 1 AKI, at year 1.74), lower left: ID=46 (stage 1 AKI, at year 1.73), lower right: ID=187 (stage 3 AKI, at year 4.15). Dots denote repeated log(eGFR) measurements, straight lines denote the point predictions (middle) and 95% prediction intervals for t distribution, and dashed lines denote the point predictions (middle) and 95% prediction intervals for normal distribution. Time of AKI occurrence is denoted in the $x-axis$.

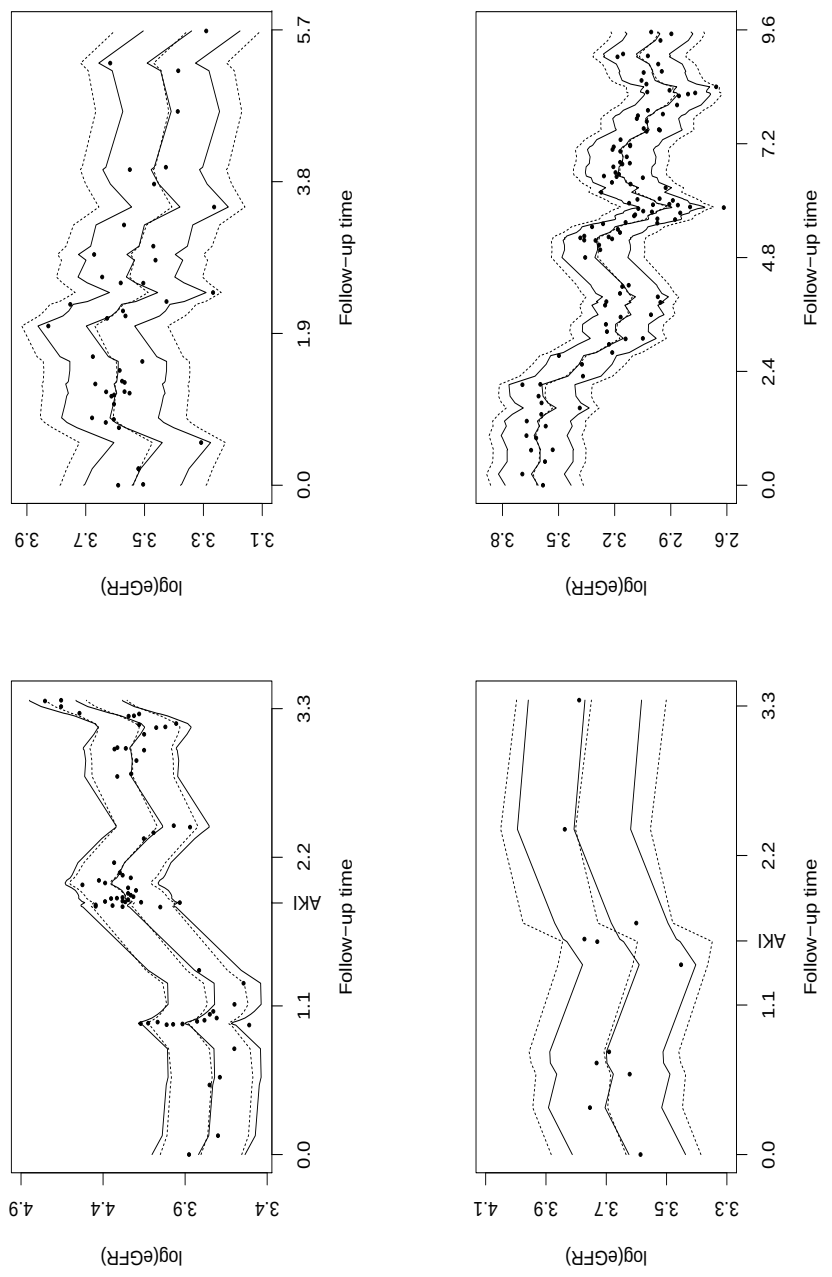

FIGURE 4.3: Observed data and predictions for four subjects. Upper left panel: patient with ID=430 (stage 2 AKI, at year 1.86), upper right panel: ID=474 (no AKI), lower left: ID=875 (stage 2 AKI, at year 1.75), lower right: ID=1220 (no AKI). For details, see Figure 4.2.

and known. On the other hand, the first change-points are assumed to be at a time somewhere before AKI events occur. These change-points correspond to accelerations in the loss of kidney functions that precede occurrence of AKI events. The rationale for them are in line with the statement of Bellomo, Kellum and Ronco (2012) for replacing the term "acute renal failure" by "acute kidney injury":

"*... to emphasise that a continuum of kidney injury exists that begins long before sufficient loss of excretory kidney function...*"

The first change-points are assumed to be different in terms of experiencing stage 1 and stage 2 or 3 AKI, but are assumed to be common for the patients within these groups. In other words, whilst $c_i - a_1$ is the first change-point for the patients who experienced stage 1 AKI, $c_i - a_2$ is the first change-point for the patients who experienced stage 2 or 3 AKI. The values of a_1 and a_2 are unknown and to be decided by the following exploratory methods. Data are aligned at the AKI occurrences and LOWESS curves (Cleveland, 1979) with different smoothing parameters are drawn. Also, profile likelihood based on linear model with independent errors is considered. The third change-points, on the other hand, are assumed to occur after AKI events. These change-points correspond to the end of recovery in kidney function, which the patients are assumed to have after an AKI event. Here, these change-points are based on the experiences of the physicians who run the CRISIS cohort. Similar to the first change-points, the third change-points are assumed to be different for stage 1 and stage 2 or 3 AKI experience, but are assumed to be same for the patients within these groups. We denote these change-points with $c_i + b_1$ and $c_i + b_2$ for the stage 1 and stage 2 or 3 AKI groups, respectively. b_1 and b_2 are also assumed to be unknown and to be decided by the exploratory methods discussed above.

We call the change-point before AKI occurrence as "acceleration point", and the one after AKI occurrence as "end-of-recovery point". We then label four periods based on these change-points as, pre-acceleration: from study entry to acceleration point, acceleration: from acceleration point to AKI occurrence, recovery: from AKI occurrence to end-of-recovery point, post-recovery: after end-of-recovery point.

4.3 Model

4.3.1 Formulation

We consider a class of linear mixed effects models for repeated blood samples data. The general form of the model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{U}_i + \boldsymbol{W}_i + \boldsymbol{Z}_i. \tag{4.2}$$

Here, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ denotes log-transformed eGFR measurements for subject i $(i = 1, \ldots, m)$ at time points $\boldsymbol{t}_i = \{t_{i1}, \ldots, t_{in_i}\}$. $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in_i})^T$ with $\boldsymbol{X}_{i\cdot} = (1, X_{i11}, \ldots, X_{ip1})^T$ are explanatory variables that might include time-independent and/or time-dependent variables, and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_p)^T$ are the corresponding fixed effects parameters to be estimated. $\boldsymbol{U}_i = (U_i, \ldots, U_i)^T$ (with n_i identical elements) are patient-specific random intercepts that account for the between-patient heterogeneity in kidney function at study entry. $\boldsymbol{W}_i = (W_i(t_{i1}), \ldots, W_i(t_{in_i}))^T$ are subject-and-time-specific random-effects that account for serial correlation arising from the repeated measurements belonging to the same patients. Finally, $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{in_i})^T$ are mutually independent random measurement error.

The conditional distributions of $\boldsymbol{U}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$ given γ_i, an unobserved random variable, and the marginal distribution of γ_i are specified as

$$[\boldsymbol{U}_i|\gamma_i] = MVN\left(\boldsymbol{0}, \frac{\omega^2}{\gamma_i}\boldsymbol{J}_i\right), \tag{4.3}$$

$$[\boldsymbol{W}_i|\gamma_i] = MVN\left(\boldsymbol{0}, \frac{\sigma^2}{\gamma_i}\boldsymbol{R}_i\right), \tag{4.4}$$

$$[\boldsymbol{Z}_i|\gamma_i] = MVN\left(\boldsymbol{0}, \frac{\tau^2}{\gamma_i}\boldsymbol{I}_i\right), \tag{4.5}$$

$$[\gamma_i] = Gamma\left(\nu/2, \nu/2\right), \tag{4.6}$$

where $[\cdot]$ denotes the "distribution of", $\boldsymbol{J}_i$ is an $n_i \times n_i$ matrix of 1s, $\boldsymbol{R}_i$ is an $n_i \times n_i$ matrix, details of which to be discussed below, $\boldsymbol{I}_i$ is the $n_i \times n_i$ identity matrix, and $\nu/2$, is the shape and rate parameter of the gamma distribution, with $\nu > 0$. The specifications in (4.3)-(4.6) equivalently imply that

$$[\boldsymbol{U}_i] = MVt\left(\boldsymbol{0}, \omega^2\boldsymbol{J}_i, \nu\right), \tag{4.7}$$

$$[\boldsymbol{W}_i] = MVt\left(\boldsymbol{0}, \sigma^2\boldsymbol{R}_i, \nu\right), \tag{4.8}$$

$$[\boldsymbol{Z}_i] = MVt\left(\boldsymbol{0}, \tau^2\boldsymbol{I}_i, \nu\right). \tag{4.9}$$

Here, $\boldsymbol{U}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$ are uncorrelated, but not independent, when $\nu < \infty$.

We structure $\boldsymbol{R}_i$ by the Matérn correlation family (Matérn, 1960), i.e. (r,s)th element of $\boldsymbol{R}_i$ is

$$\{R_i\}_{r,s} = \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1} (|t_{ir} - t_{is}|/\phi)^{\kappa} K_{\kappa}(|t_{ir} - t_{is}|/\phi), \tag{4.10}$$

where $\Gamma(\kappa) = \int_0^{\infty} x^{\kappa-1}\exp(-x)dx$ is the gamma function, K_κ is a modified Bessel function with order κ, $\phi > 0$ is a scale parameter, $\kappa > 0$ is a shape parameter which controls the smoothness of the process, $W(\cdot)$. Matérn correlation family indicates stationarity for $W(\cdot)$, covers different degrees of smoothness depending on the value of κ and indicates that $W(\cdot)$ is ceiling$(\kappa) - 1$ times mean-square differentiable. Special cases include, amongst others, exponential correlation function when $\kappa = 0.5$ as $\{R_i\}_{r,s} = \exp(-|t_{ir} - t_{is}|/\phi)$ and Gaussian correlation function when $\kappa \to \infty$ as $\{R_i\}_{r,s} \to \exp(-|t_{ir} - t_{is}|^2/\phi)$. These cases correspond to mean-square continuous but non-differentiable and infinitely mean-square differentiable processes, respectively. Diggle and Ribeiro (2007) discusses that κ is poorly estimated as ϕ and κ are not orthogonal, and proposes to set κ to some specific values covering a range of smoothness. In this study, we follow their suggestion and fix κ at, e.g. $0.5, 1.5, 2, \infty$. These values respectively correspond to mean-square continuous but non-differentiable, once-differentiable, twice-differentiable and, infinitely mean-square differentiable processes. Explicit calculations for Matérn correlation family are not available, but approximate calculations can be obtained, e.g. as implemented in the `matern` function of the R package `geoR` (Diggle and Ribeiro, 2007).

The distributional properties of U_i, W_i, Z_i and γ_i induce $\boldsymbol{Y}_i$ to have the following conditional and marginal distributions,

$$[\boldsymbol{Y}_i|\gamma_i] = MVN\left(\boldsymbol{X}_i\boldsymbol{\alpha}, \frac{\boldsymbol{V}_i}{\gamma_i}\right), \tag{4.11}$$

$$[\boldsymbol{Y}_i] = MVt(\boldsymbol{X}_i\boldsymbol{\alpha}, \boldsymbol{V}_i, \nu), \tag{4.12}$$

where $\boldsymbol{V}_i = \omega^2\boldsymbol{J}_i + \sigma^2\boldsymbol{R}_i + \tau^2\boldsymbol{I}_i$. Here, the variance-covariance matrix of $\boldsymbol{Y}_i$ is $\frac{\nu}{\nu-2}\boldsymbol{V}_i$, for $\nu > 2$. When $\nu > 1$, $\boldsymbol{X}_i\boldsymbol{\alpha}$ is the mean of the multivariate t distribution, hence the interpretations of the estimates of α are same as with the normal model, i.e. the interpretations are for average changes in the response variable. As $\nu \to \infty$, $[\boldsymbol{Y}_i] \xrightarrow{distr.} MVN(\boldsymbol{X}_i\boldsymbol{\alpha}, \boldsymbol{V}_i)$. In other words, when $\nu \to \infty$, $\gamma_i \to 1$, the model (4.2) approximates to the multivariate normal model. For further details of the multivariate t distribution, see Kotz and Nadarajah (2004). Multivariate slash or multivariate contaminated normal distributions can also be obtained by specifying the distribution of γ_i other than gamma, in (4.3)-(4.6), e.g. see (Rosa, Padovani and Gianola, 2003).

4.3.2 Inference

We estimate the model parameters, $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \omega^2, \sigma^2, \phi, \tau^2, \nu\}$, by ML estimation with an E-M algorithm, following Liu and Rubin (1995) and Pinheiro, Liu and Wu (2001). $\boldsymbol{\theta}$ does not include κ, since we set it to known values. Data belonging to different patients are assumed to be independent. We specifically consider that the random-effects are integrated out and the complete data for patient i consist of $\boldsymbol{Y}_i$ and γ_i, where whilst $\boldsymbol{Y}_i$ are observed, γ_i is unobserved and treated as missing data. Then, the complete data likelihood for subject i is

$$[\boldsymbol{Y}_i, \gamma_i] = [\boldsymbol{Y}_i|\gamma_i][\gamma_i], \tag{4.13}$$

where $[\boldsymbol{Y}_i|\gamma_i]$ and $[\gamma_i]$ are given as in (4.11) and (4.6), respectively. The E-M algorithm then includes the following steps:

E-step: Given $\hat{\boldsymbol{\theta}}$, calculate $\hat{\gamma}_i = E(\gamma_i|\boldsymbol{Y}_i)$ based on

$$[\gamma_i|\boldsymbol{Y}_i] = Gamma\left(\frac{\nu + n_i}{2}, \frac{\nu + \delta_i}{2}\right), \tag{4.14}$$

where $\delta_i = (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\alpha})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\alpha})$.

M-step 1: Given $\hat{\gamma}_i$, maximise the expected-log-likelihood, given $\boldsymbol{Y}_i$, based on $[\boldsymbol{Y}_i|\gamma_i]$ with respect to $\boldsymbol{\theta}_{-\nu}$.

M-step 2: Given $\hat{\boldsymbol{\theta}}_{-\nu}$, obtain $\hat{\nu}$, based on (4.12), as

$$\arg\max_{\nu} \sum_{i=1}^{m} \left\{ \log\left[\Gamma\left(\frac{\nu + n_i}{2}\right)\right] - \log\left[\Gamma\left(\frac{\nu}{2}\right)\right] + \frac{\nu}{2}\log(\nu) - \frac{\nu + n_i}{2}\log\left(\nu + \delta_i\right) \right\}.$$

Parameter estimates in M-step 1 can be obtained by using available software for multivariate normal mixed models with Matérn correlation family, since when γ_i is known, the multivariate t model in (4.2) reduces to $\boldsymbol{Y}_i^* = \boldsymbol{X}_i^*\boldsymbol{\alpha} + \boldsymbol{U}_i + \boldsymbol{W}_i + \boldsymbol{Z}_i$, where $\boldsymbol{Y}_i^* = \boldsymbol{Y}_i\sqrt{\hat{\gamma}_i}$ and $\boldsymbol{X}_i^* = \boldsymbol{X}_i\sqrt{\hat{\gamma}_i}$, and $[\boldsymbol{U}_i]$, $[\boldsymbol{W}_i]$ and $[\boldsymbol{Z}_i]$ are all multivariate normal. We use the `R` package `lmenssp` for this purpose. On the other hand, `R` function `optimize` is used for M-step 2. In M-step 2, $\hat{\nu}$ might also be obtained by maximising the expected-log-likelihood, given $\boldsymbol{Y}_i$, based on $[\gamma_i]$, but the resulting procedure involves a term that is not available in a closed-form and this might result the algorithm to be extremely slow (Liu and Rubin, 1995; Pinheiro, Liu and Wu, 2001). The main advantage of exploiting (4.12) in M-step 2 is that it is free of γ_i.

At convergence, large-sample variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$, using expected Fisher information matrix, based on (4.12), are calculated by

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{m} \frac{\nu + n_i}{\nu + n_i + 2} \boldsymbol{X}_i{}^T \boldsymbol{V}_i{}^{-1} \boldsymbol{X}_i \right)^{-1}, \tag{4.15}$$

at $\hat{\boldsymbol{\theta}}_{-\alpha}$. On the other hand, variance-covariance matrix of $\boldsymbol{\theta}_{-\alpha}$ are calculated as the inverse of the negative Hessian matrix, which is approximated numerically by the central-difference method.

The model in (4.2) can be equivalently written as, $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{D}_i\boldsymbol{B}_i + \boldsymbol{Z}_i$, where $\boldsymbol{D}_i = (\boldsymbol{1}_i^T, \boldsymbol{I}_i)$ with $\boldsymbol{1}_i$ is vector of 1s of length n_i, and $\boldsymbol{I}_i$ as before and $\boldsymbol{B}_i = (U_i, W_{i1}, \ldots, W_{in_i})^T$. The conditional distribution, $[\boldsymbol{B}_i|\boldsymbol{Y}_i, \gamma_i]$, can then be obtained by using multivariate normal theory (Anderson, 1984), as $[\boldsymbol{Y}_i, \boldsymbol{B}_i|\gamma_i]$ is multivariate normal,

$$[\boldsymbol{B}_i|\boldsymbol{Y}_i, \gamma_i] = MVN\left(\boldsymbol{\psi}_i \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}), \frac{\boldsymbol{\psi}_i}{\gamma_i} \left(\boldsymbol{J}_i - \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \boldsymbol{\psi}_i \right) \right), \tag{4.16}$$

where $\boldsymbol{\psi}_i/\gamma_i$ is the variance-covariance matrix of $[\boldsymbol{B}_i|\gamma_i]$, $\boldsymbol{V}_i$ and $\boldsymbol{J}_i$ are as before. We plug-in in (4.16) $\hat{\boldsymbol{\theta}}_{-\nu}$, and $\hat{\gamma}_i$. We opt to proceed by replacing $\hat{\gamma}_i$ to calculate $[\boldsymbol{B}_i|\boldsymbol{Y}_i, \gamma_i]$, since explicit forms of conditional distributions based on multivariate t distribution is not available (Arellano-Valle and Bolfarine, 1995).

4.4 Results

We analyse the data for all the patients, i.e. the ones without any AKI events and the ones with AKI events as presented in Section 4.2.3, by using the model given in Equation 4.2. There are a total of 38,829 repeated measurements belonging to 2,289 patients. The main explanatory variables used in the analysis are listed in Table 4.3. Piecewise linear terms in the slopes based on the change-points are considered as additional explanatory variables.

We build mixed models with varying shape parameter for the Matérn correlation family, $\kappa = 0.5, 1.5, 2.5, \infty$, whilst keeping the covariate sets same. Respective maximised log-likelihood values are 14,534.43, 13,343.70, 13,046.89, and 12,635.39. Therefore, we proceed with the results of the model with $\kappa = 0.5$. The change-points are identified as $a_1 = 0.3$ years (≈ 110 days), $b_1 = 0.1$ (≈ 37) $a_2 = 0.4$ (≈ 146), and $b_2 = 0.15$ (≈ 55). The ML estimates for both the multivariate t and normal models are displayed in Table 4.4. Multivariate t distribution assumption clearly provides a better fit for the CRISIS data set compared to the multivariate normal distribution assumption, as indicated by the maximised log-likelihoods, 14,534.43 vs. 8,892.63.

TABLE 4.3: Descriptions of the variables.

Variable	Explanation
t_{ij}	Follow-up time: age at measurement - age at study entry
Stage1	1 = if the patient experienced stage 1 AKI, 0 = no AKI or stage 2 or 3 AKI
Stage23	1 = if the patient experienced stage 2 or 3 AKI, 0 = no AKI or stage 1 AKI
Baseline age	Age at study entry
Base hospital	1 = if the base hospital is SRFT, 0 = if the base hospital is different than SRFT
Gender	1 = male, 0 = female
Smoking	1 = current or former smoker at study entry, 0 = never smoked
Alcohol	1 = alcohol consumer at study entry, 0 = abstinent from alcohol
Diabetes	1 = type I or type II diabetes at study entry, 0 = no diabetes
Co-morbidity	1 = history of myocardial infarction and/or coronary revascularization procedure at study entry, 0 = none

TABLE 4.4: Maximum likelihood estimates of the model parameters based on $\kappa = 0.5$ for the multivariate Normal and t models.

		Normal			t		
Variable	Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	α_0	3.4299	0.0544	<0.0001	3.5674	0.0539	<0.0001
t_{ij}	α_1	-0.0392	0.0027	<0.0001	-0.0366	0.0021	<0.0001
$\max(0, t_{ij} - (c_i - 0.3))$ * Stage1	α_2	-0.5695	0.0333	<0.0001	-0.4070	0.0251	<0.0001
$\max(0, t_{ij} - c_i)$ * Stage1	α_3	1.3868	0.1091	<0.0001	0.7088	0.0773	<0.0001
$\max(0, t_{ij} - (c_i + 0.1))$ * Stage1	α_4	-0.8482	0.1001	<0.0001	-0.3595	0.0712	<0.0001
$\max(0, t_{ij} - (c_i - 0.4))$ * Stage23	α_5	-0.8377	0.0316	<0.0001	-0.7264	0.0241	<0.0001
$\max(0, t_{ij} - c_i)$ * Stage23	α_6	2.1402	0.1451	<0.0001	1.3221	0.1057	<0.0001
$\max(0, t_{ij} - (c_i + 0.15))$ * Stage23	α_7	-1.3700	0.1499	<0.0001	-0.6421	0.1093	<0.0001
Baseline age	α_8	-0.0031	0.0008	0.0001	-0.0042	0.0008	<0.0001
Base hospital	α_9	0.2194	0.0242	<0.0001	0.1871	0.0239	<0.0001
Gender	α_{10}	0.0372	0.0238	0.1174	0.0475	0.0236	0.0438
Smoking	α_{11}	-0.0217	0.0242	0.3696	-0.0177	0.0240	0.4592
Alcohol	α_{12}	0.0593	0.0229	0.0096	0.0533	0.0227	0.0189
Diabetes	α_{13}	-0.1033	0.0243	<0.0001	-0.1169	0.0240	<0.0001
Co-morbidity	α_{14}	0.0140	0.0292	0.6309	0.0065	0.0289	0.8214
Random intercept	ω^2	0.0676	0.0091		0.1275	0.0132	
Stationary process	σ^2	0.2442	0.0047		0.1327	0.0115	
Stationary process	ϕ	8.0843	0.0003		7.1196	0.6437	
Measurement error	τ^2	0.0160	0.0002		0.0055	0.0001	
Degrees-of-freedom	ν				3.7065	0.1307	
Max. log-likelihood			8,892.63			14,534.43	

4.4.1 Population-averaged results

The parameter estimate $\hat{\alpha}_0 = 3.5674$ denotes average kidney function level (in log-scale) at study entry. $\hat{\alpha}_1 = -0.0366$ denotes mean loss of kidney function per year at a rate of approximately 3.6% ($= (\exp(-0.0366) - 1) * 100$) for patients who did not have any AKI. At the same time, it corresponds to the slope of the pre-acceleration period for the patients who experienced stage 1 and stage 2 or 3 AKI, and we call $\hat{\alpha}_1$ as $\hat{\alpha}_{pre-acc-1}$ and $\hat{\alpha}_{pre-acc-2}$ for these group of patients, respectively. $\hat{\alpha}_2$, $\hat{\alpha}_3$ and $\hat{\alpha}_4$ denote the changes in the slope at 0.3 years prior to stage 1 AKI occurrence, at stage 1 AKI occurrence and 0.1 years after stage 1 AKI occurrence, respectively. Together with $\hat{\alpha}_{pre-acc-1}$, they collectively establish the slopes for the acceleration period, $\hat{\alpha}_{acc-1} = -0.4436$ ($SE = 0.0249$), for the recovery period, $\hat{\alpha}_{rec-1} = 0.2653$ ($SE = 0.0684$), and for the post-recovery period, $\hat{\alpha}_{post-rec-1} = -0.0942$ ($SE = 0.0110$). Similarly, $\hat{\alpha}_5$, $\hat{\alpha}_6$ and $\hat{\alpha}_7$ denote the changes in slope at 0.4 years prior to stage 2 or 3 AKI occurrence, at stage 2 or 3 AKI occurrence and 0.15 years after stage 2 or 3 AKI occurrence, respectively. Together with $\hat{\alpha}_{pre-rec-2}$, they collectively establish the slopes for the acceleration period, $\hat{\alpha}_{acc-2} = -0.7630$ ($SE = 0.0239$), for the recovery period, $\hat{\alpha}_{rec-2} = 0.5591$ ($SE = 0.1010$), and for the post-recovery period, $\hat{\alpha}_{post-rec-2} = -0.0830$ ($SE = 0.0262$). These parameter estimates are depicted in Figure 4.4, for example, for a (hypothetical) female patient with baseline age of 67 years, whose base hospital was not SRFT, was not smoking, was not consuming alcohol, did not have diabetes mellitus and co-morbidity history at study entry and was followed for 2 years. In addition to the results of the multivariate t distribution (in black), the results of the multivariate Normal distribution (in grey) as presented in Figure 4.4.

Key findings regarding the influences of AKI occurrences on kidney function are summarised below:

- The rate of kidney function loss significantly accelerates at 0.3 years prior to stage 1 AKI occurrence; p-value is <0.0001 for $H_0 : \alpha_{pre-acc-1} = \alpha_{acc-1}$. Whilst the rate of kidney function loss was 3.6% in the pre-acceleration period, it became 35.8% in the acceleration period.
- The rate of kidney function loss significantly accelerates at 0.4 years prior to stage 2 or 3 AKI occurrence; p-value is <0.0001 for $H_0 : \alpha_{pre-acc-2} = \alpha_{acc-2}$. Whilst the rate of kidney function loss was 3.6% in the pre-acceleration period, it became 53.4% in the acceleration period.
- After experiencing stage 1 AKI, patients significantly recovered their kidney function until 0.1 years after AKI occurrence with a rate of 30.4%; p-value is <0.0001 for $H_0 : \alpha_{rec-1} = 0$.

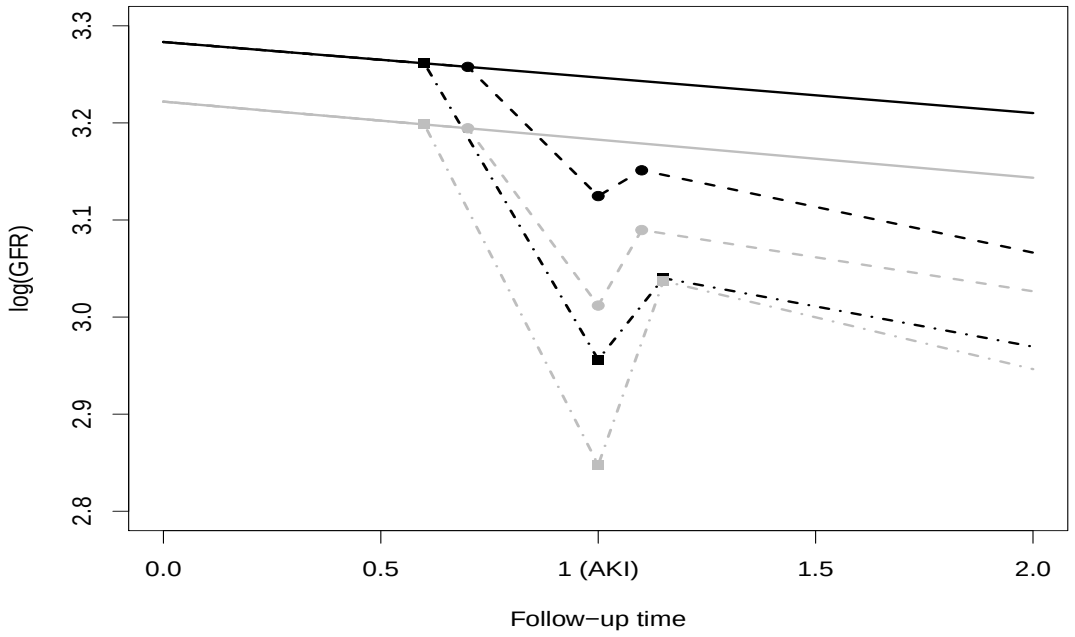

FIGURE 4.4: Average kidney function evolution for a hypothetical patient. Results based on multivariate t model are in black, results based on multivariate Normal model are in grey. Straight lines denote mean profiles for no AKI occurrence; dashed lines denote mean profiles for stage 1 AKI occurrence; dashed lines with dots denote mean profiles for stage 2 or 3 AKI occurrence. Solid dots denote change-points for stage 1 AKI, solid squares denote change-points for stage 2 or 3. Follow-up time 1 is the time at AKI occurrence for both stage 1 and stage 2 or 3 AKI.

- After experiencing stage 2 or 3 AKI, patients significantly recovered their kidney function until 0.15 years after AKI occurrence with a rate of 74.9%; p-value is <0.0001 for $H_0 : \alpha_{rec-1} = 0$.
- The slope of the post-recovery period for stage 1 AKI is negative, $\alpha_{post-rec-1} = -0.0942$, and significantly different form 0; p-value is <0.0001 for $H_0 : \alpha_{post-rec-1} = 0$. Its magnitude is significantly larger than the magnitude of the slope of the pre-acceleration period, $\hat{\alpha}_{pre-acc-1} = -0.0366$; p-value = 0.0001 for $H_0 : \alpha_{pre-acc-1} = \alpha_{post-rec-1}$. The rate of kidney function loss become 9.0%, whilst it was 3.6% at the pre-acceleration period.
- The slope for the post-recovery period for stage 2 or 3 is negative, $\hat{\alpha}_{post-rec-2} = -0.0830$, and significantly different form 0; p-value is 0.0015 for $H_0 : \alpha_{post-rec-2} = 0$. However, its magnitude is not significantly larger than the magnitude of the slope of the pre-acceleration period, $\hat{\alpha}_{pre-acc-2} = -0.0366$; p-value = 0.0769 for $H_0 : \alpha_{pre-acc-2} = \alpha_{post-rec-2}$.

Here, the decision of failing to reject the hypothesis, $H_0 : \alpha_{pre-acc-2} = \alpha_{post-rec-2}$, is surprising, because there is not enough evidence to say that the slopes at post-recovery periods for the patients with stage 1 AKI and stage 2 or 3 AKI are different; p-value for $H_0 : \alpha_{post-rec-1} = \alpha_{post-rec-2}$ is 0.6929. This might be explained by the number of observations falling into the post-recovery periods. There are 617 observations in this period for patients with stage 2 or 3 AKI. On the other hand, there are 4,199 observations in this period for patients who had stage 1 AKI. The influence of the number of observations might also be seen in the width of the confidence intervals: whilst the estimate (95% confidence interval) for $\alpha_{pre-acc-1} - \alpha_{post-rec-1}$ is 0.0576 (0.0356, 0.0797), it is 0.0464 (-0.0050, 0.0978) for $\alpha_{pre-acc-2} - \alpha_{post-rec-2}$. Based on these, testing the following hypothesis might be more suitable to test the influence of AKI on kidney function, $H_0 : \alpha_1 = (\alpha_{post-rec-1} + \alpha_{post-rec-2})/2$. Whilst the related p-value is found to be 0.0003, the estimate (95% confidence interval) of $\alpha_1 - (\alpha_{post-rec-1} + \alpha_{post-rec-2})/2$ is 0.0520 (0.0239, 0.0802). Based on these results, we can say that the influence of AKI on kidney function is significant, and the collective loss of kidney function after recovery is 8.5%, whilst it was 3.6% in the pre-acceleration period.

A one year increase in the baseline age was associated with an average loss of kidney function at an approximate rate of 0.4%. Patients whose base hospital was SRFT were found to have approximately 20.6% better kidney function compared to the ones whose base hospital was not SRFT. Males had 4.9% better kidney health compared to females at study entry. Alcohol consumers were associated with 5.5% better kidney function at study entry compared to the patients who were abstainers. This is indeed an unexpected result, yet might be considered natural as regression equations do not generally imply causation. Another explanation for this might be that the patients with worse kidney health might be abstinent from alcohol. On the other hand, some previous works also discuss better kidney function amongst alcohol consumers, e.g. see Kim, Kim and Song (2014). Patients with diabetes mellitus had 11.0% worse kidney health compared to the patients without diabetes mellitus. No significant relationships were found between kidney function and smoking or co-morbidity; respective p-values are 0.4592 and 0.8214. ML estimates of the covariance parameters were found to be $\hat{\omega}^2 = 0.1275$, $\hat{\sigma}^2 = 0.1327$, $\hat{\phi} = 7.1196$, $\hat{\tau}^2 = 0.0055$. The degrees-of-freedom parameter was estimated to be $\hat{\nu} = 3.7065$. This indicates that the outlier effect is non-negligible and $\hat{\boldsymbol{\alpha}}$ has the mean parameter interpretation.

A multivariate normal distribution assumption produces the same decisions with regards the hypothesis tested above. The only difference is on the relationship between gender and kidney function. A multivariate normal distribution indicates that the relationship is insignificant, p-value = 0.1174. There are differences in terms of the parameter estimates. For example, whilst the multivariate normal model indicates that the patients

who experienced stage 1 AKI had a loss of kidney function at a rate of 45.6% at the acceleration period, multivariate t model indicates it was 35.8%. As an another example, whilst the multivariate normal model indicates a recovery of a rate of 253.7% for the patients with stage 2 or 3 AKI, multivariate t model indicates recovery of 74.9%. Also, the standard error estimates of the fixed effects are inexorably smaller under the multivariate t model.

4.4.2 Patient-specific results

The inferences we have presented so far are population-averaged. Our model allows us to obtain patient-specific inferences as well. For example, predictions for eight patients are displayed in Figures 4.2 and 4.3. Here, we shall note that these patients are deliberately selected as the ones who reflect deviations from the population-averaged behaviours. Point predictions and point-wise prediction intervals for the multivariate t model are calculated as the mean and 2.5 and 97.5% quantiles of $[Y_{ij}|\boldsymbol{X}_{ij}, \hat{\boldsymbol{\alpha}}, \tilde{U}_i, \tilde{W}_i(t_{ij}), \hat{\tau}^2, \hat{\gamma}_i] = N(\boldsymbol{X}_{ij}\hat{\boldsymbol{\alpha}} + \tilde{U}_i + \tilde{W}_i(t_{ij}), \hat{\tau}^2/\hat{\gamma}_i)$, where whilst $\hat{\gamma}_i$ is the mean of $(4.14), \tilde{U}_i$ and $\tilde{W}_i(t_{ij})$ are the means of the conditional distributions $[U_i|\boldsymbol{Y}_i; \boldsymbol{\theta}]$ and $[W_i(t_{ij})|\boldsymbol{Y}_i; \boldsymbol{\theta}]$, based on (4.16). These are obtained for the multivariate normal model by setting $\hat{\gamma}_i = 1$. The predictions for log-transformed eGFR measurements seem to be successful, as the noise in the data are smoothed out and almost all the observations are included within the 95% prediction intervals. The patients exhibit different behaviours in the pre-acceleration, acceleration, recovery and post-recovery periods. For example, the patient with ID=46 has a longer recovery period, approximately 2 years, as opposed to the population-averaged behaviours presented before. The patient with ID=187 has a sharp recovery after which the level and rate of loss of kidney function are similar to these of the pre-AKI period. The difference between the point predictions of the multivariate t (straight lines) and multivariate normal (dashed lines) models are almost indiscernible. However, in most cases the point-wise prediction intervals of the former are narrower. There are a few observations that are covered by the prediction intervals of the multivariate normal model, e.g. patients with ID=474 and 1220, but not these of the multivariate t model. An explanation for this is that these observations are treated as outlying observations by the latter model.

4.5 Diagnostics

To check the appropriateness of the model assumptions, diagnostic tests are applied to standardised empirical residuals, $\boldsymbol{r}_i{}^* = \boldsymbol{S}_i{}^{-1}\boldsymbol{r}_i$, where $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}$ are the empirical

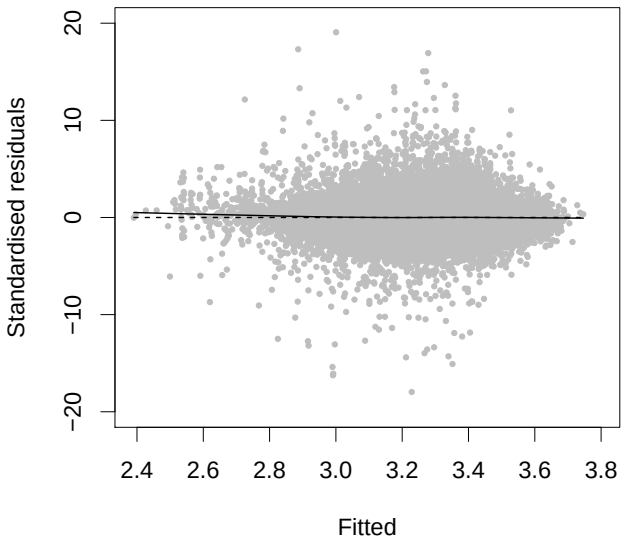

FIGURE 4.5: Fitted values vs. standardised residuals. The dashed line is the zero line (x-axis). The solid line is the LOWESS curve.

residuals, and $\boldsymbol{S}_i\boldsymbol{S}_i{}^T = \hat{\boldsymbol{V}}_i$. For a well-fitting model, theoretically, we expect $[r^*_{ij}] \overset{\text{i.i.d.}}{=} t(0, 3.7065)$.

Fitted values, $\boldsymbol{X}_{ij}\hat{\boldsymbol{\alpha}}$, are plotted against the standardised residuals in Figure 4.5, where dashed line denotes the LOWESS curve obtained by the `R` function `lowess` with the default value for the smoothing parameter. There is no discernible systematic pattern in the standardised residuals and no evidence of non-constant variance, and the fitted LOWESS curve is close to 0.

We inspect the appropriateness of the assumed covariance structure through the variance structure of the standardised residuals. For a well-fitting model they are expected to fluctuate around the theoretical variance of $t(3.7065)$, 2.1720. The empirical and theoretical variances are displayed in Figure 4.6. The empirical variances are calculated from binned residuals through time, with bin size of two weeks, but with baseline data treated separately, that is, variances are calculated separately at baseline and over follow-up measurements between 0+ and 14 days after baseline. Bins with fewer than 30 elements are omitted. Empirical variances of the standardised residuals randomly fluctuate around 2.1720, which indicates the appropriateness of the assumed covariance structure.

To further check the appropriateness of the assumed covariance structure, we use the variogram of the standardised residuals (Diggle *et al.*, 2002). The theoretical variogram

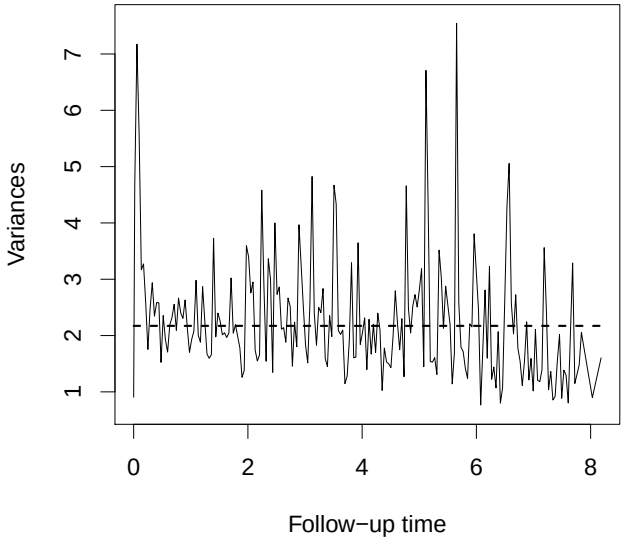

FIGURE 4.6: Empirical variances of the standardised residuals through time. The dashed line is the theoretical variance of the t distribution, 2.1720. Residuals are binned through time with bin size of two weeks. Baseline data are treated separately. Bins with fewer than 30 residuals are omitted.

function is defined as $\gamma(u_{ijk}) = \frac{1}{2}\mathrm{E}(r^*_{ij} - r^*_{ik})^2$, where $u_{ijk} = |t_{ij} - t_{ik}|$. Empirical variogram is calculated by $g_{ijk} = \frac{1}{2}(r^*_{ij} - r^*_{ik})^2$. For a well-fitting model, the empirical variogram of the standardised residuals should randomly fluctuate around the variance of $t(3.7065)$, through $u^*_{ijk} = |t^*_{ij} - t^*_{ik}|$, where $t^*_{i\cdot}$ belongs to the vector of $\boldsymbol{t}^*_i = \boldsymbol{S}_i{}^{-1}\boldsymbol{t}_i$ (Fitzmaurice, Laird and Ware, 2011). Since the observations are completely irregularly spaced in the CRISIS data set, we average g_{ijk} over successive bins with equal and pre-specified lengths through u^*_{ijk}. We specifically consider the bins have a length of two weeks, and omit the bins with fewer than 30 residuals. Variograms are displayed in Figure 4.7. Here, the dashed line at height 2.1720 denotes the variance of $t(3.7065)$. The variograms fluctuate around 2.1720, which is consistent with the assumed covariance structure.

The distributional assumption is checked by comparing theoretical and empirical quantiles. The quantile-quantile plot for the multivariate normal model as well as the multivariate t model are plotted in left and right panels of Figure 4.8, respectively. We can say that the multivariate t distribution is an appropriate choice for the CRISIS data set.

There are some discrepancies in the variogram, between lag 0-1.5 and 4.5-6, and the quantile-quantile plot for the multivariate t model, towards the tails. Simulated realisations for the variogram and quantile-quantile plot under the fitted model (not shown

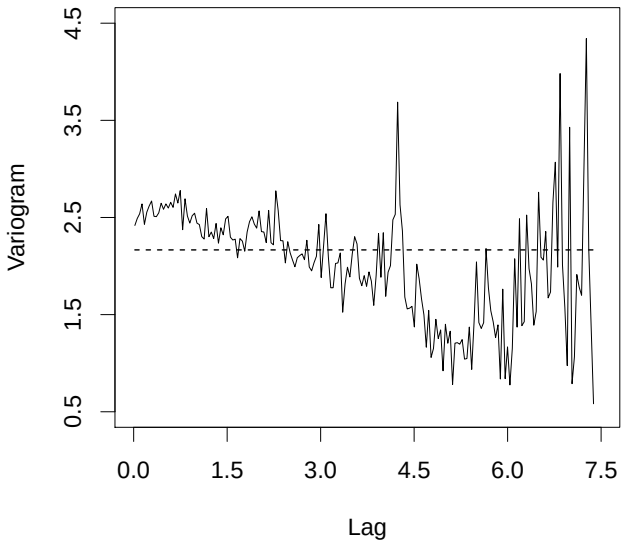

FIGURE 4.7: Empirical variogram based on the standardised residuals against the lag based on the transformed time-scale. The variogram ordinates are averaged over bins with width 14 days. Bins with fewer than 30 residuals are omitted. The dashed line at 2.1720 is the theoretical variogram of the standardised residuals under $t(3.7065)$.

here) collectively suggest that the discrepancies can well be explained by randomness.

4.6 Simulation assessment

We conduct a simulation study to inspect the finite-sample behaviours of the estimators. Data are generated under the multivariate t model, with $\kappa = 0.5$. Explanatory variables of the CRISIS data are kept same, i.e. there are 2,289 patients with 38,829 repeated measurements, 15 explanatory variables, etc. Only the random components of the model are simulated. Parameters are set to those estimated based on the analysis of CRISIS data set (Table 4.4). The simulated data sets are then analysed by the multivariate normal and t models. Means of the parameter estimates, percentage biases, standard deviations of the parameter estimates, means of the asymptotic standard error estimates, and the coverage probabilities at 95% confidence level, based on 1,000 simulation replications, are reported in Table 4.5.

Fixed effects parameter estimates for both the multivariate normal and t models are essentially unbiased. Within each of these models, standard deviations of the fixed effects parameter estimates and the means of the asymptotic standard errors are close to each other. Also, for both of these models, coverage probabilities are close to the nominal

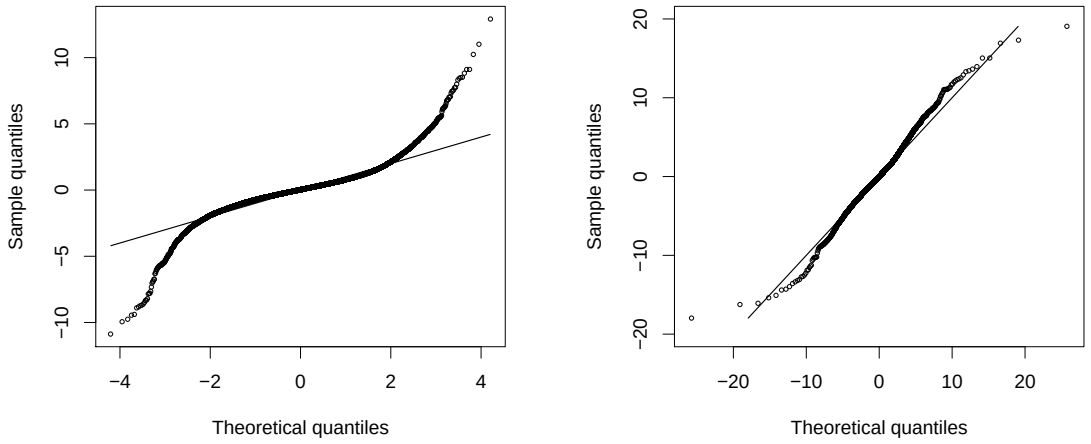

FIGURE 4.8: Quantile-quantile plot based on the standardised residuals under multivariate Normal (left panel) and multivariate t (right panel) models. Straight line on each plot is the line of equality ($y = x$).

rate of 95%. The only difference between these models in terms of estimating the fixed effects parameters is in the magnitude of the standard errors: the standard deviations of the parameter estimates, as well as the means of the asymptotic standard error estimates, are inexorably smaller for the multivariate t model. This feature is also observed in the analysis of the CRISIS data set. Estimates of the random effects parameters and degrees-of-freedom are essentially unbiased under the multivariate t model, and standard deviations and the means of the asymptotic standard errors are close to each other for these parameters. Whilst the coverage probabilities are close to the nominal rate for ω^2, τ^2 and ν, they are sightly lower than the nominal rate for σ^2 and ϕ; 91.0% and 91.2% for the latter parameters, respectively. Here, we shall note that the standard error estimates for these parameters are obtained numerically by central difference approximation to the Hessian matrix. Although the random effects parameter estimates are essentially unbiased under the multivariate normal model, standard errors of the parameter estimates are much larger than the means of the asymptotic standard error estimates. Moreover, coverage probabilities for these parameters are much lower than the nominal rate. Here, we shall note that true values of ω^2, σ^2 and τ^2 for multivariate normal model are considered to be 0.2770 ($= 0.1275 * \frac{3.7065}{3.7065-2}$), 0.2882 ($= 0.1327 * \frac{3.7065}{3.7065-2}$), and 0.0119 ($= 0.0055 * \frac{3.7065}{3.7065-2}$), respectively, since these values are the underlying variances of the random effects.

TABLE 4.5: Simulation results for 2,289 patients based on 1,000 replications for the multivariate Normal and t models. Means of the parameter estimates (Mean), percentage biases (Bias%), standard deviations of the parameter estimates (SE), means of the asymptotic standard error estimates (meSE), and percentage coverage probabilities at 95% confidence level (CP%) are reported.

		Normal					t				
Parameter	True	Mean	Bias%	SE	meSE	CP%	Mean	Bias%	SE	meSE	CP%
α_0	3.5674	3.5689	0.0424	0.0744	0.0738	94.9	3.5692	0.0515	0.0547	0.0538	94.8
α_1	-0.0366	-0.0366	-0.1079	0.0030	0.0030	94.6	-0.0366	0.0787	0.0022	0.0021	94.8
α_2	-0.4070	-0.4068	-0.0375	0.0355	0.0355	95.4	-0.4072	0.0703	0.0252	0.0251	95.3
α_3	0.7088	0.7063	-0.3639	0.1113	0.1099	93.9	0.7084	-0.0639	0.0773	0.0774	94.9
α_4	-0.3595	-0.3567	-0.7766	0.1034	0.1013	94.3	-0.3585	-0.2622	0.0718	0.0712	95.2
α_5	-0.7264	-0.7259	-0.0632	0.0341	0.0336	94.3	-0.7259	-0.0675	0.0243	0.0241	94.8
α_6	1.3221	1.3221	0.0002	0.1459	0.1497	95.4	1.3239	0.1412	0.0983	0.1058	95.9
α_7	-0.6421	-0.6406	-0.2344	0.1495	0.1551	95.5	-0.6435	0.2121	0.1018	0.1094	96.1
α_8	-0.0042	-0.0043	0.6044	0.0011	0.0011	94.4	-0.0043	0.7383	0.0008	0.0008	94.6
α_9	0.1871	0.1861	-0.5315	0.0319	0.0330	95.5	0.1865	-0.3525	0.0230	0.0238	96.6
α_{10}	0.0475	0.0486	2.3457	0.0329	0.0323	94.4	0.0480	1.1595	0.0242	0.0235	94.3
α_{11}	-0.0177	-0.0168	-5.1203	0.0331	0.0329	95.1	-0.0170	-4.0960	0.0246	0.0239	94.4
α_{12}	0.0533	0.0515	-3.3548	0.0312	0.0311	94.8	0.0523	-1.8396	0.0225	0.0226	95.7
α_{13}	-0.1169	-0.1161	-0.6715	0.0320	0.0330	95.2	-0.1156	-1.1254	0.0239	0.0240	95.5
α_{14}	0.0065	0.0064	-2.6535	0.0382	0.0397	96.1	0.0061	-6.9291	0.0287	0.0289	95.4
ω^2 †	0.1275	0.2536	-8.4335	0.0884	0.0219	41.5	0.1295	1.5455	0.0178	0.0188	92.9
σ^2 †	0.1327	0.3069	6.4888	0.0882	0.0146	29.7	0.1297	-2.2283	0.0169	0.0176	91.0
ϕ	7.1196	7.7102	8.2950	2.2597	0.2962	22.8	6.9463	-2.4354	0.9599	0.9987	91.2
τ^2 †	0.0055	0.0118	-0.7145	0.0013	0.0001	22.1	0.0055	1.1203	0.0001	0.0001	93.7
ν	3.7065						3.7161	0.2598	0.1306	0.1293	94.8

† denotes that true values for these parameters are considered to be 0.2770 ($= 0.1275 * \frac{3.7065}{3.7065-2}$), 0.2882 ($= 0.1327 * \frac{3.7065}{3.7065-2}$), and 0.0119 ($= 0.0055 * \frac{3.7065}{3.7065-2}$), respectively, for the multivariate normal model.

4.7 Discussion

In this study, we have investigated the influences of AKI on long-term kidney function using longitudinal statistical modelling. The serial dependence is specified flexibly by the Matérn correlation family. A robust distribution, specifically multivariate t distribution, is considered as the distribution of the random components of the model. Normal-gamma hierarchical representation of the distribution helps us to easily apply the E-M algorithm. We specify robust distributions for the random components, since diagnostics based on the multivariate normal model indicated a bad fit. Possible explanations for outlying observations in the CRISIS data set are low eGFR values in the recovery period, high eGFR values in the same period due to very quick recovery, abrupt losses in kidney function that occur more than seven days apart, e.g. non-acute renal problems, or simply mistakes in data-entry. We considered three change-points in the longitudinal trajectories of the patients who experienced AKI events. The change-points were identified by exploratory methods. We mainly have found that occurrence of AKI events are preceded by significant accelerations in the kidney function losses. The patients had quick recovery after having the AKI events. After the recovery-period, they started to lose kidney function with a greater magnitude compared to the pre-acceleration period. Overall, we can say that occurrence of AKI might have serious impact on long-term kidney function.

The results in this study are based on observational data. Clinical trials would be needed to have more reliable results. We considered two of the change-points are shared across patients within AKI stages, then found these by exploratory methods. A better method would be profile likelihood based on the multivariate t model. When the interest is on individual change-points Bayesian methods would be more preferable, but when the interest is on shared change-points, there is little to choose between Bayesian methods and profile likelihood (Hall *et al.*, 2003). In this study, we only considered longitudinal modelling and ignored the influence of survival data. For example, in the CRISIS data set, 305 (13.3%) of the patients had second AKI, 64 (2.8%) died due to renal related reasons, 441 (19.3%) died due to other reasons, 1,479 (64.6%) were administratively censored. Here, death due to renal reasons implies potentially informative drop-out. To test this empirically, we fitted the multivariate t model to the CRISIS data set without the patients whose reason for death is renal related. However, we observed that the resulting estimates and standard errors are not considerably different than the ones presented in the current paper. Potential future work is the incorporation of survival information to the model presented in this study, as competing risks. Also, we only investigated the influences of the first AKI events. The influence of multiple AKI events

can also be investigated. Risk of having AKI events might be explored possibly with joint modelling of longitudinal eGFR measurements and time to recurrent AKI events.

4.8 Supplementary material: R codes

```
# R codes for parameter estimation and prediction of the random effects

# loading the package, version 1.1
R> library(lmenssp)

# reading the data set into R
R> crisis.data <- read.csv("crisis.censor2.omit.txt", header = T)

# formula to be used parameter estimation and smoothing
# for the explanations of the variables see Table 4.4
R> formula = log.eGFR ~ fu + pwl1.stage1 + pwl3.stage1 + pwl4.stage1 +
+                       pwl2.stage23 + pwl3.stage23 + pwl5.stage23 +
+                       age.0 + salford + male + smoking +
+                       alcohol + diabetes + mcp

# obtaining the parameter estimates
R> fit.heavy.exp <- lmenssp.heavy(formula = formula, data = crisis.data,
+                                 id = crisis.data$id,
+                                 process = "sgp-matern-0.5",
+                                 timeVar = crisis.data$fu, init.em = 5,
+                                 tol.em = 1e-5, maxiter.em = 1000,
+                                 silent = FALSE)
R> fit.heavy.exp

# obtaining the predictive distributions of [U_i|Y_i],
# [W_i(t_{ik})|Y_i], for patient with ID = 1

R> smooth <- smoothed.heavy(formula = formula, data = crisis.data,
+                     id = crisis.data$id, process = "sgp-matern-0.5",
+                     timeVar = crisis.data$fu,
+                     estimate = fit.heavy.exp$estimate, subj.id = 1)
R> smooth
```

Appendix: Sensitivity analysis on patients with no AKI

In the previous analyses, presented in Chapter 4.4, we assumed that the data for the patients with no AKI events (1,576 of them) were all on the pre-acceleration period. Nonetheless, for these patients, some portion of the data from the end might fall into the acceleration period. Ignoring this might add biases to the parameter estimates, e.g. population-averaged slope estimates. To inspect this, we conduct a sensitivity analysis, in which we exclude some amount of data from the end of data recording, i.e. data points falling into $[t_{n_i} - e, t_{n_i}]$, for the patients with no AKI events. We specifically considered $e = 0.04$ years (≈ 15 days), 0.08 years (≈ 30 days), 0.12 years (≈ 45 days), and 0.16 years (≈ 60 days). If a patient has a total follow-up time less than these values, i.e. $t_{n_i} \leq e$, we only keep the baseline data for that patient. Respective number of observations excluded for the above values of e are 1,801 (4.6%), 2,024 (5.2%), 2,215 (5.7%), 2,429 (6.3%), out of 38,829 repeated measurements. We then re-analyse these data sets by the multivariate normal and t models. Results based on the multivariate t model are presented in Tables 4.6 and 4.7. There are negligible changes in the parameter estimates compared to the results presented in Table 4.4. Results based on the multivariate normal model are not reported here, since the changes are based on this model are also negligible.

TABLE 4.6: Maximum likelihood estimates of the model parameters based on $\kappa = 0.5$ for the t model when $e = 0.04$ & 0.08.

		e = 0.04			e = 0.08		
Variable	Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	α_0	3.5507	0.0537	<0.0001	3.5508	0.0537	<0.0001
t_{ij}	α_1	-0.0399	0.0023	<0.0001	-0.0397	0.0023	<0.0001
$\max(0, t_{ij} - (c_i - 0.3))$ * Stage1	α_2	-0.4033	0.0255	<0.0001	-0.4034	0.0254	<0.0001
$\max(0, t_{ij} - c_i)$ * Stage1	α_3	0.7079	0.0782	<0.0001	0.7081	0.0780	<0.0001
$\max(0, t_{ij} - (c_i + 0.1))$ * Stage1	α_4	-0.3586	0.0720	<0.0001	-0.3589	0.0718	<0.0001
$\max(0, t_{ij} - (c_i - 0.4))$ * Stage23	α_5	-0.7221	0.0244	<0.0001	-0.7223	0.0244	<0.0001
$\max(0, t_{ij} - c_i)$ * Stage23	α_6	1.3218	0.1070	<0.0001	1.3221	0.1067	<0.0001
$\max(0, t_{ij} - (c_i + 0.15))$ * Stage23	α_7	-0.6437	0.1107	<0.0001	-0.6439	0.1103	<0.0001
Baseline age	α_8	-0.0041	0.0008	<0.0001	-0.0041	0.0008	<0.0001
Base hospital	α_9	0.1981	0.0238	<0.0001	0.1982	0.0238	<0.0001
Gender	α_{10}	0.0460	0.0235	0.0506	0.0454	0.0235	0.0534
Smoking	α_{11}	-0.0167	0.0239	0.4858	-0.0149	0.0239	0.5325
Alcohol	α_{12}	0.0519	0.0226	0.0219	0.0512	0.0226	0.0236
Diabetes	α_{13}	-0.1141	0.0240	<0.0001	-0.1173	0.0239	<0.0001
Co-morbidity	α_{14}	0.0008	0.0289	0.9787	0.0014	0.0288	0.9615
Random intercept	ω^2	0.1157	0.0144		0.1158	0.0143	
Stationary process	σ^2	0.1406	0.0129		0.1398	0.0128	
Stationary process	ϕ	7.2914	0.7013		7.3046	0.6984	
Measurement error	τ^2	0.0055	0.0001		0.0055	0.0001	
Degrees-of-freedom	ν	3.8646	0.1397		3.8804	0.1404	

TABLE 4.7: Maximum likelihood estimates of the model parameters based on $\kappa = 0.5$ for the t model when $e = 0.12$ & 0.16.

		e = 0.12			e = 0.16		
Variable	Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	α_0	3.5459	0.0537	<0.0001	3.5462	0.0536	<0.0001
t_{ij}	α_1	-0.0396	0.0023	<0.0001	-0.0396	0.0023	<0.0001
$\max(0, t_{ij} - (c_i - 0.3))$ * Stage1	α_2	-0.4034	0.0253	<0.0001	-0.4034	0.0253	<0.0001
$\max(0, t_{ij} - c_i)$ * Stage1	α_3	0.7081	0.0778	<0.0001	0.7079	0.0776	<0.0001
$\max(0, t_{ij} - (c_i + 0.1))$ * Stage1	α_4	-0.3589	0.0716	<0.0001	-0.3589	0.0714	<0.0001
$\max(0, t_{ij} - (c_i - 0.4))$ * Stage23	α_5	-0.7222	0.0243	<0.0001	-0.7220	0.0242	<0.0001
$\max(0, t_{ij} - c_i)$ * Stage23	α_6	1.3217	0.1064	<0.0001	1.3212	0.1062	<0.0001
$\max(0, t_{ij} - (c_i + 0.15))$ * Stage23	α_7	-0.6435	0.1101	<0.0001	-0.6434	0.1098	<0.0001
Baseline age	α_8	-0.0041	0.0008	<0.0001	-0.0041	0.0008	<0.0001
Base hospital	α_9	0.1972	0.0237	<0.0001	0.1970	0.0237	<0.0001
Gender	α_{10}	0.0474	0.0235	0.0437	0.0495	0.0235	0.0347
Smoking	α_{11}	-0.0118	0.0239	0.6214	-0.0113	0.0238	0.6364
Alcohol	α_{12}	0.0502	0.0226	0.0265	0.0512	0.0226	0.0234
Diabetes	α_{13}	-0.1170	0.0239	<0.0001	-0.1179	0.0239	<0.0001
Co-morbidity	α_{14}	-0.0015	0.0288	0.9575	-0.0015	0.0288	0.9571
Random intercept	ω^2	0.1162	0.0142		0.1154	0.0142	
Stationary process	σ^2	0.1389	0.0127		0.1388	0.0128	
Stationary process	ϕ	7.3000	0.6969		7.3324	0.7044	
Measurement error	τ^2	0.0055	0.0001		0.0054	0.0001	
Degrees-of-freedom	ν	3.8853	0.1406		3.8931	0.1412	

Bibliography

Anderson T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons: New York.

Arellano-Valle R. B. and Bolfarine H. (1995). On some characterizations of the t-distribution. *Statistics & Probability Letters* **25**, 79–85.

Asar O. and Diggle P. J. (2014). lmenssp: Linear Mixed Effects Models with Non-stationary Stochastic Processes, R package version 1.1, URL: http://CRAN.R-project.org/package=lmenssp.

Bellomo R., Kellum J. A. and Ronco C. (2012). Acute kidney injury. *Lancet* **380**, 756–766.

Chavla L. S. and Kimmel P. L. (2012). Acute kidney injury and chronic kidney disease: an integrated clinical syndrome. *Kidney International* **82**, 516–524.

Chawla L. S., Eggers P. W., Star R. A. and Kimmel P. L. (2014). Acute kidney injury and chronic kidney disease as interconnected syndromes. *New England Journal of Medicine* **371**, 58–66.

Cleveland W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society - Series B (Methodological)* **39**, 1–38.

Diggle P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959–971.

Diggle P. J., Heagerty P. J., Liang K.-Y. and Zeger S. L. (2002). *Analysis of Longitudinal Data, 2nd edition*. Oxford University Press: Oxford.

Diggle PJ, Ribeiro PJ Jr. (2007). *Model-based Geostatistics*. Springer-Verlag: New York.

Eckardt K. U., Coresh J., Devuyst O., Johnson R. J., Köttgen A., Levey A. S. and Levin A. (2013). Evolving importance of kidney disease: from subspecialty to global health burden. *Lancet* **382**, 158–169.

El Nahas M. and Levin A. (2009). *Chronic Kidney Disease: A practical Guide to Understanding and Management*. Oxford: Oxford University Press.

Eddington H., Hoefield R., Sinha S., Chrysochou C., Lane B., Foley R. N., Hegarty J., New J., O'Donoghue D. J., Middleton R. J. and Kalra P. A. (2010). Serum phosphate and mortality in patients with chronic kidney disease. *Clinical Journal of the American Society of Nephrology* **5**, 2251–2257.

Finlay S., Bray B., Lewington A. J., Hunter-Rowe C. T., Banerjee A., Atkinson J. M. and Jones M. C. (2013). Identification of risk factors associated with acute kidney injury in patients admitted to acute medical units. *Clinical Medicine* **13**, 233–238.

Fitzmaurice G. M., Laird N. M. and Ware J. H. (2011). *Applied Longitudinal Analysis, 2nd edition*. John Wiley & Sons: New Jersey.

Hall C. B., Ying J., Kuo L. and Lipton R. B. (2003). Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Computational Statistics & Data Analysis* **42**, 91–109.

Hoefield R. A., Kalra P. A., Baker P., Lane B., New J. P., O'Donoghue D. J., Foley R. N. and Middleton R. J. (2010). Factors associated with kidney disease progression and mortality in a referred CKD population. *American Journal of Kidney Diseases* **56**, 1072–1081.

Jha V., Garcia-Garcia G., Iseki K., Li Z., Naicker S., Plattner B., Saran R., Wang A. Y. and Yang C. W. (2013). Chronic kidney disease: global dimension and perspectives. *Lancet* **382**, 260–272.

Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group (2012). KDIGO Clinical Practice Guideline for Acute Kidney Injury. *Kidney International, Supplementary* **2**, 1–138.

Kerr M., Bedford M., Matthews B. and O'Donoghue D. (2014). The economic impact of acute kidney injury in England. *Nephrology Dialysis Transplantation* **29**, 1362–1368.

Kim N. H., Kim S. H. and Song S. W. (2014). Is alcohol drinking associated with renal impairment in the general population of South Korea? *Kidney & Blood Pressure Research* **39**, 30–39.

Kotz S. and Nadarajah S. (2004). *Multivariate t-Distributions and Their Applications*. Cambridge University Press: Cambridge.

Lameire N. H., Bagga A., Cruz D., De Maeseneer J., Endre Z., Kellum J. A., Liu K. D., Mehta R. L., Pannu N., Van Biesen W. and Vanholder R. (2013). Acute kidney injury: an increasing global concern. *Lancet* **382**, 170–179.

Levey A. S., Bosch J. P., Lewis J. B., Greene T., Rogers N. and Roth D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of Internal Medicine* **130**, 461–470.

Levey A. S. and Coresh J. (2012). Chronic kidney disease. *Lancet* **379**, 165–180.

Liu C. and Rubin D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica* **5**, 19–39.

Matérn B. (1960). *Spatial Variation*. Stockholm: Statens Skogsforsningsinstitut.

Pinheiro J. C., Liu C. and Wu Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10**, 249–276.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, URL http://www.R-project.org/.

Rosa G. J. M., Padovani C. R. and Gianola D. (2003). Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* **45**, 573–590.

Chapter 5

lmenssp: an R package for linear mixed effects models with non-stationary stochastic processes

This chapter is based on the following paper:

Asar Ö. and Diggle P. J. (2014). lmenssp: An R Package for Linear Mixed Effects Models with Non-stationary Stochastic Processes. Submitted to *Journal of Statistical Software*.

Abstract

Linear mixed models are widely used for analysis of repeated measurement data. The classical random-intercept-and-slope model is useful for datasets with short follow-up times. Models based on stationary Gaussian processes are more useful for data-sets with long follow-up times, but the stationarity assumption is too restrictive for universal application. Models based on non-stationary Gaussian processes have been proposed in the literature, but are not implemented in widely available software. In this paper, we introduce the `R` package `lmenssp` for linear mixed models whose random effect component includes a continuous-time non-stationary ergodic component. Three such processes are considered: 1) Brownian motion, 2) integrated Brownian motion, and 3) integrated Ornstein-Uhlenbeck process. Core features of `lmenssp` are illustrated using a simulated data set.

Key words: Filtering, forecasting, maximum likelihood estimation, random effects, smoothing

5.1 Introduction

Gaussian linear mixed effects models are widely used for analysis of repeated measurement data. Two examples are the random-intercept-and-slope, or Laird-Ware model (Laird and Ware, 1982) and various stationary Gaussian process models; see, for example, Diggle (1988). Both of these model classes can be fitted in readily available software, including `R` (R Development Core Team, 2014) packages, e.g., the `nlme` package (Pinheiro *et al.*, 2014).

In the Laird-Ware model, the random effect component for subject i at time t is $U_i + V_i t$. The subject-specific variation in the response variable at study entry is captured by random intercepts, U_i, whilst subsequent subject-specific variation in the longitudinal trajectories over time is captured by the random slopes, V_i. Here, V_i is time-independent, and postulates that each subject has his/her own slope, but these slopes are straight lines. In the stationary Gaussian process model, the random effect component is $U_i + S_i(t)$, where the U_i are as before and the $S_i(t)$ are independent copies of a stationary Gaussian process whose correlation function $\rho(u)$ is typically specified to lie within one of a small number of standard families, for example the exponential, $\rho(u) = \exp(-u/\phi)$, the so-called Gaussian, $\rho(u) = \exp\left(-(u/\phi)^2\right)$, or the more general Matérn family (Matérn, 1960) that includes the exponential and Gaussian as special cases. Now the random

slopes, $S_i(t)$, depend on time, hence the straight line assumption of the Laird-Ware model is relaxed.

The Laird-Ware model is useful for data sets with short sequences of repeated measurements, but is usually too inflexible to capture the covariance structure of data sets with long follow-up times. Continuous-time ergodic Gaussian processes are more suitable for data sets with long follow-up times, but the stationarity assumption is too restrictive for universal application. Several authors have proposed non-stationary models, including Taylor, Cumberland and Sy (1994), Taylor and Law (1998), and Diggle, Sousa and Asar (2014).

Models whose random effect component includes a stationary stochastic process can be fitted in `R` using the `nlme` package. However, mixed models with continuous-time non-stationary ergodic components are not available in `R`. In this paper, we introduce the `R` package `lmenssp` that fits mixed models whose random effect component can include any one of the following continuous-time non-stationary ergodic processes: 1) Brownian motion; 2) integrated Brownian motion; 3) integrated Ornstein-Uhlenbeck process.

The paper is organised as follows. In Section 5.2, we give details of the modelling framework. In Section 5.3, we discuss the core features of the `lmenssp` package. Section 5.4 gives examples of its use, and the paper ends with a brief discussion in Section 5.5.

5.2 Modelling framework

The model we consider is of the form

$$Y_{ij} = X_{ij}\alpha + U_i + W_i(t_{ij}) + Z_{ij}. \tag{5.1}$$

Here, Y_{ij} is the jth ($j = 1, \ldots, n_i$) response of the ith ($i = 1, \ldots, m$) subject at the time point t_{ij}. X_{ij} is an associated vector of $p + 1$ covariates (including intercept) and α is the corresponding vector of fixed effects regression parameters. The U_i are random intercepts, assumed to be independent $N(0, \omega^2)$ random variables. The $W_i(t_{ij})$ are continuous-time non-stationary stochastic process, the details of which are discussed below. Finally, the Z_{ij} are mutually independent $N(0, \tau^2)$ random error terms.

We consider three specifications for $W(t)$. The first is Brownian motion (BM) for which the marginal distribution is

$$[W(t)] = \mathrm{N}(0, \sigma^2 t), \tag{5.2}$$

where we use $[\cdot]$ to mean "the distribution of". The covariance function of the process is

$$\text{Cov}\left(W(s), W(t)\right) = \sigma^2 \min(s,t). \tag{5.3}$$

The second specification is integrated Brownian motion (IBM),

$$W(t) = \int_0^t B(t), \tag{5.4}$$

where $B(t)$ is now Brownian motion. The marginal distribution and covariance function of IBM are

$$[W(t)] = \text{N}\left(0, \sigma^2 \frac{t^3}{3}\right), \tag{5.5}$$

and

$$\text{Cov}\left(W(s), W(t)\right) = \sigma^2 \frac{\min(s,t)^2}{2}\left(\max(s,t) - \frac{\min(s,t)}{3}\right). \tag{5.6}$$

The cross-covariance between IBM and its related BM is

$$\begin{aligned} \text{Cov}(B(s), W(t)) &= \sigma^2 \frac{t^2}{2}, \qquad \text{if } \; s \geq t, \\ &= \sigma^2 \left(st - \frac{s^2}{2}\right), \text{if } \; s < t. \end{aligned} \tag{5.7}$$

The third specification for $W(t)$ is an integrated Ornstein-Uhlenbeck (IOU) process,

$$W(t) = \int_0^t B(t), \tag{5.8}$$

where $B(t)$ is now an Ornstein-Uhlenbeck (OU) process. The OU process has the following marginal distribution and covariance function,

$$[B(t)] = \text{N}\left(0, \frac{\kappa^2}{2\nu}\right), \tag{5.9}$$

and

$$\text{Cov}\left(B(s), B(t)\right) = \frac{\kappa^2}{2\nu} e^{-\nu|t-s|}, \tag{5.10}$$

respectively. These in turn yield the following marginal distribution and covariance structure for IOU,

$$[W(t)] = \text{N}\left(0, \frac{\kappa^2}{\nu^3}\left(\nu t + e^{-\nu t} - 1\right)\right), \tag{5.11}$$

and

$$\text{Cov}\left(W(s), W(t)\right) = \frac{\kappa^2}{2\nu^3}\left(2\nu \min(s,t) + e^{-\nu t} + e^{-\nu s} - 1 - e^{-\nu|t-s|}\right), \tag{5.12}$$

respectively. The cross-covariance function of IOU and its related OU is

$$\begin{aligned} \text{Cov}(B(s), W(t)) &= \frac{\kappa^2}{2\nu^2} e^{-\nu s}\left(e^{\nu t} - 1\right), \qquad \text{if } \ s \geq t, \\ &= \frac{\kappa^2}{2\nu^2}\left(2 - e^{-\nu s} - e^{\nu(s-t)}\right), \text{if } \ s < t. \end{aligned} \tag{5.13}$$

Note that BM can be obtained as a limiting case of IOU when κ^2/ν^2 is constant and $\nu \to \infty$.

For the details of these stochastic processes, we refer the interested reader to Taylor, Cumberland and Sy (1994); Taylor and Law (1998); Ross (1996); Robinson (2010); Diggle, Sousa and Asar (2014).

The distributional properties of U_i, $W_i(t_{ij})$ (either specified as BM, IBM or IOU) and Z_{ij} induce a multivariate normal distribution for $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$,

$$[Y_i] = \text{MVN}(X_i\alpha, V_i(\phi)),$$

where $X_i = (X_{i1}^T, \ldots, X_{in_i}^T)^T$, α is as before, and

$$V_i(\phi) = \omega^2 J_i + R_i + \tau^2 I_i. \tag{5.14}$$

Here, $\phi = \left\{\omega^2, \sigma^2, \tau^2\right\}$ for BM and IBM and $\phi = \left\{\omega^2, \kappa^2, \nu, \tau^2\right\}$ for IOU. J_i is an $n_i \times n_i$ matrix of ones, R_i is an $n_i \times n_i$ matrix structured by Cov(W(s), W(t)) depending on the specification of $W(t)$, and I_i is an $n_i \times n_i$ identity matrix. We estimate the parameters, $\theta = \{\alpha^T, \phi\}$, by maximum likelihood (ML) estimation with a Fisher-Scoring (FS) algorithm as described in Jennrich and Schluchter (1986) and Diggle, Sousa and Asar (2014). Convergence is assessed by the following criterion: $\sqrt{\{\phi^m - \phi^{m+1}\}\{\phi^m - \phi^{m+1}\}^T} <$ tolerance, where m and $m+1$ denote successive FS steps.

The following conditional distributions are of interest: $[U_i|Y_i;\theta]$, $[W_i(t_{ik})|Y_i^k;\theta]$, $[W_i(t_{ik})|Y_i;\theta]$, $[B_i(t_{ik})|Y_i^k;\theta]$ and $[B_i(t_{ik})|Y_i;\theta]$, where $Y_i^k = (Y_{i1}, \ldots, Y_{ik})^T$. The second and fourth of these are needed for filtering, the third and fifth for smoothing and forecasting of the stochastic processes in the sense of Durbin and Koopman (2012). The general forms of these distributions are obtained from multivariate Normal theory (Anderson, 1984) as follows:

$$[U_i|Y_i;\theta] = \text{N}\left(K_i^T V_i^{-1}(Y_i - X_i\alpha), \omega^2 - K_i^T V_i^{-1} K_i\right), \tag{5.15}$$

$$[W_i(t_{ik})|Y_i^k;\theta] = \text{N}\left(F_i^{kT} V_i^{k-1}\left(Y_i^k - X_i^k\alpha\right), \text{Var}\left(W_i(t_{ik})\right) - F_i^{kT} V_i^{k-1} F_i^k\right) \tag{5.16}$$

$$[W_i(t_{ik})|Y_i;\theta] = \text{N}\left(F_i{}^T V_i{}^{-1}\left(Y_i - X_i\alpha\right), \text{Var}\left(W_i(t_{ik})\right) - F_i{}^T V_i{}^{-1} F_i\right), \tag{5.17}$$

and, for the IBM and IOU processes,

$$[B_i(t_{ik})|Y_i^k;\theta] = \mathrm{N}\left(L_i^{k^T} V_i^{k^{-1}} \left(Y_i^k - X_i^k\alpha\right), \mathrm{Var}\left(B_i(t_{ik})\right) - L_i^{k^T} V_i^{k^{-1}} L_i^k\right) \quad (5.18)$$

$$[B_i(t_{ik})|Y_i;\theta] = \mathrm{N}\left(L_i{}^T V_i{}^{-1}\left(Y_i - X_i\alpha\right), \mathrm{Var}\left(B_i(t_{ik})\right) - L_i{}^T V_i{}^{-1} L_i\right). \quad (5.19)$$

Here, $K_i = \omega^2 1_i$, where 1_i is an $n_i \times 1$ matrix and V_i^k is the variance-covariance matrix of Y_i^k. F_i^k and F_i are structured by $\mathrm{Cov}(W(s), W(t))$, and L_i^k and L_i by $\mathrm{Cov}(B(s), W(t))$, depending on the specification of $B(t)$ and $W(t)$.

5.3 The package `lmenssp`

The function `lmenssp` obtains the ML estimates of θ. The default version of the function has the following form:

```
R> lmenssp(formula, data = NULL, id, process = "bm", timeVar, init = NULL,
+    tol = 1e-5, maxiter = 100, silent = TRUE)
```

Here, `formula` is a standard `R` formula for fixed effects component of the model. `data` is a data frame from which the variables are to be extracted. `id` is a vector for subject identification. `process` is a character string for the stochastic process, where `"bm"`, `"ibm"` and `"iou"` correspond to BM, IBM and IOU, respectively. `"timeVar"` is a vector for the time variable. `"init"` is a vector of initial values to start the FS algorithm. If the user does not provide their own values, `lmenssp` obtains these by fitting a random intercept and random slope model, using the `lme` function of the `nlme` package. `tol` is the maximum tolerance value to assess the convergence. `maxiter` is the maximum number of iterations for the FS algorithm. `silent` is a logical variable; if set to `FALSE` the details of the FS steps are to be printed whilst the algorithm is running.

The distributions (5.16) and (5.18) are obtained by the function `filtered`, the distributions (5.17) and (5.19) by the function `smoothed`. In each case, estimated parameter values are used for plug-in prediction. The default forms of these functions are given below:

```
R> filtered(formula, data = NULL, id, process = "bm", timeVar,
    estimate, subj.id)
R> smoothed(formula, data = NULL, id, process = "bm", timeVar,
    estimate, subj.id = NULL, fine = NULL, eq.forec = NULL,
    uneq.forec = NULL)
```

The explanations for `formula`, `data`, `id`, `process` and `timeVar` are the same as those for the function `lmenssp`. `estimate` is a vector for the ML estimates produced by the `lmenssp` function. `subj.id` is a vector for the IDs of the subjects for whom the aforementioned distributions are to obtained. `fine` is a numerical value for smoothing at fine intervals within the follow-up period. `eq.forec` is a two-element vector for forecasting at equally spaced time intervals, where whilst the first element corresponds to the forecast time interval, the second element corresponds to the number of forecasts. `uneq.forec` is a two-column data frame or matrix for forecasting at unequally spaced time intervals, where the first column includes the IDs, the second column includes the forecast time intervals.

`lmenssp`, `filtered` and `smoothed` return their results as lists as described in Section 5.4.

5.4 Examples

The package includes a longitudinal data set, `data.sim.ibm`. The data set was simulated under a mixed effects model with IBM as the stochastic component. It includes data for 500 subjects with total of 8,462 repeated measurements. There are 6 variables: 1) `id`: identification number of the subjects, 2) `sex`: sex of the subjects taking 0 for male, 1 for female, 3) `bage`: baseline age in years, 4) `fu`: follow-up time in years, 5) `pwl`: piecewise linear term, calculated as $\max(0, \text{age at measurement} - 56.5)$, 6) `log.egfr`: the response variable, representing the logarithm of estimated glomerular filtration rate. For the details of these variables and the methods for data simulation, see `diggle14`.

The ML estimates of the parameters based on the mixed model with IBM as the process can be obtained by the following script:

```
R> library("lmenssp")
R> data("data.sim.ibm")
R> formula <- log.egfr ~ sex + bage + fu + pwl
R> fit.ibm <- lmenssp(formula = formula, data = data.sim.ibm,
+    id = data.sim.ibm$id, process = "ibm", timeVar = data.sim.ibm$fu,
+    silent = FALSE)
```

The related output includes a list of results which can be printed by

```
R> fit.ibm
```

```
$title
[1] "Mixed effects model with random intercept and integrated Brownian motion"

$date
[1] "Tue Oct 21 19:34:19 2014"

$estimates
              Estimate Standard error Z-estimate      p-value
Intercept  4.432030443   0.1468746545 30.1755974 0.0000000000
sex       -0.110998143   0.0336960921 -3.2940954 0.0009873897
bage      -0.001063977   0.0028473161 -0.3736773 0.7086444131
fu        -0.009685761   0.0078669187 -1.2312014 0.2182475462
pwl       -0.012347446   0.0039634284 -3.1153447 0.0018373009
omegasq    0.118806815   0.0084673249         NA           NA
sigmasq    0.017784990   0.0012323705         NA           NA
tausq      0.051566781   0.0008603949         NA           NA

$maxloglik
[1] 6275.454

$score
          omegasq    sigmasq     tausq
[1,] -0.008013553 -0.6852443 0.2547693
```

In the output, `maxloglik` is the value of the maximised log-likelihood, and `score` gives the values of the gradient (first partial derivatives of the log-likelihood function), evaluated at the ML estimates of the parameters.

Filtering and smoothing distributions for the subjects, e.g., the ones with ID=1 and 2, can be obtained by the following scripts:

```
R> subj.id <- c(1, 2)
R> fil.res <- filtered(formula = formula, data = data.sim.ibm,
+    id = data.sim.ibm$id, process = "ibm",  timeVar = data.sim.ibm$fu,
+    estimate = fit.ibm$estimate[, 1], subj.id = subj.id)
R> smo.res <- smoothed(formula = formula, data = data.sim.ibm,
+    id = data.sim.ibm$id, process = "ibm", timeVar = data.sim.ibm$fu,
+    estimate = fit.ibm$estimate[, 1], subj.id = subj.id)
```

These two functions produce similar output. For example, the output of `filtered` can be printed by

```
R> fil.res

$title
[1] "Filtering for the mixed model with integrated Brownian motion"

$date
[1] "Tue Oct 21 19:37:19 2014"

$u
     id        mean    variance
[1,]  1  0.03032635 0.011268446
[2,]  2 -0.27005285 0.005422657

$w
      id       time           mean     variance
 [1,]  1 0.00000000  0.000000e+00 0.000000e+00
 [2,]  1 0.15331964 -1.098346e-04 2.136091e-05
 [3,]  1 0.68993840  1.052777e-02 1.898799e-03
...
[16,]  2 0.00000000  0.000000e+00 0.000000e+00
[17,]  2 0.03011636  1.960017e-07 1.619343e-07
[18,]  2 0.03559206  6.810425e-07 2.672946e-07
...

$b
      id       time           mean     variance
 [1,]  1 0.00000000  0.000000e+00 0.0000000000
 [2,]  1 0.15331964 -1.074565e-03 0.0027262892
 [3,]  1 0.68993840  2.350504e-02 0.0120395202
...
[16,]  2 0.00000000  0.000000e+00 0.0000000000
[17,]  2 0.03011636  9.762222e-06 0.0005356184
[18,]  2 0.03559206  2.855346e-05 0.0006330028
...
```

In the output, `mean` and `variance` are the mean and variance of the corresponding distribution.

Smoothing also can be carried out at time points within the follow-up period at which measurements were not taken. For example, for subjects with ID=1 and 2, smoothing at fine intervals of 0.01 years are obtained by

```
R> smo.within <- smoothed(formula = formula, data = data.sim.ibm,
+    id = data.sim.ibm$id, process = "ibm", timeVar = data.sim.ibm$fu,
+    estimate = fit.ibm$estimate[, 1], subj.id = subj.id, fine = 0.01)
```

On the other hand, forecasting with lead time u, i.e. at $t_{in_i} + u$ can be carried out using the `smoothed` function. This might be done either at equally spaced intervals with the same number of forecasts for each subject by using the `eq.forec` option, or at unequally spaced intervals with equal or varying number of forecasts for different subjects by using the `uneq.forec` option. For example, whilst one, two and three month ahead forecasts for the subjects with ID=1 and 2 are obtained by

```
R> eq.forecast <- smoothed(formula = formula, data = data.sim.ibm,
+    id = data.sim.ibm$id, process = "ibm", timeVar = data.sim.ibm$fu,
+    estimate = fit.ibm$estimate[, 1], subj.id = subj.id,
+    eq.forec = c(1/12, 3))
```

one, two and six month ahead forecasts for the subject with ID=1 and one and three month ahead forecasts for the subject with ID=2 are obtained by

```
R> uneq.forec <- data.frame(c(1, 1, 1, 2, 2), c(1/12, 2/12, 6/12, 1/12, 3/12))
R> uneq.forecast <- smoothed(formula = formula, data = data.sim.ibm,
+    id = data.sim.ibm$id, process = "ibm", timeVar = data.sim.ibm$fu,
+    estimate = fit.ibm$estimate[, 1], uneq.forec = uneq.forec)
```

Filtering and smoothing (as well as forecasting) for new subjects whose data are not included in estimation of the model parameters can also be carried out. For example for a new (hypothetical) subject, filtering is carried out by the following scripts

```
R> data.501   <- data.frame(id = c(501, 501, 501), sex = c(0, 0, 0),
+   bage = c(50, 50, 50), fu = c(0, 0.2, 0.4),
+   pwl = c(0, 0, 0), log.egfr = c(4.3, 2.1, 4.1))
R> fil.501 <- filtered(formula = formula, data = data.501,
+   id = data.501$id, process = "ibm", timeVar = data.501$fu,
+   estimate = fit.ibm$estimate[, 1], subj.id = 501)
```

5.5 Discussion

In this paper, we introduced the `R` package `lmenssp` for linear mixed effects models with three different structures for a non-stationary stochastic component, BM, IBM and IOU. The package includes three main functions: `lmenssp` for ML estimates of the model parameters; `filtered` for filtering, `smoothed` for smoothing. Filtering and smoothing are carried out for the current level of the stochastic process, $W(t)$, as well as the rate of change, $B(t)$, since they are both of scientific interest in some studies, e.g. see Diggle, Sousa and Asar (2014). The function `variogram` in `lmenssp`, which is not discussed in the current paper, but illustrated in the package manual, can be used to calculate the empirical variogram, which is a useful tool for deciding the use of non-stationary stochastic processes, e.g., when the variogram does not level-off (Diggle *et al.*, 2002).

A potential extension is to replace random intercept U_i in (5.1) by a random effect $D_{ij}b_i$ where D_{ij} is typically a subset of X_{ij} and b_i is a corresponding vector of random effects, assumed to follow a zero-mean multivariate Normal distribution.

Bibliography

Anderson T. W. (1984). *An introduction to multivariate statistical analysis*, 2nd edition. New York: John Wiley & Sons.

Diggle P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959–971.

Diggle P. J., Sousa I. and Asar O. (1988). Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics* 1–15, doi:10.1093/biostatistics/kxu053.

Diggle P. J., Heagerty P., Liang K.Y. and Zeger S. L. (2002). *Analysis of longitudinal data*, 2nd edition. Oxford: Oxford University Press.

Durbin J. and Koopman S. J. (2012). *Time series analysis by state space methods*, 2nd edition. Oxford: Oxford University Press.

Jennrich R. I. and Schluchter M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Matérn B. (1960). *Spatial Variation*. Stockholm: Statens Skogsforsningsinstitut.

Pinheiro J., Bates D., DebRoy S., Sarkar D. and the R Development Core Team (2014). nlme: linear and nonlinear mixed effects models, R package version 3.1-117, http://CRAN.R-project.org/package=nlme.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, URL http://www.R-project.org/.

Robinson G. K. (2010). Continuous time Brownian motion models for analysis of sequential data. *Journal of the Royal Statistical Society - Applied Statistics* **59**, 477–494.

Ross S. M. (1996). *Stochastic Processes*, 2nd edition. New Jersey: John Wiley & Sons.

Taylor J. M. G., Cumberland W. G. and Sy J. P. (1994). A stochastic process model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**, 727–736.

Taylor J. M. G. and Law N. (1998). Does the Covariance Structure Matter in Longitudinal Modelling for the Prediction of Future CD4 Counts? *Statistics in Medicine* **17**, 2381–2394.

Chapter 6

General discussion and future works

In this thesis we developed statistical methods and associated software that are motivated by public health renal problems. Additionally, we did a case study on renal medicine which we presented as an educational material that aims at both statisticians and non-statisticians.

In Chapter 2, we presented the educational material on joint analysis of longitudinal and survival data. Methods were illustrated on a longitudinal data set on chronic kidney disease patients. Such a study is important in terms of encouraging the use of new statistical methods amongst practitioners of statistics.

In Chapter 3, we offered a probabilistic solution to a prediction problem concerning referral of primary care patients to secondary care. The methods consider the following clinical criterion: patients losing kidney function at least at a relative rate of 5% should be referred to secondary care. We considered a large data set which belongs to patients who were flagged as having pre-disposing conditions for renal failure. This is an important study, since if the patients can be appropriately referred to secondary care, the time to serious renal problems, e.g. renal failure, and the need for serious treatments, e.g. renal replacement therapy, might be postponed.

In Chapter 4, we investigated the influence of AKI on long-term kidney health amongst chronic kidney disease patients. AKI is a medically well studied phenomenon, but its influence on kidney function is still unknown. We considered data-driven statistical methods to inspect the influence. The model is a linear mixed effects model with a stationary stochastic process based on multivariate t-distributed repeated measures. This is an important study, since understanding the impact of it might lead doctors to take necessary precautions to prevent its recurrence and other related complications.

In Chapter 5, we developed the `R` package `lmenssp` (version 1.0) to estimate the model parameters for mixed effects models with non-stationary stochastic process components. Predictions based on filtering and smoothing distributions and forecasting are also considered. The core features of the package are illustrated by a simulated data set that is similar to the primary care data set used in Chapter 3. We have added to `lmenssp` in the version 1.1 parameter estimation, prediction and forecasting for mixed models with stationary processes, and mixed models with stationary and non-stationary process with a multivariate t distribution.

The methods that are developed in Chapter 3 and 4 ignore the survival information. A natural extension of the methods presented in these chapters is joint modelling, which is on-going. Inference for joint models with stochastic processes is computationally cumbersome because of the high-dimensional integrals in the E-steps when the E-M scheme is opted for parameter estimation, and the requirement of the continuous path

of the stochastic processes in the survivor function integral. Monte-Carlo methods are typically needed to approximately solve the former integrals. The latter is not a problem when the baseline hazard is left unspecified, since the value of the process is only needed at the event times. However, some form of approximation, e.g. discretisation, is needed for parametric baseline hazard options. Similarly, we assumed in these studies that the follow-up time is non-informative. This might not be true for observational studies, since the timing of the follow-up times are not pre-specified and might depend on the underlying kidney function. We are currently working on an extension of these methods to informative follow-up times with a joint modelling approach for simultaneous analysis of the observation and follow-up processes (Ryu *et al.*, 2007; Diggle *et al.*, 2010). To convey our message in Chapter 3 to physicians, a more detailed case-study on a more up-to-date data is needed. We are currently working on such a project. We aim to develop a real-time surveillance system to be used by general practitioners as a result of this study. In Chapter 4, we detected AKI events retrospectively based on SCr measurements using a clinical guideline. Observed SCr data are noisy and might add uncertainty to the AKI detection. However, under the current circumstances, this is the best we can do, since 'real' AKI events are typically unrecorded. In most of the cases only the primary disease that drives AKI are recorded, and the rough record rate of AKI events is one in three. Currently, we have been checking the correctness of the AKI events we detected and the underlying causes of these events, through the paper records of the SRFT. In the same chapter, we only investigated the influences of the first AKI events on kidney function. Future work regarding the influnce of AKI on long-term kidney function is the incorporation of these new features together with the underlying cause of CKD to our analyses. In addition to these studies, which are related to renal medicine, we have started working on a new project on osteoarthritis. Our aim in this study is to develop spectral analysis methods for replicated non-stationary high-frequency time series data (Diggle and Al Wasel, 1997). Two projects on mental health randomised clinical trials are also other on-going projects of ours.

Bibliography

Diggle P. J. and Al Wasel I. (1997). Spectral analysis of replicated biomedical time series (with discussion). *Journal of the Royal Statistical Society - Applied Statistics* **46**, 31–71.

Diggle P. J., Menezes R. and Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society- Applied Statistics* **59**, 191–232.

Ryu D., Sinha D., Mallick B., Lipsitz S. R. and Lipshultz S. (2007). Longitudinal studies with outcome-dependent follow-up: models and Bayesian regression. *Journal of the American Statistical Association* **102**, 952–961.

Curriculum Vitae

(Restricted to October 2012 - June 2015)

Personal Information

Name	:	Özgür Asar
Gender	:	Male
Date of Birth	:	28 November 1986
Office Address	:	B 73, CHICAS, Lancaster Medical School Faculty of Health and Medicine, Lancaster University Lancaster, LA1 4YG, United Kingdom
Office phone	:	+44 (0) 1524-593519
E-mail	:	o.asar@lancaster.ac.uk
Personal web page	:	www.lancaster.ac.uk/pg/asar/

Publications

1. Diggle PJ, Sousa I, **Asar Ö** (2015) Real time monitoring of progression towards renal failure in primary care patients. Biostatistics, 16, 522–536.

2. **Asar Ö**, Ritchie J, Kalra P, Diggle PJ (2015) Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. International Journal of Epidemiology, 44, 334–344.

3. **Asar Ö**, Diggle PJ (2014) lmenssp: an R package for linear mixed effects models with non-stationary stochastic processes. Under review in Journal of Statistical Software.

4. **Asar Ö**, Diggle PJ, Ritchie J, Kalra P (2015) Acute kidney injury amongst chronic kidney disease patients: a case-study in statistical modelling. Under review in Statistics in Medicine.

Software

1. **Asar Ö**, Diggle PJ (2014) lmenssp: an R package for linear mixed effects models with non-stationary stochastic processes. R package version 1.0. CRAN link: http://CRAN.R-project.org/package=lmenssp.

Oral Presentations

1. **Asar Ö**, James Ritchie, Philip Kalra, Diggle P (2015) Acute kidney injury amongst chronic kidney disease patients: a case-study in statistical modelling. The 2015 International Conference of the Royal Statistical Society, September 7 - 10, Exeter, UK.

2. Diggle P, Sousa I, **Asar Ö** (2014) Real-time monitoring of progression towards renal failure in primary care patients. The Farr Institute PhD Symposium 2015, June 9 - 10, Manchester, UK.

3. **Asar Ö**, Diggle P, James Ritchie, Philip Kalra (2015) A longitudinal modelling case study in renal medicine and an associated R package. Eastern North American Region of International Biometric Society (ENAR-IBS) 2015 Spring Meeting, Miami, USA, 15-18 March.

4. Diggle P, Sousa I, **Asar Ö** (2014) Real-time monitoring of progression towards renal failure in primary care patients. The 2014 International Conference of the Royal Statistical Society, September 1 - 4, Sheffield, UK.

5. Diggle P, Sousa I, **Asar Ö** (2014) Real-time monitoring of progression towards renal failure in primary care patients. XXVII International Biometric Conference (IBC), Florence, Italy, 5-11 July.

6. **Asar Ö**, Diggle P, Sousa I (2014) Investigating longitudinal measurement of kidney function to predict hazard for renal replacement therapy and death. Survival Analysis for Junior Researchers Conference, 3-4 April, University of Warwick, Coventry, UK.

7. **Asar Ö**, Diggle P, Ritchie J, Kalra P (2014) Dynamic modelling of kidney function with interventions at acute kidney injury occurrences. Medical Research Council Conference on Biostatistics, 24-26 March, pp. 6, University of Cambridge, Cambridge, UK.

8. Diggle P, Sousa I, **Asar Ö** (2014) Real-time monitoring of progression towards renal failure in primary care patients. Eastern North American Region of International Biometric Society (ENAR-IBS) 2014 Spring Meeting, pp. 181, Baltimore, USA, 16-19 March.

9. **Asar Ö**, Diggle P (2013) Another joint model for longitudinal and survival data. The 36th Research Students Conference in Probability, Statistics, and Social Statistics, pp. 53, 25-28 March, Lancaster, UK.

Poster Presentations

1. **Asar Ö**, James Ritchie, Philip Kalra, Diggle P (2015) Acute kidney injury amongst chronic kidney disease patients: a case-study in statistical modelling. Workshop on Flexible Models for Longitudinal and Survival Data with Applications in Biostatistics, July 27 - 29, University of Warwick, Coventry, UK.

2. Diggle P, Sousa I, **Asar Ö** (2014) Real-time monitoring of progression towards renal failure in primary care patients. Statistical Analysis of Multi-Outcome Data Workshop, 30 June - 1 July, Cambridge, UK.

3. **Asar Ö**, Ritchie J, Kalra P, Diggle P (2014) Joint analysis of longitudinal eGFR measurements and time to renal replacement therapy and death in a CKD cohort. UK Kidney Week 2014, 29 April - 2 May, 2014, Glasgow, UK.

4. Diggle P, Sousa I, **Asar Ö** (2014) Real-time monitoring of progression towards renal failure in primary care patients. UK Kidney Week 2014, 29 April - 2 May, 2014, Glasgow, UK.

5. **Asar Ö**, Ritchie J, Diggle P, Kalra P, Alderson H, O'Donoghue D (2014) Investigating the influence of acute kidney injury on longitudinal eGFR measurements. UK Kidney Week 2014, 29 April - 2 May, 2014, Glasgow, UK.

6. Ritchie J, **Asar Ö**, Green D, Alderson H, Chiu D, Middleton R, Diggle P, O'Donoghue D, Kalra P (2014) Rates and predictors of admission in a referred CKD population. UK Kidney Week 2014, 29 April - 2 May, 2014, Glasgow, UK.

Awards

- Travel grant from the Faculty of Health and Medicine, Lancaster University to attend Eastern North American Region of International Biometric Society 2015 Spring Meeting, held in March 15-18 2014, Miami, USA (£500)
- Financial support from the Royal Statistical Society to attend The 2014 International Conference of the Royal Statistical Society, September 1 - 4, Sheffield, UK. (£250)
- Financial support from Graduate College of Lancaster University to attend the Statistical Analysis of Multi-Outcome Data Workshop, 30 June - 1 July, Cambridge, UK. (£100)
- Best poster award for "**Asar Ö**, Ritchie J, Kalra P, Diggle P (2014) Joint analysis of longitudinal eGFR measurements and time to renal replacement therapy and death in a CKD cohort." in the Epidemiology session of the UK Kidney Week 2014, 1 May, 2014, Glasgow, UK.
- Travel grant from the Faculty of Health and Medicine, Lancaster University to attend Eastern North American Region of International Biometric Society 2014 Spring Meeting, held in March 16-19 2014, Baltimore, USA (£500)
- PhD studentship by Health e-Research Centre at the overseas level covering fees, stipend and a budget for conference and short courses attendances etc.

Professional Activities

- Refereed for Statistics in Medicine, Cancer Informatics, Statistical Methods in Medical Research, Biometrics, BMJ Open, Hemodialysis International
- Tutored "Statistical Inference" in the Department of Mathematics & Statistics, Lancaster University
- Judge for "The ENAR Regional Advisory Board (RAB) Poster Competition Award", in Eastern North American Region of International Biometric Society (ENAR-IBS) 2015 Spring Meeting, 15-18 March 2015, Miami, USA.
- Chaired the session "Methods in Causal Inference: Instrumental Variable, Propensity Scores and Matching", in Eastern North American Region of International Biometric Society (ENAR-IBS) 2015 Spring Meeting, 15-18 March 2015, Miami, USA.

- Chaired the session "Outside the Standard Setting" in Survival Analysis for Junior Researchers Conference, 3-4 April 2014, University of Warwick, Coventry, UK.
- Chaired the session "Tools for Longitudinal Data Analysis" in Eastern North American Region of International Biometric Society (ENAR-IBS) 2014 Spring Meeting, 16-19 March 2014, Baltimore, USA.
- Chaired the session entitled "Statistical Modelling" in Research Student's Conference in Probability, Statistics and Social Statistics, 25-28th March 2013, Lancaster, UK.

PhD Training

- Personalized Medicine and Dynamic Treatment Regimes, at Eastern North American Region of International Biometrics Society (ENAR-IBS) 2015 Spring Meeting, 15 March 2014, Miami, USA.
- Methods for longitudinal data analysis in the social sciences workshop, London School of Economics, 8 September 2014, London, England.
- Joint modelling of longitudinal and survival data, at Eastern North American Region of International Biometrics Society (ENAR-IBS) 2014 Spring Meeting, 16 March 2014, Baltimore, USA.
- Practical statistical computing: A workshop on advanced R, by RSS Lancashire Local Group, 21 May 2013, Lancaster University.
- Statistical Modelling, Statistical Asymptotics, Spatial and Longitudinal Data Analysis and Nonparametric Smoothing. Academy for PhD Training in Statistics (APTS), 2013.
- Bayesian Inference, Computationally Intensive Methods, Principles of Epidemiology, Survival and Event History Analysis, Environmental Epidemiology, Methods for Missing Data, Bootstrap Methods and Their Applications. Department of Mathematics & Statistics, Lancaster University.

Professional Association Membership

- The Royal Statistical Society
  - Student fellow
  - Committee member of the Young Statistician Section
  - Young Statistician Section representative in the Electronic Publication Working Group
- The International Biometric Society
  - British and Irish Region
  - East - North American Region
  - Eastern - Mediterranean Region
- The Institute of Mathematical Statistics