UNIVERSITY OF SOUTHAMPTON

Impact of Factors Associated with Short-Term Transplant Outcomes

by

Luke A. Day

Doctor of Philosophy in Mathematics
Thesis

Faculty of Social Sciences
School of Mathematical Sciences

August 2020

Contents

List of Figures

List of Tables

Acknowledgements

I am extremely grateful for the support and guidance that I have received from my main supervisor Dr Alan Kimber, from as far back as my masters dissertation up to the completion of my doctorate degree. He has been a fantastic mentor in terms of academic support, while giving me the independence and freedom to explore new methods of analysis that caught my interest. He has also been an excellent support at a personal level. Having lost my brother, uncle and grandmother throughout the course of this project, there were times that completion of the project would have been impossible without Alan's support.

I am also grateful for the guidance from my secondary supervisor Professor David Collett and the whole team at the NHS Blood and Transplant, who are changing the world with their work and have made all of this research possible.

I thank Dr Dominic Summers, a kidney transplant surgeon from Addenbrooke's Hospital Cambridge, for his input in terms of the clinical side of the project. His enthusiasm has been a strong inspiration, which has undoubtedly influenced the direction of this work.

I would also like to acknowledge a few pioneers in the field of joint modelling that have taken the time to respond to my queries and have shown interest in my work, namely: Professor Peter Diggle, Professor Dimitris Rizopoulos and Professor Joseph Ibrahim.

Finally, I would like to acknowledge my housemates for besides the countless good times, the very deep conversations that have been a huge inspiration, particularly in terms of machine learning and artificial intelligence.

This project was funded by the NHS Trust Fund (award TF047) and the University of Southampton Vice Chancellor's Scholarship.

Dedicated to the loving memory of Liam Day, Peter Day and Elsie McGinn.

Part I

General Context

Chapter 1

Introduction

Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world.

Atul Butte
Stanford University

Transplantation remains the best, and often only, treatment for patients with severe kidney, liver, heart and lung disease, but there is a great shortage of organs available for transplantation in the UK and around the world. Deceased donor donation rates in the UK have increased by 75% (NHSBT 2018) in the last decade, and this has largely been due to an increase in the use of organs from donation after circulatory death donors (DCD). DCD donors are typically patients who have suffered a catastrophic injury and are managed in the intensive care unit, but who do not fulfil the criteria for *brain-death* (Andrews et al. 2014) and so cannot become heart-beating, donation after brain death donors (DBD). Organ recovery in DCD donors may only proceed following the withdrawal of life-supporting treatment and confirmation of death following circulatory arrest. This process is notoriously difficult to predict, and around a third of potential DCD donors do not proceed to organ donation following the withdrawal of treatment.

The unpredictability of the withdrawal period has important clinical implications. Firstly, at present a full organ recovery team, including at least two surgeons, a scrub nurse, a perfusionist and a specialist nurse for organ donation are deployed away from their base hospitals to wait for the donor's circulatory arrest, which prevents them from attending other donors, and performing transplants. On the other hand, donors are scarce and so there is a great cost associated with missing donors when the retrieval teams return to their base hospitals. This uncertainty has led to an active area of research, where predictive models are being used to predict the time of asystole or death time of DCD donors (de Groot et al. 2012, Suntharalingam et al. 2009, Wind et al. 2012, Davila

et al. 2012, Kotsopoulos et al. 2018, Rabinstein et al. 2012). If a model were able to accurately predict these outcomes, clinicians would be able to allocate resources much more effectively. However, the expense of an incorrect prediction is extremely costly. For this reason, authors have stated for a predictive model to be useful in practice, an improvement in predictive ability is required compared to what has been achieved up to now (Pugin et al. 2017).

The increased tendency in practice of collecting data such that a set of patient recordings are gathered over time while simultaneously monitoring survival status has resulted in a wealth of research in the biostatistics community relating to the field of *joint modelling*. By combining longitudinal and survival data in a single model, one is able to take advantage of the rich nature of longitudinal data when performing prediction. Joint modelling offers the ability to dynamically predict survival probabilities throughout the period that the longitudinal recordings are taken. This methodology complements the current trend in medical research towards personalised medicine, and the spread of this methodology's application to novel clinical challenges is still in its infancy.

Restrictive protocols are in place in the UK, and mostly worldwide, that prohibit organ retrieval from proceeding for DCD donors when the duration of the treatment withdrawal period is prolonged. A prolonged withdrawal is thought to detriment the quality of the organs as they become starved of adequate oxygenation. However, clinical studies (Bradley et al. 2013) have criticised the relevance of the duration's impact on recipient transplant outcome and claim that the behaviour of the physiological profiles throughout this vital period play a more important role. The little evidence to support this conjecture provides scope for this research. If it can be proven that the duration itself is not associated with recipient transplant outcome, it may be possible to increase the conversion rate of potential to actual donors by relaxing restrictive protocols, which could ultimately increase the number of successful transplants.

Addressing these clinical problems through analysis poses many interesting statistical challenges that are discussed throughout this work. First, transplant data is rarely gathered by clinicians according to a well designed experiment, which has strong implications. In this case it is not always possibly to rely on the strict assumptions made by conventional methods of analysis. The retrospective analysis of observational data often requires flexibility, which comes at the cost of interpretability. Moreover, the temporal nature of the physiological profiles is strongly associated with the survival process of the donor (as the donor must be alive for taking readings to make sense), which raises the concern of informative missingness. In this work we discuss various use cases for the joint model, that is elegantly able to cope with various complications that are present in this work.

The joint model is employed in this work to dynamically predict survival probabilities of DCD donors in the treatment withdrawal to death period based on multiple physiological variables and also various demographic variables. We also investigate whether the joint model can be used in a two-stage approach, to extract important information from the trajectories of physiological variables throughout the withdrawal phase, to see how characteristics of interest are related to short-term transplant outcome. Stage two of the considered approach involves using the summaries of the physiological variable characteristics derived from the joint model as covariates in another regression model whose response relates to transplant outcome.

The main objectives of this work are now defined. We seek to apply sophisticated methodology to derive statistical models to predict DCD donor event times in the treatment withdrawal period. We also aim to improve our understanding of how characteristics of the treatment withdrawal phase are associated with short-term transplant outcomes based on the motivating datasets that are described in Section 1.1. The secondary aim of this work is to benchmark various predictive modelling methods to predict recipient transplant outcome while simultaneously gaining insight into the predictive structure of the data (by ranking the importance of variables used to train the models). We thereby aim to determine whether these methods are likely to be useful for the NHSBT in future related applications.

1.1 The Motivating Data

We begin by introducing the main dataset, which is referred to throughout this thesis as the *novel dataset*. This dataset is analysed in Chapters 4 and 5, where Chapter 4 is only concerned with the donor data and the latter relates to both donors and recipients. The novel dataset has been given its name to highlight the novelty arising from having access to trajectories of physiological variables, making this analysis to the best of our knowledge the only one of its kind.

The novel dataset consists of 227 controlled DCD organ donors from January 2013 to April 2015. 146 of these patients for which we have data were kidney donors. 113 of these patients donated both kidneys and 33 donated a single one. A categorical response variable delayed graft function (DGF) is available for 215 of the 259 kidney recipients that correspond to the donors just described (as 44 were missing). This variable's categories consist of: immediate function, delayed function, and primary non-function; which we dichotomise to represent an immediately or non immediately functioning graft. A delayed graft function indicates that the recipient had to return to dialysis within one week of transplant (which corresponds to a negative transplant outcome). An event time variable (relating to time in minutes until either death or censoring) is available with a corresponding censoring indicator. As only 22 of the 257 subjects incurred an

event (91.5% censoring), we are likely to be limited in our ability to detect a significant association should it be present when using this response variable. For this reason, it is more appropriate to consider the response DGF, which has a 67% success rate.

We also have access to repeated measures of various donor physiological variables taken throughout the treatment withdrawal to death phase. These physiological variables include: systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), heart rate (HR), oxygen saturation (O_2) and respiration rate (Resp). Although longitudinal data of this type can provide a rich insight into the nature of the problem at hand, analysis of temporal data such as this is complex and must be handled with care. As this is an observational study, the repeated measures were not recorded according to a pre-specified experimental design and are thought to be taken at random intervals. We are therefore faced with highly unequally spaced and unbalanced data, which our chosen statistical methods must be able to deal with. Specifically, there is a minimum of 2, mean of 12 and maximum of 52 repeated measures. The measurement times are highly skewed ranging between 0 and 406 minutes, with a median of 25 and an interquartile range of 9 to 70 minutes.

Various potentially confounding variables are present with regards to both the donor and recipient including: age, gender, blood group, ethnicity, height and weight. Potential confounding variables unique to recipients include: was the recipient on dialysis at time of transplant, transplant centre and primary renal disease. Cause of death is the only potentially confounding variable unique to the donor, consisting of the categories: road traffic accident, cerebrovascular accident, other trauma and miscellaneous. Cold ischaemic time (CIT) is the only potentially confounding variable available relating to the graft.

In Chapter 3 a more standard NHSBT data extract is analysed (which avoids hierarchical complications arising from repeated measures of physiological variables), for this reason we hereby refer to this data as the *NHSBT dataset*. These data correspond to all DCD donors that proceeded in the UK between 1 April 2010 and 31 March 2015 (extracted from the UK transplant registry on 3 August 2015). In particular, there are 1906 kidney recipients that correspond to 1120 DCD donors. 825 donated both kidneys and 256 donated a single kidney.

The NHSBT dataset has 14 baseline characteristic variables in total, seven relating to donors and seven relating to recipients (age, gender, blood group, cause of death, ethnicity, height and weight). Eight variables correspond to the surgery process at the donor level (time from treatment withdrawal to death, surgery time, time until blood pressure drops to 70, 60 and 50mmHg; time until oxygen saturation drops to 90, 80, 70%). Other variables at the recipient level include primary renal disease, whether the recipient was on dialysis at the time of transplant, CIT (in minutes) and which of the 24 transplant centres they had attended. Three response variables relating to short term

transplant outcomes are present (a binary indicator representing DGF, survival time in days and a censoring indicator).

Table A.1, provided in the appendix, displays the variable code names with the corresponding data types and descriptions for the variables that are present in the NHSBT dataset. The variable code names are referred to throughout this thesis interchangeably with the written name (for example, `dage` and donor age are the same). It can be seen that there are many variables in common in the two different datasets described in this section. The same variable code names are referred to for variables in the other dataset, as the variable description is the same for the variables in common.

1.2 Outline of Thesis

In Chapter 2 we discuss the relevant background information relating to both the clinical and methodological sides of the project. We begin by describing end stage renal disease, the transplantation process, types of kidney donors, and the relevance of the treatment withdrawal to death (also referred to as the agonal) phase. Chapter 2 then proceeds by providing the methodological background, beginning by making the clear distinction between explanatory and predictive modelling. Relevant statistical learning theory is then covered, which is particularly relevant for benchmarking machine learning methods in Chapter 3. Chapter 2 concludes with a discussion relating to missing data in longitudinal studies and its relation to the joint modelling framework, where the joint model is introduced.

In Chapter 3 we introduce various machine learning methods (adaptive boosting, extreme gradient boosting, random forests and conditional random forests). A simulation study is then performed to assess the ability of these methods to rank the importance of predictor variables based on various importance metrics when faced with data complications present in the motivating dataset (such as multicollinearity, variables with many categories and a hierarchical structure). The NHSBT dataset is then analysed using the methods introduced, and their performance is benchmarked against random intercept logistic regression. Finally, we propose the use of various methods of visualisation that can be used when multiple imputation is performed.

The multivariate Bayesian joint model (MBJM) is introduced in chapter 4 and extensions that are relevant in our application are described. In particular, we discuss the handling of multiple longitudinal covariates, allowing a more elaborate parametrisation of the longitudinal covariates in the joint model (allowing the hazard to depend on functions of the current biomarker value, such as the current gradient) and also relaxing the assumption of a constant association between the (possible function of the) biomarker value and the risk of event over time. We formally describe how the joint model can be used to perform dynamic prediction and explain measures of discrimination and

calibration that account for the dynamic nature of the problem at hand. Finally, we conduct the analysis and apply the discussed methods to the novel dataset to predict DCD donor event times in the treatment withdrawal period.

In Chapter 5 a two-stage approach for deriving summaries of characteristics of the physiological variables in the treatment withdrawal period that are in turn used as covariates in another regression model to predict recipient transplant outcome is investigated. A simulation study is conducted to study how inferential properties of this approach compare to those of the alternative two-stage approach using a linear mixed effects model (LMEM) instead of the joint model. We also investigate how both of these approaches compare to simply using functions of the observed physiological trajectories, which we expect to suffer from bias resulting from measurement error. The novel dataset is then formally analysed and the model derived to improve our understanding of how characteristics of the treatment withdrawal period relate to recipient outcome is interpreted.

This thesis concludes with a final discussion in Chapter 6, where final conclusions are drawn, assumptions and limitations are discussed and potential scope for future work is outlined.

Chapter 2

Background

Throughout this chapter relevant background information is provided from both the clinical and methodological perspectives of the problem at hand. In Section 2.1, the clinical setting is introduced and relevant concepts that are necessary for understanding the clinical aspects of this work are covered. In particular, end-stage renal disease (ESRD) is explained and the various types of organ donors are discussed. Details of the organ transplantation process for DCD donors are given.

In Section 2.2, many of the methods of analysis employed throughout this thesis are introduced. We distinguish between algorithmic and explanatory modelling, and consider when the use of either approach is appropriate. We then delve into more formally defining theoretical concepts in the *statistical learning theory* section (Section 2.2.2).

Once the statistical learning theoretical concepts have been discussed, this chapter draws focus to topics of fundamental importance in Chapters 4 and 5. In particular, Section 2.2.4 covers the challenging problem of missing data in longitudinal studies, with a focus on missingness that arises as a result of longitudinal variables that constitute endogenous time-dependent covariates (endogenous meaning that the occurrence of an event, such as a patient dying, causes the remaining follow up readings to be missing). As a joint modelling approach is employed in this work to address this issue (note that this approach does not involve imputing missing values, but rather inference is performed once a model is explicitly specified and incorporated into the likelihood function for the missingness process), the following Section (2.2.5) gives a natural introduction to the joint model for longitudinal and time-to-event data. These sections are necessary to understand the connection between joint modelling and the missing data framework, which is explained in detail in Section 2.2.5.1. This chapter concludes with Section 2.2.5.2, which discusses the different scenarios that a joint modelling approach could be beneficial.

2.1 Clinical Background

2.1.1 End Stage Renal Disease

ESRD is also known as kidney (renal) failure, which is a life-threatening condition where patients' kidneys fail to function normally. It is the final stage of Chronic Kidney Disease (CKD), which involves a gradual worsening of the kidneys' ability to function. Patients that have ESRD require either dialysis or a kidney transplant in order to survive.

Dialysis is a process that involves the use of a machine to remove waste products from the blood. It effectively carries out the required function of the failing kidney. Depending on the type of dialysis, patients undergo treatment between seven hours and eighteen hours per week. Patients often suffer from various side effects (such as vomiting, nausea, cramps and light-headedness) and are at risk of developing potentially fatal complications such as sepsis. Given the lengthy nature of the treatment the quality of life of patients may be substantially compromised. Moreover, patients on dialysis in general have a reduced chance of survival compared to those that undergo the alternative treatment of a kidney transplant. This is due to the fact that dialysis can only partially compensate for the loss of kidney function.

A kidney transplant is the most effective treatment for patients that have ESRD. The kidney transplantation process is complex and involves allocating the donated kidneys to recipients deemed to be the most suitable, while knowing that due to the scarcity of kidneys available, many patients will die from ESRD while on the waiting list. For this reason, much of the research that is discussed in the literature that we review has focused on expanding the pool of available kidneys, which has led to the acceptance of various types of donors, some of whom have an inherent higher risk of transplant failure.

2.1.2 Types of Kidney Donors

The flowchart in Figure 2.1 displays the total number of kidney transplants in the UK for the year beginning April 2016, and the number of various types of donors that characterise this population. Approximately one third of kidney transplants that year were from living donors. This type of donation is the most successful with a national rate of graft survival five years after the transplant being 93%, in contrast to that of 87% for deceased donors (NHSBT 2018). Tissue matches between relatives are generally optimal, which is one reason that donation by living donors have such high success rates.

According to Figure 2.1, approximately 59% of transplants from deceased donors between 1 April 2016 and 31 March 2017 were by donation after brain death (DBD). In order for a donation to proceed under the category of DBD, an irreversible brain injury that fulfils the criteria for brain-stem death must have occurred. For this type of donor,

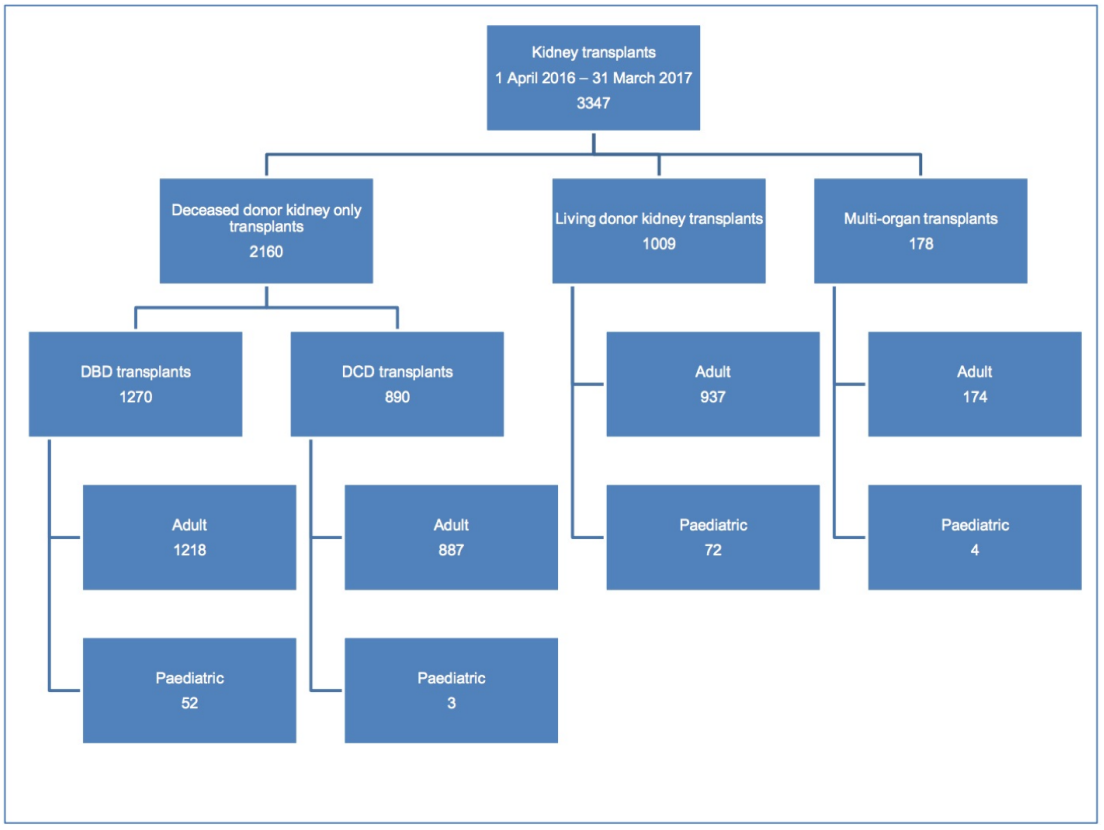


Figure 2.1: *Number of kidney transplants from different types of donors from 1 April 2016 to 31 March 2017 (source: Annual Report on Kidney Transplantation 2016/2017 NHSBT).*

the heart is still beating when death is certified. DBD kidneys are allocated using a national organ sharing scheme, which aims to minimise inequity of access and to allocate organs to the most suitable matches. This is a points based scoring system, which takes into consideration the time spent on the waiting list, HLA (Human Leukocyte Antigen) match and age (Johnson et al. 2010).

Donation after circulatory death (DCD) corresponds to donors that do not fulfil the criteria of DBD. DCD can either be uncontrolled (where death occurs outside of the hospital or on admission to hospital) or controlled where patients are in hospital with an irreversible brain injury but do not fulfil the brain-stem death criteria. A controlled DCD donor may also correspond to a patient that has been diagnosed by brain-stem criteria but suffers cardiac arrest while awaiting the removal team (Andrews et al. 2014). Currently in the UK, uncontrolled DCD donation is rare.

2.1.3 The Agonal Phase

DCD donors experience a time period in which they are withdrawn from their life-support machine, which is typically ventilatory support (Suntharalingam et al. 2009).

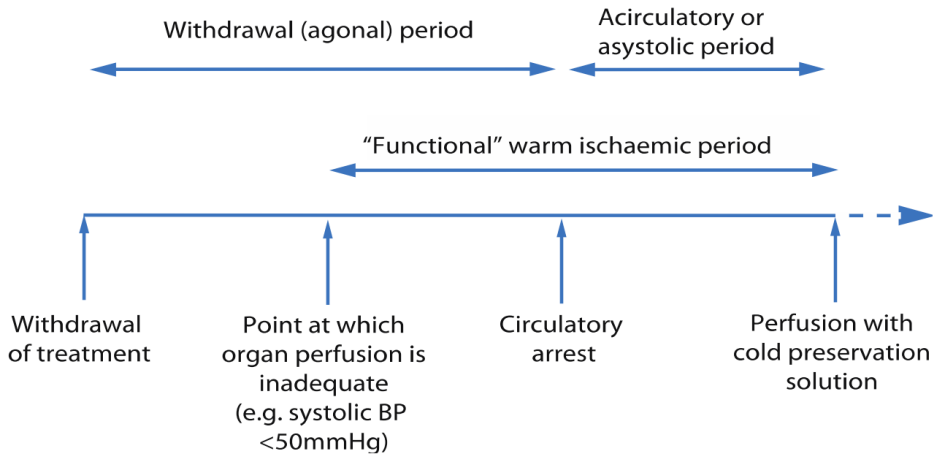


Figure 2.2: *A timeline of the kidney transplantation process for DCD donors, source: British Transplantation Society (2013).*

Figure 2.2 displays a timeline of events of clinical interest throughout DCD donation, beginning at treatment withdrawal. It can be seen that the time until the patient suffers circulatory arrest is referred to as the agonal phase. During this period haemodynamic variables (such as blood pressure, oxygen saturation, heart rate and respiration rate) are often highly volatile until the patient inevitably suffers a cardiac arrest.

Figure 2.2 also shows the functional *warm ischaemic time* (WIT) period, which refers to the time the organ is exposed to warm ischaemia. Warm ischaemic exposure is known to cause cellular decay in the kidneys. This is thought to begin once the systolic blood pressure drops below a certain subjective threshold, though less than 50mmHg seems to be the general consensus. This exposure is the reason that DCD organs are thought to be less favourable than DBD organs. Research shows that recipients whose donors were DCD donors have twice the risk of having a delayed graft function (DGF), compared to DBD donors (Summers et al. 2015). DGF is a complication where the graft fails to function within seven days. Fortunately this can easily be supported by dialysis and therefore is not life-threatening. Despite the fact that DCD donors inherently have a higher risk of experiencing DGF compared to DBD donors, the two types of donors' organs have a comparable 5 year survival to that of DBD donor organs.

Once the circulatory arrest has occurred, the asystolic period begins. Part of this period is made up a mandatory 5 minute stand off time, in order to confirm that the patient is actually deceased. Despite the asystolic period being composed of the 5 minute stand off time, this period in the UK is very short with a median of 14 minutes (interquartile range of 11 to 17 minutes) (Summers et al. 2013). After the stand off period, this period refers to the time to remove the organ and place it into the cold preservation solution.

After the organ is placed into the cold preservation solution the *cold ischaemic time* (CIT) period begins. This time-frame spans until the organ is transplanted into the

recipient. Longer CIT is also thought to correspond to suboptimal transplant outcomes. Moreover, donations from DCD donors are particularly vulnerable to CIT graft injury (Summers et al. 2013).

The agonal phase is of great importance for clinicians, as the duration and nature of this period are a major determinant of whether organ donation proceeds and of the quality of the graft (Bradley et al. 2013). Investigating the impact of factors relating to the treatment withdrawal to death period on graft quality and subsequently recipient transplant outcome is a major focus of this thesis.

2.2 Methodological Background

Statistics has traditionally been defined as collecting, summarising and analysing data where probability is used for the purpose of statistical inference. Highly influential scholars have claimed that commitment to this definition restricts the field of statistics from adapting to an exciting new era of data analysis, where the nature of data being presented by clients has evolved substantially (Breiman 2001*b*). Nevertheless, these problems are exhilarating. An extreme example is where physicists are continually adding data to astronomical databases that are made up of billions of data points representing the galaxy. Standard statistical methods cannot be applied to analyse data of such magnitude, however, scalable algorithmic models are being applied with great success. For this reason it is beneficial to be open minded to a wide range of methods, including non-standard methods that have not necessarily originated from the field of statistics.

Explanatory models are a standard class of statistical methodology that assume the underlying data generating mechanism is sufficiently approximated by a stochastic model. The dataset obtained is assumed to be a set of independent realisations from this generating mechanism, which is used to estimate the parameters that explain the dependence between the explanatory and response variables. The beauty of these models lies in their straightforward interpretation coupled with access to various apparatus of inference, such as significance tests and confidence intervals.

Explanatory models are most suited to data that is collected according to a well designed experiment. For example, they have typically been applied in agriculture experiments to determine the optimal use of fertiliser to maximise produce. The applicability of these models, however, is dependent on the problem at hand and the nature of the data. Shmueli et al. (2010) conjecture that models fitted with the aim of testing causal hypotheses are most often association based explanatory models, applied to observational data. The limitations of explanatory modelling become particularly evident when applied to observational data, occurring as a result of its complex nature.

Breiman (2001*b*) states that the a priori assumption that observational data is generated according to a parametric model, selected by the researcher is questionable. But this is rarely criticised in published works. As data becomes more complex an explanatory model's ability to portray nature's generating mechanism through a clear straightforward interpretation is lost. A typical example of this is when non-linear associations are present. In this case, an explanatory modelling approach can appropriately deal with this by including non-parametric spline terms. However, this is at the cost of the parameter estimates no longer being interpretable. Worryingly, a common practice is to categorise continuous variables in order to retain interpretability, which is widely accepted by the statistics community as dangerous, inducing bias and a loss in accuracy (Royston & Altman 1994, Royston et al. 2006, Harrell 2015).

Many of these limitations are a result of the aim to find an interpretable model, which requires simplifying the process to the degree that when faced with complex data, the explanatory model lacks flexibility to provide reliable conclusions. Breiman (2001*b*) argues that the aim should be to obtain accurate information. In order to do so, a model need not be simple nor an explanatory model.

2.2.1 Explanatory or Algorithmic Modelling?

Algorithmic models are primarily concerned with predicting an unseen output given a set of inputs[1]. In contrast, although explanatory models can also be used for the purpose of prediction, they are often preferred in practice for determining causal relationships. This is because model selection for explanatory models requires many combinations of models to be explicitly compared, particularly when a lot of flexibility is required to capture non-linearities and interactions present in the data. Instead of assuming a stochastic model, algorithmic models use the data to determine nature's generating mechanism. Treating nature as a black box has understandably led to a large amount of scepticism, particularly in the medical community (Wyatt 1995, Plate 1999). However, as one will see from this section, algorithmic models provide an invaluable set of tools for the analysis of complex data.

Since the mid-twentieth century there has been a rapid development in machine capability in relation to computational speed and capacity. This resulted from improvements in hardware, theoretical computer science, statistics and parallel computing. Having computational power that was previously incomprehensible has led to a revolution in the development of methodology for data analysis, particularly for the purpose of prediction. Consequently, clients' problems and the data that is being gathered for analysis are evolving rapidly, increasing in both size and complexity.

[1]To clarify the contrasting terminology between statistics and machine learning: inputs, features, attributes correspond to explanatory variables. Label, target, output refers to the response variable.

An astounding amount of progress has been made by the *machine learning* (a branch of *artificial intelligence*) community in terms of developing methods capable of dealing with complications arising from these complex problems presented by clients. The majority of researchers in this community is composed primarily of computer scientists, however, in the last two decades a coercion between machine learning and statistics has progressed. In recent years many statistics journals have published papers relating to machine learning methods. Furthermore, many of the revolutionary machine learning developments occurred as a result of contributions from the statistics community. Grace Wahba's theoretical research leading to the proposal of *smoothing splines* (Wahba 1975), as well as Trevor Hastie and Robert Tibshirani's *generalised additive models* (GAMs) (Hastie & Tibshirani 1987) sparked a new culture of predictive modelling.

A subset of algorithmic models, namely *deep learning* methods, have become extremely popular in the last three decades and have a multitude of exciting applications. Such applications include *computer vision*, where images are classified based on a training set containing thousands (or possibly millions) of images. These methods have been shown to outperform clinicians for detecting brain tumours from MRI scans (Havaei et al. 2017). Such methods include the *artificial neural network*, which was designed in the 1980's to mimic the way that the human brain passes information between neurons. These methods are simply layered non-linear statistical models, whose purpose is solely for prediction when trained on large datasets. As we consider relatively small sized datasets in this research, no further attention is given to deep learning methods.

The huge amount of data required to make deep learning methods applicable often causes confusion as to whether machine learning methods can at all be used for small to medium sized datasets. Various algorithmic models such as *tree based methods* (Breiman et al. 1984) can be successfully applied to any size (or shape) data. Perhaps the most famous application of machine learning (that was not deep learning) was the use of boosted trees (Roe et al. 2005), which led to the detection of the Higgs boson at CERN (Chatrchyan et al. 2012).

One may question the relevance of predictive modelling approaches when our interest is in how a set of explanatory variables are associated with a response. In this section we justify their use and argue that conflation is beneficial.

Shmueli et al. (2010) lists various reasons as to why predictive modelling can be of assistance in the context of a practical application, such as that with the NHSBT data:

- Nowadays, data is routinely gathered that contains relationships that are too complex to test simple hypotheses. Predictive modelling such data can uncover potential causal mechanisms and lead to new hypotheses.

- The empirically rigorous approach of predictive modelling can serve as a reality check that findings from alternative approaches are reasonable, and that assumptions hold. For example, GAMs can be used to check an alternative explanatory model's assumption of linearity (we do this in our analysis in Section 4.4.0.2).
- Predictive power assessment (such as cross-validation) provides a simple way to compare competing models.
- A predictive model whose predictive ability is close to that of an explanatory approach may suggest our understanding of that phenomenon can only be improved marginally. If they are very different, this may suggest there are many potential gains to be made.
- Novel importance measures for predictive models can provide a means of interpretation, thus offering insight into the predictive structure of the model while maintaining flexibility to capture complex relationships.

Many of the revolutionary improvements made by the machine learning community relate to three important principles listed below (Breiman 2001*b*), which we aim to take advantage of the analysis of the NHSBT dataset.

- **Rashomon**[2]: *The multiplicity of explanatory models.* Various competing models may be found by the analyst to be equally as good, but each gives a very different interpretation. This is a commonly occurring phenomenon during a model selection procedure.
- **Occam's Razor**: *The trade-off between simplicity and accuracy.* In general, the aim is to choose the simplest model that achieves a satisfactory level of accuracy. This relates to the phenomenon known as *over-fitting*, where a model that is too complex models the noise present in the training data.
- **Bellman**: *The curse (or blessing) of dimensionality*[3]. Including too many variables in a explanatory model leads to over-parametrisation. In this case the model does not have enough statistical power to detect true associations. This can also result in a contest for information. By contrast, algorithmic models are able to cope with large numbers of variables, and reducing dimensionality may result in a loss of information. In this case, high dimensional representations of covariates is advised for small datasets.

Although algorithmic models suffer from multiplicity, *ensemble learning* (discussed in Chapter 3), is a method analogous to model averaging that is a potential fix to the

[2]Rashomon is a Japanese film from the year 1950, where four people recount different versions of the story of a man's murder in court.

[3]A famous quote by the applied mathematician Richard E. Bellman when considering problems in dynamic optimisation.

problem. The idea is to utilise the information drawn from each of the competing models that have a similar level of accuracy (but tell different stories) and take a meaningful measure, such as the average or a majority vote of their output. In this case one can take advantage of the fact that various models capture different aspects of the data generating mechanism.

In explanatory modelling the trade-off between simplicity and accuracy can be managed by hypothesis testing to determine the better fit between nested models (for example, by using the *likelihood ratio test*). Alternatively one can use a metric that combines a goodness-of-fit measure with a penalty for complexity, such as the Akaike information criterion (AIC) (Ojo et al. 1973) (this is a relative goodness-of-fit measure); or the cross validation error as in predictive modelling. However, when faced with complex data that contains non-linear associations, it is not possible to achieve the desired level of simplicity such that a straightforward interpretation of the parameter estimates can be performed without sacrificing an unreasonable amount of accuracy.

In contrast, some algorithmic models suffer from the same problem. The decision tree (Morgan & Sonquist 1963) provides a very straightforward interpretation that can be easily understood by clinicians, but suffers from poor predictive accuracy. This led to the proposal of *random forests* (Breiman et al. 1984) (details given in Chapter 3) which essentially involves fitting trees to bootstrap samples of the data and averaging the predicted values. Hastie et al. (2008) and James et al. (2013) give fantastic explanations of how averaging models is a highly effective method of reducing variance and improving predictive accuracy.

As a result of having multiple trees, interpretation becomes distorted. Random forests have a novel method of ranking variable importance, providing insight into the predictive structure of the data (Breiman et al. 1984). This is based on a permutation test for each variable, where the mean decrease in accuracy on a validation set can be calculated. Breiman (2001*b*) provides examples of how one can interpret a random forest and interactions of interest by looking at the predicted probabilities. He claims that they are able to reveal hidden aspects within the data that standard methods cannot, for example by applying a random forest based clustering method to show the proximity between samples.

A real advantage of algorithmic models relates to the *blessing of dimensionality*, as opposed to the "curse", referring to the fact that unlike standard methods they are able to handle potentially thousands of predictor variables without the need for a subjective model selection procedure. A notable example being applied to the microarray dataset (Dudoit et al. 2002), where more than 4000 gene expressions were used as variables that only had 81 samples to predict the presence of a tumour. Díaz-Uriarte & De Andres (2006) found high accuracy even though most of the variables were random noise. This

extreme example demonstrates why the phrase *blessing of dimensionality* is used instead for algorithmic models.

In the case of explanatory modelling, suppose the design matrix $\boldsymbol{X}$ has dimension $n \times p$. In order to obtain the least square estimates, for example, $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ must first be evaluated. This inversion will lead to a singularity when p is larger than n. This is one incentive for dimensionality reduction besides the aim to obtain a simple interpretable model.

Diaconis & Efron (1983) claim that it is unwise to fit an explanatory model to a dataset that has, for example, 155 observations and 19 variables. This is similar to the size of the data that we analyse in this thesis. To select a model with 5 of these variables, there are $11,628$ possible combinations of models that can be compared, and many of those that fit the data equally well may tell very different stories.

It is clear that algorithmic models provide an effective set of tools for the applied statistician whom has the pragmatic aim to solve clients' problems. In many cases, an explanatory modelling approach is more appropriate, and in others an algorithmic modelling approach is more appropriate (particularly when data is complex and observational).

The art of data analysis requires the subjective decision of which methods to apply to the problem at hand. In determining what would be an appropriate method, we consider the following points:

- *Accuracy*: Does the model capture the true underlying data generating mechanism to a satisfactory level?
- *Interpretability*: How much insight does the model give to the relationship between the inputs and outputs of the model?
- *Efficiency*: How much memory and time is used to run the method? Bearing in mind that many machine learning methods are highly computational (although usually applied to large datasets).
- *Robustness*: How robust is the model to complications present in the dataset, and do they impact the accuracy, interpretability or efficiency of the model?

To summarise this section, it is clear that both explanatory and algorithmic modelling approaches provide distinct advantages, yet both methods have their caveats. The approach that should be employed depends very much on the nature of the problem. For example, if hypothesis testing is required an explanatory model should be used, but in most cases using an algorithmic model in conflation is beneficial for establishing hypotheses and checking modelling assumptions such as a linearity. In contrast, where the aim is prediction, an explanatory model should still be used when there is not a large number of variables and it is reasonable to assume a smooth relationship between the predictors

and response (but comparing finding with an algorithmic modelling approach can serve as a reality check, as a large departure in findings suggests the explanatory modelling assumptions may have been violated). When $p > n$ or when p is large, and prediction is the purpose of the study, an algorithmic model should be considered. Moreover, when there is a prior belief of multiple interacting non-linear effects, an algorithmic modelling approach is likely to be more appropriate.

Acknowledging the criteria listed in the above bullet points, in this work we consider both explanatory and algorithmic modelling approaches, such that they are used to complement each other. To avoid the dangers of employing black box methods, we cover relevant theory to understand the mechanics of these algorithms, that are used in the analysis. Moreover, simulation studies are conducted to test the performance of such methods when applied to the NHSBT data in Chapter 3. In the remainder of this chapter we introduce important theoretical and practical concepts that are relevant throughout the remainder of this thesis. We also give necessary definitions and introduce mathematical notation.

2.2.2 Statistical Learning Theory

Suppose we have covariates[4] $X = (X_1, \ldots, X_p) \in \mathcal{X}$ and a response $Y \in \mathcal{Y}$. X and Y jointly take their values from $X \times \mathcal{Y}$ with respect to the (usually unknown in practice) joint probability distribution $f_{X,Y}(X, Y)$.

A *learning set* is an n-tuple of samples drawn randomly and independently from the sample space Ω according to the joint distribution $f_{X,Y}(X, Y)$. This set of realisations is denoted by $\mathfrak{D} = (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.

The relationship between the inputs and outputs is characterised using the following formula of conditional probability:

$$\mathbb{P}(Y \mid X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)}, \text{ if } \mathbb{P}(X) > 0. \tag{2.1}$$

Our aim is to derive the function that is the best approximation of an underlying phenomenon of interest. *Supervised learning* corresponds to estimating (also referred to as learning) the function, or *decision rule* as in Equation 2.2, using the learning set $\mathfrak{D}$.

$$\varphi : \mathcal{X} \to \mathcal{A} \tag{2.2}$$

Here $\mathcal{A}$ denotes the action space that plays an important role in decision theory (Young et al. 2005) representing the set of allowable actions that may be taken in terms of making the prediction $a = \varphi(\boldsymbol{x})$. Therefore, we map each input $\boldsymbol{x} \in \mathcal{X}$ to a predicted value $a \in \mathcal{A}$. This notation is used instead of $\hat{y}$ to emphasise that it is not necessarily the case that $\varphi(\boldsymbol{x}) \in \mathcal{Y}$.

Solving the learning problem requires defining the **risk functional** (also known as the generalisation error). The risk of the function φ trained on $\mathfrak{D}$ is denoted by $R(\varphi_{\mathfrak{D}})$. The *loss function* $L(\cdot, \cdot)$ measures the discrepancy between its two arguments (Geurts 2002).

$$R(\varphi_{\mathfrak{D}}) = \mathbb{E}_{X,Y}\{L(Y, \varphi_{\mathfrak{D}}(X))\} \tag{2.3}$$

- In the case of explanatory modelling the loss function is a function of the parameters θ, i.e, $L : \Theta \times \mathcal{A} \to \mathbb{R}^+$, where Θ represents the parameter space. Here the action space corresponds to the set of possible parameter values.
- For a predictive model the loss function is a function of the predicted value a: $L : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}^+$.

[4] Throughout this thesis both statistics and machine learning terminology are used interchangeably. In general we use the machine learning terminology when statistical learning theory is relevant, although overlap is unavoidable. There are many names for covariates in both fields (explanatory, predictor variables, inputs, features, attributes). There are also numerous names for the response variable (target, labels, output).

Therefore, in the case of predictive modelling, the loss function measures the discrepancy between the observed and predicted values based on the candidate model, which is trained on the learning sample. Before a prediction is made the loss function serves as a random variable because the true outcome is unknown. Once the prediction has been made the loss function becomes a measure of accuracy.

- In the **regression case**, a common loss function corresponds to the *quadratic loss*: $L(Y, \varphi_{\mathcal{D}}(X)) = \mathbb{E}_{X,Y}\{Y - \varphi_{\mathcal{D}}(X)\}^2$, also known as the *mean squared error (MSE)*.
- Alternatively in the **classification case**, occurring when $Y = \{0, 1\}$ (as we restrict our attention to binary classes), a common choice of loss function is the misclassification error, represented by[5] $L(Y, \varphi_{\mathcal{D}}(X)) = \mathbb{1}(Y \neq \varphi_{\mathcal{D}}(X))$.

The risk for the classification problem can be expressed as in Equation 2.4, which is by definition equivalent to the probability of misclassification.

$$R(\varphi_{\mathcal{D}}) = \mathbb{E}_{X,Y}\{\mathbb{1}(Y \neq \varphi_{\mathcal{D}}(X))\} = \mathbb{P}(Y \neq \varphi_{\mathcal{D}}(X)) \tag{2.4}$$

Statistical learning can be viewed as a function estimation problem. Denoting the function space $\mathcal{A}^{\mathcal{X}}$ as the set of all possible functions that map $\mathcal{X}$ to $\mathcal{A}$, the target function is defined as:

$$\mathring{\varphi} = \arg\min_{\varphi \in \mathcal{A}^{\mathcal{X}}} \mathbb{E}_{X,Y}\{L(Y, \varphi_{\mathcal{D}}(X))\} \tag{2.5}$$

$$= \arg\min_{\varphi \in \mathcal{A}^{\mathcal{X}}} R(\varphi_{\mathcal{D}}) \tag{2.6}$$

Thus, the target function is the risk minimiser corresponding to the distribution of the population from which the learning set was obtained. However, when $f_{X,Y}(X, Y)$ is unknown and the cardinality of $\mathcal{X}$ is infinite, this problem is ill-posed (Tikhonov & Arsenin 1977). In this case, $\mathcal{D}$ would be used in order to attempt to estimate an infinite number of parameters. It is therefore necessary to impose additional assumptions on the solution, as discussed in Section 2.2.2.2.

2.2.2.1 Bayes Model and Irreducible Error

In the rare case where $f_{X,Y}(X, Y)$ is known, the target of the learning function i.e. the best possible model, can be derived analytically, which corresponds to that which minimises the risk. This is known as *Bayes' model* and is denoted by φ_B. The risk for

[5] Here $\mathbb{1}(\cdot)$ represents an indicator function that takes on the value 1 when the logical statement in the parentheses is satisfied and 0 otherwise.

Bayes' model corresponds to the *irreducible error*[6] expressed as:

$$Err(\varphi_B) = \mathbb{E}_{X,Y}\{L(Y, \varphi_B(X))\} \tag{2.7}$$
$$= \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y, \varphi_B(X))\}\} \tag{2.8}$$

Bayes' model is the function that minimises the inner expectation in Equation 2.8 at each point X of the input space, that is:

$$\varphi_B(x) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{Y|X=\boldsymbol{x}}\{L(Y, a)\} \tag{2.9}$$

Again focussing on the classification setting, this is referred to as *Bayes' classifier*, which predicts for each input value the most likely class:

$$\varphi_B(x) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{Y|X=\boldsymbol{x}}\{\mathbb{1}(Y \neq y)\} \tag{2.10}$$
$$= \arg\min_{y \in \mathcal{Y}} \mathbb{P}(Y \neq y | X = \boldsymbol{x}) \tag{2.11}$$
$$= \arg\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y | X = \boldsymbol{x}) \tag{2.12}$$

Although Bayes' model and the irreducible error are useful concepts for studying the capabilities of learning algorithms, in practice $f_{X,Y}(X, Y)$ is rarely known. It may seem intuitive to solve the learning problem by estimating $f_{X,Y}(X, Y)$ by counting the frequency of observations over the predictor space, and then defining the model according to Equation 2.9. However, in order to be a reasonable approximation this would require a very large sample whose points are sufficiently dense across each point of the predictor space $\mathcal{X}$. To make things worse, the required number of samples would grow exponentially with the number of predictors (Geurts 2002). A better approach is to estimate the risk in Equation 2.3 using the learning set $\mathfrak{D}$.

2.2.2.2 Empirical Risk Minimisation

The *empirical risk of function* φ is an estimate of the true risk. Let $\mathfrak{D}'$ denote the dataset that is used to assess the risk of the model, which may or may not be $\mathfrak{D}$. The empirical risk can be calculated by averaging over each sample's contribution to the loss based on the model $\varphi_{\mathfrak{D}}$ fitted on the learning set, as in Equation 2.13. Note that n' is the number of observations in dataset $\mathfrak{D}'$.

$$\hat{R}(\varphi_{\mathfrak{D}}, \mathfrak{D}') = \frac{1}{n'} \sum_{(\boldsymbol{x}_i, y_i) \in \mathfrak{D}'} L(y_i, \varphi_{\mathfrak{D}}(\boldsymbol{x}_i)) \tag{2.13}$$

[6] Also known as the residual error.

The *strong law of large number* tells us that almost surely:

$$\lim_{n\to\infty} \hat{R}(\varphi) = R(\varphi). \tag{2.14}$$

The task of determining the target function as in Equation 2.5 can be made well-posed for reasonable sample sizes (with a possible large number of covariates) by restricting the function space to $\mathcal{F} \subset \mathcal{A}^{\mathcal{X}}$ (Evgeniou et al. 2000), where the choice of $\mathcal{F}$ is very important. Linear models are a common choice of model class, i.e:

$$\mathcal{F} = \{\varphi : \varphi(\boldsymbol{x}) = \theta_0 + \sum_{j=1}^{p} \theta_j x_j, \forall x \in \mathcal{X}\}. \tag{2.15}$$

The function found to minimise the empirical risk out of the functions belonging to the restricted function space $\mathcal{F}$, defines the task of statistical learning. This is known as the *empirical risk minimisation principle* (Vapnik 1999). This results in the approximation to the target function as required:

$$\hat{\varphi} = \arg\min_{\varphi\in\mathcal{F}} \hat{R}(\varphi). \tag{2.16}$$

A learning algorithm is therefore characterised by the choice of model class $\mathcal{F}$ and loss function, and can be regarded as an optimisation procedure (via empirical risk minimisation) that accepts a dataset $\mathcal{D}$ and returns an estimated function $\hat{\varphi}$. These choices affect which optimisation procedure is appropriate, although few can be solved analytically and most require numerical methods.

Most model classes are defined by a set of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that are estimated based on the data as $\hat{\boldsymbol{\theta}}$, in which case $\hat{\varphi}(\boldsymbol{x}) = \varphi(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$. For example, when $\mathcal{F}$ is chosen to be the class of linear models, the problem simplifies to finding the parameter vector that minimises Equation 2.13.

Maximum likelihood estimation can be viewed as an empirical risk minimisation problem when an appropriate loss function is chosen. However, loss functions based on parameters rather than predicted values are not optimal for prediction.

2.2.2.3 Maximum Likelihood Estimation

Maximum likelihood estimation is the most common method for estimating parameters in a parametric model. A known probability density function is assumed based on the nature of the data. For example, a Gaussian density is used for the class of linear models. This would only be chosen when $Y \in \mathbb{R}$.

The likelihood function represents the joint density of the data:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_Y(y_i; \boldsymbol{\theta}), \tag{2.17}$$

whose arguments correspond to the observed responses $\boldsymbol{y} = \{y_i\}_{i=1}^n$ and the parameters. The maximum likelihood estimate (MLE) is derived using Equation 2.18.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l(\boldsymbol{\theta}; \boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} \log f_{Y|X}(y_i; \boldsymbol{\theta}(x_i)) \tag{2.18}$$

This directly corresponds to the loss function in Equation 2.19, which is minimised with respect to the parameters $\boldsymbol{\theta}$ in order to derive the parameter estimates.

$$L(y, \boldsymbol{\theta}(x)) = -\log f_{Y|X}(y; \boldsymbol{\theta}(x)) \tag{2.19}$$

In the case of *linear regression* a conditional Gaussian distribution (Equation 2.20) is selected as the conditional probability distribution function contained in the loss function in Equation 2.19.

$$[Y|X] \sim \mathcal{N}(\mu(X), \sigma^2) \tag{2.20}$$

This corresponds to the loss function displayed in Equation 2.21.

$$L(y, \mu(x)) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - \mu(x))^2 \tag{2.21}$$

An analytical solution is available for the optimisation problem in Equation 2.18 in the case of linear regression. This is obtained by the least squares estimates (Equation 2.22), where $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^\top$ is the design matrix.

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{2.22}$$

In the classification case where $Y = \{0, 1\}$, perhaps the most common approach is to employ *logistic regression* (Cox 1958). The conditional probability mass function for this method is the Bernoulli distribution:

$$[Y|X] \sim \text{Bernoulli}(p(X)), \tag{2.23}$$

where applying the sigmoid function to the linear predictor gives:

$$p(X) = \frac{\exp(\theta_0 + \sum_{j=1}^{p} \theta_j X_j)}{1 + \exp(\theta_0 + \sum_{j=1}^{p} \theta_j X_j)}. \tag{2.24}$$

For logistic regression the loss function corresponds to:

$$L(y, \varphi(x)) = -y \log(p(x)) - (1 - y) \log(1 - p(x)) \tag{2.25}$$

and the target function is:

$$\mathring{p}(x) = \mathbb{P}(Y = 1 | X = x). \tag{2.26}$$

As $p(x) = [0, 1]$ is bounded, a link function $g : \Theta \to \mathcal{A}$ is used to give $\varphi(x) = g(p(x)) \in \mathcal{A}$. Logistic regression can therefore be expressed as Equation 2.27:

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \theta_0 + \sum_{j=1}^{p} \theta_j X_j. \tag{2.27}$$

Here the response is referred to as the *log odds* and the right hand side of the equality in Equation 2.27 is the linear predictor. This formulation has the advantage of a straightforward interpretation. For example, a unit increase in X_1 corresponds to a multiplicative change in the odds of success by $\exp(\theta_1)$, while adjusting for the other covariates.

Statistical models whose likelihood function is based on the exponential family of distributions (Pitman 1936), such as logistic regression, come under the class of *generalised linear models* (GLMs). These models most often do not have analytical solutions, in which case they can be fit using the *Iteratively Reweighted Least Squares (IRLS)* (Nelder & Wedderburn 1972) algorithm, which is a variation of Newton's method for optimising $\boldsymbol{\theta}$.

2.2.2.4 Performance Evaluation

It is often necessary to evaluate the performance of a model by assessing its ability to generalise and accurately predict on unseen data. This can be useful for model selection, validation and tuning hyper-parameters of a learning algorithm. This is most often used with models built for the purpose of prediction but can also be applied to explanatory models.

In most practical applications, an extra dataset is not available for validation, in which case it may be tempting to replace $\mathfrak{D}'$ in Equation 2.13 with $\mathfrak{D}$, which results in the training estimate of the generalisation error (the ability of the model to generalise the unseen data). This is well known to give an over-optimistic estimate of the true risk, because the same data is used to estimate the risk that is used to fit the model. This is denoted as follows:

$$\hat{Err}^{train}(\varphi_{\mathfrak{D}}) = \hat{R}(\varphi_{\mathfrak{D}}, \mathfrak{D}). \tag{2.28}$$

A better approach is to randomly partition the data into two sets, creating $\mathfrak{D}_{train}$ and $\mathfrak{D}_{test}$. For example, 70% of the data becomes $\mathfrak{D}_{train}$ and 30% as $\mathfrak{D}_{test}$. Although this reduces the bias regarding the estimate of the generalisation error, a caveat is that the model becomes more biased due to a reduced sample size. This would be costly in our application due to the relatively small sample size. Losing this much data would seriously impact accuracy. The test estimate of the generalisation error is denoted as:

$$\hat{Err}^{test}(\varphi_{\mathcal{D}}) = \hat{R}(\varphi_{\mathcal{D}_{train}}, \mathfrak{D}_{test}). \tag{2.29}$$

A preferable approach, particularly for data sizes similar to what we have in our application, is to use *K-fold cross-validation* (Hastie et al. 2008). This involves choosing a value of K that determines how many folds the data is partitioned into. Common values are 5 or 10. In this case a single fold is used as a testing set $\mathfrak{D}_k$ and the other folds are combined to form the training set $\mathfrak{D}^{-k}$. The performance is evaluated and the process is repeated K times and then averaged, so that each fold is used as a training set sequentially.

$$\hat{Err}^{CV}(\varphi_{\mathcal{D}}) = \frac{1}{K}\sum_{k=1}^{K}\hat{R}(\varphi_{\mathcal{D}^{-k}}, \mathfrak{D}_k) \tag{2.30}$$

This approach has the advantage of using the whole dataset, at the expense of being more computationally demanding. A large value for K would result in a more accurate estimate of the generalisation error compared to a low value, but more computational resources would be required for its computation.

A common fault in practice is to use K-fold cross-validation for the use of both model selection (or tuning the hyper-parameters) and estimating the generalisation error. It is often not recognised that this also results in an over-optimistic estimate of the generalisation error as the learned model is not independent of the test set. This is because the criterion used to select the best model was based on minimising of the test error, which could result in a large under estimation of the true generalisation error.

Alternatively, the data could be split to give another component $\mathfrak{D}_{valid}$, namely the validation set, which is used to tune the hyper-parameters without using the test set. However, this exacerbates the problem of reducing the training sample size, as even more data is being given away than splitting into training and testing sets. One method for achieving an unbiased estimate of the generalisation error, while using the whole dataset but keeping the model selection and validation procedures separate is *nested K-fold cross-validation*.

Figure 2.3 demonstrates how nested cross-validation is performed. In this example there are three outer loops and four inner loops. During each outer loop the full dataset is partitioned into training and testing datasets. This is done such that each segment (shaded in light green) is used as the test set once, after all three outer loops are

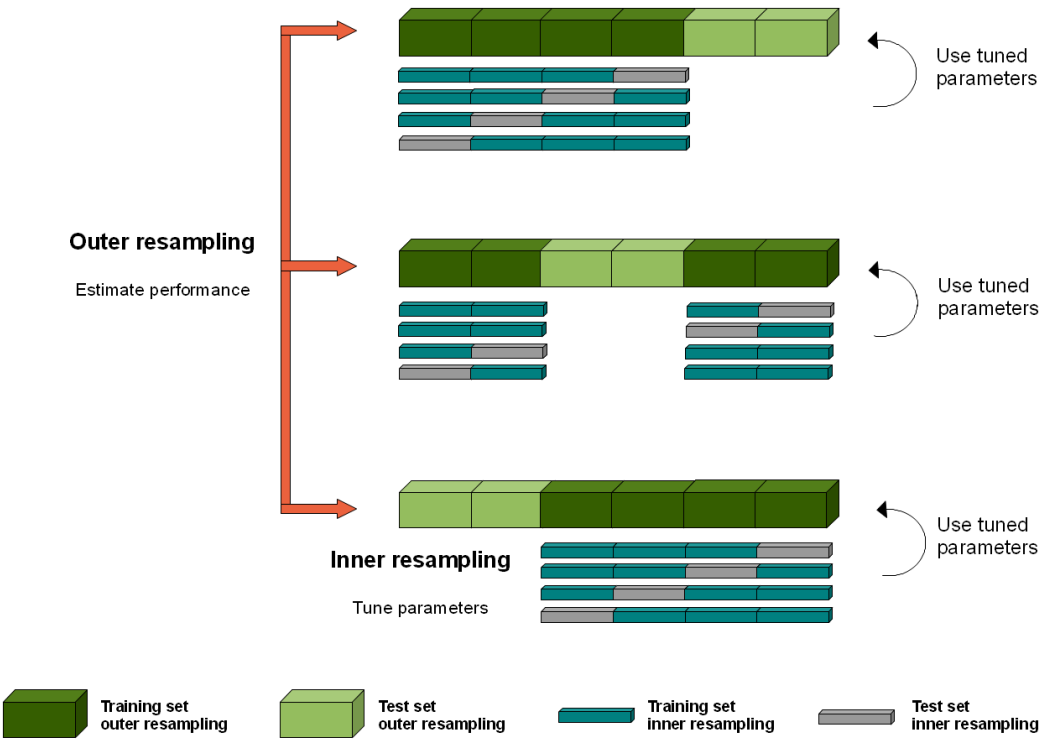


Figure 2.3: *An illustration of how the dataset is split for nested cross-validation for tuning hyper-parameters. This graphic presents a 3 fold cross-validation in the outer loop and a 4 fold cross-validation in the inner loop (Figure taken from Schiffner et al. (2016)).*

performed. In each inner loop the training data is split into four segments, three of which become the new training set and the left out segment is the test set. Again, after each inner loop has been performed each segment has its turn as the inner test set. This enables a hyper-parameter selection procedure to be performed by minimising the test error within the inner loop. The generalisation error is then calculated as the outer test error of averaged across the three loops. This ensures that the test set for estimating the generalisation is not used a priori to determine the hyper-parameters. Nested cross-validation is formally given in Algorithm 1 (taken from Louppe (2014)).

Although nested cross-validation is a highly effective method for simultaneously determining the hyper-parameters and estimating the generalisation error, it is the most computationally demanding approach out of those discussed in this chapter.

Algorithm 1 *Nested K-fold cross-validation for tuning and evaluation.*

1: Split $\mathcal{D}$ into K folds $\mathcal{D}^1, \ldots, \mathcal{D}^K$ each of size n/K
2: **for** $i = 1, \ldots, K$ **do**
3: Split $\mathcal{D} \setminus \mathcal{D}^i = \mathcal{D}^{-i}$ into K folds $\mathcal{D}_1^{-i}, \ldots, \mathcal{D}_K^{-i}$
4: Tune model using subset $\mathcal{D}^{-i}$ using K-fold CV estimates, i.e., calculate:
$\hat{\theta}_i^* = \arg\min_\theta \frac{1}{K} \sum_{l=1}^{K} \hat{R}(\varphi(\theta, \mathcal{D}^{-i} \setminus \mathcal{D}_l^{-i}), \mathcal{D}_l^{-i})$
5: Evaluate generalisation error of sub-model as: $\hat{R}(\varphi(\hat{\theta}_i^*, \mathcal{D}^{-i}), \mathcal{D}^i)$
6: **end for**
7: Derive the unbiased generalisation error estimate of selected model, as the average error of sub-models over folds:
$\frac{1}{K} \sum_{i=1}^{K} \hat{R}(\varphi(\hat{\theta}_i^*, \mathcal{D}^{-i}), \mathcal{D}^i)$
8: Tune hyper-parameters on full set $\mathcal{D}$, i.e
$\hat{\theta}_i^* = \arg\min_\theta \frac{1}{K} \sum_{i=1}^{K} \hat{R}(\varphi(\theta, \mathcal{D}^{-i}), \mathcal{D}^i)$
9: Learn final model on entire dataset $\mathcal{D}$.

2.2.3 Performance and Information Metrics

It is important to distinguish between the meaning of *performance* and *information* metrics. At a high level, performance metrics reflect how *correct* the statistic is, and information metrics describe how *precise* it is. What is referred to here as 'correct' represents how far the estimated value is from the true underlying value being estimated. Precision in this context reflects the estimated value's variation. For example, if the experiment conducted to derive the estimated value were repeatable, how clustered would the estimated value be when derived over multiple repeats. A high precision corresponds to a low variation (clustered values) and low precision is high variation (randomly scattered values). It is important to note that it is possible for the following seemingly counter-intuitive cases to arise: an estimated value can be precise but not accurate, or accurate yet not precise.

In the case of explanatory modelling, the performance of a parameter estimate is often described by its bias: $\mathbb{E}(\tilde{\theta}) - \theta$; and its precision by the standard error $\sqrt{Var(\hat{\theta})}$ (the standard deviation of the sampling distribution). The remainder of this section is focussed on metrics in the case of predictive modelling. Metrics that are used throughout this thesis are defined accordingly.

A natural starting point for describing performance metrics is to consider the *confusion matrix* (Table 2.1), which is a table of predicted and actual outcomes for each class (here we restrict our attention to binary classification). The diagonals highlighted in green contain the total number of correctly predicted cases and those highlighted in red display the total number of incorrectly predicted cases. The cell values in the confusion matrix (TP, FN, FP, TV) are important values that many metrics of predictive performance are derived from. FP corresponds to type 1 error and FN corresponds to type 2 error.

Table 2.1: *The confusion matrix displaying the number of predicted and actual outcomes. Correct classifications are highlighted in green and incorrect classifications are highlighted in red.*

	Actual Positive	Actual Negative
Predict Positive	True Positive (TP)	False Negative (FN)
Predict Negative	False Positive (FP)	True Negative (TN)

Accuracy is perhaps the most intuitive performance metric, which gives the proportion of values that were correctly predicted. This can be derived mathematically as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.31}$$

Accuracy is a useful metric when TP and TN are more important than their false counterparts. However, accuracy is a poor measure of predictive performance when classes out unbalanced. For example, if the training data has an event rate of 95%, a model

that predicts all values to have an event would have a misleadingly strong predictive performance. Accuracy is strongly connected to the generalised error discussed in the previous section, as it is its complement, i.e., accuracy = 1 - generalisation error.

Certain cases arise in practice where the aim is to keep FP and or FN low, requiring metrics designed on that basis. In this case *precision* and *recall* may offer value to the analyst. Precision is derived as $TP/(TP + FP)$, which is high when FP are low. Recall (also known as *sensitivity*) is the true positive rate i.e., $TP/(TP + FN)$. This corresponds to the fraction of predicted successes that are correct. Recall is a useful measure when the cost of committing a type 2 error (false negative) is high, as recall is high when the number of false negatives are low.

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{2.32}$$

Specificity represents the proportion of the negative classes predicted correctly. This metric is large when the number of false positives are low. Specificity is therefore a useful measure when the penalty of committing a type 1 error (false positive) is high.

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{2.33}$$

We now introduce performance metrics that are conditional on sensitivity and specificity. The receiver operating characteristic curve (ROC) is a plot of sensitivity against 1 - specificity, varied over all possible threshold values for a classification rule. The area under the curve (AUC) of the ROC curve is possibly the most commonly reported performance measure for classification in practice. This corresponds to the concordance index and represents the ability of a model to discriminate between the two classes. The following AUC ROC values can be used as a guide for determining discriminatory ability:

- 0.5: Equivalent to 'tossing a coin'.
- 0.5 - 0.6: poor discriminatory ability.
- 0.6 - 0.7: reasonable ability to discriminate.
- 0.7 - 0.8: good discrminiatory ability.
- 0.8 - 0.9: strong discriminatory ability.
- 0.9 - 1: exceptional to perfect ability to discriminate.

Finally, the balanced error rate (BER) is the average error of each class, given as follows:

$$\text{BER} = \frac{1}{2}\left(\frac{FN}{FN+TP} + \frac{FP}{FP+TN}\right) \tag{2.34}$$

This metric is useful when interest is being able to accurately predict both positive and negative outcomes.

It is necessary at this stage to give a high level description of *variable importance* metrics, which are described in more detail in Chapter 3. In addition for the interested reader, an excellent description of variable importance metrics is given by Fisher et al. (2018). Although the term 'variable importance' can be used in the case of inference, in this work we restrict our attention to the case of prediction. In this case, a variable's importance relates to how much a model's accuracy relies on the given covariate. This definition highlights the distinction between 'performance' and 'importance' metrics, as the latter is a proxy for a specific covariate's contribution to the model's overall performance (captured by the accuracy performance metric). An example of a variable importance metric that is often used in practice with random forests (described in Chapter 3) is the decrease in prediction accuracy after permuting a covariate, i.e., the permuted importance metric (Breiman 2001*b*, Strobl et al. 2008).

Although importance metrics are not directly interpretable and do not provide information relating to the structure of the corresponding covariates in the model, they are useful for ranking the overall contribution of individual covariates to the performance of the model. This is useful for informing model selection and deciding whether a given variable should be included at all, once accounting for the confounding effects of the other predictor variables.

2.2.4 Missing Data in Longitudinal Studies

Missing data is highly prone to occur in longitudinal studies, due to the nature of the data being recorded and the way that it is collected. For instance, repeatedly measuring biomarkers on the same set of subjects over time, implies that the subject must be alive and present when the follow-up recordings are planned for the measurement to exist. In many cases, patients drop-out as result of death or censoring, which may be related to a treatment effect. In either case, missing data occurs such that it certainly is not completely at random (as it is related to the data being observed).

Following Rubin (1976), we informally outline the three mechanisms of missing data:

- *Missing Completely at Random (MCAR)*: The missingness process does not depend on the observed or unobserved data. The observed and missing data distributions are the same.

- *Missing at Random (MAR):* The missingness process is associated with the observed data. However, conditional on the observed data the missingness is unrelated to the unobserved data.
- *Missing not at Random (MNAR):* The missingness process directly depends on at least a subset of the unobserved (and possibly also the observed) data.

It is important to distinguish between the two types of patterns of missingness, namely *monotone* and *nonmonotone* missingness. The former relates to attrition (or drop-out), which occurs when a patients' planned longitudinal recordings are no longer recorded after the first missing data point occurs. Nonmonotone missingness relates to intermittent missing data, where a missing value occurs, more recordings are then observed and missingness again commences (Ibrahim & Molenberghs 2009).

The distinction between the two patterns of missingness is important because appropriate treatment of the problem depends on the type of pattern. Monotone missingness is more straightforward to address, as the likelihood function can be factorised into the conditional density terms (Ibrahim & Molenberghs 2009). In the case of nonmonotone missingness, factorising the likelihood function is most often not possible. Despite this, in the case of ignorability (MAR), longitudinal models that cope with unbalanced data (the linear mixed effects model) achieve satisfactory results (Rizopoulos 2012*b*).

We now define the following notation to formalise the concepts introduced so far. Let $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{n_i}$ be the set of all planned longitudinal measurements for the ith subject. The missing data process is denoted by $\boldsymbol{\omega}$, where $\omega_{ij} = \mathbb{I}(\exists\, y_{ij})$ (i.e, takes a value of 1 if y_{ij} exists and 0 otherwise). This allows us to partition the data into the observed component $\boldsymbol{y}_i^{Obs} = \{y_{ij}, \text{if } \omega_{ij} = 1\}_{j=1}^{n_i}$ and missing component $\boldsymbol{y}_i^{Mis} = \{y_{ij}, \text{if } \omega_{ij} = 0\}_{j=1}^{n_i}$.

The missingness mechanism is the conditional density of the missing data process given the complete response $\boldsymbol{y}_i$. This is denoted as: $p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}_\omega)$, where $\boldsymbol{\theta}_\omega$ is the vector of parameters corresponding to the missing data mechanism.

The above notation can be used to define the three missing data mechanisms mathematically (Ibrahim & Molenberghs 2009):

- Under MCAR, the missing data process is independent of both $\boldsymbol{y}_i^{Obs}$ and $\boldsymbol{y}_i^{Mis}$, i.e., $p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}_\omega) = p(\boldsymbol{\omega}; \boldsymbol{\theta}_\omega)$.
- Under MAR, the missing data process is independent of the unobserved data given the observed data, i.e., $p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}_\omega) = p(\boldsymbol{y}_i^{Obs}, \boldsymbol{\omega}; \boldsymbol{\theta}_\omega)$.
- Under MNAR, the missingness mechanism depends on at least some values of the unobserved response (and possibly also observed response values), i.e., $p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}_\omega) = p(\boldsymbol{y}_i^{Mis}, \boldsymbol{\omega}; \boldsymbol{\theta}_\omega)$ or $p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\omega}; \boldsymbol{\theta}_\omega)$.

The missing data mechanism is MAR when data is missing at random or when drop-out is random (also referred to as non-informative missingness or drop-out). As the distribution of $\boldsymbol{y}_i$ depends on $\boldsymbol{y}_i^{Obs}$, the observed data alone is not a random sample from the target population. However, when the missing data mechanism is MAR and assuming that the longitudinal model is correctly specified, likelihood-based analysis using the observed data alone can provide valid inference. Rizopoulos (2012*b*) demonstrates how this is the case by factorising the likelihood of the longitudinal model as follows:

$$\mathcal{L}_i(\boldsymbol{\theta}) = \int p(\boldsymbol{y}_i, \boldsymbol{\omega}_i; \boldsymbol{\theta})\ d\boldsymbol{y}_i^{Mis} \tag{2.35}$$

$$= \int p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}_y)\ p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}_\omega)\ d\boldsymbol{y}_i^{Mis} \tag{2.36}$$

$$= \int p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}_y)\ p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}; \boldsymbol{\theta}_\omega)\ d\boldsymbol{y}_i^{Mis} \tag{2.37}$$

$$= p(\boldsymbol{y}_i^{Obs}; \boldsymbol{\theta}_y)\ p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}) \tag{2.38}$$

$$= \mathcal{L}_i(\boldsymbol{\theta}_y) \times \mathcal{L}_i(\boldsymbol{\theta}_\omega). \tag{2.39}$$

Note that Bayes' rule is used to derive Equation 2.36 and the conditional independence assumption is used to derive Equation 2.37. If the product of the parameter spaces for $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\omega$ is equal to the parameter space of the full parameter vector $\boldsymbol{\theta} = [\theta_y^\top, \boldsymbol{\theta}_\omega^\top]$ (i.e., the two individual parameter vectors are disjoint), then it is legitimate to conduct inference based on the marginal observed density $p(\boldsymbol{y}_i^{Obs}; \boldsymbol{\theta}_y)$ (Rizopoulos 2012*b*). This is referred to as *ignorability*, as the likelihood of the missingness process can be ignored without compromising the validity of inference on the observed data.

When the missing data mechanism is MNAR, the missingness process is nonignorable (as the conditional independence assumption used to derive Equation 2.37 cannot be used) and a parametric model for the missingness process must be specified in conjunction with the longitudinal outcome. The model for the missingness process be incorporated into the complete data likelihood function to provide valid inference in the case of MNAR.

Various methods have been proposed to incorporate the missingness process into the likelihood function. The family of models proposed to model this joint distribution include: *selection models*, *pattern-mixture models* and *shared-parameter models*. Ibrahim & Molenberghs (2009) provide a review of these approaches that we now briefly describe.

Selection models (Diggle & Kenward 1994, Little 1995) decompose the joint distribution using the following factorisation:

$$p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\omega}_i; \boldsymbol{\theta}) = p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}_y)\ p(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}_\omega). \tag{2.40}$$

The first term on the right hand side of the equality in Equation 2.40 is the marginal density of the longitudinal process. The term that follows corresponds to the density of

the missingness process conditioned on the longitudinal process.

The reverse factorisation is used by pattern-mixture models (Little 1993), which is given by Equation 2.41.

$$p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\omega}_i; \boldsymbol{\theta}) = p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis} \mid \boldsymbol{\omega}_i; \boldsymbol{\theta}_y)\ p(\boldsymbol{\omega}_i; \boldsymbol{\theta}_\omega) \tag{2.41}$$

Here the joint likelihood is composed of the conditional distribution of the longitudinal process given the missingness process and the distribution of the missingness process.

Finally, we introduce shared-parameter models (Wu & Carroll 1988, Wu & Bailey 1989, 1988) which are the focus of this work. These models use latent variables such as random effects to account for the dependence between the longitudinal and missingness processes.

$$p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\omega}_i; \boldsymbol{\theta}) = \int p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis} \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y)\ p(\boldsymbol{\omega}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y)\ p(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\ d\boldsymbol{b}_i \tag{2.42}$$

The shared-parameter formulation (Equation 2.42) postulates that the longitudinal and missingness processes are independent given the latent random effects. We restrict our attention to these models due to their ability to handle both monotone and non-monotone missing data patterns. Moreover, software implementation of shared-parameter models is well established and implementing extensions to standard shared-parameter models (as discussed in Chapter 4) requires relatively minimal ad-hoc programming.

2.2.5 Joint Modelling Longitudinal and Time-to-Event Data

We now give a brief introduction to the *joint model for longitudinal and time-to-event data* (JM). More details are given in Chapters 4 and 5, including parameter estimation, performance evaluation, and various extensions that we employ to analyse the motivating data. In this chapter we present the standard JM formulation as discussed by Rizopoulos (2012*b*) and consider the various scenarios where a joint modelling approach may be beneficial to the analyst. We also discuss the connection between the JM and the missing data framework. We aim to exploit this important connection in our analysis in Chapter 5, which is the driver of our proposed approach that we investigate.

The JM comes under the family of shared-parameter models discussed in Section 2.2.4. In particular, parameter estimation is performed using the joint likelihood of the longitudinal and time-to-event processes. The likelihood is factorised by assuming conditional independence between the longitudinal and time-to-event processes, given the latent random effects. By including the time-to-event process in the likelihood function, parameter

estimation is performed conditioned on the missing data process. This approach therefore does not involve imputation of missing values, but allows for unbiased estimation when the longitudinal process involved endogenous time-dependent covariates (which corresponds to informative missingess)

Suppose that we are faced with the task of analysing a dataset that consists of n patients, where an individual patient is indexed by i. A biomarker (such as the patient's systolic blood pressure) is recorded repeatedly at specific (possibly irregular and or unequally spaced) time intervals, such that $y_{ij} = \{y_i(t_{ij})\}_{j=1}^{n_i}$. We compact this notation, denoting $y_i(t)$ as the observed biomarker value at t. Moreover, we have access to these patients' survival information. We denote the observed event time by T_i^* and the true event time $T_i = \min(T_i^*, C_i)$, where C denotes the censoring time. There is also a censoring indicator $\delta_i = \mathbb{I}(T_i^* < C_i)$.

The longitudinal component of the JM is modelled by a linear mixed effects model (LMEM) (Laird & Ware 1982), as given by Equation 2.43. This is used to model the evolution of the longitudinal trajectory over time.

$$y_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \epsilon_i(t) \tag{2.43}$$

$$y_i(t) = m_i(t) + \epsilon_i(t) \tag{2.44}$$

Here, $\boldsymbol{x}^\top(t)$ represents the design vector of the fixed effects, which corresponds to the fixed effects parameter vector $\boldsymbol{\beta}$. The random effects' design vector $\boldsymbol{z}_i^\top(t)$ corresponds to the random effects $\boldsymbol{b}_i$, which describe the ith individual's deviation from the mean profile. The measurement error $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ and the random effects $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$, where q represents the dimension of the random effects and $\boldsymbol{D}$ is the variance-covariance matrix of the random effects. Equation 2.44 highlights that the observed biomarker at t is the sum of the true unobserved biomarker value $m_i(t) = x_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i$ and the measurement error.

As the biomarker is generated by the same patient that the survival information is recorded, it constitutes an endogenous time-dependent covariate (Kalbfleisch & Prentice 2002). The time-dependent Cox model results in biased parameter estimates when the time-dependent covariate is endogenous, and the full likelihood must be used to perform parameter estimation (Wulfsohn & Tsiatis 1997, Rizopoulos 2012*b*).

The time-to-event component of the JM postulates the following relative risk model, where the instantaneous risk depends on the history of longitudinal measurements $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ and the design matrix for the confounding variables is $\boldsymbol{w}_i$, which corresponds to the parameter vector $\boldsymbol{\gamma}$.

$$h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) = \lim_{dt \to 0} \frac{1}{dt} P(t \leq T_i^* < t + dt \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i) \tag{2.45}$$

$$= h_0(t) \exp(\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha[\boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i]), \ t > 0 \tag{2.46}$$

The parameter α contained within the frailty term represents the association between the longitudinal and survival processes. When $\alpha = 0$ the missingness mechanism corresponds to MCAR, because it does not depend on either $\boldsymbol{y}_i^{Obs}$ or $\boldsymbol{y}_i^{Mis}$ (in which case there is no benefit from employing a joint modelling approach). Due to the fact that both Equations 2.43 and 2.46 share the same random effects, the JM comes under the family of shared-parameter models.

2.2.5.1 Connection with the Missing Data Framework

It can be seen that the joint model for longitudinal and time-to-event data has a distinct connection with missing data in longitudinal studies. This is not by any means of imputation, but rather by enabling valid inference (even in the case of MNAR, resulting from the endogenous nature of the time-dependent covariates), by incorporating the missingness process into the joint likelihood function by explicitly specifying a survival model. This section formally describes how this is so.

The time-to-event component of the JM explicitly specifies a survival model for the missingness process. Although the goodness-of-fit of this model is not testable from the data, this formulation corresponds to the MNAR missingness mechanism (Equation 2.47). Rizopoulos (2012*b*) shows that by making implicit assumptions for the complete longitudinal profile $[\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}]$, the JM missingness mechanism depends on $\boldsymbol{y}_i^{Mis}$ through the posterior distribution of the random effects conditioned on the complete longitudinal response, and therefore corresponds to MNAR. Note that Equation 2.49 is derived by assuming that the survival and longitudinal processes are independent given the random effects.

$$p(T_i^* \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}, \boldsymbol{\theta}) = \int p(T_i^*, \boldsymbol{b}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}) d\boldsymbol{b}_i \tag{2.47}$$

$$= \int p(T_i^* \mid \boldsymbol{b}_i, \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}) \ p(\boldsymbol{b}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}) d\boldsymbol{b}_i \tag{2.48}$$

$$= \int p(T_i^* \mid \boldsymbol{b}_i; \boldsymbol{\theta}) \ p(\boldsymbol{b}_i \mid \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}) d\boldsymbol{b}_i \tag{2.49}$$

Note that the JM censoring mechanism corresponds to MAR, as the missingness process is assumed to depend on the observed covariates and longitudinal responses (but not the unobserved values). We restrict our attention in this work to right censoring that is non-informative.

Rizopoulos (2012*b*) also illustrates how shared-parameter models have an elegant and computationally efficient way of dealing with both intermittent missingness and attrition (as opposed to the computationally demanding approaches of selection models and pattern-mixture models). This can be seen by writing the likelihood of the observed data under the complete data model $\{\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}\}$ for the longitudinal outcome (as displayed in Equations 2.50 to 2.53).

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int p(T_i, \delta_i, \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis}; \boldsymbol{\theta}) \; d\boldsymbol{y}_i^{Mis} \tag{2.50}$$

$$= \prod_{i=1}^{n} \int \int p(T_i, \delta_i, \boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis} \mid \boldsymbol{b}_i; \boldsymbol{\theta}) \; p(\boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{y}_i^{Mis} \; d\boldsymbol{b}_i \tag{2.51}$$

$$= \prod_{i=1}^{n} \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}) \left[\int p(\boldsymbol{y}_i^{Obs}, \boldsymbol{y}_i^{Mis} \mid \boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{y}_i^{Mis} \right] p(\boldsymbol{b}_i; \boldsymbol{\theta}) \; d\boldsymbol{b}_i \tag{2.52}$$

$$= \prod_{i=1}^{n} \int p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}) \; p(\boldsymbol{y}_i^{Obs} \mid \boldsymbol{b}_i; \boldsymbol{\theta}) \; p(\boldsymbol{b}_i; \boldsymbol{\theta}) \; d\boldsymbol{b}_i. \tag{2.53}$$

Equation 2.52 is derived by assuming conditional independence between the longitudinal and survival processes given the random effects. Clearly, the missing data here is only related to the longitudinal component of the likelihood function. The second conditional independence assumption (that the longitudinal measurements are independent of each other given the random effects) can be exploited so that the integral in Equation 2.52 is easily dropped to derive Equation 2.53.

2.2.5.2 When could a Joint Modelling Approach be Beneficial?

We have already covered various possible motivations behind employing a joint modelling approach. In particular, the JM can be used to provide valid inference when a longitudinal covariate constitutes an endogenous time-dependent covariate and is subject to measurement error. Although it is not possible to distinguish between MAR and MNAR from the data alone (this can only be known from expert knowledge of the problem at hand), the JM can produce valid inference under MNAR. It is sensible to employ the JM when the missingness mechanism is unknown, as a means to conduct a sensitivity analysis. If the parameter estimates are sensitive between the LMEM and JM approaches, this indicates that the mechanism is likely to be MNAR.

Henderson et al. (2000) outlines various possible interests of the analyst when a joint modelling approach is appropriate:

- **Interest is in the survival outcome**. The analyst may be interested in the distribution of the event times conditional on an endogenous time-dependent covariate that is contaminated with measurement error. This also comes under the

case where the purpose of the study is to exploit the rich nature of longitudinal data to perform dynamic prediction. This is our motivation for employing the joint modelling framework in Chapter 4.

- **Interest is in the longitudinal outcome**. It is often the case where the analyst is concerned with a longitudinal outcome that is subject to possible outcome dependent drop-out. This corresponds to adjusting for the implicit outcome of missing data. This relates to our motivation for employing this framework in Chapter 5.

- **Interest is in the association between the survival and longitudinal processes**. In this case the parameter α is of interest as it describes the association between the two processes.

In this work we employ the JM for various purposes. In particular, in Chapter 4 we wish to take advantage of the rich nature of DCD donor physiological profiles throughout the treatment withdrawal to death period (while adjusting for baseline covariates) in order to dynamically predict survival probabilities between medically relevant time frames. In chapter 5 we study whether the JM's ability to deal with informative missingness and measurement error simultaneously can be exploited to extract information from an endogenous time-dependent covariate, which becomes a covariate in a predictive model for predicting transplant outcome DGF.

Part II

An Application of Machine Learning Methods

Chapter 3

Ensemble Learning for Predicting Failed Kidney Transplants

Summary

The prevalence of delayed graft function (DGF) is rising due to a substantial increase in the number of kidney transplants from deceased after circulatory death (DCD) donors. Being able to accurately predict DGF for recipients from DCD donors at an individual level would enable more personalised and effective medical care. In this chapter we benchmark the ability of various machine learning ensemble methods (adaptive boosting, extreme gradient boosting, random forests and conditional random forests) for predicting DGF against a semi-parametric random intercept logistic regression that contains spline terms.

The NHSBT dataset is inherent of various features that typically complicate analysis. In particular, having a hierarchical structure, non-linear associations, categorical variables with a large number of categories, and highly correlated variables; can cause standard methods of analysis to break down. We consider whether ensemble methods have the potential to be an invaluable set of tools for the NHSBT in related, future applications.

By the means of a simulation study we assess the ensemble methods' ability to identify important variables, when faced with the complications present in the NHSBT dataset. Finally, we propose a novel approach for visualising variable importance when missing data is present and multiple imputation is performed.

3.1 Introduction

For patients with ESRD there are two treatment options. Dialysis can be undertaken, where a machine is used to perform the functioning of the failing kidney. However this option is not always sustainable and in most cases the alternative option of a kidney transplant is the preferred (and only long-term) treatment. A kidney transplant does not always come without complication. In some cases, once the kidney is transplanted into the recipient it fails to function immediately. Delayed graft function (DGF) is most often defined as the failure of the kidney to function within seven days of transplant (Yarlagadda et al. 2008) and is a bottleneck in the transplantation process.

Previously there had been a consensus that transplant outcomes for donation after circulatory death (DCD) donors are inferior to donation after brain-stem death (DBD) donors, which are the most common type of donor. However, a shift in attitude occurred as a result of research findings suggesting that the two types of donors have comparable long-term transplant outcomes (Summers et al. 2015). This led to a substantial increase in the number of DCD donor transplants in the UK, as the NHSBT implemented a transplant program specifically for controlled DCD. Nevertheless, recipients from DCD donors are more susceptible to developing DGF, which is thought to be due to ischaemic injury being inflicted during the treatment withdrawal to death phase. As research is limited with regards to what characteristics of the withdrawal phase are associated with the chances of a recipient experiencing DGF, we investigate this research question in Chapter 5.

Clinicians record data relating to the physiological profiles during the treatment withdrawal phase, as well as donor, recipient and graft characteristics. Being able to predict which patients are likely to develop DGF and determine which factors contribute towards the occurrence of DGF would enable clinicians to make more informed decisions, such as which patients the removal team should attend. Deriving a predictive model for transplant outcome DGF based on the motivating data, and ranking the importance of variables in the derived model are the primary aims of this chapter.

Due to the nature of the kidney transplantation process, these data were not gathered according to a well designed experiment. One should acknowledge that this is observational data that is to be analysed retrospectively. In this case, we conjecture that it is not possible to determine causal effects and that results from any such analysis should be used as no more than a guide. However, it is worth noting that in some cases observational data is preferred to "overly clean" experimental data when the aim is prediction, as it reflects the realistic context of the problem due to the presence of uncontrolled factors and noise (Shmueli et al. 2010).

Machine learning methods are appealing for any application that involves the analysis of observational data. This is due to their ability to elegantly cope with various possible

complications that can arise with real data, that are problematic for standard statistical methods. Many of these possible complications are present within the NHSBT dataset (including non-linear associations, correlated variables, a large number of variables). However, little research to this author's knowledge has been conducted to study their predictive performance, and ability to rank the importance of variables, when a hierarchical structure is present and categorical variables exist that have many categories yet a relatively small sample size.

Popular machine learning methods include the artificial neural network (ANN), the support vector machine (SVM) (Vapnik 1999), decision trees (DT) (Breiman et al. 1984) and random forests (RF) (Breiman 2001*a*). Despite the large amount of scepticism that these methods have quite rightly faced in the past due to being treated as "black-box" algorithms, they have proven to be tenable and their application has now become mainstream in a multitude of applications. The majority of these applications involve the analysis of large datasets, however, this is not always the case and it has become more common to analyse relatively small datasets, such as the NHSBT dataset, with machine learning methods.

Recall from Section 2.2.1 the characteristics that determine an appropriate method for the problem at hand. A model should be sufficiently accurate, interpretable, efficient and robust to complications in the data. Based on these criteria we consider tree based ensemble methods, while acknowledging that interpretability is limited. This choice was in part inspired by these methods having achieved an exceptionally high predictive ability when applied to various prestigious machine learning competitions, such as Kaggle[1]. A notable example was the use of ensemble methods to win the Netflix prize (Bell et al. 2010).

In order to be aware of the dangers that arise from using black box algorithms we cover relevant methodological concepts that underlie these methods. A secondary aim of this chapter is to assess the applicability of these methods to the NHSBT dataset. This is examined by benchmarking the discriminatory ability of various machine learning algorithms to standard statistical methods, and by a simulation study designed to test the algorithms' ability to rank the importance of covariates at various levels of sample size.

The remainder of this chapter is structured as follows. In Section 3.2 a review of relevant literature is given, justifying this work by exposing grey areas in current research and providing scope for novelty in this chapter, in terms of both application and methodology. Subsequently, the ensemble methods are introduced in Section 3.3, providing relevant concepts and intuition as to how they function. This is followed by a simulation study in Section 3.4 to assess the ability of ensemble methods to identify important variables.

[1]Kaggle is an online community of data scientists owned by google, that runs machine learning competitions.

Finally, we conduct the analysis in Section 3.5, where the objective is to achieve the primary aims stated above.

3.2 Review of Literature

Specialists have claimed that being able to accurately predict kidney transplant outcomes would be invaluable for the clinical decision making process (Bergler & Hutchinson 2017). The US already (albeit controversially) use post-transplant predicted survival time as a basis for allocating kidneys. The proportional hazards model that was used to form this basis, was found to have a reasonably good discriminatory ability, achieving 0.69 Harrell's concordance index (C-index) (Clayton et al. 2014).

There are many other studies that employ standard statistical methods (logistic or proportional hazards regression) in order to predict patient or survival outcomes (Thorogood et al. 1991, Hernández et al. 2005, Tiong et al. 2009, Foucher et al. 2010, Lin et al. 2008). A relatively recent predictive analysis was conducted by Molnar et al. (2017). Using proportional hazards models, they found many variables to be significantly associated with graft survival (recipient age, cause and duration of ESRD, haemoglobin, albumin, selected co-morbidities, race, type of insurance, donor age, diabetic status, number of HLA-mismatches). This model achieved a 0.63 C-index, demonstrating reasonable discriminatory ability.

The NHSBT have conducted studies attempting to predict transplant outcomes for kidney recipients from deceased donors. Watson et al. (2012) analysed 7,620 adult recipients from deceased donors between 2000 to 2007. Recipient age, ethnicity, primary renal disease, CIT, HLA mismatch, donor age, donor hypertension, donor weight, hospital duration and use of adrenaline were found to be significantly associated with graft survival. A C-index of 0.62 was achieved. Watson et al. (2012) compare their analysis to that of Rao et al. (2009), claiming that their selected model is more parsimonious yet has a comparable predictive ability (where Rao and colleagues achieved 0.63 C-index).

The study by Watson et al. (2012), however, could be criticised for categorising the continuous covariate age. Modelling a variable that is continuous by nature with a step function is unnatural and will often result in a loss of predictive accuracy (Harrell 2015). As Shmueli et al. (2010) explain, predictive models are fitted with a different procedure and aim to that of explanatory models and when the former are employed the aim should not be to simplify the model in order to improve interpretability, or else a loss in predictive ability is inevitable.

Many of the same authors from the study Watson et al. (2012) were involved in another study (Li et al. 2016) that aimed to predict long-term survival outcome for recipients from deceased donors. This was based on a dataset consisting of 12,000 recipients

from 2003 to 2013. In this study, non-linearity was dealt with by employing a flexible proportional hazards regression with restricted cubic spline terms. However, model selection becomes significantly more complicated under this approach, as many more potential combinations of models become apparent. Testing interactions with all possible non-linear combinations of continuous variables is not computationally feasible.

Predicting DGF can also be useful for clinical decision making (Schröppel & Legendre 2014). An attempt to predict DGF was made by Irish et al. (2003), performing logistic regression on 13,846 recipients from deceased donors between 1995 and 1998. In this study the incidence of DGF was 23.7% and an AUC of the ROC was 0.70, indicating a good discriminatory ability. Irish et al. (2010) performed another 'new era' logistic regression analysis on 24,337 recipients from deceased donors between 2003 and 2006, achieving an identical discriminatory ability of AUC of the ROC. They found many variables to be significantly associated with DGF (ethnicity, gender, dialysis at transplant, single organ donor, previous transplant, panel reactive antibody, repeat blood transfusion, HLA mismatch, donor age, CIT, hypertension, heart beating, donor cause of death). In the latter study warm ischaemic time and recipient BMI were also included in the analysis and were found to be significant.

External validation of the study by Irish et al. (2010) found an AUC of the ROC equal to 0.69 (Michalak et al. 2017). These studies could be criticised for not having accounted for the multilevel structure of the data as a result of multiple kidneys being donated from a single donor. Use of random intercept logistic regression, for example, may have improved the predictive ability.

The application of machine learning to kidney transplant data is an emerging area of research. Supervised learning algorithms, such as the ANN, SVM, DT and RF, have been used to predict graft or patient survival of kidney recipients (Greco et al. 2010, Shaikhina et al. 2017, Akl et al. 2008, Lasserre et al. 2011, Vijayarani & Dhayanand 2015). The predictive ability achieved in these studies varies substantially. Lasserre et al. (2011) claim that this variation is attributed to the varying quality of the available data. They suggest that the data origin, the number of samples, a few extra features and the chosen outcome make up most of the difference. A closer inspection reveals that the studies including both pre and post-transplant clinical indicators achieved higher accuracy (Greco et al. 2010, Akl et al. 2008). Furthermore, the studies whose datasets consisted of living donors had a higher accuracy compared to those with deceased donors (Akl et al. 2008).

Krikov et al. (2007) analysed a dataset relating to a large multi-centre study consisting of 94,844 patients. Based on 31 pre-transplant predictor variables (relating to the donor, recipient and graft characteristics) they used DT to predict graft survival at 1, 3, 5, 7 and 10 years. They achieved AUC of the ROC of 0.63, 0.64, 0.71, 0.82 and 0.90 respectively. The unintuitive increase in accuracy that occurs as events are being predicted further

into the future is a result of the imbalance between classes becoming more prominent over time. Graft losses and drop-out at earlier stages result in a high success rate for those that remain alive in the distant future, causing this imbalance. If there is a 90% success rate a predictive model will appear to perform very well if a success is predicted for every observation.

The phenomenon of a high predictive accuracy as a result of imbalanced classes is also present in studies that use machine learning to predict DGF using all types of donors. This is because its incidence is only relatively high for DCD donors. Shoskes et al. (1998) used a dataset consisting of 100 recipients from deceased donors (from a single transplant centre) to predict DGF by employing a neural network. In this study, only 24% of patients developed a DGF and by randomly partitioning the dataset such that 80 of the transplants were used to train the neural network and 20 were used to determine the predictive ability, they reported 80% accuracy. This result is questionable having used such little data to estimate so many parameters, and by a poor choice of evaluation criteria. A stratified K-fold cross-validation would have enabled the whole dataset to be used to estimate accuracy while accounting for the unbalanced classes.

Brier et al. (2003) compared the performance of an ANN to logistic regression for predicting DGF using a dataset consisting of 304 kidney transplants from deceased donors. In this single centre study there was a higher incidence of 38% DGF. The ANN was found to be more sensitive to predicting the presence of DGF (56% against 37%), while logistic regression was more sensitive to predicting the absence of DGF (91% against 70%). They found the interaction between donor and recipient ethnicity to be the only variable that had a highly significant association with the development of DGF. The ANN had a marginally better predictive ability overall with an AUC of the ROC 0.67, in contrast to 0.64 for logistic regression.

Decruyenaere et al. (2015) compared the performance of logistic regression to eight machine learning algorithms for predicting DGF. The dataset consisted of 497 transplants from deceased donors (10% of which were DCD donors), which was obtained from a single transplant centre. There was a low incidence of DGF in this dataset of 12.5%. 24 predictors were used by the learning algorithms DT, RF, stochastic gradient boosting (SGB), SVM (with linear, polynomial and radial basis functions), linear and quadratic discriminant analysis (LDA and QDA). AUC of the ROC (using stratified K-fold cross-validation) achieved 0.53, 0.74, 0.77, 0.84, 0.80, 0.83, 0.82, 0.80 respectively. However, the estimate of the generalisation error is likely to be optimistic as a result of the same data being used to tune the hyper-parameters as was used to estimate the generalisation error. A nested cross-validation approach (Algorithm 1) would have been a less biased estimate despite it being much more computationally demanding.

To the best of this author's knowledge, no study exists that attempts to predict DGF based on a dataset consisting of only DCD donors. Due to the increased incidence of

DGF for DCD donors, a binary outcome indicating the development of DGF is much less likely to suffer from a biased estimate of the generalisation error as a result of highly imbalanced classes. Although Yoo et al. (2017) used tree based ensemble methods to predict kidney patient survival, there are many variants of ensemble methods that have not yet been applied to transplant data (particularly for predicting DGF). For example, adaptive gradient boosting (AdaBoost) and extreme gradient boosting (XGBoost) are some of the most popular boosting methods providing further scope for novelty in this research.

3.3 Methods

In this section various methods are discussed that are employed in the analysis section later in this chapter. The machine learning methods considered are variations of tree based ensembles, that have been used in a wide range of applications. In the context of classification (which we restrict our attention to), ensemble methods provide a means of combining a set of *weak learners*[2] to create a powerful classifier. The idea is analogous to model averaging in statistics and two powerful classes of algorithms, namely *bagging* and *boosting*, dominate in practice as a means of combining weak learners. Both of these methods have deep roots in statistics that we consider later in this section.

Regression trees have a number of properties (discussed in Section 3.3.1) that make them highly effective weak learners for an ensemble. For this reason, they are the most popular choice of weak learner to form an ensemble in practice. As the methods we employ are all tree based, regression trees are now introduced for the sake of completeness.

3.3.1 Regression Trees

Regression trees are a class of tree based methods for non-parametric regression and classification. In either case, the model takes a vector of inputs $(\boldsymbol{x}_i, y_i)$ and returns an output $\hat{y}_i$ (which is a prediction) for each observation $(i = 1, \ldots, n)$.

The feature space $\mathcal{X}$ is recursively partitioned into T rectangular areas $R_1, \ldots, R_T$. The idea is that observations with similar outcomes are grouped by these distinct partitions (interchangeably referred to as regions, leaves or nodes). The regression tree that fits a constant model to each region is displayed in Equation 3.1, where w represents the leaf weights. In this case, the decision tree is in effect a piecewise constant function.

$$\varphi(\boldsymbol{x}) = \sum_{j=1}^{T} w_j \mathbb{I}(\boldsymbol{x} \in R_j), \;\; w_j \in \mathbb{R} \tag{3.1}$$

[2]Weak or base learners refer to simple predictive models that have limited predictive accuracy, i.e., they perform only slightly better than random guessing.

Regression trees can also be seen as adaptive basis functions. In this case there are T basis functions $\phi_j(x) = \mathbb{I}(x \in R_j)$ (which represent the nodes). The number of nodes is a hyper-parameter that controls the complexity of the model. This integer value is best determined using cross-validation.

Minimising the empirical risk given by Equation 3.2 requires choosing an appropriate loss function L and determining both the regions and weights. This is an ill-posed problem due to a possibly infinite number of regions. This is circumvented by approximating the target function with a *greedy top-down*[3] binary partitioning algorithm. Various algorithms exist for this purpose. However, the CART (classification and regression trees) (Breiman et al. 1984) algorithm appears to be the most developed in available software, and is used in the ensemble methods that we consider in this chapter.

$$\hat{R}(\varphi) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{T} w_j \mathbb{I}(\boldsymbol{x}_i \in R_j)\right) \tag{3.2}$$

The *gain* is calculated as the difference between the empirical risk of the model evaluated before and after a new split is made. Having tried each possible split at each possible node, the split is made such that the gain is maximised. The splitting continues until a specified stopping criterion is achieved. In practice, this is usually chosen to be until a certain number of observations are contained within each region.

The complexity of decision trees can be controlled in various ways. As previously discussed, specifying the number of terminal nodes and a stopping criterion (the number of observations in a given region) can stop the model from over-fitting the data. Furthermore, a "tall" tree can be grown and pruned in a bottom-up fashion, a process named *cost-complexity pruning* (Hastie et al. 2008). This refers to growing a complex tree and removing terminal nodes that have a negative gain. This is equivalent to penalising the empirical risk function to give the penalised objective function $\tilde{\mathcal{L}}(\varphi)$ as displayed in Equation 3.3.

$$\tilde{\mathcal{L}}(\varphi) = \hat{R}(\varphi) + \lambda T \tag{3.3}$$

Hastie et al. (2008) outline various advantages and disadvantages of regression trees. As previously stated, they are able to capture interactions and non-linear associations and are robust to outliers. They have an in-built model selection procedure and are invariant to monotone transformations of the input. They have a natural approach to dealing with missing data and are scalable to large datasets.

Regression trees are not competitive in terms of predictive ability compared to other methods as they have a high variance and are unstable. Due to fitting rectangular

[3]This refers to the splits being made based on what appears the best at the time (that minimises the empirical risk), without looking ahead of the current split, which otherwise may have further reduced the empirical risk. See Friedman et al. (2001) for more details.

regions, they often lack smoothness. Strobl et al. (2007) show that they are biased towards selecting categorical variables that have many categories (that tend to result in over-fitting) and towards numerical variables that have a large number of distinct values.

3.3.2 Adaptive Boosting (AdaBoost)

Kearns (1988) first posed the question of whether multiple weak learners can be combined to create a powerful classifier. Schapire (1990) showed that this was possible, which led to the first practical boosting algorithm AdaBoost (Freund & Schapire 1997). This remains one of the most popular boosting algorithms.

Suppose that a sequence of weak learners $\{\varphi_m(\boldsymbol{x})\}_{m=1}^{M}$ return predictions coded such that $\hat{y} \in \{-1, 1\}$, where in our case $\varphi_m(\boldsymbol{x})$ is given by Equation 3.1. Each of these weak learners (regression trees) correspond to basis functions of the boosted tree model. The tree weights $\{\alpha_m\}_{m=1}^{M}$ allow trees with a higher predictive accuracy to assign a larger weight towards the final predicted values in the voting process. The AdaBoost algorithm (Algorithm 2) (Hastie et al. 2008) works by iteratively weighting the data, such that on each iteration more attention is given to observations that were previously misclassified. Line 10 of Algorithm 2 shows how the final predictions are determined via a majority vote, which explains why $\hat{y}$ is coded as $\in \{-1, 1\}$.

Algorithm 2 AdaBoost (Adaptive Boosting)

1: Initialise the observation weights $\{\tilde{w}_i\}_{i=1}^{n} = \frac{1}{n}$
2: **for** $m = 1, \ldots, M$ **do**
3: Fit the weak classifier $\varphi_m(\boldsymbol{x})$ to the training data using the weights w_i.
4: Compute:
$$Err_m = \frac{\sum_{i=1}^{n} \tilde{w}_i \mathbb{I}(y_i \neq \varphi_m(\boldsymbol{x}))}{\sum_{i=1}^{n} \tilde{w}_i}$$
5: Compute the tree weights $\alpha_m = \log((1 - Err_m)/Err_m)$
6: **for** Update the observation weights $i = 1, \ldots, n$ **do**
7: $\tilde{w} \leftarrow \tilde{w}_i \exp(\alpha_m \mathbb{I}(y_i \neq \varphi_m(\boldsymbol{x})))$
8: **end for**
9: **end for**
10: Output the strong classifier $\varphi(\boldsymbol{x}) = \text{sign}(\sum_{i=1}^{n} \alpha_m \varphi_m(\boldsymbol{x}))$

The remarkable results that this algorithm has achieved in practice, as a result of reducing both bias and variance of the weak learners, was at one time a mystery. Friedman et al. (2000) proved that this phenomenon can be explained by the well-known statistical principle of additive modelling. In particular, they showed that AdaBoost can be interpreted as a stage-wise estimation procedure for fitting an additive logistic regression model. AdaBoost uses the *forward stage-wise additive modelling* procedure (Hastie et al. 2008) which iteratively fits the following:

$$\{\hat{\alpha}_m, \hat{\gamma}_m\} = \arg \min_{\{\alpha_m, \gamma_m\}} \sum_{i=1}^{n} L\left(y_i, \hat{\varphi}^{(m-1)}(\boldsymbol{x}) + \alpha_m \varphi_m(\boldsymbol{x}; \gamma_m)\right), \tag{3.4}$$

where γ parametrises the basis functions, which in this case refers to the split variables and split points at the nodes. The number of iterations should be determined by cross-validation.

3.3.3 Extreme Gradient Boosting (XGBoost)

XGBoost (Chen & Guestrin 2016) is a highly scalable machine learning algorithm. It employs a Newton gradient boosting method proposed by Friedman et al. (2000), for optimisation in the function space. It has been implemented in many programming languages and is used in this work in R (Chen et al. 2015). XGBoost has won many machine learning competitions, notably the Higgs Boson discovery challenge (Chen & He 2015).

We proceed with the same task as in Section 3.3.2, of finding the function φ that best fits the data using boosted tree models:

$$\hat{y} = \varphi(\boldsymbol{x}) = \sum_{m=1}^{M} \varphi_m(\boldsymbol{x}), \;\; \varphi_m \in \Phi \tag{3.5}$$

where Φ is the function space containing all possible classification trees. As this is an infinite function space, the solution is approximated with a greedy additive modelling approach. The regularised objective function is given in Equation 3.6, where the loss function is chosen to be convex and differentiable. Here the second term represents the regularisation term penalising the complexity of the model.

$$\tilde{\mathcal{L}}^{(m)}(\varphi) = \sum_{i=1}^{n} L\Big(y_i, \hat{y}^{(m-1)} + \varphi_m(\boldsymbol{x}_i)\Big) + \sum_{j=1}^{m} \Omega(\varphi_j) \tag{3.6}$$

In this application the logistic loss function is used:

$$L(y_i, \hat{y}_i) = y_i \log(1 + e^{-\hat{y}_i}) + (1 - y_i) \log(1 + e^{\hat{y}_i}). \tag{3.7}$$

Optimisation for this objective function must be performed in the function space, which means that traditional methods of optimisation that are performed in the Euclidean space are not appropriate. A second order Taylor expansion is used to approximate the

loss function.

$$\tilde{\mathcal{L}}^{(m)}(\varphi) = \sum_{i=1}^{n} L\Big(y_i, \hat{y}_i^{(m-1)} + g_i\varphi_m(\boldsymbol{x}_i) + \frac{1}{2}h_i\varphi_m(\boldsymbol{x}_i)^2\Big) + \sum_{j=1}^{m}\Omega(\varphi_j) \tag{3.8}$$

where the gradient $g_i = \partial_{\hat{y}^{(m-1)}} L(y_i, \hat{y}^{(m-1)})$ and the Hessian $h_i = \partial^2_{\hat{y}^{(m-1)}} L(y_i, \hat{y}^{(m-1)})$. The XGBoost regularisation term penalises both the number of leaves with hyper-parameter γ and imposes an L2 norm penalty on the leaf weights that is controlled by the hyper-parameter λ, such that:

$$\Omega(\varphi_m) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2. \tag{3.9}$$

With some manipulation Equation 3.8 can be written as Equation 3.10, which is the sum of T independent polynomial functions of w (Chen & He 2015).

$$\tilde{\mathcal{L}}^{(m)} = \sum_{j=1}^{T}\left[\Big(\sum_{i\in I_j} g_i\Big)w_j + \frac{1}{2}\Big(\sum_{i\in I_j} h_i + \lambda\Big)w_j^2\right] + \gamma T, \tag{3.10}$$

where I_j represents the indices for those observations that are contained within leaf j. Now compacting notation by setting $G = \sum_{i\in I_j} g_i$ and $H = \sum_{i\in I_j} h_i$, the optimal weights $\tilde{w}$ for a fixed structure can be derived as:

$$\tilde{w}_j = \frac{G_j}{H_j + \lambda} \tag{3.11}$$

In determining the structure at the current iteration $\{R_j\}_{j=1}^{T}$, a greedy approach is employed to maximise the gain defined as:

$$Gain = \frac{1}{2}\Big[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\Big] - \gamma \tag{3.12}$$

where indices L and R represent the daughter nodes of an attempted split. Once the regions and optimal weights are determined, the function is estimated for the current iteration as given in Equation 3.13. Here η is the step-size (or shrinkage), in practice usually set to 0.1 (Chen & He 2015).

$$\hat{\varphi}_m(\boldsymbol{x}) = \hat{y}^{(m-1)}(\boldsymbol{x}) + \eta\sum_{j=1}^{T}\tilde{w}_j\mathbb{I}(\boldsymbol{x}\in\hat{R}_j) \tag{3.13}$$

Other hyper-parameters in the XGBoost algorithm include:

- *Maximum depth*: the maximum depth to grow the regression trees.

- *Stopping criterion*: the threshold specifying the minimum sample size within a leaf such that splits stop.
- *Row subsampling*: random subsample of rows (without replacement) (Friedman 2002), fitting trees to samples at each boosting iteration. This hyper-parameter is determined as a fraction of subsampling.
- *Column subsampling*: the fraction of variables to include at each boosting iteration (Ho 1998).

3.3.4 Random Forests

RF are conceptually more straightforward than boosting based ensembles. Although both methods in the classification context involve a committee of trees that cast a vote for a predicted outcome, the pivotal difference is that boosted trees evolve over time whereas RF build a set of de-correlated trees. RF use bagging (bootstrap aggregation) (Breiman 1996), which involves taking bootstrap samples of the data, fitting a regression tree to each sample and subsequently combining the output. By inducing random variability through sampling with replacement, this approach reflects the inherent nature of the sampling process. RF have improved predictive accuracy compared to weak learners by reducing variance through averaging multiple trees. However unlike boosted trees, RF are unable to reduce bias as the trees are identically distributed because they are fitted on bootstrap samples of the original dataset (Hastie et al. 2008).

RF tend to be favoured in practice compared to bagged trees. This is due to a few modifications in the algorithm that increases the potential for an improved predictive accuracy. In particular, by pre-specifying a number of features by which to randomly subset the original set of features (to be tried at each node split), RF create more diverse trees. This gives features that would otherwise be consistently outplayed by a dominating feature a chance to determine the tree structures. This in essence "levels the playing field". Algorithm 3 displays the algorithm for RF.

RF have the additional benefit of an alternative and convenient approach for estimating the generalisation error. This is referred to as the out-of-bag error, which uses the observations that were not included in the bootstrap sample as an in-built test set. This provides an unbiased estimate of the generalisation error (Breiman 2001*a*, Strobl et al. 2009).

Part of the popularity of RF in practice is attributed to its ability to provide insight into the predictive structure of the model. This insight is achieved by using variable importance measures, most of which are a weighted mean of impurity[4] reduction across

[4]In this context an impure node is a rectangular region in the feature space that contains a large number of misclassified observations.

Algorithm 3 Random Forests

1: **Input**: $\mathfrak{D}_{train}$, number of trees (ntree), number of variables tried at each split $mtry$, minimum node size.
2: **for** $i = 1, \ldots, ntree$ **do**
3: Take a bootstrap $\mathfrak{D}_i^*$ sample of size n from $\mathfrak{D}_{train}$.
4: Grow a tree using $\mathfrak{D}_i^*$, by the following steps for each node until the minimum node size is reached:

- Randomly select $mtry$ features from $\mathfrak{D}_i^*$.
- Select the best feature and splitting point from the $mtry$ features.
- Create a binary partition to create two daughter nodes.

5: **end for**
6: **Output**: Aggregate the predicted values over the $ntree$ trees (the mean or mode for regression or classification respectively).

all nodes that contain the feature of interest. Louppe et al. (2013) show that the mean decrease in impurity measures are zero if and only if the corresponding variable is irrelevant and that the importance of a relevant variable is invariant to the addition or removal of irrelevant variables.

Empirical studies show that some measures of impurity, such as the Gini importance index[5] (Friedman et al. 2001), can be biased in certain situations favouring variables composed purely of noise (Strobl et al. 2007). In particular, this measure is optimistic towards categorical variables with a large number of categories and also towards continuous variables with many distinct values (both types are present in the NHSBT dataset).

A more robust measure of performance that considers the impact of each variable while adjusting for the others is the permutation importance. The intuition behind this method involves evaluating the out-of-bag accuracy for a model including variable x_j among others, evaluating it again having permuted variable $x_{j'}$ and assessing the difference in accuracy. In the case that x_j is associated with the target variable, the prediction accuracy will drop significantly.

Following Strobl et al. (2007), let $B^{(t)}$ be the out-of-bag sample for tree t, where $t = 1, \ldots, ntree$. The permuted variable importance for x_j on the tth tree is given as follows:

$$VI^{(t)}_{permute}(x_j) = \frac{\sum_{i \in B^{(t)}} \mathbb{I}(y_i = \hat{y}_i^{(t)})}{|B^{(t)}|} - \frac{\sum_{i \in B^{(t)}} \mathbb{I}(y_i = \hat{y}^{(t)}_{i,\pi_j})}{|B^{(t)}|}, \tag{3.14}$$

where $\hat{y}^{(t)}$ is the predicted class of the ith individual from tth tree based on the full feature vector and $\hat{y}^{(t)}_{i,\pi_j}$ is the same except the jth feature is randomly permuted. It can be seen that this is simply the mean accuracy of a given tree's predictions before and

[5]The Gini index can be expressed as: $\sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$, where $\hat{p}_{jk}$ are the number of observations in the kth class in region j.

after x_j is permuted. The overall permuted variable importance for x_j is:

$$VI_{permute}(x_j) = \frac{1}{ntree} \sum_{t=1}^{ntree} VI^{(t)}_{permute}(x_j). \tag{3.15}$$

3.3.5 Conditional Inference Random Forests

Strobl et al. (2007) introduced CRF as an attempt to address some of the inherent limitations of RF. In particular, they claim that RF use a biased variable selection procedure. As previously discussed, this bias is a result of the optimal binary split search that favours certain variable types. CRF circumvent this bias by employing a method with a strong statistical foundation.

CRF use an alternative set of weak learners, namely conditional inference trees (CI trees) (Hothorn et al. 2006), which are based on the well-defined statistical principle of conditional inference. This method was proposed as a response to the demand for a method that accounts for the distributional properties of variables when selecting variables to perform splits, so that an unbiased model selection procedure can be performed (i.e., certain variable types are not favoured in the splitting process).

A defining characteristic of CI trees is that the procedures of selecting and splitting variables are performed separately. Over-fitting is addressed by model selection based on significance tests (chi-squared tests), which also provide the stopping criterion. The significance level α is a hyper-parameter that is used to control the complexity of the CI trees, where a larger value will result in a taller and more complex tree.

Hothorn et al. (2006) provide the generic algorithm given below for the conditional inference approach to recursive binary partitioning, which is implemented in the `ctree` package in R. The `party` package is used to build CRF, which is built on the `ctree` package. This approach can be applied when both the target and the features are measured at arbitrary scales. It is therefore not problematic to include variables that are continuous, categorical (with possibly many categories), multivariate or censored.

Let the non-negative integer observation weights $\tilde{\boldsymbol{w}} = \{\tilde{w}\}_{i=1}^{n}$. Each node contains a vector $\tilde{\boldsymbol{w}}$ which is non-zero if the observations are elements of the node and zero otherwise.

1. Test a global hypothesis of independence in the form of p partial hypothesis tests: $H_0^j : p(Y|X_j) = p(Y)$, where the global hypothesis to be tested has the form $H_0 = \cap_{j=1}^{p} H_0^j$. Terminate step if cannot reject H_0 at α, otherwise, choose X_j with the strongest association.

2. Choose the split point using optional split criterion, such as that used by CART or the permutation based approach described by Hothorn et al. (2006) to produce

$R \subset \mathcal{X}_j$. The observation weights for the daughter nodes are $\tilde{\boldsymbol{w}}_{left,i} = \tilde{\boldsymbol{w}}_i \cdot \mathbb{I}(X_j \in R)$ and $\tilde{\boldsymbol{w}}_{right,i} = \tilde{\boldsymbol{w}}_i \cdot \mathbb{I}(X_j \notin R)$ for $i = 1, \ldots, n$.

3. Repeat both steps until termination, with modified case weights $\tilde{\boldsymbol{w}}_{left}$ and $\tilde{\boldsymbol{w}}_{right}$.

Another important difference between the most common implementation of CRF (the `party` package) as opposed to the conventional RF, is that subsampling (sampling without replacement) is used instead of bootstrap sampling. In general, the subsampling fraction is 63.2% of the training set sample size, which corresponds to the percentage of non-duplicate observations when the bootstrap is used. This results in approximately the same out-of-bag sample size. Strobl et al. (2007) show empirically that bootstrap sampling biases the model selection procedure and variable importance measures that are based on selection frequencies, which led to a subsampling implementation.

3.4 A Simulation Study

We conducted a simulation study to assess the ability of the RF, CRF and XGBoost to identify important variables when applied to data with the inherent complications faced in the analysis section of this chapter. Note that AdaBoost is excluded here as at the time of this analysis, AdaBoost's implementation in R did not have functionality to return importance indices. This is done by taking bootstrap samples of the NHSBT dataset and simulating a binary response from a random intercept logistic regression (RILR). This enables us to closely mimic the NHSBT dataset and therefore to study properties of interest. In particular, it is of interest how applicable these methods are to multilevel data, which to this author's knowledge has not been studied empirically in terms of ensemble method variable importance measures.

As previously discussed, the NHSBT dataset exhibits various complications, including:

- A multilevel structure, arising from most donors (level 2) having donated both kidneys, for which we analyse the response at the recipient level (level 1).
- The presence of highly correlated variables, in particular the donor variables relating to time, which one would expect to be correlated by definition (such as `o2sat90time`[6], `o2sat80time`, `o2sat70time`, and possibly `deathtime`).
- A mixture of continuous and categorical variables, where some categorical variables have a large number of categories (approximately 20 categories), and continuous variables with skewed distributions often containing many outliers.
- Possible non-linear associations between the predictor variables and the response.

[6]It is convenient here to use the variable code names. A description relating to these names can be found in appendix Table A.1.

3.4.1 Simulation Design

As the intention of this section is not to interpret the data, the dataset is restricted to complete cases and feature engineering (described in Section 3.5.1) is not performed. A RILR model was fit to the dataset including four variables found to be predictive of DGF in the literature (`CIT_MINS`, `dage`, `DIAL_AT_TX`, `o2sat90time`). This provided a wide scope for research, by including both categorical and continuous variables and one that is highly correlated (`o2sat90time`) with (what we specify to be) non-significant variables (`o2sat80time` and `o2sat70time`). Non-linearity was introduced by fitting `CIT_MINS` with a natural spline with two interior knots. Furthermore, a log transformation was applied to `o2sat90time`. A single random intercept term was fitted relating to the donor ID. More specifically, we have:

$$y_{ij} \mid \pi_{ij} \sim \text{Bernoulli}(\pi_{ij}) \tag{3.16}$$

$$\pi_{ij} = \mathbb{P}(y_{ij} = 1 | \boldsymbol{x}_{ij}, \zeta_j), \tag{3.17}$$

where the random intercept term $\zeta_j \sim \mathcal{N}(0, \Psi)$. The simulation is based on the following model:

$$\begin{aligned} \text{logit}(\pi_{ij}) = &\beta_0 + \beta_1 \texttt{dage}_{1ij} + \beta_2 \mathsf{B}(\texttt{CIT\_MINS}_{ij}, \upsilon_1) \ + \beta_3 \mathsf{B}(\texttt{CIT\_MINS}_{ij}, \upsilon_2) + \\ &\beta_4 \mathsf{B}(\texttt{CIT\_MINS}_{ij}, \upsilon_3) + \beta_5 \texttt{DIAL\_AT\_TX}_{ij} + \beta_6 \log(\texttt{o2sat90time}_{ij} + 1) \\ &+ \zeta_j, \end{aligned} \tag{3.18}$$

where $\{\mathsf{B}(\texttt{CIT\_MINS}, \upsilon_k)\}_{k=1}^{3}$ represents the B-spline basis matrix for a natural spline with two internal knots placed at the 33.3% and 66.7% percentiles (in the range of the observed data for variable `CIT_MINS`). υ_k indicates which of the three segments separated by the two internal knots the piecewise polynomial is fit to.

The specified model is fit to the data and the coefficients are extracted, which become the set parameters of the simulation study. $\Psi = 1.8$ and the coefficients for the fixed effects $\boldsymbol{\beta} = (\beta_0 \dots \beta_6)^\top$ are:

$$\boldsymbol{\beta} = \begin{pmatrix} 3.58 \\ -0.24 \\ -1.05 \\ -4.12 \\ -2.14 \\ -1.86 \\ 0.22 \end{pmatrix}.$$

The RILR found each variable to be associated with DGF, where `DIAL_AT_TX`, `dage`, `CIT_MINS` and `o2sat90time` had the strongest to the weakest association respectively. Each bootstrap sample was drawn by sampling donor ID (which in most cases corresponded to two recipients). For every sample the donor was assigned a new ID and each recipient was assigned a simulated response. This way the hierarchical structure of the data at the donor level was preserved while recognising duplicate bootstrap samples as different donors.

1000 simulations were performed at each of the three investigated levels of sample size (100, 200, 1000 donors). Each bootstrap dataset was split into a 70% training and 30% testing set, where the training set was used to tune the hyper-parameters over a pre-specified grid of points. The algorithms were subsequently fit on the full dataset with the tuned hyper-parameters. The Gini and permuted importance measures were extracted for the RF and the permuted importance for the CRF. The gain importance was extracted for XGBoost.

3.4.2 Results

Figure 3.1 displays a matrix of box-plots corresponding to the simulation results for RF and CRF. Due to XGBoost requiring all features as dummy variables, the importance for each category is returned, which made a comparison in the form of this figure impossible. The left column corresponds to the Gini importance for RF, the middle column gives the permuted importance for RF, and the right column gives the permuted importance for CRF. The rows correspond to $n = 100, 200, 1000$ bootstrap samples of donors (top to bottom respectively). The x-axis gives the variable code names, where the important variables are colour coded by their level of importance (brown, red, beige and yellow; most important to least).

The left column is consistent with findings by Strobl et al. (2007) who found that the Gini index is biased towards variables with many categories (`REC_UNIT` and `recip_prd` have 24 and 21 levels respectively) and continuous variables. The Gini index had barely any ability at all to determine important variables. This is confirmed by Table 3.1 which displays how many times out of the 1000 simulations per sample size each important variable was included in the top four ranks and how many times they were correctly ranked. Notably, based on the Gini importance the most important variable was not once correctly identified, for any sample size. Only `CIT_MINS` was correctly identified a reasonable number of times, which is likely to be a result of the Gini importance bias towards continuous variables. Reporting this measure in our analysis would clearly lead to misleading conclusions.

In contrast to the Gini importance, the RF permutation importance shows a substantial improvement, not only in identifying the top four most important variables, but also

their correct rank. For low sample sizes, one of the categorical variables with many levels (`REC_UNIT`) was identified as the third most important variable (ranked by the median importance index across the simulations). Low cell counts for many of this variable's categories may explain this overoptimistic importance index for this variable, as perfect separation of classes is more likely to occur when only a couple of observations occurred.

At the 1000 donors sample size, the top two variables are correctly ranked by the median importance. Three of the four variables are correctly ranked to be in the top four most importance variables.

Strobl et al. (2008) found that the permutation importance measure is biased towards correlated variables. Our results are consistent with this finding, as `sbp60time` and `sbp50time` were found to be important (by median importance) for each sample size and are very highly correlated (correlation 0.99). Although it was not computationally feasible to implement in the context of a simulation, we investigated the performance of the most recently proposed importance measure (Strobl et al. 2008) (a conditional permutation measure that accounts for correlated variables). Having run a few iterations of the simulation with this measure, the results were comparable to the Gini index. This suggests that this measure is not appropriate for multilevel data.

Finally, for each sample size the CRF permutation importance correctly ranked the true important variables to be in the top four ranks (by median importance). More impressively, for the largest sample size they were ranked by median importance in the correct order. It can be seen from Figure 3.1 that as sample size increases there are much less outliers that lead to false conclusions, compared to the lower sample sizes. This suggests that asymptotically, the permutation importance for CRF is appropriate for the NHSBT dataset. However, even for the largest sample size considered in this study, outliers from non-important variables make it difficult to distinguish between important variables. For example, `o2sat90time` was correctly ranked in the top four 45% of the time). This is opposed to the most important variable `DIAL_AT_TX`, which was correctly ranked 100% of the time. XGBoost importance ranking was found to give competitive results to RF, but was outperformed by CRF.

It is important to note that this study was designed to be conservative, in the sense that tree models are known to give a poor approximation when the true association is linear or smooth. As we simulate this to be the case, the results from this simulation study reflect the worst case scenario of employing a non-parametric approach when the true relationship is linear and suggest that even in this case ensemble methods provide an effective means of identifying important variables. This result provides useful information, as the opposed case of incorrectly assuming linearity can be very costly.

These results may have been improved by subsampling, leaving scope for future work. However, in our case it was of interest to understand how these methods perform when homogeneity exists between observations.

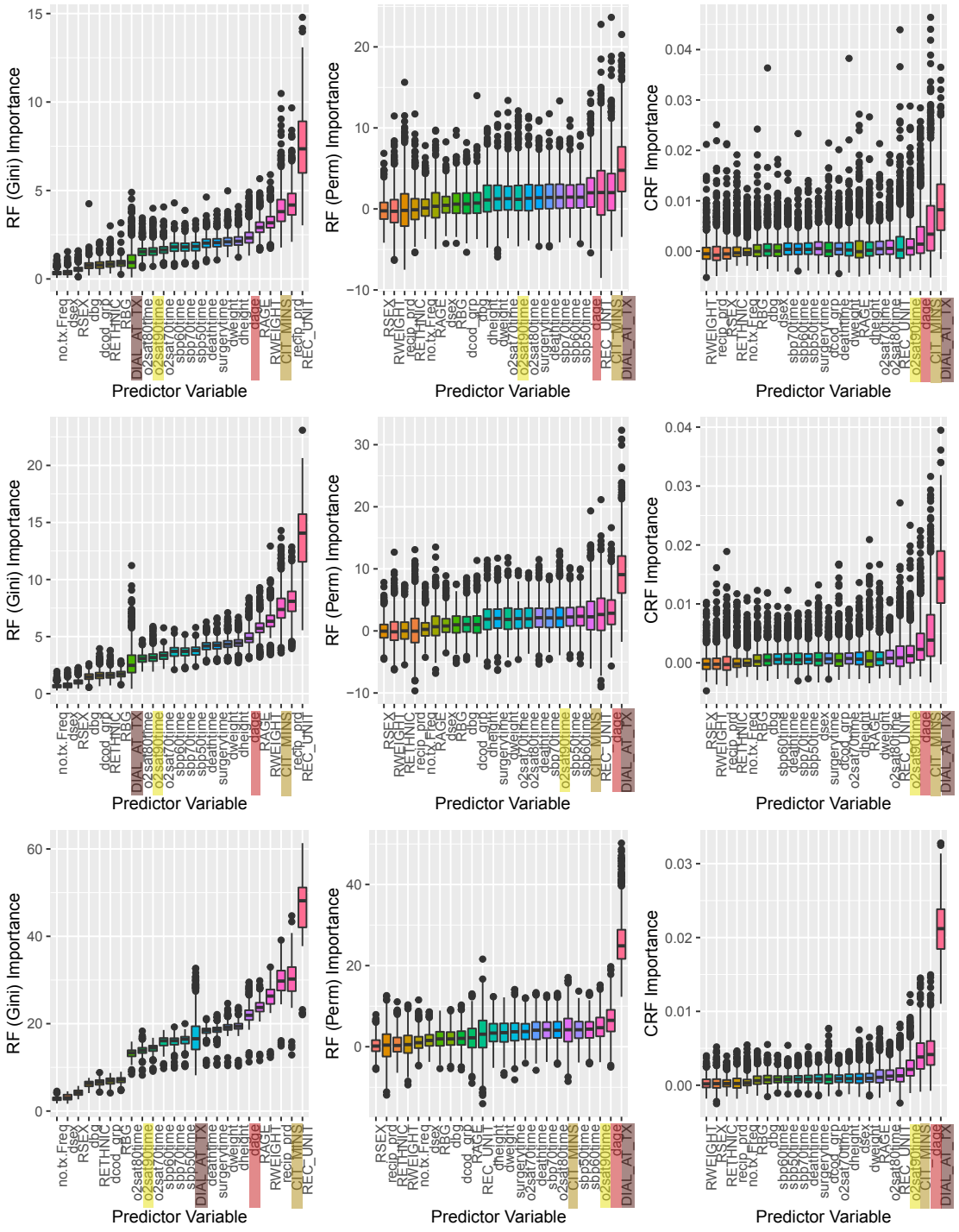


Figure 3.1: *An ordered box-plot displaying the model selection simulation results, presenting the importance indices of the Random Forest related methods for 1000 simulations. The rows top to bottom correspond to sample size* $n = 100, 200, 1000$*, and the names are coloured by true significant where the darkest colour is the most significant (i.e* `DIAL_AT_TX`*).*

Table 3.1: *Simulation results for ensemble methods as a means of model selection. Values represent the number of simulation out of 1000 that were either correctly ranked or included within the top four most importance variables. This is presented for four variables of varying importance for* $n = 100, 200, 1000$ *bootstrap sampled individuals.*

Ensemble Method	RF (Gini)			RF (Permute)			CRF			XGBoost		
n	100	200	1000	100	200	1000	100	200	1000	100	200	1000
`DIAL_AT_TX`												
Top four ranks	10	18	25	648	905	1000	803	979	1000	350	680	1000
Correct rank	0	0	0	399	711	1000	422	772	1000	70	247	974
`dage`												
Top four ranks	83	45	9	282	330	648	325	418	836	362	398	654
Correct rank	9	3	0	82	111	368	108	142	444	95	112	253
`CIT_MINS`												
Top four ranks	924	974	999	352	323	298	519	559	728	671	682	846
Correct rank	422	511	445	70	63	85	109	121	235	98	121	207
`o2sat90time`												
Top four ranks	2	0	0	187	186	323	187	221	450	208	197	297
Correct rank	1	0	0	57	56	85	45	79	193	48	62	125

3.5 Application to the NHSBT Dataset

Recall the NHSBT dataset introduced in Section 1.1. This dataset contains all of the proceeding DCD donors in the UK between 1 April 2010 and 31 March 2015 (extracted from the UK transplant registry on 3 August 2015). This is made up of 1906 kidney recipients, corresponding to 1120 DCD donors. 825 of these donors donated both kidneys and 256 donated one kidney each.

There are 14 baseline characteristic variables in total, seven relating to donors and seven relating to recipients (age, gender, blood group, cause of death, ethnicity, height and weight). Eight variables relate to the surgery process at the donor level (time from treatment withdrawal to death[7], surgery time, time until blood pressure drops to 70, 60 and 50mmHg; time until oxygen saturation drops to 90, 80, 70%). Other variables at the recipient level include primary renal disease, whether the recipient was on dialysis at the time of transplant, CIT (in minutes) and which of the 24 transplant centres they attended. Three response variables relating to short term transplant outcomes were present (a binary indicator representing DGF, survival time in days and a censoring indicator).

The aim of this application is to benchmark various machine learning algorithms described in Section 3.3 (RF, CRF, AdaBoost, XGBoost) and RILR in terms of discriminatory ability, in order to determine the applicability of these methods to the NHSBT dataset and to determine how well we are able to predict DGF based on only DCD donor data. To this author's knowledge this has not been done before. We subsequently use ensemble importance measures to rank the importance of the variables in the NHSBT dataset for predicting outcome DGF.

3.5.1 Preprocessing and Feature Engineering

Variables thought to carry no information and those with a high proportion of missingness (that were not known in the literature to be predictive of DGF) were dropped. This led to the removal of donor and recipient ID (except when donor ID was used to merge datasets and to specify the random intercept term for RILR). Recipient height was also dropped due to a large number of missing values.

The censoring indicator was found to give a very high proportion of censoring (95%). This is due to a relatively short follow up time, where the maximum event time is just over two years. Analysis using a response with such a large imbalance in the proportion of classes would suffer from low statistical power. For this reason the survival response was dropped and the remainder of this work focuses on the binary DGF response.

[7]Note that this corresponds to `deathtime`, the donor survival time in minutes *once withdrawn from life-supporting treatment*. The importance of this variable is compared against the physiological variables, to determine whether there is a 'duration of withdrawal' or a'physiological variable variation' effect.

Categorical variables that were found to have a low cell count (of less than five) once stratified by DGF were combined to avoid complications in the analysis and as an attempt to improve predictive power. For this reason, donor and recipient ethnicity variables were converted to binary indicators, which represent whether the patient was white or not white (a large majority of the donors were white). Many levels of `recip_prd` had a count below five, all of which were combined to give a single category that represents uncommon types of primary renal disease.

For XGBoost, categorical variables were transformed with *one-hot encoding*[8], which means that each category is represented by dummy variables. For example, transplant centre was transformed from a single variable with 24 categories to 24 variables with two categories. As tree based methods are invariant to changes of scale, normalisation was not necessary.

In order to improve predictive ability, feature engineering (creating new predictor variables from the data) was carried out. A binary variable was created using donor ID to represent whether a single kidney or both kidneys were donated. Furthermore, a binary indicator was created to represent blood group compatibility (compatible or identical) between the donor and the recipient. Variables were created to represent a mismatch between gender (i.e., a male organ was donated to a female or vice versa) and also for ethnicity mismatch, where the same principle applies. Thus, some of the variables in Table A.1 were created from the dataset.

3.5.2 Missing Data

The appropriate treatment of missing data between predictive and explanatory modelling approaches differ. The missing data literature is monopolised by explanatory modelling and it is a regular occurrence in predictive modelling studies that missing data is handled inappropriately. A common mistake is often made by imputing values based on statistics (such as the mean or values randomly generated from the empirical distribution) obtained from the full dataset, rather than just the training set, resulting in *leakage*[9].

Appropriate treatment begins by considering both the proportion of missing data in the full dataset and the types of variables that have data missing. Table 3.2 displays the percentage and total number of missing values in the NHSBT dataset. The table includes only the variables that have missing values and is sorted in ascending order of missingness. It appears that most of the variables have a relatively low number of missing values, with the large exception of `recip_prd`.

[8] A common terminology in the machine learning community.

[9] Leakage refers to an indirect access to information in the test set that may lead to an optimistic generalisation error.

Table 3.2: *The total number and percentage of missing values in the NHSBT dataset, including only variables that have missing values. Sorted in ascending order of proportion missing.*

Variable	Type	Total Missing	% Missing
RSEX	factor	1	0.05
GMM	factor	1	0.05
dheight	integer	4	0.21
RETHNIC	factor	8	0.42
CIT_MINS	integer	25	1.31
sbp70time	integer	27	1.42
sbp50time	integer	33	1.73
o2sat70time	integer	52	2.73
o2sat90time	integer	57	2.99
dethnic	factor	58	3.04
o2sat80time	integer	58	3.04
EMM	factor	66	3.46
sbp60time	integer	72	3.78
surgerytime	integer	83	4.35
deathtime	integer	85	4.46
RWEIGHT	numeric	141	7.40
recip_prd	factor	732	38.40

Six of the seventeen variables with missing values are factors, two of which have only a single missing value. Three of the other factor variables relate to ethnicity. The NHSBT have implemented schemes aimed at urging the non-white population to become organ donors due to this being a highly under represented group (4% non-white patients in this dataset). Acknowledging this large imbalance, it may be reasonable to replace these missing values with the mode. In predictive modelling missing values can sometimes be a blessing in disguise, as the pattern of missingness may reveal features that cannot be drawn from the observed data alone. For this reason, a reasonable treatment for the large number of missing values in `recip_prd`, is to create an extra category of "missing". Case-wise deletion for this variable would have resulted in a loss of 522 observations, which is which is approximately 27% of the full dataset.

Single imputation can be a highly effective method for improving predictive accuracy. However, when the aim of a study is explanation, multiple imputation (MI) is preferred due to its ability to appropriately treat uncertainty arising as a result of drawing random numbers from posterior distributions for the purpose of inference. Although MI is applicable in the predictive modelling context it is far more computationally demanding and in many cases simple methods provide comparable predictive accuracy. In this work, we use a single imputation for the dataset used to tune the hyper-parameters, and fit the tuned model to multiple datasets.

Our employed approach for imputing missing data is as follows:

Table 3.3: *Learning algorithm hyper-parameter names, types and bounded range.*

Algorithm	Hyper Parameter	Type	Bounded Space
Random Forest	ntree	Integer	Between 50 and 2000
	mtry	Integer	Between 1 and 30
	Node size	Integer	Between 1 and 100
CRF	ntree	Integer	Between 50 and 2000
	mtry	Integer	Between 1 and 30
AdaBoost	Number of boosting iterations	Integer	Between 50 and 1000
	Shrinkage	Numerical	$10^x, -1 \leq x \leq 0$
XGBoost	Booster	Discrete	Linear or tree based
	Max depth of tree	Integer	Between 1 and 10
	Minimum child weight	Integer	Between 1 and 10
	Subsample	Numerical	Between 0.5 and 1
	Eta (shrinkage)	Numerical	Between 0.1 and 0.5
	Lambda (L2 Regularisation term)	Numerical	$10^x, -1 \leq x \leq 0$
	Col sample by tree	Numerical	Between 0.5 and 1

- For `recip_prd` create an extra category for missing values. Missing values corresponding to factor variables are imputed using the mode. Numerical and integer values are drawn at random from their empirical distribution. Dummy variables are created for each variable that was imputed, representing which values were imputed, so that patterns of missingness can be modelled explicitly.

3.5.3 Model Selection and Hyper Parameter Tuning

In Section 3.3 the methods were described such that the algorithm hyper-parameters could be understood and therefore dealt with appropriately. By performing a nested CV procedure (see Section 2.2.2.4 for an explanation of nested CV) with 5 outer loops and 3 inner loops, we have a means of simultaneously selecting hyper-parameters and obtaining an unbiased estimate of the generalisation error in the benchmarking experiment.

Table 3.3 shows the relevant hyper-parameters that are optimised for each of the machine learning algorithms. Those that are not displayed were deemed to be unnecessary for tuning and set to their default values in the corresponding software packages. The specified bounds for the ranges of values for those that are continuous, and the possible values for those that are discrete are given in Table 3.3. Selection is performed by randomly drawing values within the ranges of the continuous hyper-parameters and possible categories for those that are categorical. 50 random combinations of hyper-parameters are generated and their performance evaluated (using the evaluation criteria explained in Section 3.5.4) on the inner test sets for each of the outer loops of the nested CV procedure.

We aim to benchmark the machine learning methods as well as the standard statistical method RILR. However, it is not practically feasible to perform a manual model selection

procedure for the standard statistical methods for each fold of the nested cross-validation scheme. For this reason, these models are pre-specified based on variables found to be predictive of DGF in the literature and also based on the expert knowledge of a kidney transplant surgeon. Moreover, based on Harrell (2015)'s recommendation for gauging predictive potential, we *prespecify complexity without later simplification*. This refers to fitting a model with as much complexity as the effective sample size will allow and by allowing greater complexity for variables thought to be the most important (for example by including more knots for the spline terms). We adopt a slightly more conservative rule of the 15:1 rule of thumb[10] (being 20:1) to account for having to estimate random effects. Therefore, we conjecture that estimating more than 40 parameters would make the model's validity questionable.

3.5.4 Evaluation Criteria

In this work, various measures of predictive performance are considered. In particular, the hyper-parameters are optimised according to the mean misclassification error (MMCE). Equation 3.19 is the misclassification error (MCE), which is averaged across the inner folds to give the MMCE. The optimal set of hyper-parameters is chosen on this basis.

$$\text{MCE} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i \neq \hat{y}_i) \tag{3.19}$$

Recall the confusion matrix introduced in Chapter 2 (Table 2.1). Here, positive refers to the default event (positive outcome), which in this case is an immediately functioning graft. Negative therefore represents a delayed graft function.

A useful measure in our case is the *accuracy* which is the proportion those correctly classified as given in Equation 2.31. This is a useful metric when the outcomes are reasonably balanced. In our case there are 798 delayed graft functions (42%) and 1108 immediately functioning grafts (58%), which is considered to be well balanced data.

We consider the metrics defined in Section 2.2.3, in particular the BER and AUC ROC. In our case, as the data outcomes are reasonably balanced and the false positive and false negative outcomes could be considered equally costly, the accuracy is a reasonable measure. However, BER is a better measure because we are interested in the ability of the model to predict both an immediate and a delayed graft function.

[10]This 15:1 rule means that we can afford to estimate $p < \frac{m}{15}$ parameters, where m is the limiting sample size. The limiting sample size is the minimum number of cases in the binary response (which is 798 in the NHSBT dataset).

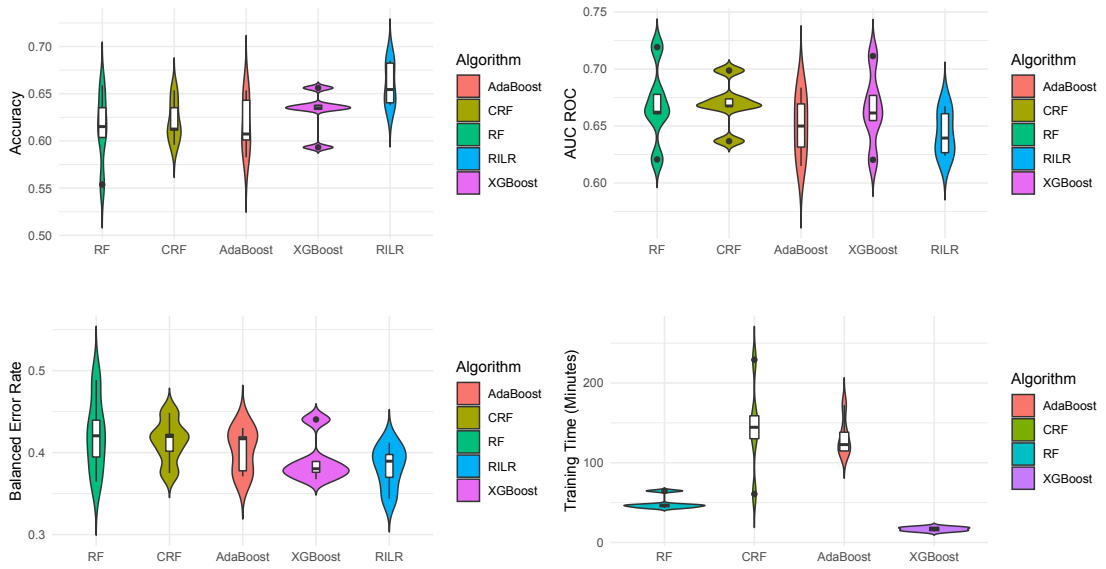


Figure 3.2: *Violin plots displaying the density and summary statistics of various performance measures across the outer folds of the nested cross validation, performed in order to tune and evaluate the performance of learning algorithms for predicting DGF.*

3.5.5 Results

This analysis was performed in R version 3.5.1 using the `mlr` package (Bischl et al. 2016) for building machine learning pipelines. Tuning and the nested cross-validation was run on the University of Southampton high performance computing facility Iridis 4, enabling a much more rigorous search for the optimal hyper-parameters in the specified hyper-parameter space, by using parallel computing.

Note that the RILR model that represented the benchmark, was pre-specified in order to avoid complications arising from multicollinearity by including a single one of the set of highly correlated variables (`o2sat70time`). This particular variable chosen as it was found to be the strongest predictor in this set based on the variable importance metrics of the machine learning methods. The pre-specified model included: `dage`, `CIT_Mins`, `o2sat70time`, `DIAL_AT_TX` (continuous variables were represented with B-spline terms with four knots). Recipient centre and donor ID were both included as random intercept terms.

The results for the benchmarking experiment are presented in Figure 3.2. Each violin plot presents the distribution of a corresponding performance measure obtained across the five outer loops of the nested cross-validation. Box-plots can be seen inside each of the violin plots displaying summary statistics. These are presented for each learning algorithm fitted with the tuned hyper-parameters and for the pre-specified RILR model.

It is immediately obvious that for each performance measure, RF and AdaBoost had a wide range in performance across the five folds (except RF training time). These were

arguably the worst performers overall, having both large variances on each plot and comparatively low median values for measures that we aim to maximise (accuracy and AUC ROC). They also appeared to have comparatively large values for those that the aim is to minimise (BER and training time, except training time for RF).

Despite the reduced variance in general compared to RF and AdaBoost, CRF performed similarly. This is with the exception of being the best performer in terms of AUC ROC. For this measure, it does not only have the highest median value but also the least variance. CRF was however the slowest algorithm to train, taking approximately eight times longer than XGBoost.

The best performers were arguably XGBoost and RILR, particularly in terms of accuracy. However, counter-intuitively RILR had the lowest AUC ROC, suggesting it performs poorly for certain decision rules. XGBoost remained competitive for this measure, however suffered from a large variance. In this work we consider BER to be the most important measure, as we are interested in the prediction error for each class. In terms of this measure XGBoost was the winner despite being multi-modal, due to having the lowest median and variance. This can be seen by the wide section of the violin, representing a high density in that corresponding area. This violins multi-modal behaviour is likely to be due to the number of outer loops of the nested cross-validation being low (5 loops only).

As previously mentioned, XGBoost is highly scalable algorithm, being by far the fastest to train and tune. This means that in practice a much more thorough search in the hyper-parameter space can be performed compared to the 50 iterations performed here. Note that RILR is not included in the training time graph as it did not require tuning.

We decided to see how these performance measures vary for algorithms with tuned hyper-parameters across multiple datasets, where continuous variables were imputed by randomly drawing from the empirical distribution.

Figure 3.3 displays the variation of the mean performance measures (where the mean is taken across five CV folds) between ten imputed datasets. Algorithms are run with tuned hyper-parameters using the same indices used in the nested CV outer loop to estimate performance.

After taking into account the uncertainty induced from imputing missing data, RILR and XGBoost become even more prominent as the best predictive models. However, RILR has one of the lowest AUC ROC. XGBoost, RF and CRF have similar performances for this metric (RF being marginally better than the others).

Despite being the most sensitive to missing data (having the highest variance in Figure 3.3), XGBoost could be argued to be the best performing machine learning algorithm overall in terms of predictive ability. However, the results from the simulation study in Section 3.4 suggest that CRF is the most appropriate algorithm for determining variable

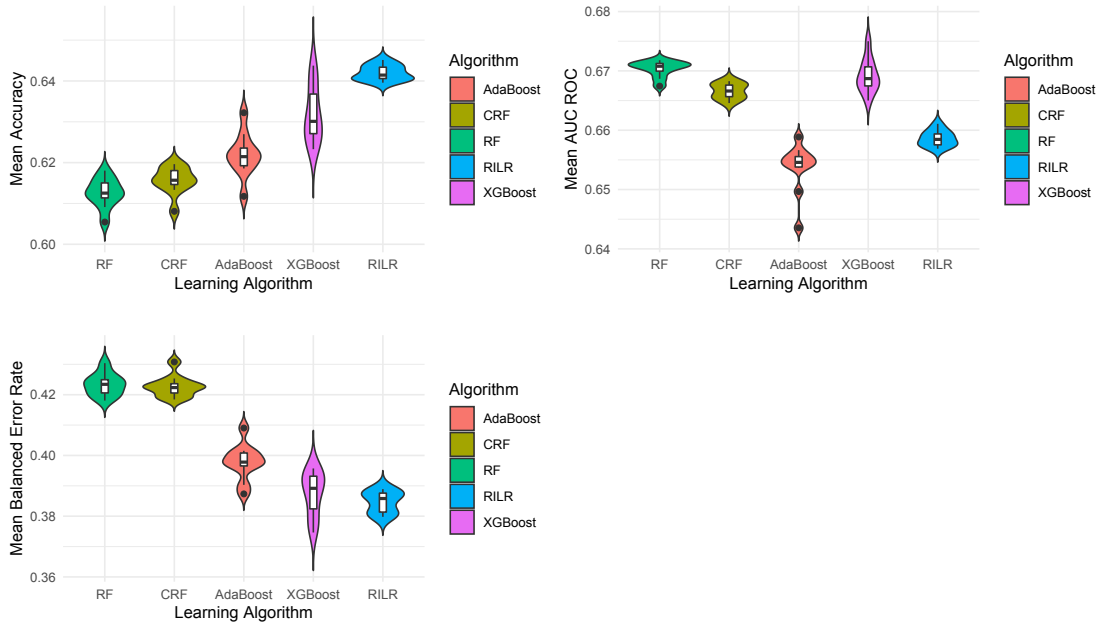


Figure 3.3: *Violin plots displaying the variance of various performance measures over ten generated dataset, where continuous variables were drawn at random from their empirical distribution.*

importance in the NHSBT dataset. As CRF had a high mean AUC ROC, we proceed to rank importance using this algorithm.

At this stage of the analysis the CRF is fitted to the full dataset. We propose a novel method for analysing importance when missing data is present that is similar in spirit to how random forests tactically induce randomness to understand distributional properties. Our method involves imputing multiple (in our case ten) datasets by randomly drawing from the empirical distributions (now from the full dataset as testing is not required). The imputation can be treated as a sampling process, exposing which variables have large importance indices by random chance.

Ranking the variables by their median importance index enables us to incorporate uncertainty from missing data at the same time as studying the distribution of importance indices. This can be seen in Figure 3.4, which is a box-plot of the CRF importance measure (mean decrease in accuracy) for each variable across the ten imputed datasets.

Clearly, `DIAL_AT_TX` and `REC_UNIT` are important variables, being far separated from the rest. It can be seen that the variation in the importance indices across the imputed datasets is lowest (in general) for what are deemed to be unimportant variables. However, it appears that the importance ranking is little impacted by the imputation due to a small variance for all of the variables. This is confirmed by Figure 3.5 which displays a heat-map of the ranking totals across the ten datasets. It can be seen that `DIAL_AT_TX`,

`REC_UNIT`, `CIT_MINS`, `dage` and `RETHNIC`, were ranked in the same place for each imputation (ranked 1, 2, 3, 4 and 5 respectively). However, the less important variables tend to have more mixed ranks. This is with the exception of the least important variable `no.tx.Freq` which was ranked the least important almost every time.

An interesting clinical result has emerged from these findings. Bradley et al. (2013) conjecture that although transplants do not proceed when death time is prolonged, death time is not actually as related to transplant outcome as originally thought and that the physiological variables throughout the treatment withdrawal to death phase play a more important role. They express the need for further research on the matter and that there is a need to find a better surrogate relating to transplant outcome. Figure 3.4 suggests that many of the physiological variables are predictive of DGF, but death time is very close to zero. It can be seen that `o2sat70time` and `sbp60time` are better surrogates for predicting transplant outcome.

These results are all consistent with the literature. It would be a cause for concern in terms of the reliability of these results if `DIAL_AT_TX`, `CIT_MINS` or `dage` were found to be not important variables.

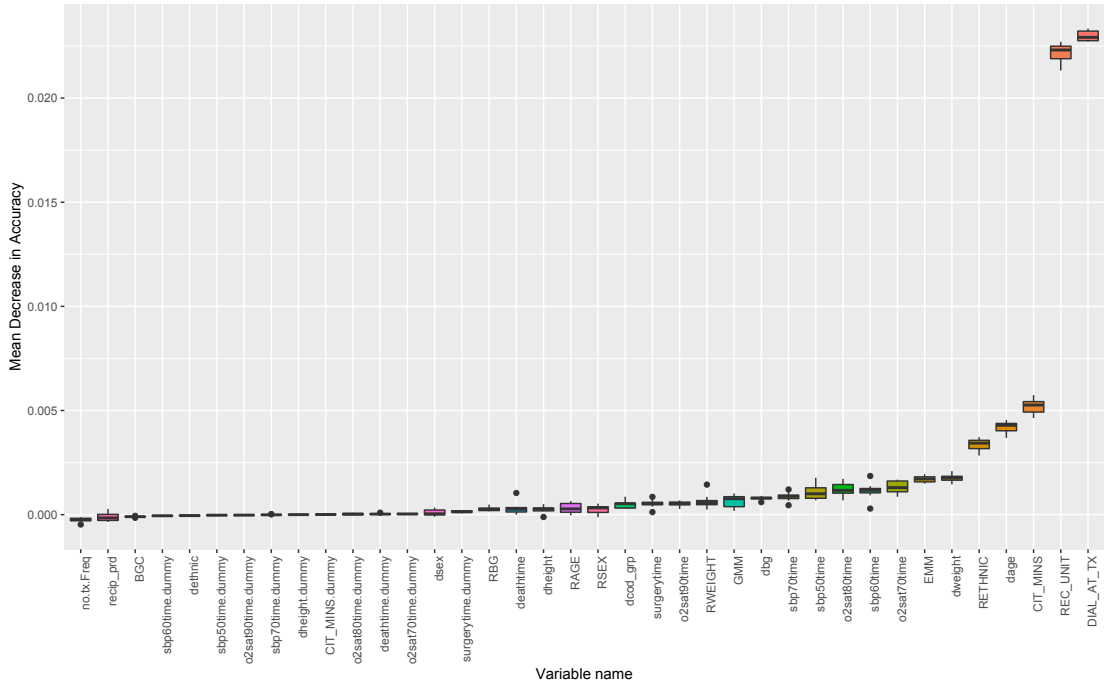


Figure 3.4: *Box-plots showing the variation in the CRF permuted importance measure across the ten datasets. Variable names ending .dummy correspond to binary indicator variables that represent imputed values.*

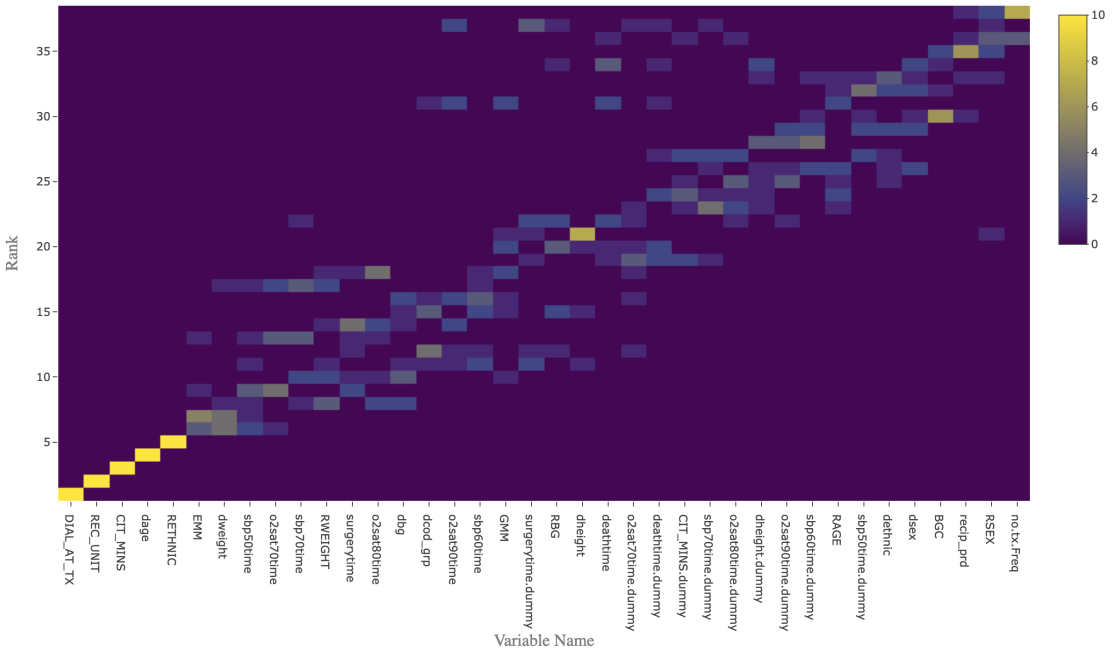


Figure 3.5: *Heat-map displaying the number of times each variable obtained each rank across the ten imputed datasets.*

3.6 Discussion

This chapter has covered a wide range of topics. We began by introducing algorithmic models as a possible alternative means of predicting an outcome and identifying important variables, as opposed to standard statistical methods. The theoretical background for these methods was covered such that the hyper-parameters could be better understood and tuned effectively in our application.

A simulation study was run to determine the ability of various ensemble methods to identify and rank important variables. This was assessed for two importance measures for the RF (Gini and permuted) and the permuted importance for CRF. The importance measure "gain" was also used for XGBoost. CRF was by far the best algorithm for determining important variables. This was surprisingly effective given that the data was generated from a RILR. The rectangular regions fitted by the tree based methods would in general be a poor approximation to these data, which suggests even in the worst case scenario in terms of the data generating mechanism, machine learning methods can provide benefit by identifying important variables.

We conducted a benchmarking experiment to determine the applicability of machine learning methods for predicting DGF when faced with various complications that the NHSBT dataset presents (multilevel structure, categorical variables with many categories, correlated variables). We found that these methods performed similarly on the dataset used to tune the hyper-parameters (where RILR performed the best followed by XGBoost), but a larger disparity emerged when assessing the performance measures over multiple datasets. RILR had the highest accuracy but the lowest AUC ROC. XGBoost was arguably the best performer overall when considering each measure.

We proposed a new (to this author's knowledge) approach of ranking variable importance when missing data is present. This involves performing multiple imputation, drawing each dataset from the empirical distribution of the variables with missing data. Importance measures are extracted from models fitted to each dataset and presented as box-plots (Figure 3.3). A heat-map displaying the number of times each rank was chosen can also provide insight into the distribution of the importance measures (Figure 3.5). Inducing randomness through imputation provides insight into the sampling distribution of the importance measures and accounts for the uncertainty resulting from missing data.

We were able to empirically investigate a conjecture form the NHSBT (Bradley et al. 2013) as to whether death time or physiological variables are related to transplant outcomes. Our findings suggest that death time is not an important variable for predicting DGF, but the time until oxygen saturation and SBP drop to certain levels in the treatment withdrawal phase are.

Part III

Joint Modelling Applications

Chapter 4

Dynamically Predicting Donor Death Time from Treatment Withdrawal

Summary

Although there exists a large disparity in the number of kidneys available for transplantation and the number of patients on the transplant waiting list, many organs from DCD donors are discarded every year. Although this is mostly due to suboptimal graft quality, it also occurs for logistical reasons, such as a prolonged withdrawal resulting in the removal team having to be elsewhere. Being able to accurately predict when donors are going to experience an event (such as death once withdrawn from life support) would be invaluable for clinicians, enabling a more efficient allocation of what are currently highly stretched resources.

In the last few years substantial progress has been made in the field of joint modelling. Important extensions to the standard joint model include the incorporation of multiple longitudinal covariates, various functional forms of the longitudinal process and flexible association structures. These extensions provide potential improvements in predictive accuracy. Moreover, these extensions allow for real time dynamic prediction of survival probabilities. This chapter aims to exploit these extensions to dynamically predict the time a donor will experience a cardiac arrest once withdrawn from life support.

4.1 Introduction

The duration of the withdrawal from life-sustaining treatment to asystole is a major determinant of whether organ procurement for DCD donors proceeds. The main reason for this is that warm ischaemic injury is thought to damage the graft during this period (D'Alessandro et al. 2004). In the UK the maximum time before the transplant ceases to proceed is 3 hours. In Australia and the USA the limit is 1.5 and 2 hours respectively (Manara et al. 2012, Bellingham et al. 2011). Although it is unclear whether the duration of the agonal phase is directly associated with the quality of the graft (Reid et al. 2011, Bradley et al. 2013), demand remains for a predictive model that can determine whether the donor will die within a given period of time (Pugin et al. 2017). Such a model would enable more efficient and effective use of the organ procurement team. At the very least hospital staff would be better prepared to deal with family expectations (Souter & Van Norman 2010).

In this chapter we analyse observational data relating to 231 DCD donors from various centres across the UK. Our aim is to provide a new prediction tool that can dynamically predict the time of an event throughout the withdrawal period, conditioning on the fact that the patient is still alive at the current time that the prediction is being made. This prediction tool is based on the joint model (JM) and its extensions, which is an approach that takes advantage of both characteristic variables and the rich nature of longitudinal data. In our case the longitudinal data refers to the physiological variables repeatedly measured over time (SBP[1], DBP, MAP, HR and O2). Note that recipient information is not relevant in this chapter.

By exploiting various extensions to the JM, in particular by including multiple longitudinal covariates (Rizopoulos & Ghosh 2011), alternative functional forms of the physiological variables (Rizopoulos 2012*b*) and positing a flexible association structure (Andrinopoulou et al. 2018), we aim to determine the discriminatory ability of these physiological variables while adjusting for both censoring and what is found to be predictive out of the baseline covariates (age, cause of death, BMI, gender and ethnicity).

This chapter is structured as follows. In Section 4.2 we give a review of literature relating to both the application and the methods that are employed. This is followed by a more formal description of the methods in Section 4.3 (the JM for longitudinal and time-to-event data and relevant extensions). We then analyse the *novel dataset* in Section 4.4 with the chosen methods to dynamically predict survival probabilities of donors throughout the treatment withdrawal period. In Section 4.5 we conclude with a discussion.

[1] Abbreviations: SBP, systolic blood pressure, DBP, diastolic blood pressure, MAP, mean arterial pressure, HR, heart rate, O2, oxygen saturation.

4.2 Review of Literature

Various studies have been conducted with the aim of predicting whether an organ donor will die within a given time-frame from treatment withdrawal (a consensus of the length of time being 60 minutes). de Groot et al. (2012) used logistic regression with bootstrapping to adjust for overoptimism and achieved reasonable discrimination of 0.74 (CI 0.69-0.90) AUC ROC. This study included as covariates: corneal reflex, absent cough reflex, oxygenation index and extensor or absent motor response.

Wind et al. (2012) conducted a prospective multi-centre study of observational data that consisted of 211 patients. Controlled mechanical ventilation, norepinephrine administration and absence of cardiovascular co-morbidity were found to be risk factors for death time and achieved 0.73 AUC ROC.

In a retrospective study of 621 DCD liver donors, Davila et al. (2012) found various features predictive of a transplant proceeding (age, BMI, duration of time in intensive care and warm ischaemic time). They also found age, use of inotropes and absence of gag reflexes predictive of time to asystole. The age and BMI variables in this study were dichotomised, which is likely to have resulted in a loss of predictive ability. Nevertheless, this study achieved a discrimination for predicting the time to asystole, with ROC AUC of 0.84 and for proceeding to transplant 0.75. However, having access to the warm ischaemic time variable would require the full trajectory to have been obtained.

A prospective study of 191 patients from nine different transplant centres across the UK was conducted by Suntharalingam et al. (2009) using Cox proportional hazards regression. Univariate analysis showed categorised age, SBP, O2, cause of death, ventilation mode, inotrope use and arterial pH were associated with death time. The multivariate analysis retained the variables age, O2 and ventilation mode as significant.

Kotsopoulos et al. (2018) provided an external validation with a single-centre retrospective study of 92 patients, on the following studies that performed the logistic regression finding: de Groot et al. (2012) AUC 0.86, CI 0.77-0.95; Wind et al. (2012) 0.62, CI 0.49-0.76; Davila et al. (2012) 0.8, CI 0.71-0.90; and for the Cox regression study Suntharalingam et al. (2009) Harrell's C-index 0.63.

The JM consists of a longitudinal sub-model (usually a linear mixed effects model) and a survival sub-model which is often a Cox model. By considering the joint likelihood of the longitudinal and outcome processes, simultaneous estimation is conducted conditional on a shared latent random effects term. Various studies (Ibrahim et al. 2010, Ratcliffe et al. 2004, Sweeting & Thompson 2011) have shown that when the two processes are associated, the JM provides more efficient and reliable estimates than achieved by a two-stage approach.

The two-stage approach, first proposed by Tsiatis et al. (1995), involves fitting a linear mixed effects model (LMEM) to the longitudinal data and substituting the trajectory into the Cox model as a time-dependent covariate. Although this approach is much less computationally demanding than the JM, Wang et al. (2000) showed it results in significant bias. They proposed a regression calibration approach which uses the best linear unbiased predictors (BLUP) as covariates in the Cox model. This approach postulates that the hazard of event is associated with the current biomarker value. Wang et al. (2000) found that although this reduced the bias, it did not eliminate it. Various other two-stage approaches have been proposed (Rice & Wu 2001, Ye et al. 2008, Albert & Shih 2010, Murawska et al. 2012) to attempt to correct for bias occurring as a result of measurement error and informative drop-out. None of these approaches were able to eliminate the bias.

The most well-known application of the JM is modelling the trajectories of biomarker CD4 count in the AIDS dataset (Wulfsohn & Tsiatis 1997, Lavalley & DeGruttola 1996, Henderson et al. 2000, Hogan & Laird 1997). This became the standard dataset that is used to demonstrate developments in methodology.

By assuming conditional independence, the JM handles endogenous (internal) time-dependent covariate (as recordings are taken on the same subject whom the event outcome relates to) (Kalbfleisch & Prentice 2002). In this case the time-dependent Cox model does not suffice as it assumes the time-dependent covariates are exogenous (external). An example in which they are exogenous is the case where they represent environmental factors (Rizopoulos 2012*b*).

Many novel extensions have been applied to the JM over the last few decades. In particular, extending the JM to the case of multiple longitudinal covariates has been a popular topic in the joint modelling literature (Rizopoulos & Ghosh 2011, Brown et al. 2005, Chi & Ibrahim 2006, Hwang et al. 2015). The framework behind this extension is well covered in both the frequentist and Bayesian context. However, the multivariate case involves an increased computational burden as a result of integrating with respect to high dimensional random effects. A large part of current research in this area aims to make the JM scalable (Rizopoulos 2012*a*, Soleimani et al. 2018, Mauff et al. 2018). This is a challenging problem for which developments are still in their infancy and availability is limited in standard statistical software.

4.3 Methods

4.3.1 The Multivariate Bayesian Joint Model (MBJM)

The MBJM is employed to predict the occurrence of an event given multiple longitudinal covariates that may be both unbalanced and unequally spaced. Moreover, samples are subject to measurement error as a result of biological variation.

Denote a sample from a target population as $\mathcal{D} = \{T_i, \delta_i, \boldsymbol{y}_{1i}, \ldots, \boldsymbol{y}_{Ki}\}_{i=1}^{n}$, where the index i represents the organ donor. T_i corresponds to the observed event time, which is the minimum of the true event time T_i^* and the censoring time C_i. δ is a censoring indicator, that is 0 when an observation is censored[2] and 1 otherwise. $\boldsymbol{y}_{ki}$ corresponds to the vector of repeated measures on the kth longitudinal covariate, where $\{y_{kij}\}_{j=1}^{n_{ki}}$. $y_{ki}(t)$ is often used to represent the longitudinal outcome at time t rather than using the index j, to emphasise the biomarker is a function of time, i.e., $t \geq 0$.

4.3.1.1 The Longitudinal Sub-Model

We begin by specifying a model that describes the subject-specific trajectories for each of the physiological variables. As all of the biomarkers in this application are continuous, we restrict our attention to this case and therefore specify a multivariate LMEM (Equation 4.1)[3]. Here $m_{ki}(t)$ is the true unobserved biomarker value and $y_{ki}(t)$ is the biomarker contaminated with measurement error $\epsilon_{ki}(t) \sim \mathcal{N}(0, \sigma_k^2)$.

$$y_{ki}(t) = m_{ki}(t) + \epsilon_{ki}(t) \tag{4.1}$$

$$y_{ki}(t) = \boldsymbol{x}_{ki}^{\top}(t)\boldsymbol{\beta}_k + \boldsymbol{z}_{ki}^{\top}(t)\boldsymbol{b}_{ki} + \epsilon_{ki}(t) \tag{4.2}$$

Equation 4.2 shows (highlighted in blue for clarity) how the unobserved biomarker value is composed of a (possibly time-dependent) design vector $\boldsymbol{x}_{ki}^{\top}(t)$ for the fixed effects $\boldsymbol{\beta}_k$ and a time-dependent design vector $\boldsymbol{z}_{ki}^{\top}(t)$ for the random effects $\boldsymbol{b}_{ki}$. The full vector of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \ldots, \boldsymbol{b}_{Ki})^{\top} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$, where the $q \times q$ variance-covariance matrix of the random effects is represented by $\boldsymbol{D}$. Denoting the dimension of random effects corresponding to the kth longitudinal covariate as q_k, then $q = \sum_{k=1}^{K} q_k$. The fixed effects component in Equation 4.2 represents the mean longitudinal profile and the latent component represents the deviation of a given individual from the mean profile.

[2] In this work we restrict our attention to right censoring, i.e., $T_i^* > T_i$.

[3] Note that any member of the exponential family of distributions can be used in practice.

4.3.1.2 The Time-to-Event Sub-Model

The time-to-event component of the JM expresses the hazard of an event conditional on the true unobserved biomarker values (or some function of them) at a specified time. More formally, the instantaneous risk at time t is expressed as:

$$h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) = \lim_{dt \to 0} \frac{1}{dt} P(t \leq T_i^* < t + dt \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i)$$

$$= h_0(t) \exp(\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \sum_{k=1}^{K} \sum_{l=1}^{L_k} f_{kl}\{\lambda_{kl}(t), \mathcal{M}_{ki}(t)\}), \tag{4.3}$$

where the baseline hazard function is denoted by $h_0(t)$. The complete history of the true unobserved longitudinal process up to (but not including) time t is represented by $\mathcal{M}_i(t) = \{m_i(s),\ 0 \leq s < t\}$. The vector of parameters $\boldsymbol{\gamma}$ correspond to the survival component's confounding effects design matrix $\boldsymbol{w}_i$. Note that $\boldsymbol{w}_i$ may contain exogenous time-dependent covariates, but in our application contains only baseline covariates.

The functions $f_{kl}(\cdot)$ and $\lambda_{kl}(t)$ each allow for a different extension to the standard JM specification. In particular, $f_{kl}(\cdot)$ allows alternative functional forms for $m_{ki}(t)$ to be specified rather than the current biomarker value. A more elaborate specification of the biomarker, such as the current gradient or area under the curve up to time t may result in a better surrogate for predicting the outcome of interest (Rizopoulos 2012*b*). Note that l indexes the functional form of the L_k forms specified for the kth longitudinal covariate.

The α parameters contained within $\lambda(t)$ determine the association between the two processes. $\lambda_{kl}(t)$ relates to an extension to the standard JM (Andrinopoulou et al. 2018) that relaxes the assumption of a constant association between $m_{ki}(t)$ and the hazard at time t. A flexible representation of the association parameters $\boldsymbol{\alpha}_{kl}$ may more accurately capture the underlying process, although this depends on the problem at hand.

To illustrate the above extensions, three possible choices of $f_{kl}(\cdot)$ are presented below. We compact notation by dropping the subscripts k and l, although they are still assumed.

$$f(\lambda(t), \mathcal{M}_i(t)) = \lambda(t) m_i(t), \tag{4.4}$$

$$f(\lambda(t), \mathcal{M}_i(t)) = \lambda(t) \frac{dm_i(t)}{dt}, \tag{4.5}$$

$$f(\lambda(t), \mathcal{M}_i(t)) = \lambda(t) \int_0^t m_i(s) ds, \tag{4.6}$$

Equation 4.4, 4.5 and 4.6 correspond to the current, gradient and cumulative parametrisations respectively, where the parameters are a smooth function of time contained within $\lambda(t)$. Note that Equation 4.4 corresponds to the standard JM when $\lambda(t) = \alpha$.

When $\lambda(t) = 0$, the longitudinal and survival processes are independent and no benefit is gained from a joint modelling approach compared to separate time-to-event and mixed effects models.

The association structure term can be flexibly modelled over time by representing it with a linear combination of B-spline basis functions evaluated at time t, i.e., $\boldsymbol{B} = \{B_g(t)\}_{g=1}^{G}$ and parameters $\boldsymbol{\alpha} = \{\alpha_g\}_{g=1}^{G}$.

$$\lambda(t) = \sum_{g}^{G} \alpha_g B_g(t) \tag{4.7}$$

Here the use of P-splines (Eilers & Marx 1996) to represent $\lambda(t)$ is appealing. With this approach the knot selection problem can be avoided by specifying a modestly large number of knots placed at the percentiles of the data and by penalising them according to the differences of the adjacent B-spline coefficients. This can be done naturally in the Bayesian context by choosing an appropriate prior. We delay this discussion and return to it in Section 4.3.2.

Due to the presence of random effects, parameter estimation in the joint modelling framework requires the complete likelihood function. In this case, the partial-likelihood approach used in proportional hazards regression does not suffice. For this reason we are required to specify the baseline hazard function $h_0(t)$ to complete the specification of Equation 4.3. In this work we estimate the baseline hazard function in the same spirit as the flexible association structure with P splines, i.e.,

$$\log(h_0(t)) = \sum_{u=1}^{U} \gamma_{h_0,u} B_u(t), \tag{4.8}$$

where $\gamma_{h_0,u}$ is the uth parameter of the baseline hazard function corresponding to the uth B-spline basis function.

4.3.2 Bayesian Parameter Estimation

When parameter estimation of the JM is conducted in the Bayesian paradigm, the Markov chain Monte Carlo (MCMC) class of algorithms is most commonly used to sample from the posterior distribution of the parameters, which is conditioned on the observed data. In this research, we use the `JMbayes` package to fit the MBJM which calls the sampling software Stan (Stan Development Team 2018). This implements a Hamiltonian Monte Carlo (HMC) algorithm for sampling from the posterior distribution.

Deriving the likelihood function of the MBJM relies on three assumptions of conditional independence. First, that the event times and longitudinal outcomes are independent given the random effects (Equation 4.9). Second, given the random effects the repeated

measures for a given subject are independent of each other for each longitudinal outcome (Equation 4.10). Finally, the longitudinal outcomes are independent of each other given the random effects (also Equation 4.10).

$$p(T_i, \delta_i, \boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_t)\ p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) \tag{4.9}$$

$$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y) = \prod_{k=1}^{K} \prod_{j=1}^{n_{ki}} p(y_{ki}(t_{kij}) \mid \boldsymbol{b}_i; \boldsymbol{\theta}_y), \tag{4.10}$$

where $\boldsymbol{\theta} = (\ \boldsymbol{\theta}_y^\top, \boldsymbol{\theta}_t^\top, \boldsymbol{\theta}_b^\top\ )^\top$ denotes the full parameter vector, containing the parameters relating to the longitudinal component $\boldsymbol{\theta}_y$, the time-to-event component $\boldsymbol{\theta}_t$, and the random effects component $\boldsymbol{\theta}_b$. Using Bayes' rule with the conditional independence assumptions above, the posterior distribution is derived as:

$$p(\boldsymbol{\theta}, \boldsymbol{b}_i \mid T_i, \delta_i, \boldsymbol{y}_i) \propto \prod_{k=1}^{K} \prod_{j=1}^{n_{ki}} p(y_{kij} \mid b_{ki}, \boldsymbol{\theta}_y)\ p(T_i, \delta_i \mid b_{ki}, \boldsymbol{\theta}_t)\ p(b_{ki} \mid \boldsymbol{\theta}_b)\ p(\boldsymbol{\theta}) \tag{4.11}$$

The likelihood contribution of the ith subject for the time-to-event, longitudinal and random effects components are respectively given by Equations 4.12, 4.13 and 4.14. The conditional survival function evaluated at time T_i is denoted by $S_t(T_i \mid \mathcal{M}_i(T_i))$.

$$p(T_i, \delta_i \mid \mathcal{M}_i(\cdot), \boldsymbol{\theta}_t) = h_i(T_i \mid \mathcal{M}_i(T_i))^{\delta_i} \underbrace{\exp\left\{ -\int_0^{T_i} h_i(s \mid \mathcal{M}_i(s)) ds \right\}}_{S_i(T_i \mid \mathcal{M}(T_i))} \tag{4.12}$$

$$p(y_{kij} \mid b_{ki}; \boldsymbol{\theta}_y) = \frac{1}{(2\pi\sigma_k)^2} \exp\left\{ \frac{-(y_{kij} - \boldsymbol{x}_{kij}^\top \boldsymbol{\beta}_k - \boldsymbol{z}_{kij}^\top b_{ik})^2}{2\sigma_k^2} \right\} \tag{4.13}$$

$$p(\boldsymbol{b}_i | \boldsymbol{\theta}) = (2\pi)^{-\frac{\sum_{k=1}^{K} q_k}{2}} \det(\boldsymbol{D})^{-\frac{1}{2}} \exp\left\{ \frac{-\boldsymbol{b}_i^\top \boldsymbol{D}^{-1} \boldsymbol{b}_i}{2} \right\} \tag{4.14}$$

The non-closed form integral contained within the survival function is numerically approximated using a 15 point Gauss-Kronrod quadrature rule (De Boor 2001).

4.3.3 Selecting Prior Distributions for Parameters in the Joint Model

Although the MBJM contains many parameters, the literature has well covered the problem of selecting priors for joint models (Brown & Ibrahim 2003, Ibrahim et al. 2010, Rizopoulos & Ghosh 2011). However, the choice of the most appropriate prior can depend on the type of data being analysed. Consistent with the studies noted above,

fixed effect coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are chosen to have standard non-informative priors. In both cases the variance is chosen to be 1000.

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_{\beta}^2)$$

$$\gamma \sim \mathcal{N}(0, \sigma_{\gamma}^2)$$

We employ a Bayesian P-spline approach that imposes a penalty on the complexity of the functions $h_0(t)$ and $\lambda(t)$ by assigning a set of hierarchical priors. In particular, the prior for the baseline hazard coefficients can be expressed as the following global smoothness prior:

$$p(\gamma_{h_0} \mid \tau_h) \propto \tau_h^{\rho(\boldsymbol{M}_{\gamma h_0})/2} \exp\Big(-\frac{\tau_h}{2} \gamma_{h_0}^{\top} \boldsymbol{M}_{\gamma h_0} \gamma_{h_0} \Big), \tag{4.15}$$

where the smoothing parameter $\tau_h \sim \Gamma(1, \tau_{h\delta})$ has a hyper-prior $\tau_{h\delta} \sim \Gamma(0.001, 0.001)$ ensuring that the posterior distribution of γ_{h_0} is proper (Lang & Brezger 2004). $\boldsymbol{M}_{\gamma h_0} = \Delta_r^{\top} \Delta_r + 10^{-6} \boldsymbol{I}$, where Δ_r is the r-th difference penalty matrix and the rank of $\boldsymbol{M}_{\gamma h_0}$ is represented by $\rho(\boldsymbol{M}_{\gamma h_0})$. The scaled identity matrix $\boldsymbol{I}$ ensures that the covariate matrix is positive-definite.

The smoothness of the time varying association structure $\lambda(t)$ can be controlled by selecting the following set of hierarchical priors for the coefficients:

$$\boldsymbol{\alpha} \mid \tau_{\alpha} \sim \mathcal{N}_G(\mathbf{0}, \tau_{\alpha} \boldsymbol{M}_{\alpha})$$

$$\tau_{\alpha} \sim \Gamma(c_1, c_2),$$

where $\boldsymbol{M}_{\alpha} = \boldsymbol{M}_{\gamma h_0}$, and in both cases we assume that the penalty is second order based on the recommendation by Eilers & Marx (1996). Furthermore, we follow the recommendation made by Lang & Brezger (2004) to set c_1 and c_2 to 1 and 0.005.

The precision parameters of the longitudinal components have an inverse gamma distribution with shape and scale parameters a and b respectively, i.e.,

$$\sigma_k^2 \sim \text{Inv-}\Gamma(a, b).$$

The most challenging prior to assign corresponds to the variance-covariance matrix of the random effects $\boldsymbol{D}$. Barnard et al. (2000) proposed the use of an inverse Wishart prior due to its conjugacy, which is the most commonly selected prior for multivariate normal data. However, various authors (Gelman 2006, Gelman et al. 2013, Alvarez et al. 2014) note detrimental issues relating to this prior, such as the marginal distribution for the variances having low density in a region near zero. In many cases users are restricted when using Bayesian software (such as JAGS) due to the inverse Wishart prior being the

only option for the latent covariance matrix. This is due to the Gibbs sampler approach suffering from the computational burden of alternative possible priors.

By employing a HMC algorithm, Stan does not suffer from this burden and its manual (Stan Development Team 2018) recommends a re-parametrisation of the covariance matrix of the random effects in terms of the correlation matrix $\boldsymbol{\Omega}$ and vector σ_d. An LKJ-correlation prior (Lewandowski et al. 2009) is assigned and the scale parameter is chosen to be $\zeta = 1.5$. A half-Student's t prior is used for each element of σ_d with 3 degrees of freedom.

4.3.4 Dynamic Prediction

The endogenous nature of the biomarkers in the problem at hand implies that in order for a longitudinal measurement to be taken at time t, survival is implied up to that point in time (Kalbfleisch & Prentice 2002). For this reason, prediction must be performed in a dynamic manner conditioning on both the observed data up to t and the fact that the subject has a survival probability of 1 at t.

Suppose we have a set of observed biomarker values for a new subject i', that are represented by $\mathcal{Y}_{ki'}(t) = \{y_{ki'}(t_{i'j}); 0 \leq t_{i'j} \leq t\}_{j=1}^{n_{i'}}$. We compact notation by letting $\mathcal{Y}_{i'}(t) = \{\mathcal{Y}_{1i'}(t), \ldots, \mathcal{Y}_{Ki'}(t\}$. The probability of surviving up to time u given survival up to t can be expressed mathematically as follows:

$$\pi_{i'}(u \mid t) = \mathbb{P}(T_{i'}^* \geq u \mid T_{i'}^* > t, \mathcal{Y}_{i'}(t), w_i(t), \mathfrak{D}; \boldsymbol{\theta}^*), \tag{4.16}$$

where $\mathfrak{D}$ is the training set that the JM was fitted on. $\boldsymbol{\theta}^*$ represents the vector of true parameters. We now compact notation further by dropping $w_i(t)$, although it is still assumed. The conditional independence assumption (Equation 4.9) allows us to rewrite Equation 4.16 as follows (Rizopoulos 2011):

$$\mathbb{P}(T_{i'}^* \geq u \mid T_{i'}^* > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta})$$

$$= \int \mathbb{P}(T_{i'}^* \geq u \mid T_{i'}^* > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta}) \times p(\boldsymbol{b}_{i'} \mid T_{i'}^* > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta}) d\boldsymbol{b}_{i'} \tag{4.17}$$

$$= \int \mathbb{P}(T_{i'}^* \geq u \mid T_{i'}^* > t; \boldsymbol{\theta}) \times p(\boldsymbol{b}_{i'} \mid T_{i'}^* > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta}) d\boldsymbol{b}_{i'} \tag{4.18}$$

$$= \int \frac{S_{i'}\{u \mid \mathcal{M}_{i'}(u, \boldsymbol{b}_{i'}, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{S_{i'}\{t \mid \mathcal{M}_{i'}(u, \boldsymbol{b}_{i'}, \boldsymbol{\theta}); \boldsymbol{\theta}\}} \times p(\boldsymbol{b}_{i'} \mid T_{i'}^* > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta}) d\boldsymbol{b}_{i'} \tag{4.19}$$

Estimating $\pi_{i'}(u \mid t)$ proceeds using asymptotic Bayesian arguments and rewriting Equation 4.16 in terms of its posterior predictive distribution:

$$\pi_{i'}(u \mid t) = \int \underbrace{\mathbb{P}(T^*_{i'} \geq u \mid T^*_{i'} > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta})}_{(=\textit{Equation } 4.19)} \; p(\boldsymbol{\theta} \mid \mathfrak{D}) d\boldsymbol{\theta} \tag{4.20}$$

$$= \int \int \frac{S_{i'}\{u \mid \mathcal{M}_{i'}(u, \boldsymbol{b}_{i'}, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{S_{i'}\{t \mid \mathcal{M}_{i'}(u, \boldsymbol{b}_{i'}, \boldsymbol{\theta}); \boldsymbol{\theta}\}} p(\boldsymbol{b}_{i'} \mid T^*_{i'} > t, \mathcal{Y}_{i'}(t); \boldsymbol{\theta}) \; p(\boldsymbol{\theta} \mid \mathfrak{D}) db_{i'} d\boldsymbol{\theta} \tag{4.21}$$

Note that the parameter vector $\boldsymbol{\theta}$ does not update with new subject i', which makes dynamic prediction computationally feasible. Following Rizopoulos (2011) we devise a Monte Carlo scheme to determine estimates of survival probabilities using Algorithm 4, which breaks the three terms in Equation 4.21 into separate steps.

Algorithm 4 *Monte Carlo Scheme for Dynamic Prediction in the MBJM*

1: **for** $l = 1, \ldots, L$ **do**
2: Draw $\tilde{\boldsymbol{\theta}}^{(l)}$ from the posterior distribution of the parameters $[\, \boldsymbol{\theta} \mid \mathfrak{D} \,]$
3: Draw $\tilde{\boldsymbol{b}}^{(l)}_{i'}$ from the random effects posterior distribution $[\, \boldsymbol{b}_{i'} \mid T^*_{i'} > t, \mathcal{Y}_{i'}(t), \boldsymbol{\theta}^{(l)}]$
4: Compute the survival probabilities ratio $\frac{S_{i'}(u \mid \mathcal{M}_{i'}(u, \tilde{\boldsymbol{b}}^{(l)}_{i'}), \tilde{\boldsymbol{\theta}}^{(l)})}{S_{i'}(t \mid \mathcal{M}_{i'}(t, \tilde{\boldsymbol{b}}^{(l)}_{i'}), \tilde{\boldsymbol{\theta}}^{(l)})}$
5: **end for**
6: Estimate the conditional survival probabilities by: $\frac{1}{L} \sum_{l=1}^{L} \frac{S_{i'}(u \mid \mathcal{M}_{i'}(u, \tilde{\boldsymbol{b}}^{(l)}_{i'}), \tilde{\boldsymbol{\theta}}^{(l)})}{S_{i'}(t \mid \mathcal{M}_{i'}(t, \tilde{\boldsymbol{b}}^{(l)}_{i'}), \tilde{\boldsymbol{\theta}}^{(l)})}$

We note that line 2 and 4 are straightforward, however line 3 requires the use of the Metropolis Hastings algorithm with multivariate t proposals. Rizopoulos et al. (2008) show that as n_i increases this posterior converges to a multivariate normal distribution.

The standard deviation over the Monte Carlo samples gives the standard errors which are used to derive confidence intervals. This sample scheme assumes that n is sufficiently large such that $[\, \boldsymbol{\theta} \mid \mathfrak{D} \,]$ can be well approximated.

4.3.5 Assessment of Predictive Performance

In this work we focus on performance measures that test a model's ability to correctly discriminate between patients (i.e., assign a higher survival probability to a patient that survives longer than another patient, who is assigned a lower survival probability). We also consider measures that assess the ability of the model to correctly predict the observed event rates. These phenomena are referred to **discrimination** and **calibration** respectively (Harrell 2015). In Chapter 3 we used the AUC ROC as a measure of discriminatory ability for a binary classifier. In this chapter we are required to extend this concept to be able to account for dynamic prediction and censoring. We also want to use a measure that allows the probability of censoring to depend on the observed biomarker values and baseline covariates.

4.3.5.1 Discrimination

In order to assess the discriminatory ability of the physiological variables recorded up to time t ($\mathcal{Y}(t)$) and baseline covariates $\boldsymbol{w}(t)$ in dynamically predicting survival probabilities, we must specify a medically relevant time frame $(t, t+\Delta t]$. As discussed in the literature review, a consensus for predicting donor death time is 60 minutes. Assuming we predict from 10 minutes to an hour, Δt would represent the difference being 50 minutes. In this case $t + \Delta t = u$.

Being able to discriminate between two patients, where one incurs an event within the medically relevant time frame and another that does not, would enable the organ procurement team to be in the right place at the right time. We define a prediction rule such as Equation 4.16, where the probability of an event $\pi_{i'}(t + \Delta t \mid t)$ depends on the history of biomarkers $\mathcal{Y}_{i'}(t)$. A threshold value $c \in [\, 0, 1\,]$ is chosen such that:

$$\begin{cases} \pi_{i'}(t + \Delta t \mid t) \leq c, & \text{Predict Event between } t \text{ and } \Delta t \\ \pi_{i'}(t + \Delta t \mid t) > c, & \text{Predict No Event between } t \text{ and } \Delta t \end{cases}$$

We can define sensitivity and specificity such that they depend on the specified time frame as given by Equations 4.22 and 4.23.

$$\text{Sensitivity} = \mathbb{P}\{\pi_{i'}(t + \Delta t) \leq c \mid T^*_{i'} \in (t, t + \Delta t]\ \} \tag{4.22}$$

$$\text{Specificity} = \mathbb{P}\{\pi_{i'}(t + \Delta t) > c \mid T^*_{i'} > t + \Delta t\} \tag{4.23}$$

Denote a randomly chosen pair of subjects as i and i'. The AUC ROC can be derived using Equation 4.24, which reflects the overall discriminatory ability of the assumed model across a varied threshold c. Intuitively speaking, this formula reflects the chances of the model correctly assigning a higher probability of survival for patient i' who incurs an event after the medically relevant time frame compared to subject i who incurs an event within the time frame. This formulation is based on similar arguments to Harrell's C-index.

$$\text{AUCROC}(t, \Delta t) = \mathbb{P}[\, \pi_{i'}(t+\Delta t \mid t) < \pi_{i'}(t+\Delta t \mid t) \mid \{T^*_i \in (t, t+\Delta t]\ \} \cap \{T^*_{i'} > t+\Delta t\}] \tag{4.24}$$

The AUC ROC is estimated by the frequency of concordant[4] pairs of subjects. However, due to the presence of censoring many pairs are not directly comparable. The AUC ROC is decomposed into sets of comparable pairs (i.e., patients whose event times can

[4] A pair is said to be concordant if a randomly selected pair of patients is selected and the patient that experiences the event first is assigned a higher event probability.

be ordered) (Rizopoulos et al. 2017). That is,

$$\widehat{\text{AUCROC}} = \sum_{a=1}^{4} \widehat{\text{AUCROC}}_a(t, \Delta t), \tag{4.25}$$

where index a corresponds to a single component of the decomposition. The components can be derived as follows,

$$\widehat{\text{AUCROC}}_a(t, \Delta t) = \frac{\sum_{i=1}^{n} \sum_{i'=1; i' \neq i}^{n} \mathbb{I}\{\hat{\pi}_i(t + \Delta t \mid t) < \hat{\pi}_{i'}(t + \Delta t \mid t)\} \times \mathbb{I}\{\Psi_{i,i'}^{(a)}(t)\} \times \hat{v}_{i,i'}^{(a)}}{\sum_{i=1}^{n} \sum_{i'=1, i' \neq i}^{n} \mathbb{I}\{\Psi_{i,i'}^{(a)}(t)\} \times \hat{v}_{i,i'}^{(a)}}, \tag{4.26}$$

where the four distinct sets of comparable pairs are defined by,

$$\Psi_{i,i'}^{(1)}(t) = [\ \{T_i \in (t, t + \Delta t]\ \} \cap \{\delta_i = 1\}\ ]\ \cap \{T_{i'} > t + \Delta t\} \tag{4.27}$$

$$\Psi_{i,i'}^{(2)}(t) = [\ \{T_i \in (t, t + \Delta t]\ \} \cap \{\delta_i = 0\}\ ]\ \cap \{T_{i'} > t + \Delta t\} \tag{4.28}$$

$$\Psi_{i,i'}^{(3)}(t) = [\ \{T_i \in (t, t + \Delta t]\ \} \cap \{\delta_i = 1\}\ ]\ \cap \{T_i < T_{i'} \leq t + \Delta t\} \cap \{\delta_{i'} = 0\}] \tag{4.29}$$

$$\Psi_{i,i'}^{(4)}(t) = [\ \{T_i \in (t, t + \Delta t]\ \} \cap \{\delta_i = 0\}\ ]\ \cap \{T_i < T_{i'} \leq t + \Delta t\} \cap \{\delta_{i'} = 0\}] \tag{4.30}$$

and $\{\hat{v_a}\}_{a=1}^{4}$ weight the relative frequency of concordant pairs to account for censoring. In particular, $\hat{v}_{i,i'}^{(1)} = 1$, because the pairs of subjects in the set $\Psi_{i,i'}^{(1)}(t)$ are all comparable as they all experience an event. In this case, Equation 4.26 directly corresponds to the relative frequency of concordant pairs. $\hat{v}_{i,i'}^{(2)} = 1 - \hat{\pi}_i(t + \Delta t \mid T_{i'})$, $\hat{v}_{i,i'}^{(3)} = \hat{\pi}_{i'}(t + \Delta t \mid T_{i'})$ and $\hat{v}_{i,i'}^{(4)} = \{1 - \hat{\pi}_i(t + \Delta t \mid T_i)\} \times \hat{\pi}_{i'}(t + \Delta t \mid T_{i'})$ are the weights for the components where censoring occurs, representing the probability that the concordant subjects are comparable. This provides a measure of discriminatory ability that accounts for censoring and the dynamic nature of the predictions. Note that $\hat{\pi}(t + \Delta t \mid t)$ is estimated by Equation 4.21.

4.3.5.2 Calibration

A standard measure of predictive accuracy in both the survival and joint modelling frameworks is the expected prediction error, also referred to as calibration. This can be expressed as follows:

$$\text{PE}(u \mid t) = \mathbb{E}\{L[\ N_i(u) - \pi_i(u \mid t)\ ]\ \}, \tag{4.31}$$

where time dependent event status is given by $N_i(u) = \mathbb{I}(T_i^* > u)$, and the loss function (that we choose to be the square loss in this work) is represented by $L(\cdot)$. Henderson et al. (2002) acknowledge that censoring causes bias in estimating such quantities as Equation

4.31, and that simply removing censored observations is an inadequate approach. They proposed the following unbiased estimator that takes censoring into account,

$$\widehat{PE}(u \mid t) = \frac{1}{\mathcal{N}(t)} \sum_{i:T_i \geq t} \underbrace{\mathbb{I}(t_i \geq u) L[\, 1 - \hat{\pi}_i(u \mid t)]}_{A} + \underbrace{\delta_i \mathbb{I}(T_i < u) L[\, 0 - \hat{\pi}_i(u \mid t)]}_{B} + \underbrace{(1 - \delta_i)\mathbb{I}(T_i < u)\{\hat{\pi}_i(u \mid T_i) L[\, 1 - \pi_i(u \mid t)] \; + \; \{1 - \hat{\pi}_i(u \mid t) L[\, 0 - \hat{\pi}_i(u \mid t)] \, \}\}}_{C},$$

where $\mathcal{N}(t)$ gives the number subjects at risk at time t. The above formula is best understood by breaking it into three components. The component labelled A represents the patients that were remaining after $t + \Delta t$. Component B refers to those who died before $t + \Delta t$. Component C denotes those that were censored within the medically relevant time frame.

4.3.6 Model Selection for Time-Independent Baseline Variables

In order to identify important time independent baseline variables for predicting the time of asystole, a random survival forest (RSF) (Ishwaran & Kogalur 2007, Ishwaran et al. 2008), as described in section 4.3.6.1, was fitted to the time-to-event response with only time-independent baseline variables included as the predictors. A stepwise selection procedure (Rizopoulos 2009) that makes use of bootstrap sampling was used as a confirmation that the findings from the RSF were reasonable. This involved taking bootstrap samples of the data and performing a sequential replacement selection procedure (which is a combination of forward and backwards) to each sample, choosing the model with the lowest AIC. The sequential procedure involved starting with a null proportional hazards model and sequentially including the most important predictors. Every time a new variable was added, any variable that was found to no longer improve the goodness-of-fit was removed.

4.3.6.1 Random Survival Forests (RSF)

The RSF is an extension of the random forest described in detail in Chapter 3, adapted to be used for a survival outcome. In the case of this work, this relates to the bivariate outcome that includes an event time (time of asystole) and a censoring indicator.

The steps of the RSF algorithm are analogous to the random forest algorithm given in Chapter 3. They are given by Ishwaran et al. (2008) as follows:

- Draw B samples with replacement from the dataset.

- For each bootstrap sample, grow a survival tree, randomly selecting p variables at each node. The split at a node is made based on the strongest candidate variable, i.e., that which maximises the survival difference between the daughter nodes.
- Grow a tall tree (to full size) using a constraint on the number of unique deaths.
- Calculate a cumulative hazard function for each tree, based on the Nelson-Aalen estimator: $\hat{H}_h(t) = \sum_{t_l \leq t} \frac{d_{l,h}}{Y_{l,h}}$, where $d_{l,h}$ and $Y_{l,h}$ represent the number of deaths and individuals at risk at time $t_{l,h}$ respectively. Average to obtain the ensemble CHF.
- Use data not included in bootstrap sample to calculate prediction error for the ensemble CHF.

4.4 Analysis of the novel dataset

In this section we conduct a predictive analysis with data consisting of 232 DCD organ donors. Two datasets are merged, one containing only donor baseline covariates and the other containing a set of physiological variables repeatedly measured over time. As five of these donors only have a single measure of the longitudinal covariates, we restrict the dataset to 227 donors.

The same donor baseline covariates as the dataset analysed in Chapter 3 are present (`dage`, `dheight`, `dweight`, `dsex`, `dbg`, `dcod_grp`, `dethnic`)[5]. These data have the additional variables `sod`, `kiddonor` and `Proceed`, representing whether the patient was a solid organ donor, kidney donor, or proceeded to transplant respectively. As these variables were only recorded once the treatment withdrawal phase had terminated, these variables were not used to predict death time (which otherwise would result in leakage). We create the additional variable `dbmi` from `dheight` and `dweight`, which represents the donor body mass index (BMI)[6].

The longitudinal dataset contains various physiological variables (SBP, DBP, HR, O2, Resp Rate) that are measured repeatedly from the moment the donor is withdrawn from treatment until asystole or censoring. The date, time (in the format of the 24 hour clock) and ID variables were used create a new time variable representing the number of minutes from treatment withdrawal for each patient. This as well as the event indicator variable were used to calculate the survival or censoring times for each patient.

The event time variable and the censoring indicator were collapsed from the long format to the wide format, and were merged with the dataset containing the baseline covariates. Model selection was performed on this dataset to determine which baseline covariates

[5]Recall that a description of these variables can be found in Appendix Table A.1.
[6]Where BMI = Weight (Kg) / {Height(metres)}^{2}

are important for predicting the time of asystole in the withdrawal period employing the methods described in Section 4.3.6.

Subsequently, we investigate which longitudinal covariates are most predictive of the event times, by comparing null[7] univariate[8] JMs with 5-fold cross-validation (for both linear and flexible association structures). We also performed 5-fold cross-validation on all possible combinations of bivariate null JMs (for linear, flexible and a mixture of the two association structures). The baseline covariates were then included that were deemed to be important in the previous stage. The chosen model is assessed with a 5-fold cross-validation repeated 15 times (each time the data is randomly split into folds). Dynamic predictions were made at time-frames 5-20, 15-60 and 30-75 minutes. As discussed in the literature review, one hour is a medically relevant timepoint that is of interest to clinicians. 15 minutes was chosen as the starting point to allow a sufficient number of readings to be taken. Time-frames 5-20 and 30-75 minutes were chosen arbitrarily as a basis of comparison.

As only three values are missing for the baseline covariates in this dataset (as a result of `dcod` codes not being available), we perform a single imputation on this variable using predictive mean matching (see Little (1988) for details).

4.4.0.1 Exploratory Data Analysis

Tables 4.1 and 4.2 display descriptive statistics for the categorical and continuous baseline covariates respectively, stratified by the censoring indicator. It can be seen that the majority of patients were white males, for most of whom their death time was known.

Donor age appears to play an important role in whether the patient becomes censored or not. In particular, the average age of the censored patients is larger than that of those who died (60 and 49 respectively). This was highly statistically significant (P-value < 0.001 with a Mann-Whitney U test), which conforms with intuition, as younger donors are known to have higher success rates than older donors, and are thus less likely to be dropped out from the organ donation process.

Although ethnicity has a significant P-value of 0.03, this can be attributed to the small counts for some of the categories (zero censored donors whose ethnicity was unknown). Dropping the category "unknown" led to a P-value of 0.99, which is still questionable due to low count numbers for "not white" donors. The large imbalance of ethnicity in these data impedes the predictive potential of this covariate.

[7] In this Chapter we adopt a slight variation to the standard meaning of 'null' and use this terminology to relate it to a JM that includes no confounding variables (including only time as a fixed and random effect in the longitudinal component and only an intercept in the fixed component of the survival model).

[8] In this Chapter, when referring to 'univariate JMs', we mean JMs with a single longitudinal covariate (and bivariate for two longitudinal covariates)

Table 4.1: *Description of categorical baseline variables relating to organ donors that are also present in the longitudinal dataset. Totals and percentages are given per event status and a chi-squared test of independence p-value is displayed.*

	All patients n (%)	Censored n (%)	Died n (%)	P-Value (χ^2 test)
Total	227	72 (32)	155 (68)	
Gender				0.86
Male	139 (61)	43 (19)	96 (42)	
Female	88 (39)	29 (13)	59 (26)	
Ethnicity				0.03
White	205 (90)	69 (30)	136 (60)	
Not White	8 (4)	3 (1)	5 (2)	
Unknown	14 (6)	0 (0)	14 (6)	
Blood Group				0.90
O	102 (45)	31 (14)	71 (31)	
A	90 (40)	28 (12)	62 (27)	
B	24 (11)	9 (4)	15 (7)	
AB	11 (5)	4 (2)	7 (3)	
Solid Organ				< 0.001
Yes	165 (73)	14 (6)	151 (67)	
No	62 (27)	58 (26)	4 (2)	
Kidney Donor				< 0.001
Yes	157 (69)	13 (6)	144 (63)	
No	70 (31)	59 (26)	11 (5)	
Death Cause				0.47
CVA	96 (43)	35 (16)	61 (27)	
Miscellaneous	109 (49)	30 (13)	79 (36)	
RTA	12 (5)	3 (1)	9 (4)	
Other Trauma	7 (3)	3 (1)	4 (2)	
Proceed				< 0.001
Yes	167 (74)	16 (7)	151 (67)	
No	60 (26)	56 (25)	4 (2)	

Table 4.2: *Mean and standard deviation (sd) of the continuous variables, unstratified and stratified for censoring. A Mann-Whitney U test is performed to determine whether the two groups (censoring) are from the same population.*

	Unstratified	Censored	Died	Mann-Whitney U
Variable	Mean (sd)	Mean (sd)		P-Value
Age	52 (18)	60 (15)	49 (18)	< 0.001
Weight	78 (19)	78 (19)	78 (20)	0.99
Height	169 (16)	167 (17)	170 (15)	0.06
BMI	27 (6)	28 (6)	27 (6)	0.31

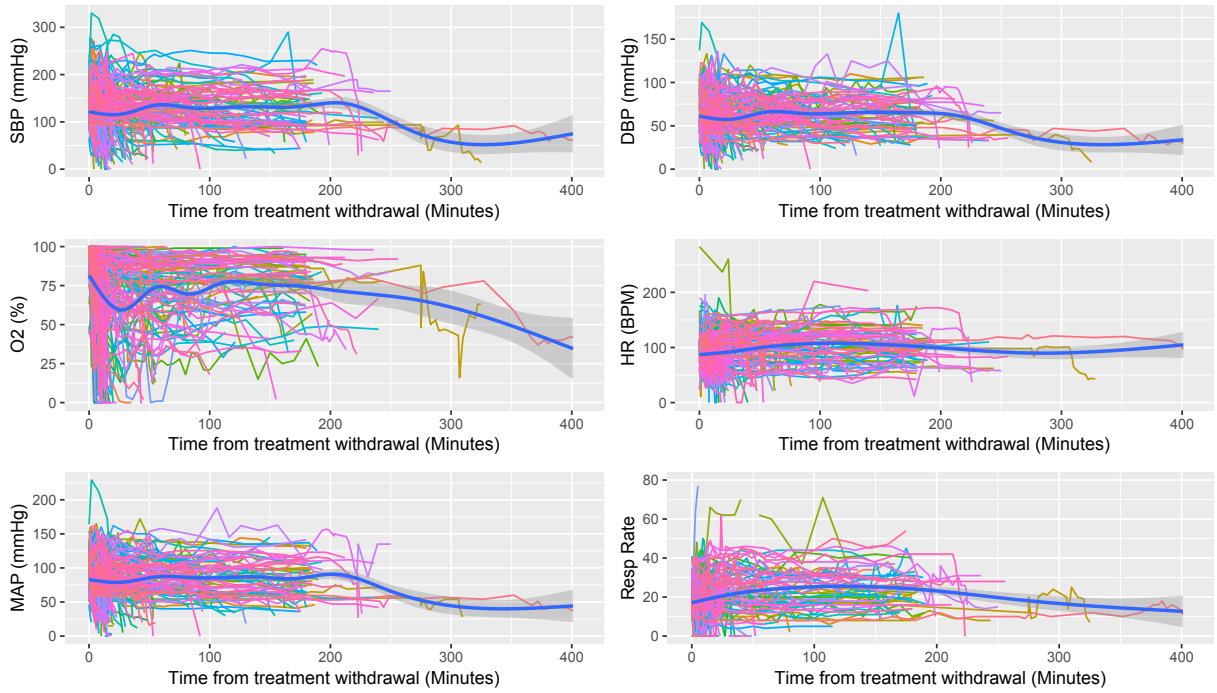


Figure 4.1: *Trajectory plots displaying the evolution of each trajectory over time for each donor (multiple coloured lines) for each longitudinal covariate. The solid blue line shows the flexible mean profile fitted by a GAM, with its 95% confidence interval (the shaded region).*

The trajectories of the physiological variables are displayed in Figure 4.1. Each line (except the thick blue line) represents the evolution of an individual donor's physiological outcome from treatment withdrawal until either death or censoring. Many of the trajectories are highly volatile for each of the variables, which is expected as these patients were at this time experiencing the dying phase.

A high density of the observations occur within first half an hour, as many patients experience a rapid decline. This can be seen in the left plot in Figure 4.2, which is a box plot of the event times stratified by censoring status. As expected the majority of patients that were censored experienced an event later than those that died, as a prolonged withdrawal period is thought to compromise graft quality, thereby leading to the decision to remove patients from the study.

Figure A.1 displays a box plot showing the within individual mean and standard deviation of the longitudinal outcomes. This confirms the presence of a large amount of volatility, that is obvious from Figure 4.1. The most variation is present for SBP, in terms of both the within individual mean and standard deviation.

The solid blue line in Figure 4.1 represents the mean profile with the shaded area representing a 95% confidence interval, which is modelled by a generalised additive model (GAM) (Hastie & Tibshirani 1987). This suggests that many of the mean profiles have a non-linear functional form and should be modelled flexibly in the MBJM. The shaded

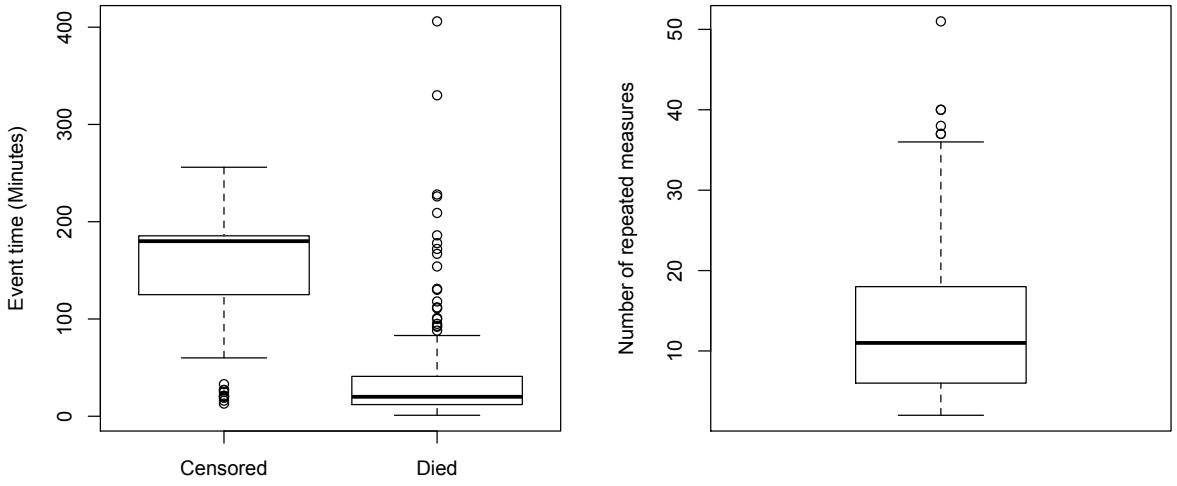


Figure 4.2: *Box plots displaying summary statistics of the donor event times (left) and also the number of repeated measures per donor (right).*

region becomes wider as the data becomes relatively sparse past approximately the 200 minute mark.

The right plot in Figure 4.2 displays a box plot representing the number of repeated measures for each donor. It is clear that the data is highly unbalanced having a minimum of 2, median of 12 and a maximum of 50 repeated measures. Furthermore, it can be seen from the trajectory plots that the follow-up times are also unequally spaced. A mixed effects modelling approach is required to deal with such complications, which is used in the longitudinal component of the MBJM.

Figure 4.3 presents Kaplan-Meier curves displaying the observed probability of survival throughout the treatment withdrawal period for each of the continuous variables categorised at their quantiles. A log-rank test (Mantel 1966, Cox 1972) is performed to test for a significant difference between survival curves. Age is the only significant variable in this plot at the 5% significance level (P-value < 0.001). However, differences between the survival curves only becomes prominent once past the median event time (27 minutes). Once past this point, young donors are the fastest to deteriorate, which is desirable for organ procurement. Unintuitively, the slowest group to deteriorate was the second youngest age group. This suggests age should be modelled flexibly in the MBJM.

Figure 4.4 provides Kaplan-Meier curves for each of the categorical covariates that are of potential use for predicting donor event time in the treatment withdrawal period. According to the log-rank test (p-values provided in the title of each plot), none of

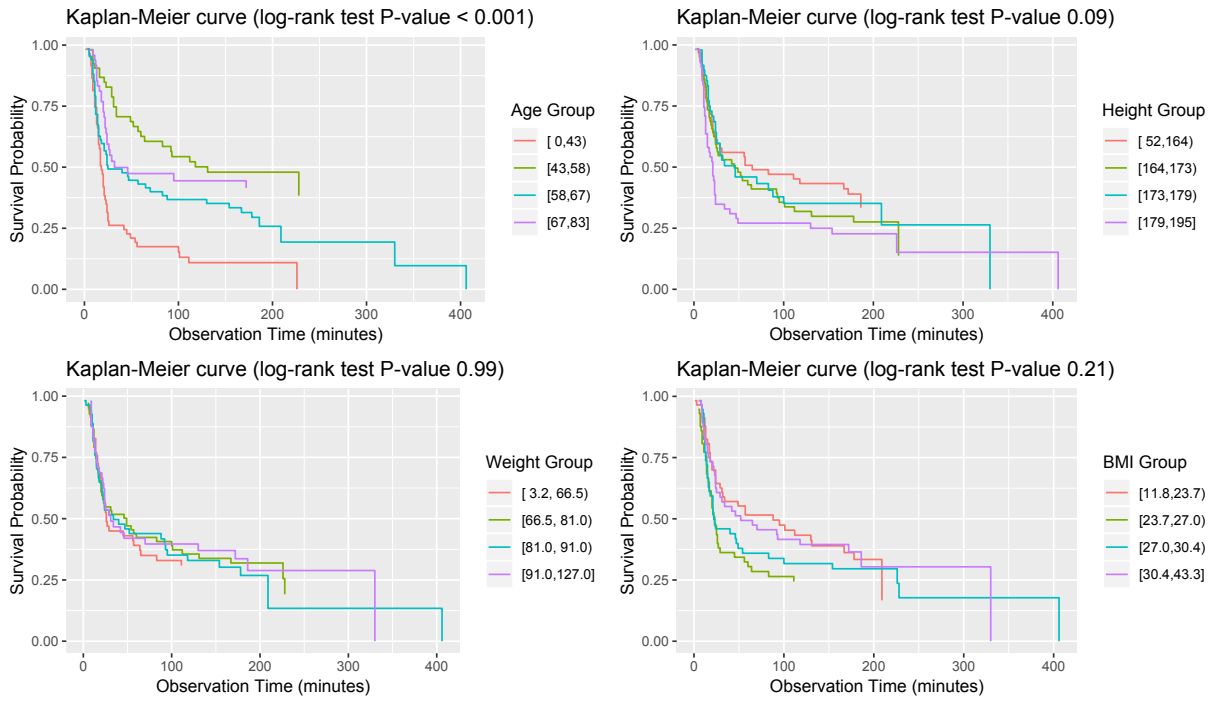


Figure 4.3: *Kaplan-Meier curves for the continuous variables categorised at their quantiles, with p-values relating to the log-rank test for equal survival curves.*

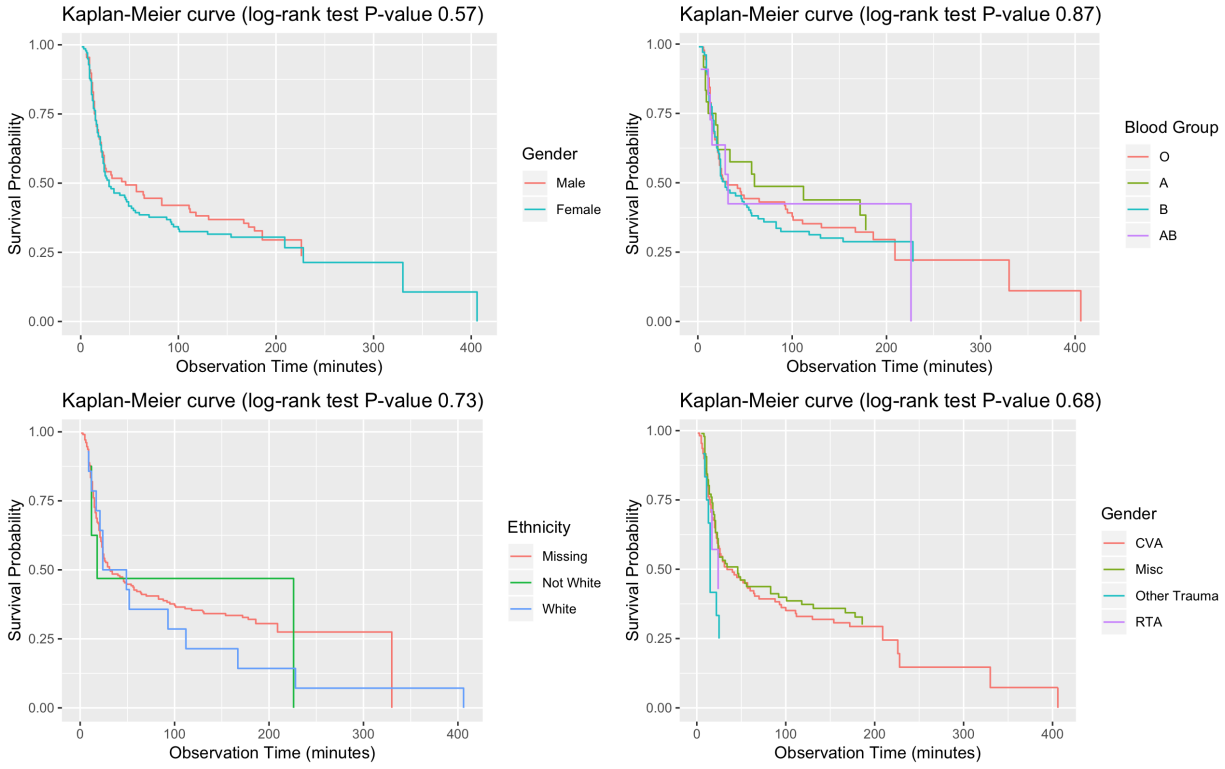


Figure 4.4: *Kaplan-Meier curves for the categorical covariates, with p-values relating to the log-rank test for equal survival curves.*

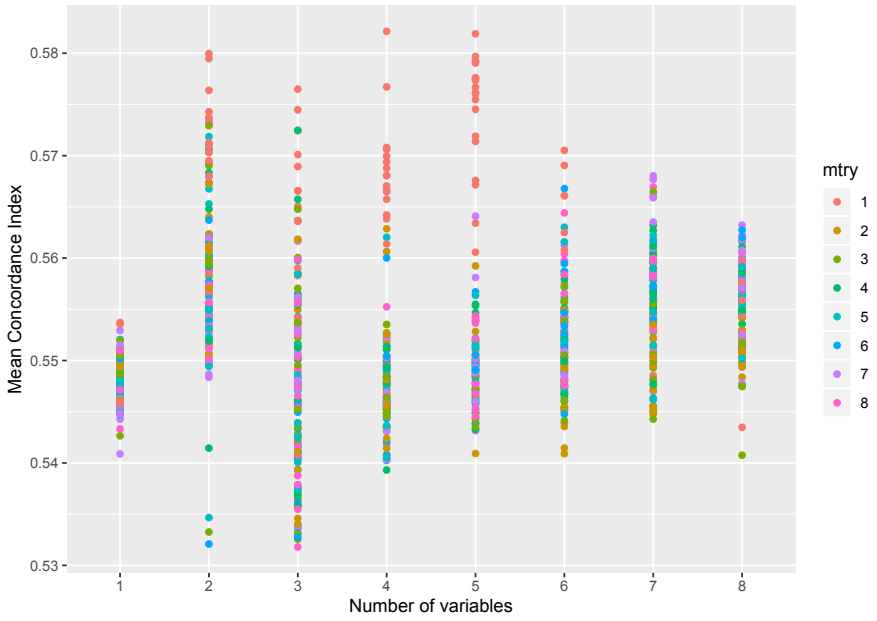


Figure 4.5: *Hyper-parameter tuning the random survival forest to maximise the mean C-index across cross-validation folds. Hyper-parameters include the number of variables to try at splits (mtry), number of relevant variables and the number of trees (omitted from figure).*

these covariates' categories have significantly different survival curves. This suggests that these variables are not likely to be predictive of the donor event time.

4.4.0.2 Identifying Important Baseline Covariates

RSFs are used to identify which baseline covariates may be predictive of the donor event times, and thus which should be included in the JM. The following hyper-parameters are tuned by randomly sampling 1000 times from the specified bounds of the hyper-parameters (as given in the following bullet points) and performing 5-fold cross-validation to calculate the mean C-index:

- `mtry`: Randomly sample `mtry` variables from the set of predictors, these are used to find the optimal split of the feature space. Specified bound between 1 and 8 (minimum and maximum number of covariates in the dataset).
- `ntree`: The number of regression trees to fit (corresponding to the number of bootstrap samples taken). Specified bound between 50 and 2000.
- `Number of variables`: The number of variables ranked in order of importance to keep in the model. Specified bound between 1 and 8.

Figure 4.5 presents the hyper-parameter tuning process. Each point represents the C-index achieved for a RSF with a single random draw from the specified hyper-parameter

Table 4.3: *Proportion of bootstrap model selection iterations variables are retained in the model. Both forward and backwards selection are performed using the AIC.*

Variable	Retained (%)
dage	100.0
dheight	86.1
dethnic	57.1
dsex	50.3
dweight	36.8
dbmi	34.3
dbg	26.5
dcod	21.7

space. As data visualisation is limited by the number of dimensions to be plotted, the number of trees is omitted from the plot. This figure is given for illustratory purposes.

The optimal set of hyper-parameters are: mtry = 1, ntree = 1633, number of variables = 4. This results in a mean test C-index of 0.582. The top four variables ranked by the permutation importance index are: age, height, gender, blood group. It is important to be aware that although the optimal parameter set retains all four variables, the improvement in accuracy by including gender and blood group is 0.002. This difference is small enough such that it could only improve accuracy by random chance.

We seek assurance by employing an additional model selection procedure based on bootstrap sampling of the dataset and fitting proportional hazards models by both forward and backwards stepwise selection (dropping variables that result in a larger AIC). The results for this procedure are given in Table 4.3 for 1000 bootstrap samples, indicating that the model containing only age and height has the best fit. The RSF containing only age and height resulted in an almost equivalent accuracy to the RSF that contained the variables in (what was determined to be) the optimal set of variables. This can be seen in Figure 4.5, as the largest mean concordance index for two variables is similar to that of four.

We proceed by fitting a GAM suitable for a survival response to determine whether it is sufficient to model the baseline covariates that were found to be important by assuming that the associations are linear, or whether a more elaborate specification is required. Figure 4.6 presents the functional form of height and age corresponding to the GAM fitted that contains only these two variables as covariates. Although the points appear to be somewhat randomly scattered, with a possibly clustered structure, a flexible representation was not found to improve the fit compared to a linear fit. Therefore, we proceed by assuming linearly in the model for these variables.

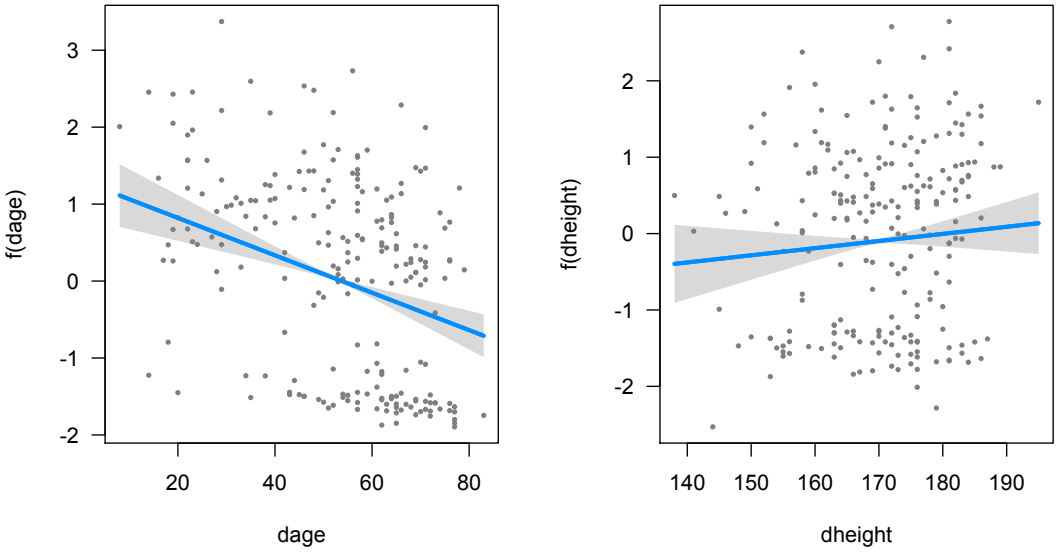


Figure 4.6: *Testing the assumption of linearity by modelling covariates with smoothing splines in a GAM. The shaded region corresponds to a 95% confidence interval.*

4.4.0.3 Joint Modelling

Joint models were compared using the discrimination and calibration measures described in Section 4.3.5, with 5-fold cross-validation. For every null[9] JM fitted, the longitudinal component corresponds to Equation 4.32, where in the multivariate case $K > 1$ (where $k = 1, \ldots, K$).

$$y_{ki}(t) = (\beta_{k0} + b_{k0}) + (\beta_{k1} + b_{k1})B_{ki1}(t) + (\beta_{k2} + b_{k2})B_{ki2}(t) + \epsilon_{ki}(t) \tag{4.32}$$

Here the B-spline basis functions for the natural spline, given by B, allow a flexible representation of both the fixed and random effects. The formulation of Equation 4.32 shows how the random intercept and slope terms represent a deviation from the mean profile, which is recovered when the random effects are set to zero.

We compared many different combinations of null JMs to determine the most appropriate longitudinal response (or combinations of responses). In particular, we compared univariate null JMs for each physiological variable for both linear and flexible association structures. Equation 4.33 displays the form of the survival component fitted. In

[9] A null JM corresponds to specifying no variables for $\boldsymbol{w}$ or any fixed effects not relating to time in the longitudinal component (as given by Equation 4.32).

Table 4.4: *Joint model selection using 5-fold cross-validation. For each univariate JM with linear association structure, the mean and standard deviation (in brackets) for the AUC ROC and PE are given across the 5 folds for three different time frames in the agonal phase.*

	Mean AUC ROC (Standard Deviation)			Mean PE (Standard Deviation)		
	5-20 mins	15-60 mins	30-75 mins	5-20 mins	15-60 mins	30-75 mins
SBP	0.67 (0.15)	0.82 (0.06)	0.73 (0.24)	0.22 (0.03)	0.17 (0.05)	0.12 (0.05)
DBP	0.68 (0.09)	0.74 (0.09)	0.72 (0.17)	0.24 (0.08)	0.22 (0.07)	0.16 (0.06)
O2	0.61 (0.08)	0.83 (0.06)	0.74 (0.17)	0.24 (0.02)	0.20 (0.05)	0.15 (0.05)
HR	0.44 (0.07)	0.64 (0.12)	0.42 (0.25)	0.25 (0.05)	0.23 (0.03)	0.18 (0.04)
MAP	0.67 (0.10)	0.76 (0.05)	0.73 (0.22)	0.24 (0.05)	0.20 (0.02)	0.13 (0.05)
Resp	0.69 (0.08)	0.77 (0.20)	0.55 (0.33)	0.19 (0.03)	0.18 (0.06)	0.15 (0.07)

Table 4.5: *Joint model selection using 5-fold cross-validation. For each univariate JM with flexible association structure, the mean and standard deviation (in brackets) for the AUC ROC and PE are given across the 5 folds for three different time frames in the agonal phase. Variable names ending with an apostrophe assume a flexible association structure.*

	Mean AUC ROC (Standard Deviation)			Mean PE (Standard Deviation)		
	5-20 mins	15-60 mins	30-75 mins	5-20 mins	15-60 mins	30-75 mins
SBP'	0.43 (0.16)	0.75 (0.14)	0.47 (0.33)	0.30 (0.05)	0.30 (0.07)	0.32 (0.10)
DBP'	0.47 (0.15)	0.47 (0.24)	0.40 (0.27)	0.29 (0.09)	0.33 (0.12)	0.34 (0.12)
O2'	0.56 (0.11)	0.90 (0.07)	0.83 (0.13)	0.23 (0.02)	0.17 (0.04)	0.12 (0.03)
HR'	0.12 (0.05)	0.09 (0.04)	0.05 (0.05)	0.54 (0.06)	0.58 (0.05)	0.67 (0.10)
MAP'	0.45 (0.13)	0.54 (0.15)	0.42 (0.21)	0.28 (0.06)	0.28 (0.05)	0.27 (0.07)
Resp'	0.31 (0.13)	0.28 (0.10)	0.12 (0.07)	0.33 (0.07)	0.43 (0.04)	0.47 (0.03)

the univariate case $k = 1$ and a linear association structure corresponds to $\lambda(t) = \alpha$.

$$h_i(t, \mathcal{M}_i(t), \boldsymbol{w}_i) = h_0(t) \exp\left(\gamma^\top \boldsymbol{w}_i + \sum_{k=1}^{K} \lambda_k(t) m_{ki}(t)\right) \tag{4.33}$$

Table 4.4 displays the results for the univariate null JMs with linear association structures. The mean AUC ROC and mean PE are given with the standard deviation (displayed in brackets) across the 5 cross-validation folds. It can be seen that `SBP` and `O2` are overall the most predictive longitudinal covariates across the three medically relevant time frames. Although `O2` has the highest discrimination between 15-60 minutes it has the second lowest discrimination between 5-20 minutes. Both of these models performed well in terms of calibration, and `Resp` has a competitive predictive performance.

The cross-validation results displayed in Table 4.5 correspond to the same fitted models as those displayed in Table 4.4, except a flexible association structure is specified. Comparing Table 4.4 and 4.5, it can be seen that a flexible specification of the association structure impedes the predictive ability of many of the physiological variables. In particular, mean AUC ROC decreases and the mean PE increases for every time frame

Table 4.6: *5 fold cross-validation for identifying the best pair of longitudinal responses, where association structure is modelled linearly.*

	Mean AUC ROC (Standard Deviation)			Mean PE (Standard Deviation)		
	5-20 mins	15-60 mins	30-75 mins	5-20 mins	15-60 mins	30-75 mins
SBP+DBP	0.53 (0.14)	0.61 (0.19)	0.54 (0.26)	0.27 (0.07)	0.28 (0.09)	0.29 (0.08)
SBP+O2	0.73 (0.07)	0.88 (0.05)	0.72 (0.15)	0.24 (0.06)	0.18 (0.05)	0.13 (0.03)
SBP+HR	0.76 (0.11)	0.88 (0.11)	0.78 (0.17)	0.22 (0.07)	0.16 (0.08)	0.14 (0.06)
SBP+MAP	0.75 (0.09)	0.84 (0.09)	0.77 (0.15)	0.22 (0.04)	0.17 (0.05)	0.13 (0.04)
SBP+Resp	0.72 (0.18)	0.78 (0.09)	0.76 (0.18)	0.20 (0.06)	0.16 (0.05)	0.13 (0.03)
DBP+O2	0.70 (0.09)	0.85 (0.06)	0.70 (0.15)	0.25 (0.11)	0.21 (0.05)	0.12 (0.03)
DBP+HR	0.67 (0.06)	0.76 (0.07)	0.71 (0.17)	0.24 (0.03)	0.21 (0.05)	0.14 (0.05)
DBP+MAP	0.70 (0.08)	0.82 (0.09)	0.77 (0.15)	0.24 (0.04)	0.18 (0.07)	0.14 (0.06)
DBP+Resp	0.72 (0.19)	0.80 (0.13)	0.70 (0.22)	0.20 (0.08)	0.19 (0.10)	0.15 (0.04)
O2+HR	0.52 (0.12)	0.79 (0.08)	0.69 (0.17)	0.23 (0.04)	0.21 (0.04)	0.19 (0.05)
O2+MAP	0.71 (0.17)	0.87 (0.10)	0.82 (0.08)	0.22 (0.10)	0.17 (0.04)	0.13 (0.04)
O2+Resp	0.71 (0.19)	0.88 (0.07)	0.71 (0.41)	0.25 (0.08)	0.19 (0.06)	0.13 (0.07)
HR+MAP	0.70 (0.10)	0.84 (0.09)	0.73 (0.19)	0.22 (0.04)	0.17 (0.05)	0.14 (0.05)
HR+Resp	0.45 (0.31)	0.46 (0.39)	0.51 (0.23)	0.31 (0.16)	0.33 (0.22)	0.39 (0.33)
MAP+Resp	0.72 (0.12)	0.80 (0.11)	0.70 (0.26)	0.21 (0.07)	0.18 (0.07)	0.14 (0.08)

for each variable with the large exception of `O2`. With the exception of the 5-20 minute time frame, there is a substantial improvement in AUC ROC for `O2`. Moreover, the PE decreases across each time frame and the standard deviation also decreases. Despite largely decreasing the predictive ability of all but one physiological variables, a flexible association structure for `O2` gives the highest predictive accuracy out of the models considered so far.

Table 4.6 displays the cross-validation results for bivariate null JMs with a linear association structure. Notably, many of the weak variables in the univariate case see a large improvement in predictive ability when combined with `SBP` (`HR`, `MAP`, `Resp`). Including `SBP` with `O2` resulted in an improvement in discriminatory ability but somewhat weaker calibration performance. Arguably, the best performing model overall in Table 4.6 is the bivariate JM containing `SBP` and `HR`.

Table 4.7 presents the 5-fold cross-validation results for each combination of null JMs that have a bivariate longitudinal response, where the time-to-event component assumes a flexible association structure between the instantaneous risk and one of the two longitudinal outcomes. Comparing Table 4.7 to Table 4.6, it can be seen that only four models have a notable improvement by modelling the association structure flexibly. These models include: `SBP + O2'`, `DBP + SBP'`, `DBP + O2'`, `MAP + O2'` and `Resp + O2'`. Clearly, modelling the association structure of `O2` flexibly results in a substantial improvement in predictive performance. The model highlighted in red in Table 4.7 is deemed to be the best model so far (`SBP + O2'`), which has an exceptional discriminatory ability for

Table 4.7: *5 fold cross-validation for identifying the best pair of longitudinal responses, where the association structure between the hazard and a single longitudinal outcome is modelled flexibly. Variable names ending with an apostrophe assume a flexible association structure. The chosen model is highlighted in red.*

	Mean AUC ROC (Standard Deviation)			Mean PE (Standard Deviation)		
	5-20 mins	15-60 mins	30-75 mins	5-20 mins	15-60 mins	30-75 mins
SBP+DBP'	0.53 (0.14)	0.61 (0.19)	0.54 (0.26)	0.27 (0.07)	0.28 (0.09)	0.29 (0.08)
SBP+O2'	0.85 (0.05)	0.97 (0.03)	0.93 (0.05)	0.22 (0.05)	0.11 (0.02)	0.09 (0.03)
SBP+HR'	0.62 (0.17)	0.82 (0.12)	0.66 (0.20)	0.23 (0.07)	0.18 (0.07)	0.17 (0.07)
SBP+MAP'	0.65 (0.24)	0.73 (0.10)	0.56 (0.14)	0.25 (0.11)	0.23 (0.08)	0.25 (0.12)
SBP+Resp'	0.49 (0.15)	0.46 (0.18)	0.30 (0.14)	0.25 (0.07)	0.28 (0.05)	0.29 (0.07)
DBP+SBP'	0.56 (0.17)	0.65 (0.17)	0.61 (0.10)	0.27 (0.07)	0.26 (0.11)	0.30 (0.11)
DBP+O2'	0.80 (0.09)	0.96 (0.03)	0.93 (0.02)	0.24 (0.09)	0.15 (0.02)	0.09 (0.03)
DBP+HR'	0.34 (0.15)	0.45 (0.18)	0.39 (0.18)	0.32 (0.06)	0.35 (0.08)	0.33 (0.09)
DBP+MAP'	0.60 (0.22)	0.71 (0.11)	0.54 (0.11)	0.26 (0.09)	0.24 (0.07)	0.25 (0.10)
DBP+Resp'	0.43 (0.14)	0.47 (0.22)	0.31 (0.29)	0.26 (0.06)	0.34 (0.10)	0.34 (0.09)
O2+SBP'	0.47 (0.06)	0.62 (0.07)	0.39 (0.22)	0.28 (0.09)	0.26 (0.04)	0.29 (0.09)
O2+DBP'	0.41 (0.15)	0.60 (0.22)	0.39 (0.22)	0.31 (0.08)	0.30 (0.10)	0.36 (0.12)
O2+HR'	0.27 (0.15)	0.56 (0.22)	0.57 (0.32)	0.29 (0.08)	0.28 (0.09)	0.28 (0.12)
O2+MAP'	0.36 (0.07)	0.52 (0.11)	0.39 (0.17)	0.31 (0.09)	0.31 (0.04)	0.30 (0.07)
O2+Resp'	0.26 (0.11)	0.29 (0.11)	0.27 (0.25)	0.32 (0.04)	0.36 (0.05)	0.37 (0.10)
HR+SBP'	0.57 (0.11)	0.77 (0.11)	0.62 (0.13)	0.28 (0.07)	0.19 (0.06)	0.22 (0.03)
HR+DBP'	0.39 (0.18)	0.51 (0.09)	0.50 (0.14)	0.31 (0.08)	0.31 (0.08)	0.32 (0.09)
HR+O2'	0.32 (0.08)	0.79 (0.07)	0.75 (0.20)	0.27 (0.07)	0.22 (0.06)	0.14 (0.07)
HR+MAP'	0.54 (0.08)	0.70 (0.14)	0.51 (0.18)	0.24 (0.02)	0.23 (0.06)	0.24 (0.08)
HR+Resp'	0.29 (0.07)	0.32 (0.16)	0.14 (0.13)	0.33 (0.06)	0.42 (0.07)	0.49 (0.08)
MAP+SBP'	0.61 (0.24)	0.66 (0.15)	0.49 (0.14)	0.27 (0.11)	0.26 (0.11)	0.29 (0.14)
MAP+DBP'	0.59 (0.24)	0.68 (0.12)	0.55 (0.16)	0.27 (0.10)	0.24 (0.08)	0.26 (0.11)
MAP+O2'	0.78 (0.07)	0.94 (0.03)	0.92 (0.06)	0.23 (0.09)	0.15 (0.02)	0.10 (0.04)
MAP+HR'	0.42 (0.16)	0.62 (0.07)	0.49 (0.16)	0.26 (0.01)	0.25 (0.02)	0.25 (0.03)
MAP+Resp'	0.37 (0.22)	0.39 (0.12)	0.20 (0.13)	0.30 (0.06)	0.39 (0.05)	0.40 (0.05)
Resp+SBP'	0.45 (0.08)	0.51 (0.27)	0.28 (0.22)	0.27 (0.08)	0.33 (0.12)	0.36 (0.10)
Resp+DBP'	0.39 (0.16)	0.48 (0.25)	0.26 (0.18)	0.35 (0.10)	0.41 (0.16)	0.44 (0.17)
Resp+O2'	0.67 (0.19)	0.93 (0.09)	0.94 (0.05)	0.24 (0.07)	0.14 (0.04)	0.09 (0.04)
Resp+HR'	0.19 (0.07)	0.29 (0.23)	0.18 (0.11)	0.37 (0.04)	0.44 (0.09)	0.53 (0.05)
Resp+MAP'	0.40 (0.18)	0.43 (0.08)	0.31 (0.17)	0.33 (0.05)	0.39 (0.05)	0.42 (0.10)

predicting the donor event time. It can also be seen that for many of the models in Table 4.7, the predictive performance is substantially worse than when a linear association structure is assumed, indicating over-parametrisation.

Having selected the longitudinal responses (`SBP + O2`) and the association structure (`O2'`), we proceed to see if the predictive ability can be improved further by including baseline covariates, transformations of the longitudinal responses and alternative functional forms for the longitudinal responses. Table 4.8 displays notable models that were fitted testing various combinations of log transformations, (with 0.1 added to avoid infinity being returned where values are 0) the derivative functional form and the inclusion

Table 4.8: *5-fold cross-validation results for the final few combinations of bivariate JMs, testing for improvements in predictive accuracy by including baseline covariates, response transformations and alternative functional forms.*

	Mean AUC ROC (Standard Deviation)			Mean PE (Standard Deviation)		
	5-20 mins	15-60 mins	30-75 mins	5-20 mins	15-60 mins	30-75 mins
log(SBP+0.1)+O2'	0.86 (0.07)	0.96 (0.03)	0.91 (0.10)	0.24 (0.10)	0.12 (0.03)	0.09 (0.02)
SBP+log(O2'+1)	0.94 (0.02)	0.98 (0.02)	0.99 (0.03)	0.19 (0.05)	0.08 (0.04)	0.08 (0.05)
DerivSBP+log(O2'+0.1)	0.98 (0.02)	0.99 (0.01)	0.98 (0.03)	0.22 (0.07)	0.08 (0.04)	0.07 (0.04)
log(SBP+0.1)+log(O2'+0.1)	0.97 (0.03)	0.99 (0.01)	0.97 (0.03)	0.18 (0.07)	0.09 (0.04)	0.14 (0.18)
log(SBP+0.1)+log(O2'+0.1)+(dweight+dheight)	0.95 (0.05)	0.99 (0.01)	0.97 (0.04)	0.17 (0.07)	0.08 (0.05)	0.14 (0.18)
SBP+O2'+(dheight+dweight)	0.83 (0.06)	0.97 (0.02)	0.94 (0.05)	0.22 (0.06)	0.11 (0.02)	0.10 (0.02)
SBP+log(O2'+0.1)+(dheight+dweight)	0.92 (0.04)	0.98 (0.02)	0.98 (0.03)	0.19 (0.05)	0.09 (0.04)	0.07 (0.05)
SBP+Derivlog(O2'+0.1)+(dheight+dweight)	0.67 (0.09)	0.83 (0.11)	0.73 (0.24)	0.21 (0.05)	0.18 (0.03)	0.14 (0.06)

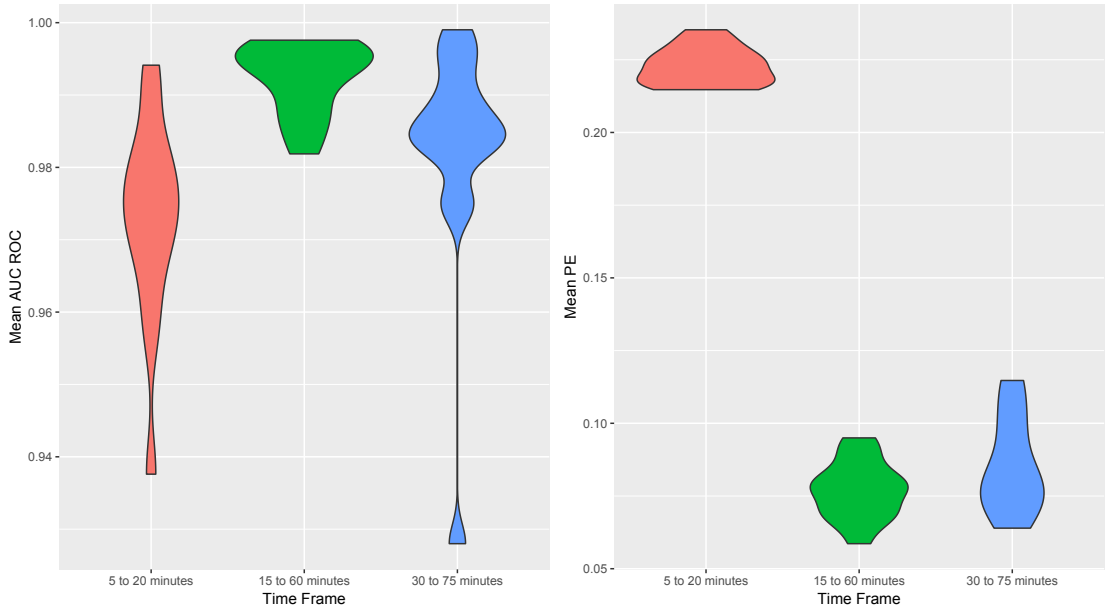


Figure 4.7: *Results for the 15 times repeated 5-fold cross-validation, presenting the AUC ROC and PE for three medically relevant time frame.*

of baseline covariates that were identified as important.

Once again a large improvement in predictive ability can be seen in Table 4.8. In particular, using the value of the slope of `SBP` and assuming a flexible association structure of `O2` with a log transformation, results in an almost perfect discriminatory ability. This indicates that for almost every pair of patients in the dataset, a higher probability of survival was allocated to the patient that survived longer than the other patient in the selected pair. The change in calibration performance is less noticeable, however a marginal improvement indicates a reduced prediction error overall. The model highlighted in red is selected as the best performing model. However, further validation is required to ensure this was not a result of a fortunate allocation of patients to the cross-validation folds.

We further validate the chosen model by performing a 15 times repeated 5-fold cross-validation (such that in each repeat, patients are randomly allocated to folds). Figure 4.7 presents violin plots displaying the distribution of the mean AUC ROC and mean PE across the 5-folds for the 15 repeats. The predictive ability is consistently high, although a relatively large variance and higher PE can be seen for the 5 to 20 minute time frame for which the predictions are made. The model validation is satisfactory.

We fit the chosen model to the full dataset using the Hamiltonian Monte Carlo algorithm implemented by the Bayesian sampling software "Stan". A warm-up period of 500 iterations is specified, to which Stan applies dual averaging (see Nesterov (2009)) to determine the optimal leapfrog step-size. This requires the target Metropolis acceptance rate parameter, which we set to 0.8 (forcing the adaption to take small step-sizes so that

reasonable acceptance rates are used). We specify 30000 iterations after the warm-up period, half of which are discarded as the burn-in period. Thinning is applied to the remaining 15000 iterations, such that one in every 50 samples are retained, to reduce autocorrelation between samples from the conditional posterior distribution. Two chains are run in parallel with different starting values to avoid convergence to isolated modes and to assist during assessment of convergence.

Figure 4.8 displays the trace and smoothed density plots for the longitudinal fixed parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. The two chains can be distinguished by the colours (red and black) and appear to be exploring the same conditional posterior distribution. The density plots appear to be approximately normally distributed for each parameter. The autocorrelation plots are omitted to avoid extensive output, however, the autocorrelation rapidly drops to zero as lag increases for each of these parameters (and for each chain). The horizontal red and black lines represent the means of each chain, which remain flat and do not diverge, indicating convergence to the stationary distribution. There were no divergent transitions and the plots are satisfactory in the sense that there is no evidence of a lack of convergence or poor mixing.

The same plots are displayed for the standard deviation of the measurement error parameters in Figure 4.9 and for the association parameters α_{11} and $\{\alpha_{2,i}\}_{i=1}^{10}$ (Figures 4.10 and 4.11). The same conclusions can be drawn from these plots. There is no evidence of a lack of convergence for any of the parameters in the chosen fitted MBJM.

We proceed by presenting the chosen model in Table 4.9, which includes the posterior mean, standard deviation, standard error, 95% credible interval and Bayesian p-values.

Table 4.9: *The chosen model (null bivariate JM with the gradient functional form of the SBP response and a flexible representation of the association parameters of log O2). The posterior means, standard deviation, standard error and credible interval are presented alongside the Bayesian p-values. The potential scale reduction factor (Gelman et al. 1992) ($\hat{R}$) is presented with the upper confidence interval limit to assess convergence.*

Parameter	PostMean	StDev	StErr	2.5%	97.5%	P	$\hat{R}$	Upper C.I ($\hat{R}$)
β_{01}	118.47	3.73	0.22	111.40	125.97	0.00	1.00	1.03
β_{11}	-15.29	9.91	0.57	-34.51	3.70	0.14	1.00	1.00
β_{21}	4.15	10.14	0.59	-15.72	24.50	0.69	1.00	1.00
σ_1	20.05	0.39	0.02	19.31	20.85	0.00	1.00	1.00
β_{02}	4.42	0.04	0.00	4.35	4.49	0.00	1.01	1.05
β_{12}	-0.09	1.01	0.07	-2.21	1.69	0.92	1.00	1.00
β_{22}	-0.31	0.33	0.03	-0.87	0.36	0.32	1.00	1.00
σ_2	0.56	0.01	0.00	0.54	0.58	0.00	1.00	1.00
$\alpha_{1,1}$	-0.15	0.05	0.00	-0.25	-0.06	0.01	1.00	1.00
$\alpha_{1,2}$	0.36	0.24	0.02	-0.13	0.79	0.15	1.00	1.00
$\alpha_{2,2}$	-0.34	0.08	0.00	-0.50	-0.16	0.00	1.00	1.00
$\alpha_{3,2}$	-1.54	0.19	0.01	-1.96	-1.15	0.00	1.00	1.00
$\alpha_{4,2}$	-2.49	0.38	0.01	-3.45	-1.83	0.00	1.00	1.00
$\alpha_{5,2}$	-3.06	0.53	0.02	-4.25	-2.10	0.00	1.00	1.00
$\alpha_{6,2}$	-3.23	0.59	0.02	-4.37	-1.98	0.00	1.00	1.00
$\alpha_{7,2}$	-3.39	0.65	0.03	-4.67	-2.15	0.00	1.00	1.00
$\alpha_{8,2}$	-3.42	0.71	0.03	-4.93	-2.08	0.00	1.00	1.00
$\alpha_{9,2}$	-3.35	0.76	0.04	-4.92	-1.84	0.00	1.00	1.00
$\alpha_{10,2}$	-3.16	0.85	0.05	-4.86	-1.54	0.00	1.00	1.00

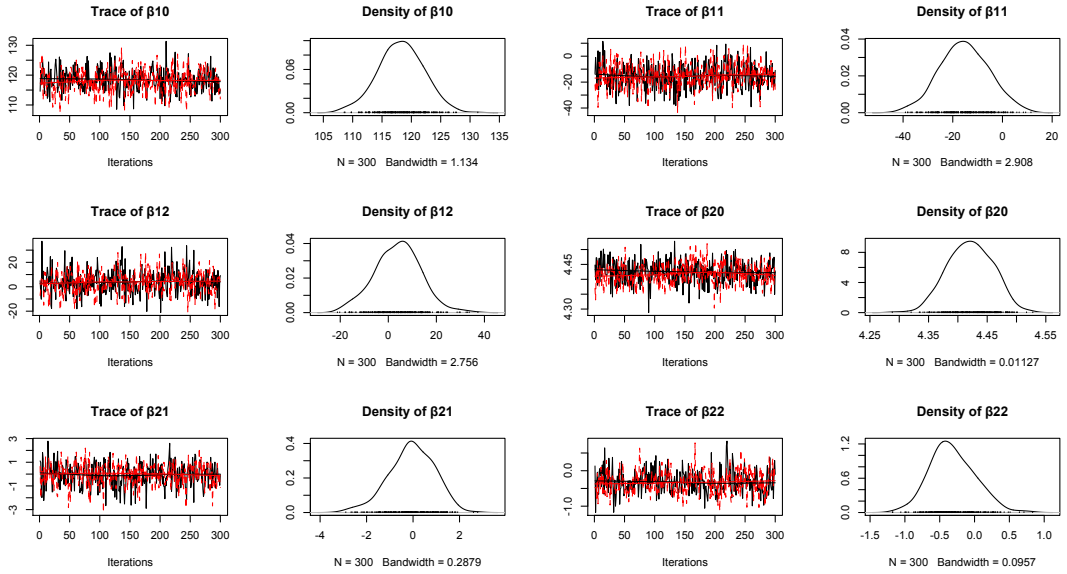


Figure 4.8: *Hamiltonian Monte Carlo diagnostic trace (for two chains - coloured in red and black) and smoothed density plots for both the β_1 and β_2 parameters.*

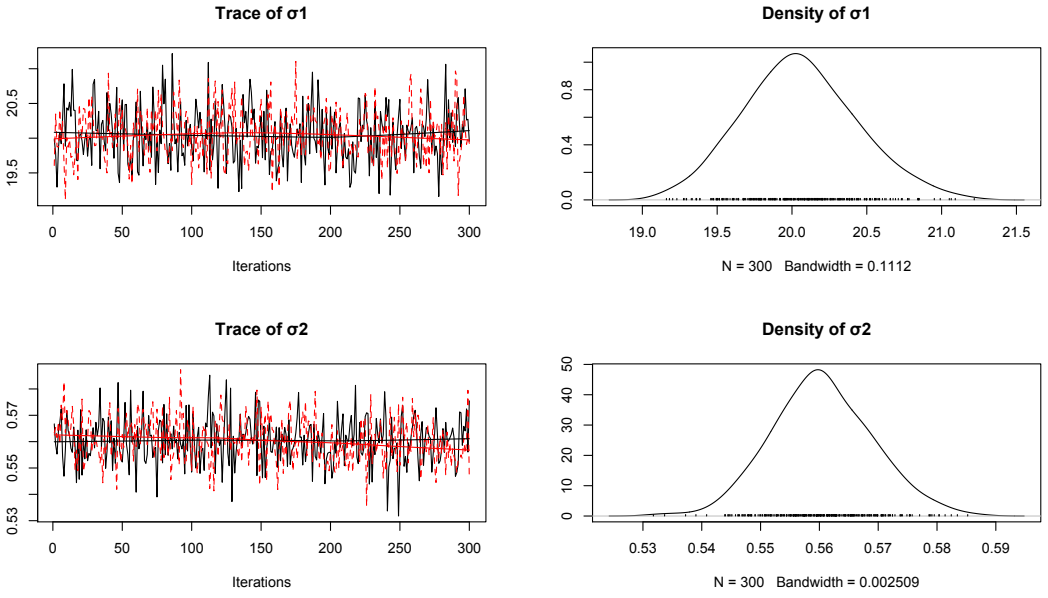


Figure 4.9: *Hamiltonian Monte Carlo diagnostic trace (for two chains - coloured in red and black) and smoothed density plots for the σ_1 and σ_2 parameters.*

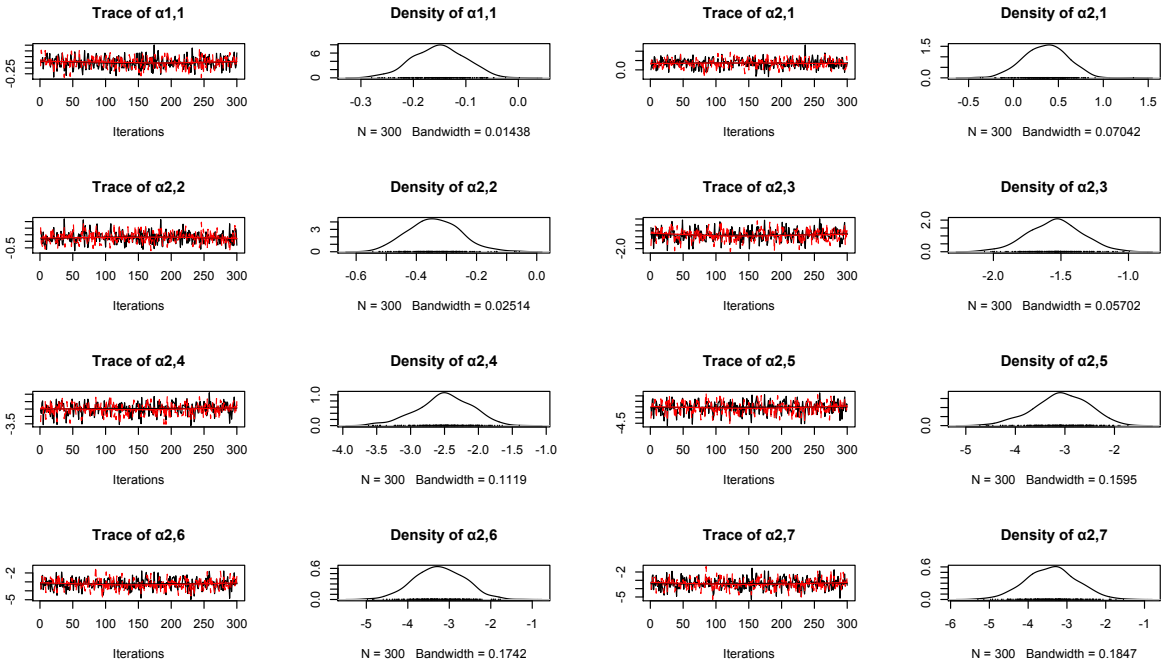


Figure 4.10: *Hamiltonian Monte Carlo diagnostic trace (for two chains - coloured in red and black) and smoothed density plots for the parameters* $\alpha_{1,1}$ *and* $\alpha_{2,i}$ *where* $i = 1, \ldots, 7$.

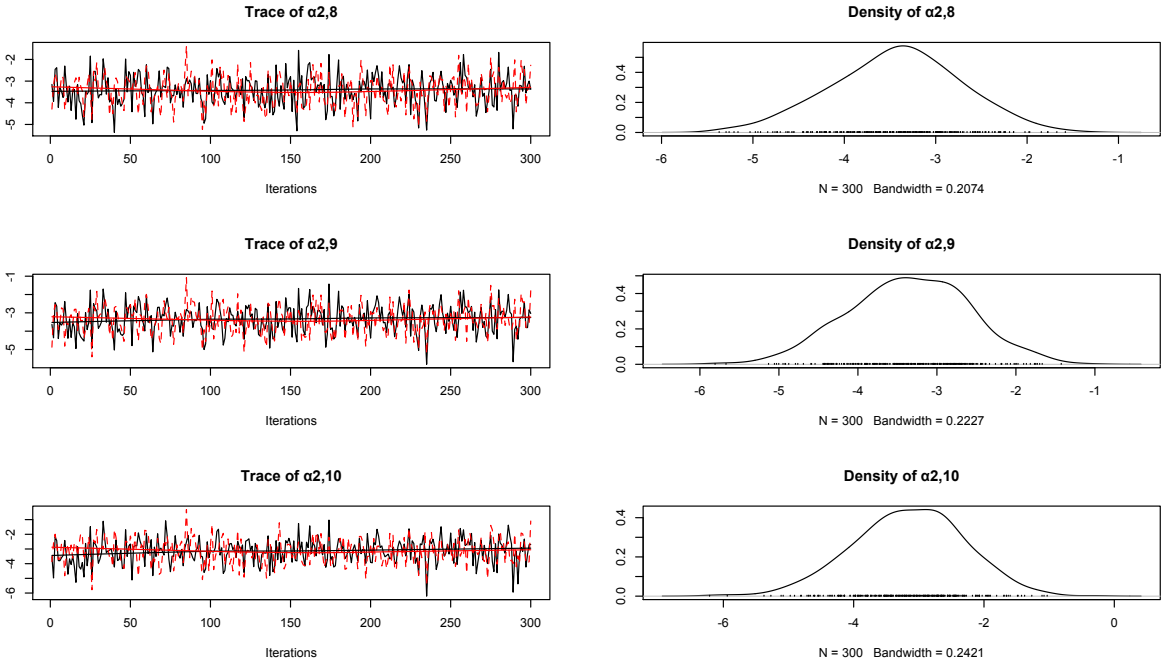


Figure 4.11: *Hamiltonian Monte Carlo diagnostic trace and density plots for the coefficients* $\alpha_{2,i}$ *where* $i = 8, \ldots, 10$ *coefficients.*

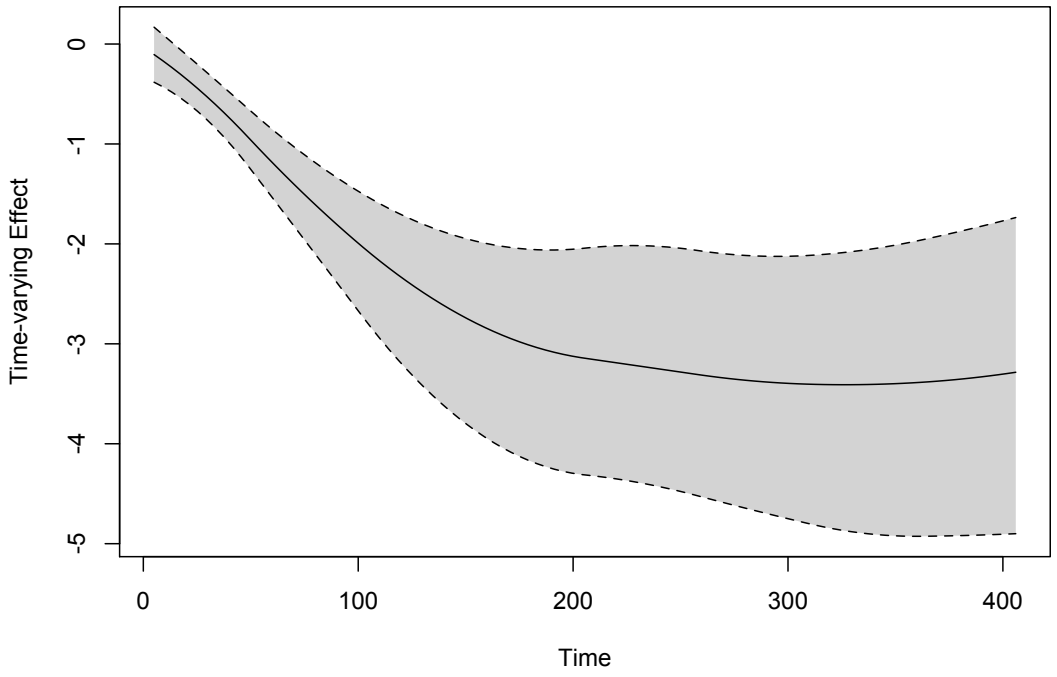


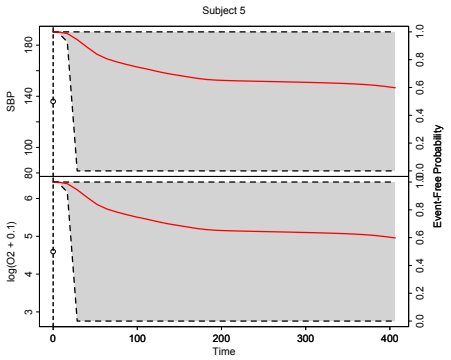
Figure 4.12: *The Time-varying effect of* $\log(O2 + 0.1)'$ *presented with the 95% credible interval (shaded region).*

The potential scale reduction factor in Table 4.9 is close enough to one such that we do not suspect convergence difficulties (consistent with our conclusions from the trace, density and autocorrelation plots).

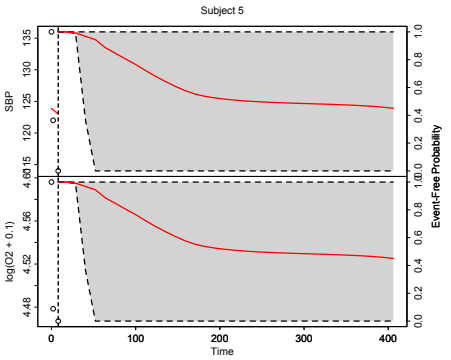
A direct interpretation of the chosen model is complicated by having specified a flexible MBJM (in both of the longitudinal and survival components). As our aim in this work is to use this model for prediction, a lack of interpretability is not problematic. However we are able to infer from Table 4.9 that the association between both of the longitudinal responses and the survival response are statistically significant. The functional form of the association structure $\lambda_2(t)$ can be seen in Figure 4.12. Clearly the link between the longitudinal and survival response is far from constant over time, which explains the large increase in predictive ability when modelling this term flexibly.

We now proceed with a subject-specific dynamic prediction to illustrate how this model can be applied in practice for a new subject who has been withdrawn from their life-support machine. Figure 4.13 and 4.14 display the dynamic predictions for "subject 5", who is chosen because they have many repeated measures (33), which is useful for the purpose of illustration.

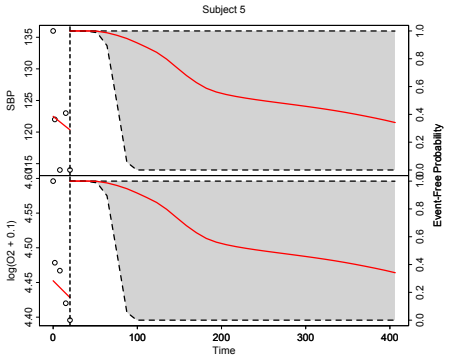
The predictions are made using all information available for the subjects remaining at time t (in minutes), represented by the vertical dashed line. To the left of this



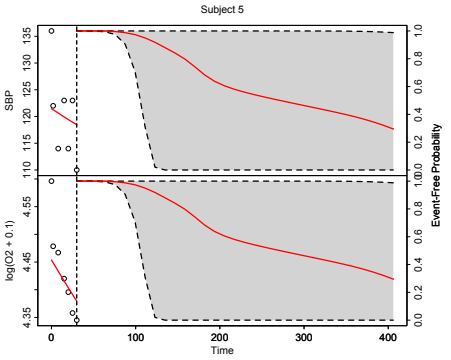
(a) After 1 repeated measure.



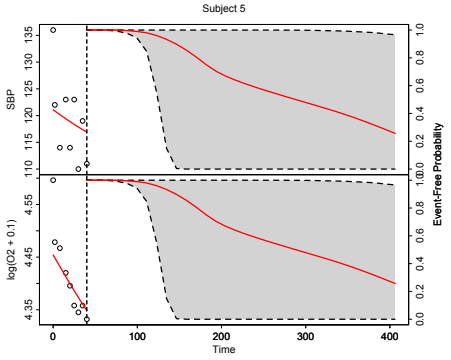
(b) After 3 repeated measures.



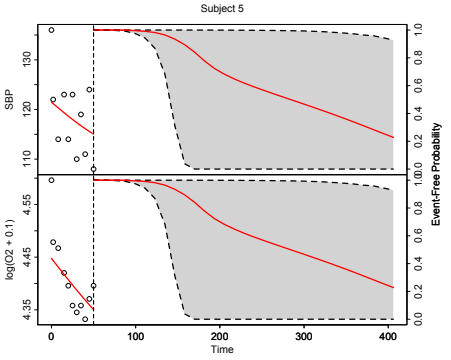
(c) After 5 repeated measures.



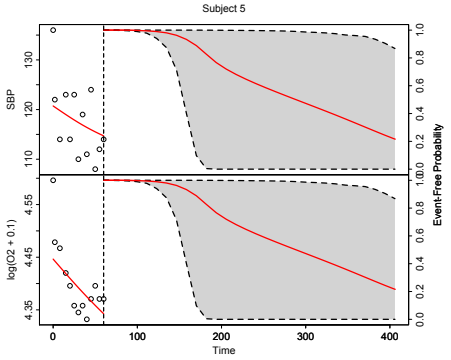
(d) After 7 repeated measures.



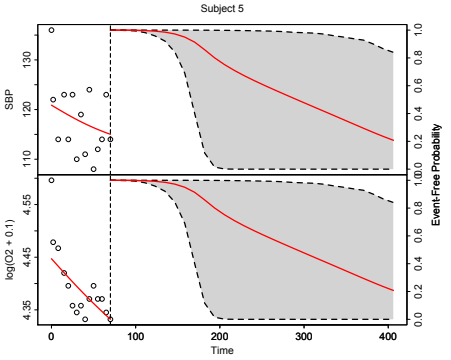
(e) After 9 repeated measures.



(f) After 11 repeated measures.

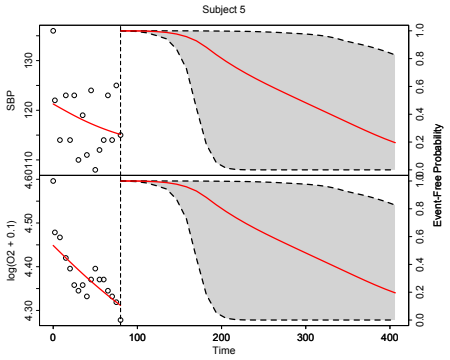


(g) After 13 repeated measures.

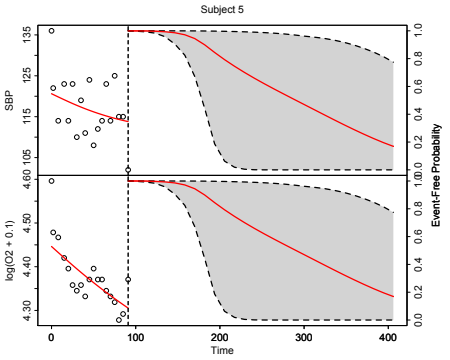


(h) After 15 repeated measures.

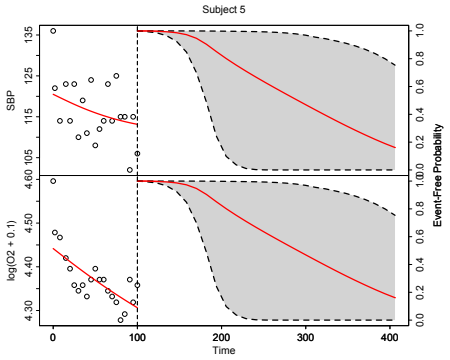
Figure 4.13: *Dynamic prediction for Subject 5, using information for all patients remaining at the time corresponding to the vertical dashed line. The subject-specific longitudinal fitted line is given in red (before the dashed line) with the raw data points as black circles. To the right of the dashed line is the estimated survival probability with 95% confidence interval given by the shaded region. Time progresses from Sub-figure (a) to Sub-figure (h).*



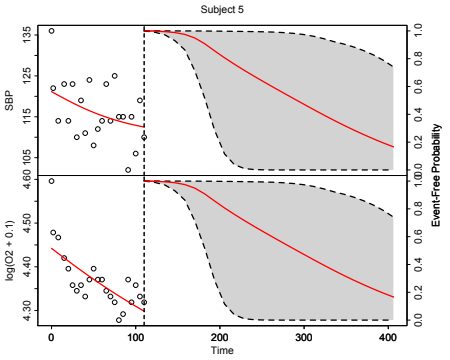
(a) After 17 repeated measures.



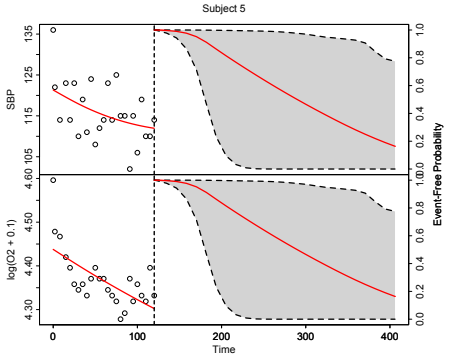
(b) After 19 repeated measures.



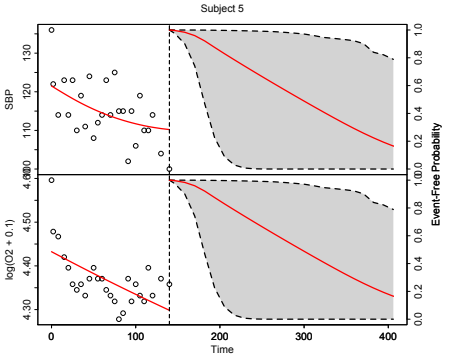
(c) After 21 repeated measures.



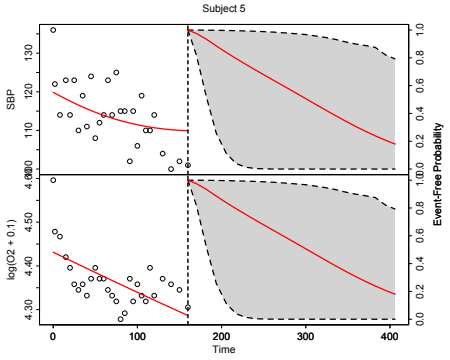
(d) After 23 repeated measures.



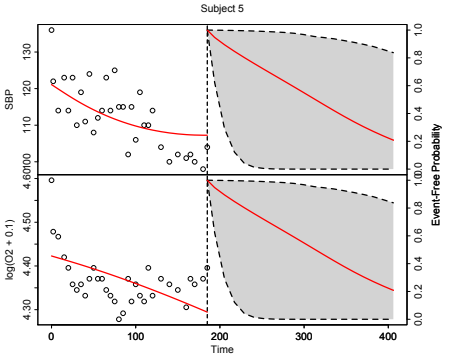
(e) After 25 repeated measures.



(f) After 27 repeated measures.



(g) After 29 repeated measures.



(h) After 33 (final number of) repeated measures.

Figure 4.14: *The remainder of dynamic prediction points for Subject 5, after Subfigure (h) in Figure 4.13.*

line is the longitudinal process (for each longitudinal outcome). The circles represent observed values and the red line represents the fitted function. To the right of the vertical dashed line is the survival process, where the red line represents the estimated survival probability for each time point until the maximum observed event time. The shaded region represents a 95% point-wise confidence interval for this predicted values, which is derived from the percentiles of the Monte Carlo samples.

It can be seen that when only a single measurement exists (taken at the time of withdrawal) the confidence interval is very wide for most of the time values. For this reason, one may question whether this method provides insight at all at this stage, however, it is important to note the skew in these data. At the time the confidence interval becomes wide, approximately 50% of the patients in the data became deceased (see Figure A.2, at approximately 20 minutes). This means that with a single observation of `SBP` and `O2`, we are able to determine with high confidence whether the patient will still be alive for a meaningful time horizon: until approximately 20 minutes. Of course, as this process is notoriously difficult to predict, the confidence interval becomes very wide at a given future time point. However before this time occurs, important and potentially life-saving insight can be drawn.

As more data points are recorded (moving from Subfigures (a) - (h) in Figure 4.13) it can be seen that estimates of survival probability are drawn with higher confidence for a longer period of time for this subject. Again this happens until the confidence interval becomes exceedingly wide (at which point, we are unsure when the event time would be). Note that simply being able to tell a clinician that we are confident that the patient will survive for another 5 minutes from a specific time, could be the difference between a transplant proceeding in practice by allowing the removal team to effectively allocate limited resources.

Notice that in Figure 4.14 as more data points are recorded, the distance from the current time point (vertical dashed line) and the lower limit of the confidence interval approach each other, suggesting the distance into the future we are able to reliably predict is shortening as time increases. This is because such few patients remain in the data around this time (approximately 200 minutes), making the data less credible. As it is uncommon for patients to enter this region of the data, this is unlikely to be a problem in practice, and could be improved by updating the model as more data becomes available in the future.

To demonstrate how this approach could be used in practice to inform clinicians, take Figure 4.13 subfigure (e) for example. Suppose we observe these trajectories of `SBP` and `O2`, and we know the patient is still alive at this point in time (represented by the vertical dashed line). We would be able to say that we are confident that the patient will still be alive for approximately another 40 minutes, thereafter we do not know when the event is likely to occur, though we can make an estimate which corresponds to the thick red

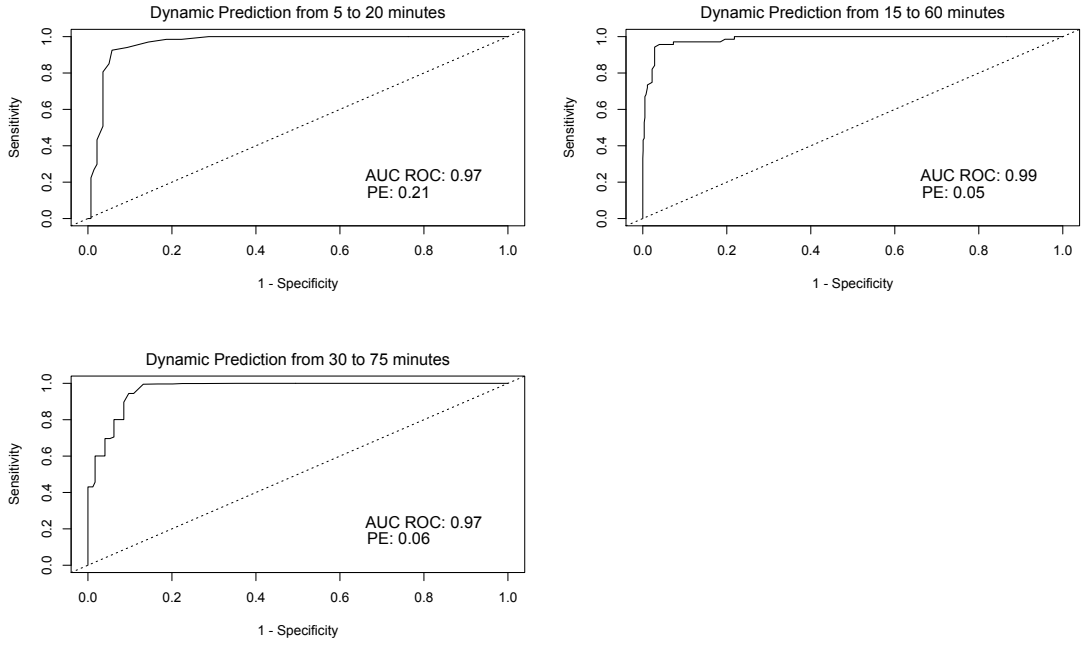


Figure 4.15: *The Time-varying effect of* $\log(O2 + 0.1)'$ *presented with the 95% credible interval (shaded region).*

line. This would allow clinicians to provide notice to the removal team to arrive before the 40 minutes have passed. Note that these estimates are based on characteristics of multiple physiological variable trajectories of the organ donors and conditioning on the fact that the patient is still alive at the current time point.

Finally, we present in Figure 4.15 the ROC curves for the same three time frames that were used for model selection (but this time the full dataset is used instead of test data). As expected the AUC ROC remains exceptionally high and prediction error is low.

4.5 Discussion

In this work we have applied sophisticated statistical methods to a unique and rich dataset, in order to solve a real practical problem. This involved dynamically predicting the probability of a controlled DCD organ donor experiencing a cardiac arrest between specific medically relevant time frames. Notably, we employed a recently proposed extension to the JM, which involved modelling the association between the longitudinal and survival processes flexibly (Andrinopoulou et al. 2018). By applying this extended JM to the longitudinal covariate O2, we obtained a substantial improvement in predictive ability. This implies that as time progresses throughout the treatment withdrawal phase the association between the current value of O2 and the risk of death varies.

There was a noticeable improvement in predictive ability when the "current slope" functional form of the longitudinal covariate SBP was used in the relative risk model, rather than the current value of SBP. This implies that the rate of change of SBP is a more important risk factor than its current value.

We were able to validate (based on the novel dataset) that the chosen model had an exceptional predictive ability, achieving to the best of our knowledge what is by far the highest predictive ability achieved. Moreover, the models in the literature that obtained a high predictive ability were based on many clinical variables that were not available in this dataset. This is one advantage of the chosen model, that although it is more computationally demanding than standard methods, only SBP and O2 were required to make such accurate predictions. This is in part attributed to the model conditioning on the fact that the patient is still alive at the time a prediction is made. We acknowledge that in practice it would be preferred to make these predictions before the treatment withdrawal period begins. However, we believe that this approach still has a wide scope for improving medical practice.

Various implicit assumptions were made in the derivation of the MBJM likelihood function. First, the model assumes that the censoring mechanism depends on the observed data but not the unobserved data, which is plausible in our application. The JM makes implicit assumptions with regards to the full vector of longitudinal recordings including the planned measurements that were missed. As the visiting process in our data is stochastic, our case is complicated. However, research shows (Rizopoulos 2012*b*) that assuming that the visiting process is not related to the event times then inference remains valid when the visiting mechanism is stochastic. Although this assumption is not testable from the data we believe it is plausible, as the British Transplantation Society guidelines state that the protocol is for nurses to take the readings once every few minutes. As no attempt is being made to resuscitate these patients, there is no incentive for nurses to take readings more regularly as the patients approach death. When expert

knowledge of the problem at hand suggests that this assumption is not valid, the visiting process must be specified explicitly and incorporated into the likelihood function (Lipsitz et al. 2002).

We also assumed that random effects follow a multivariate normal distribution. As the random effects are unobserved this is also an assumption that is not testable from the data. However, various studies have shown that as n_i increases, misspecifying the normality of the latent random effects has very little effect on the parameter estimates and standard errors (Rizopoulos et al. 2008, Huang et al. 2009). Moreover, the exceptional predictive ability achieved in our application suggests that either these assumptions hold or our chosen model in this application is robust when the aim is to perform prediction.

Chapter 5

Impact of the Treatment Withdrawal Period on Kidney Transplant Outcomes

Summary

In this chapter we utilise the novel dataset to investigate the ability of various characteristics of multiple DCD donor physiological variables in the treatment withdrawal to death period for predicting short-term recipient transplant outcome, delayed graft function (DGF). We compare the performance of various methods for summarising physiological variable characteristics that are subsequently used as predictor variables in a classification model (random intercept logistic regression). A simulation study is conducted to determine whether summaries (intercept, slope and AUC) estimated from a joint model (JM) are preferable to the observed summaries or those estimated from a linear mixed effects model (LMEM). The method deemed best is used in our analysis in the first part of the two stage approach for predicting recipient transplant outcome. The chosen predictive model is used to improve our understanding of the impact of physiological variable characteristics in the withdrawal phase on the chances of recipient's graft immediately functioning.

5.1 Introduction

The treatment withdrawal to death phase is a major determinant as to whether organ procurement proceeds. As it has been discussed throughout this thesis, the decision to withdraw a donor from transplantation can depend on both medical and logistical reasons that relate to the withdrawal phase. However, research is limited with regards to which characteristics (such as the duration, or specifically characteristics relating to physiological variables) of the withdrawal period impacts the graft quality and how, which provides scope for this research.

A common scenario that leads to the decision for a donor to be withdrawn from transplantation is when the treatment withdrawal to death phase is prolonged. Organs become deprived of oxygenation as a result of the circulation cessation, causing ischaemic injury. This is referred to as warm ischaemic exposure and is known to cause cellular decay in the organs, which is believed to increase the risk of complications including DGF.

To safeguard patients from ischaemic injury, protocols specify a limit for the duration of the treatment withdrawal to death period. In the US and the Netherlands it is one and two hours respectively. The UK protocol acknowledges the limited evidence (Reid et al. 2011, Scalea et al. 2017) indicating that the duration itself does not compromise graft quality, but more so the evolution of the physiological profiles over time. The UK allows a maximum of three hours before SBP drops below 50mmHg and also a maximum of two hours after (Peters-Sengers et al. 2018).

Bradley et al. (2013) also conjecture that the graft quality of organs from donors whose physiological profiles have certain characteristics is not compromised when the withdrawal period is prolonged. The limited evidence to support this hypothesis is attributed to restrictive protocols, meaning that the extremes of warm ischaemia have not previously been able to be explored. In this work, we analyse a novel dataset that allows these extremes to be explored (maximum death time of 406 minutes), which provides scope to investigate this conjecture.

Various attempts have been made to investigate how characteristics of the treatment withdrawal to death period are associated with recipient outcomes. However, due to the complex nature of temporal data these studies have been limited to crude statistical measures such as the observed AUC of the longitudinal trajectories.

Ho et al. (2008) found through an exploratory analysis that characteristics of the withdrawal period could be more predictive of DGF than the duration alone. Consistent findings were achieved by Scalea et al. (2017), who found using logistic generalised estimating equations that time to death was not predictive of DGF.

Allen et al. (2016) analysed various characteristics of the treatment withdrawal period (median, slope, AUC and threshold values for SBP, shock index and oxygen saturation). Only the AUC SBP categorised at the median was retained as a significant covariate in the multi-variable model (odds ratio 1.42, 95% CI 1.06 to 1.91). This implies that those with an AUC SBP above the median were more likely to experience DGF.

Peters-Sengers et al. (2018) used logistic regression with restricted cubic splines to see if the time it took physiological variables (SBP and oxygen saturation) to drop below certain threshold values (80mmHg and 60%) was predictive of DGF. They also investigated the effect of the duration of the treatment withdrawal period. They found that the duration was significantly associated with DGF as well as the time it takes SBP to drop below 80mmHg.

This clinical problem poses several interesting statistical challenges. A LMEM can be used to model the longitudinal trajectories that are recorded over time in possibly irregular and unequally spaced intervals. However, we are still faced with two major complications. First, the longitudinal recordings are contaminated with measurement error. Moreover, inference based on the LMEM is only valid under MAR (see Section 2.2.4 for a discussion on missing data in longitudinal studies), and the missingness mechanism is not testable from the data. If the true mechanism is MNAR the parameter estimates from the LMEM are likely to be biased.

We investigate a joint modelling approach to summarise characteristics of the physiological profiles, that takes full advantage of both the longitudinal and survival information of the donor. In particular, the full trajectories are modelled with a LMEM and the time-to-event process of the donor is modelled with a proportional hazards model. By performing parameter estimation based on the full joint likelihood of the two processes, a model is explicitly specified to account for the missing data implicit outcome, which may improve our ability to predict DGF. The estimated characteristics of the physiological profiles (the intercept, slope and AUC) are then used in the second part of this proposed two-stage approach to predict whether a recipient experiences DGF. To this author's knowledge, two-stage approaches have been often employed where a LMEM is fitted in the first stage, but no study has used a JM as the first stage to extract relevant information from the longitudinal process, which is then used as a covariate in a classification model (stage two).

The aim of this chapter is to improve our understanding of how withdrawal phase characteristics impact transplant outcome, and to determine the optimum combination of factors that are discovered to be important in order for a transplant to be successful.

This chapter is structured as follows. In Section 5.2 we provide a description of the data that is analysed and describe the methods employed to analyse it. In Section 5.3 a series of simulations are carried out to investigate the performance of the proposed two-stage approach. In Section 5.4.1 an exploratory data analysis is performed. This is followed

by the main analysis in Section 5.4.2, where a predictive model is derived to improve our understanding of the impact of the treatment withdrawal phase characteristics on transplant outcome. This chapter concludes with a discussion in Section 5.5.

5.2 Methods

5.2.1 Description of Data

In this chapter we re-analyse a subset of the dataset that was analysed in Chapter 4 (analysed in this chapter for a different purpose), that contains DCD donor baseline covariates and physiological variables repeatedly measured throughout the treatment withdrawal to death phase (SBP, DBP, O2, MAP, HR, Resp Rate). The difference in this chapter is that these data are merged with a dataset containing kidney recipient baseline covariates and the binary outcome DGF, indicating whether the recipient had to return to dialysis within one week of transplant.

Once the two datasets are merged, 130 DCD kidney donors remain that have an observed DGF outcome (many did not proceed to transplant or had a missing response). 85 (65%) of these donors donated both kidneys and 45 donated a single one, resulting in a total of 215 observed recipient outcomes. Out of the 45 that donated a single kidney, 19 experienced DGF (42%) and 26 had an immediately functioning graft. Out of the 85 that donated both kidneys, 47 (55%) donors resulted in two immediately functioning grafts, 25 (29%) resulted in a single functioning graft and 13 (15%) resulted in two delayed graft functions.

Various other predictor variables are available (recipient transplant unit, cold ischaemic time, sex, age, blood group, weight, height, ethnicity, type of primary renal disease, dialysis status at time of transplant). The same donor baseline variables are available that were described in Chapter 4 (age, sex, blood group, cause of death, ethnicity, weight, height).

5.2.2 Statistical Methods

We investigate a two-stage approach that can be used to determine the impact of donor treatment withdrawal phase characteristics on the probability of the recipient experiencing an immediate graft function. In particular, we use a JM for longitudinal and time-to-event data to extract important information from the longitudinal trajectories of the physiological variables. As parameter estimation in the JM is performed using the joint likelihood of the longitudinal and survival processes of the donor, we expect that the information is extracted reducing bias resulting from both measurement error and missingness as a result of drop-out.

The new covariate derived from the first stage is then used in a RILR model to predict the chances of a successful transplant, while adjusting for confounding variables. Using a multilevel modelling approach at stage two allows us to account for correlation within donors (as many donate both kidneys) and also transplant unit. This approach therefore acknowledges the hierarchical structure of the data.

More formally, suppose we observe a physiological variable at t, which is denoted by $y_i(t)$ (where donor $i = 1, \ldots, n$). In some cases it is convenient to use the alternative notation of y_{ij}, where $j = 1, \ldots, n_i$ corresponds to the jth longitudinal measurement (assuming discrete time). A crude summary of the longitudinal profile for donor i would be, for instance, $u_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, which could be used in another regression model (stage two) to see how the observed mean profile is associated with transplant outcome.

The crude approach outlined above results in biased (typically attenuated) and inefficient parameter estimates in the second modelling stage (see Section 4.2 for the literature), as a result of noise being included as part of the new covariate that we denote by v. This can be seen in Equation 5.1, which highlights that the observed value is the sum of the true physiological variable measurement $m_i(t)$ and measurement error $\epsilon_i(t)$.

The longitudinal response can be modelled explicitly with a LMEM (Equation 5.2), where the (possibly time-dependent) fixed effects design vector is denoted by $\boldsymbol{x}_i(t)$ and the corresponding vector of coefficients is given by $\boldsymbol{\beta}$. The random effects and their corresponding design vector are given by $\boldsymbol{b}_i$ and $\boldsymbol{z}_i(t)$ respectively. The measurement error term $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ and random effects $\boldsymbol{b}_i \sim \mathcal{N}_q(0, \boldsymbol{D})$ (where q represents the dimension of the random effects).

$$y_i(t) = m_i(t) + \epsilon_i(t) \tag{5.1}$$

$$y_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \epsilon_i(t) \tag{5.2}$$

We postulate the following RILR model (Equation 5.3), which assumes that the odds of a successful transplant depend on the vector of covariates $\boldsymbol{v}_{li}$ (where l indexes one of the two possible recipients that correspond to the ith donor), and the fixed effect regression coefficients are denoted by $\boldsymbol{\psi}$. u_i is the random intercept term, allowing each donor to have their own subject-specific baseline odds of success. This term allows us to incorporate the within donor correlation into the regression model.

$$\text{logit}(\mathbb{P}(B_{li} = 1)) = \boldsymbol{\psi}^\top \boldsymbol{v}_{li} + u_i, \text{ where } u_i \sim \mathcal{N}(0, \sigma_u^2), \tag{5.3}$$

where the binary indicator variable B denotes a successful transplant when $B = 1$ or a DGF otherwise. The vector of covariates $\boldsymbol{v}_{li}$ include a variable that extracts information from the longitudinal process. We denote this variable by $v_{long,li} = g(\mathcal{M}_i(T_i))$, where $\mathcal{M}_i(T_i)$ is the full true (and unobserved) history of the longitudinal measurements for

the ith subject up to the event time T_i. The choice of the function g determines which summary of the longitudinal profile is to be used as a surrogate, $g : \mathbb{R}^{n_i} \mapsto \mathbb{R}$. We hypothesise that the AUC is the best surrogate, that is: $g(\mathcal{M}_i(T_i)) = \int_0^{T_i} m_i(s)ds$. This corresponds to Equation 5.4.

$$v_{long,i} = \int_0^{T_i} \boldsymbol{x}_i^\top(s)\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i^\top(s)\hat{\boldsymbol{b}}_i \; ds \tag{5.4}$$

We expect that using the estimated value $\hat{m}_i(t)$ from the LMEM reduces bias compared to using the observed data $y_i(t)$. However, as $m_i(t)$ constitutes an endogenous time-dependent covariate (Kalbfleisch & Prentice 2002) (because the subject must be alive for the longitudinal measurement to be recorded at time t) we suspect that bias occurs as a result of drop-out. We aim to reduce this bias by performing estimation using the joint likelihood of the longitudinal and survival processes of the donor (by employing a JM approach). We expect that the estimated trajectories from the JM will reduce bias in estimating the association between the longitudinal trajectory and the recipient transplant outcome, compared to using the trajectory estimated from a LMEM.

5.3 A Simulation Study

In this section a simulation study is carried out to investigate whether the proposed two-stage approach is able to contribute any benefit compared to the standard two-stage approach in this chapter's analysis, despite being much more computationally demanding.

The mean squared error (MSE) is calculated for each estimator under three competing approaches (AUC of observed trajectory, AUC of trajectory estimated with the LMEM, AUC of the trajectory estimated with the JM). In every case the data is generated based on parameters estimated when fitting the JM to the `aids` dataset (the standard dataset for benchmarking joint modelling methodology). This dataset was chosen rather than the novel dataset for simplicity (more balanced and equally spaced data) and thus less likely to suffer from complications as a result of numerical instability.

In this experiment both the measurement error dispersion parameter σ and the association parameter α (representing the association between the longitudinal and survival process) are factors to be varied, to see if the best approach depends on the size of these parameters.

To simplify this experiment, a logistic regression is postulated in the second stage rather than a RILR, because the hierarchical structure at the recipient level is not of primary interest in this simulation study.

The binary response is simulated from a Bernoulli distribution with probability equal to the inverse logistic transformation of the linear predictor η_i, as given by Equation 5.5.

$$B_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\eta_i)}\right) \tag{5.5}$$

The linear predictor (Equation 5.6) includes the parameters ψ_0 and ψ_1, which correspond to the intercept and AUC of the longitudinal trajectory respectively.

$$\eta_i = \text{logit}(\mathbb{P}(B_i = 1)) = \psi_0 + \psi_1 v_{\text{AUC},i} \tag{5.6}$$

The AUC term v_{AUC} is simulated according to Equation 5.7, which requires the longitudinal response to be generated first (before the integral can be evaluated). The longitudinal response is created such that it is conditionally independent of the survival outcome, given the random effects (as assumed by the JM). For this reason, we hypothesise that the parameters that determine the longitudinal trajectory ($\boldsymbol{\beta}$, σ and $\boldsymbol{D}$) must be estimated simultaneously with the survival process parameters ($\alpha, \boldsymbol{\gamma}$), in order to obtain unbiased parameter estimates. The notation $\boldsymbol{\beta}_{JM}$ is used to denote that the parameters depend on the joint likelihood of the longitudinal and survival processes (and therefore all other parameters in the model are estimated using the joint likelihood).

$$v_{\text{AUC},i} = \int_0^{T_i} m_i(s)ds = \int_0^{T_i} \boldsymbol{x}_i^\top(s)\boldsymbol{\beta}_{\text{JM}} + \boldsymbol{z}_i^\top(s)\boldsymbol{b}_i \; ds \tag{5.7}$$

Although the data are simulated according to Equations 5.5 and 5.7, we fit three different models, including the observed AUC (Scenario 1), the AUC of the estimated trajectory from a LMEM (Scenario 2) and the gold standard that the data is generated from (Scenario 3). By doing so we aim to quantify the effect on performance that occurs as a result of measurement error and drop-out. In particular, the parameter ψ_1 is of primary interest. Note that $\hat{\beta}_{\text{LMEM}}$ denotes that the parameters are estimated using only the longitudinal process likelihood (Equation 4.13) (ignoring the survival process).

Scenario 1:

$$v_{\text{AUC},i} = \int_0^{T_i} y_i(s) \; ds \tag{5.8}$$

Scenario 2:

$$v_{\text{AUC},i} = \int_0^{T_i} \boldsymbol{x}_i^\top(s)\hat{\boldsymbol{\beta}}_{\text{LMEM}} + \boldsymbol{z}_i^\top(s)\hat{\boldsymbol{b}}_i \; ds \tag{5.9}$$

Scenario 3:

$$v_{\text{AUC},i} = \int_0^{T_i} \boldsymbol{x}_i^\top(s)\hat{\boldsymbol{\beta}}_{\text{JM}} + \boldsymbol{z}_i^\top(s)\hat{\boldsymbol{b}}_i \; ds \tag{5.10}$$

Under each scenario various levels of α and σ are tested. The parameters estimated by fitting a JM to the `aids` dataset were -0.28 and 1.74 for α and σ respectively. We also forced α to take on the values of -0.05 and -0.80 and σ to take on a value of 10.00 for a basis of comparison with more extreme values. This simulation study is carried out by running 1000 simulations per combination of factors under investigation.

5.3.1 Results

Tables 5.1 to 5.6 display the results from this simulation study. The true parameters displayed in Table 5.1 correspond to those estimated when fitting the JM to the `aids` dataset. Clearly the three approaches perform equivalently in this scenario, which could be due to a relatively low amount of measurement error. Note that only the ψ_0 and ψ_{AUC} can be estimated from the observed approach, and the JM is the only approach that can be used to estimate α.

In Table 5.2, α is set to have a low magnitude. This appears to have a negligible impact on the MSE, with a very small increase in bias of 0.01 across each of the three approaches.

Table 5.3 displays the results for a large relative increase in measurement error. Although the observed trajectory approach has the largest magnitude of bias, the MSE for the parameter of primary interest remains equivalent across each of the three approaches suggesting that the proposed approach would not be beneficial this analysis. It can be seen that the LMEM approach has the lowest magnitude of bias in the parameter of interest. It can be seen that although the JM has a small improvement in terms of MSE for the slope parameter β_2 (as expected), the MSE indicates a worse performance for the JM compared to the LMEM in terms of the intercept β_0 and in particular σ. This is likely to be a result convergence issues in many of the JM simulation runs, causing instability in this estimate.

Table 5.4 indicates little sensitivity to a different (less extreme) value of α. With the exception of the parameter σ, the JM and LMEM have identical MSE values, yet the JM again shows a reduction in bias for the parameter β_2. Tables 5.5 and 5.6 display the results for a more extreme value of α, for each of the two levels of σ. Once again the MSE values for the parameter of primary interest are identical across the three approaches. The intercept and treatment effect parameter (β_0 and β_1) suffer from substantially larger values of MSE for the JM approach, which is mainly attributed to a larger standard error.

Having simplified this study by using the standard joint modelling methodological development dataset (the AIDS dataset) we expect to have removed various possible complications that could arise, for example, the presence of largely unbalanced and unequally spaced data. We also studied the impact of the amount of measurement error and the effect size of the parameter α. It was found that the proposed approach did not benefit this analysis, which we expect is due to the detriment of numerical instability, as a result

of having many more parameters to estimate and integrals to approximate numerically. Based on these results we would not recommend employing the examined approach, but would encourage future work to validate our findings. It is possible that this approach under certain conditions would prove to be superior to the two-stage approach, which may relate to the effect size of the other parameters that were not varied in this study.

Table 5.1: *Simulation results displaying the mean estimates across 1000 simulations, under the three scenarios. The true parameter values for σ and α are set to 1.74 and -0.28 respectively.*

		Observed		LMEM		JM	
Parameter	True	Estimated	MSE	Estimated	MSE	Estimated	MSE
σ	1.74			1.74	< 0.01	1.74	< 0.01
β_0	6.96			6.92	0.05	6.97	0.05
β_1	7.47			7.51	0.05	7.55	0.06
β_2	-0.18			-0.17	0.00	-0.19	< 0.01
α	-0.28					-0.28	< 0.01
ψ_0	0.50	0.50	< 0.01	0.50	< 0.01	0.50	< 0.01
ψ_{AUC}	0.30	0.30	0.01	0.30	0.01	0.30	0.01

Table 5.2: *Simulation results displaying the mean estimates across 1000 simulations, under the three scenarios. The true parameter values for σ and α are set to 1.74 and -0.05 respectively.*

		Observed		LMEM		JM	
Parameter	True	Estimated	MSE	Estimated	MSE	Estimated	MSE
σ	1.74			1.74	0.00	1.74	< 0.01
β_0	6.96			7.02	0.05	7.05	0.06
β_1	7.47			7.56	0.06	7.58	0.06
β_2	-0.18			-0.18	0.00	-0.18	< 0.01
α	-0.05					-0.05	< 0.01
ψ_0	0.50	0.51	< 0.01	0.51	0.00	0.51	< 0.01
ψ_{AUC}	0.30	0.30	0.01	0.30	0.01	0.30	0.01

Table 5.3: *Simulation results displaying the mean estimates across 1000 simulations, under the three scenarios. The true parameter values for σ and α are set to 10.00 and -0.28 respectively.*

		Observed		LMEM		JM	
Parameter	True	Estimated	MSE	Estimated	MSE	Estimated	MSE
σ	10.00			10.00	0.01	9.90	0.71
β_0	6.96			6.82	0.11	6.84	0.12
β_1	7.47			7.42	0.10	7.42	0.10
β_2	-0.18			-0.10	0.01	-0.15	< 0.01
α	-0.28					-0.22	0.01
ψ_0	0.50	0.50	0.01	0.50	0.01	0.50	0.01
ψ_{AUC}	0.30	0.24	0.01	0.26	0.01	0.25	0.01

Table 5.4: *Simulation results displaying the mean estimates across 1000 simulations, under the three scenarios. The true parameter values for σ and α are set to 10.00 and -0.05 respectively.*

		Observed		LMEM		JM	
Parameter	True	Estimated	MSE	Estimated	MSE	Estimated	MSE
σ	10.00			10.00	0.01	10.04	0.63
β_0	6.96			6.95	0.11	7.00	0.11
β_1	7.47			7.52	0.11	7.52	0.11
β_2	-0.18			-0.13	0.00	-0.17	< 0.01
α	-0.05					-0.05	< 0.01
ψ_0	0.50	0.49	< 0.01	0.50	< 0.01	0.50	< 0.01
ψ_{AUC}	0.30	0.24	0.01	0.27	0.01	0.27	0.01

Table 5.5: *Simulation results displaying the mean estimates across 1000 simulations, under the three scenarios. The true parameter values for σ and α are set to 1.74 and -0.80 respectively.*

		Observed		LMEM		JM	
Parameter	True	Estimated	MSE	Estimated	MSE	Estimated	MSE
σ	1.74			1.73	0.00	1.73	0.01
β_0	6.96			6.76	0.17	6.78	0.19
β_1	7.47			7.28	0.15	7.30	0.21
β_2	-0.18			-0.08	0.01	-0.09	0.01
α	-0.80					-0.80	0.02
ψ_0	0.50	0.51	0.01	0.51	0.01	0.51	0.01
ψ_{AUC}	0.30	0.30	0.01	0.30	0.01	0.30	0.01

Table 5.6: *Simulation results displaying the mean estimates across 1000 simulations, under the three scenarios. The true parameter values for σ and α are set to 10.00 and -0.80 respectively.*

		Observed		LMEM		JM	
Parameter	True	Estimated	MSE	Estimated	MSE	Estimated	MSE
σ	10.00			9.99	0.02	9.86	0.06
β_0	6.96			6.77	0.28	6.75	0.30
β_1	7.47			7.29	0.28	7.27	0.31
β_2	-0.18			-0.04	0.02	-0.07	0.01
α	-0.80					-0.47	0.16
ψ_0	0.50	0.50	0.01	0.50	0.01	0.50	0.01
ψ_{AUC}	0.30	0.25	0.02	0.27	0.02	0.26	0.02

5.4 Statistical Analysis

5.4.1 Exploratory Data Analysis

An exploratory analysis is now carried out to identify both baseline variables and characteristics of physiological variables in the treatment withdrawal period that are potential surrogates for predicting recipient transplant outcome DGF.

Tables 5.7 to 5.10 display the patient characteristics (separately for continuous and categorical variables) at both the donor and recipient levels, while stratifying by recipient outcome. Note that at the donor level (level two) the outcome relates to how many recipients were successful that correspond to a given donor; and at the recipient level a binary indicator of success or failure.

Table 5.7 displays the total number (and percentage) of recipients for each category of all recipient characteristic factor variables. A chi-squared test of independence is performed, where the null hypothesis states that the corresponding variable is independent of recipient transplant outcome. It can be seen that *dialysis at tx*, which a binary indicator of whether the recipient was on dialysis at the time of transplantation, is the only statistically significant variable at the 5% significance level.

Table 5.8 presents the mean and standard deviations (both unstratified and stratified by DGF) for recipient continuous characteristic variables. It was found from the non-parametric Mann Whitney U test that there is not sufficient evidence to reject the null hypothesis at the 5% significance level, whose null hypothesis states that the two samples come from independent populations.

The donor characteristics that are factors are displayed in Table 5.9. There is also not sufficient evidence to reject the null hypothesis of independence (from a Chi-square test) for donor gender, ethnicity or blood group at the 5% significance level. Note that ethnicity 'not white', blood group 'B' and 'AB' have low cell counts which could make these results unreliable.

Table 5.10 displays the donor level characteristics for continuous variables (including both baseline variables and physiological variable characteristics). Variable names ending 'AUC' correspond to the AUC of the observed trajectory, 'slope' corresponds to a linear regression slope between the observed longitudinal data and names ending 'intercept' correspond to the observed value of the physiological variable (also contained in the variable name) at the time of withdrawal. It can be seen that according to the Kruskal-Wallis test the variables HR, MAP, DBP and SBP that correspond to the linear regression slopes are statistically significant at the 5% significance level across the recipient outcome groups. It is notable that donor age and BMI do not show evidence of an association with recipient outcome, but death time is significant at the 10% significance level.

Table 5.7: *Categorical recipient characteristic variable count and proportion stratified by outcome DGF. A χ^2 test of independence is performed for each outcome combination.*

	All recipients n (%)	DGF	Immediate Function	P-value (χ^2)
Total	215	70 (33)	145 (67)	
Gender				0.85
Male	144 (67)	46 (21)	98 (47)	
Female	70 (33)	24 (11)	46 (21)	
Ethnicity				0.42
White	169 (79)	52 (24)	117 (55)	
Not White	44 (21)	17 (8)	27 (13)	
Dialysis at tx				0.002
Yes	168 (78)	64 (30)	104 (48)	
No	47 (22)	6 (3)	41 (19)	
Blood Group				0.42
A	80 (37)	27 (13)	53 (25)	
B	24 (11)	11 (5)	13 (6)	
AB	22 (10)	7 (3)	15 (7)	
O	89 (42)	25 (12)	64 (29)	
Blood Compatible				0.44
Identical	202 (94)	64 (30)	138 (64)	
Compatible	13 (6)	6 (3)	7 (3)	
Gender Mismatch				0.82
Yes	117 (55)	37 (17)	80 (37)	
No	97 (45)	33 (15)	64 (31)	
Ethnic Mismatch				0.35
Yes	155 (79)	45 (23)	110 (56)	
No	42 (21)	16 (8)	26 (13)	

Table 5.8: Continuous recipient characteristic variables mean and standard deviation stratified outcome DGF. Mann-Whitney U tests of independent populations are performed.

	Unstratified	Stratified mean (sd)		Mann Whitney U
Variable	Mean (sd)	Immediate Function	DGF	P-value
Age	52.95 (13.69)	52.36 (14.23)	54.17 (14.23)	0.28
CIT	840.10 (256.08)	829.85 (240.19)	861.03 (240.19)	0.37
BMI	27.19 (4.58)	27.15 (4.05)	27.27 (4.05)	0.52

Table 5.9: *Categorical donor characteristic variable count and proportion stratified by number of successful recipients, with χ^2 test of independence.*

	All recipients n (%)	Successful Recipients n (%)			P-value (χ^2)
Variable		0	1	2	
Gender					0.30
Male	80 (62)	16 (12)	33 (25)	31 (24)	
Female	50 (38)	16 (12)	18 (14)	16 (12)	
Ethnicity					0.46
White	117 (97)	27 (22)	46 (38)	44 (36)	
Not White	4 (3)	2 (2)	1 (1)	1 (1)	
Blood Group					0.26
A	57 (44)	15 (12)	26 (20)	16 (12)	
B	12 (9)	4 (3)	5 (4)	3 (2)	
AB	7 (5)	0 (0)	4 (3)	3 (2)	
O	54 (42)	13 (10)	16 (12)	25 (19)	

Table 5.10: *Continuous donor characteristic variable mean and standard deviations, both unstratified and stratified by number of corresponding successful recipients. Observed physiological variable summaries are also included. A Kruskal-Wallis test of equal means is performed.*

	Unstratified	Successful Recipients			Kruskal-Wallis
Variable	Mean (sd)	0	1	2	P-value
Age	47.77 (17.21)	48.25 (18.47)	49.43 (18.15)	45.64 (15.32)	0.35
Death Time	40.99 (60.62)	38.53 (49.47)	49.67 (64.75)	33.26 (62.84)	0.08
BMI	26.61 (5.51)	26.68 (6.17)	27.42 (5.25)	25.68 (5.29)	0.26
SBP AUC	4448.27 (7149.42)	4442.43 (7354.13)	5547.33 (7354.13)	3203.19 (5992.88)	0.07
DBP AUC	2238.06 (3516.93)	2096.62 (2843.60)	2778.73 (3973.53)	2096.62 (2843.60)	0.09
O2 AUC	2480.09 (4370.51)	2478.95 (3868.08)	2883.12 (4390.43)	2478.95 (3638.08)	0.20
MAP AUC	2958.90 (4643.81)	2866.46 (4131.40)	3654.92 (5163.15)	2193.57 (4269.15)	0.04
HR AUC	4138.52 (7258.38)	4077.48 (6243.19)	5078.21 (7852.89)	3090.95 (7194.4)	0.23
Resp Rate AUC	925.79 (1672)	998.00 (1567.95)	1018.73 (1955.65)	751.81 (1365.41)	0.81
SBP slope	-5.99 (7.11)	-5.09 (7.13)	-4.90 (7.42)	-7.83 (6.50)	0.02
DBP slope	-2.80 (3.38)	-2.12 (3.03)	-2.26 (3.58)	-3.86 (3.15)	0.01
O2 slope	-5.12 (4.84)	-5.18 (5.87)	-4.10 (3.48)	-6.26 (5.27)	0.16
MAP slope	-4.29 (5.87)	-3.17 (4.28)	-3.38 (5.36)	-6.09 (6.94)	0.03
HR slope	-2.75 (4.10)	-1.91 (3.89)	-2.04 (3.02)	-4.16 (4.96)	0.01
Resp Rate slope	-0.28 (2.07)	-0.37 (0.99)	-0.62 (0.95)	0.22 (3.32)	0.64
SBP intercept	130.05 (39.27)	124.77 (41.14)	132.82 (39.75)	130.53 (38.00)	0.56
DBP intercept	67.56 (19.23)	62.55 (16.85)	67.75 (19.25)	70.66 (20.35)	0.14
O2 intercept	94.94 (9.13)	95.83 (5.31)	93.35 (11.46)	96.00 (8.27)	0.07
MAP intercept	87.77 (25.17)	81.13 (21.05)	89.45 (26.07)	90.27 (26.37)	0.28
HR intercept	97.93 (32.85)	95.50 (28.89)	98.67 (39.09)	98.87 (28.46)	0.78
Resp Rate intercept	18.10 (8.21)	19.00 (8.01)	17.82 (7.94)	17.52 (9.14)	0.97

Kaplan-Meier curves are presented in Figures 5.1 and 5.2 to visualise how the probability of surviving throughout the treatment withdrawal phase varies between those that experienced DGF and those that did not (based on the observed data). Figure 5.1 involves donors that only donated a single kidney, thus the survival curves are stratified by whether the corresponding recipient experienced DGF. Figure 5.2 consists of those that donated both kidneys and therefore the survival curves are stratified by the number of recipients that experienced an immediately functioning graft (`no.succ` = 0, 1 or 2). The cross-over between the confidence intervals (shaded regions) implies that there is not sufficient evidence to suggest that patients that experience DGF deteriorate faster in general than those that do not. This is confirmed by the p-value presented within each figure that corresponds to the log-rank test (p=0.38 and p=0.30).

An interesting clinical finding is immediately apparent from Figures 5.1 and 5.2. As the duration of the withdrawal period experienced by some of the patients in this dataset is longer than any (to the best of our knowledge) in the literature, it is of interest to see whether the recipients corresponding to these donors had successful transplant outcomes. It can be seen that both of the recipients corresponding to the donor who survived the longest (406 minutes) were successful. Moreover, the maximum time in the two successful recipients group is longer than the one successful recipient group, which is longer than the transplant outcome group where both recipients failed. This is consistent with the results in Figure 5.1, where the immediate graft function group is somewhat longer than the DGF group. This suggests that the duration of the withdrawal period is not a good surrogate for predicting DGF and that kidneys from donors with prolonged withdrawal periods can result in successful outcomes.

The number at risk tables and the steep survival curves show that most patients had become deceased by 100 minutes.

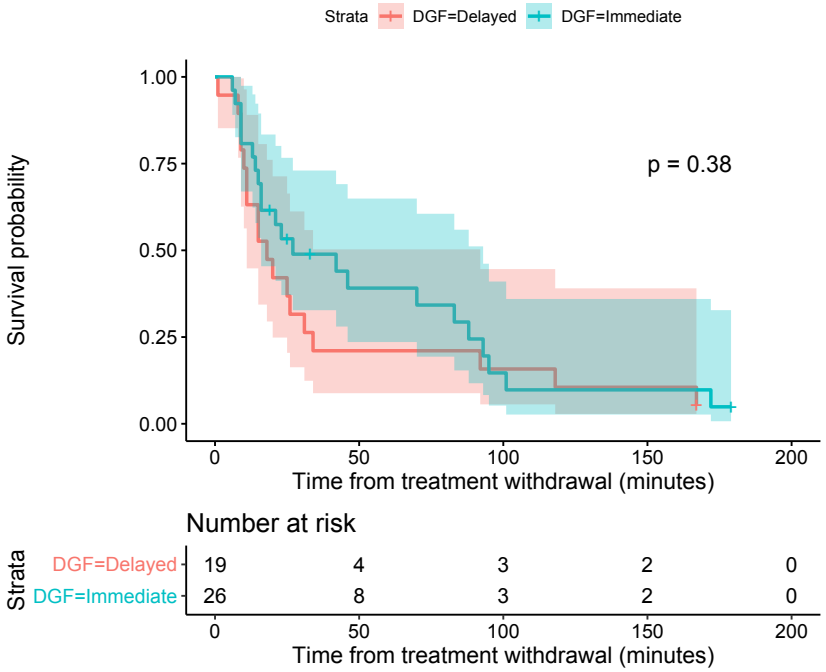


Figure 5.1: *Donor survival probability throughout treatment withdrawal (time in minutes) with 95% confidence intervals (shaded regions) and the number at risk table for donors that donated a single kidney, stratified by recipient DGF. The p-value corresponds to the log-rank test.*

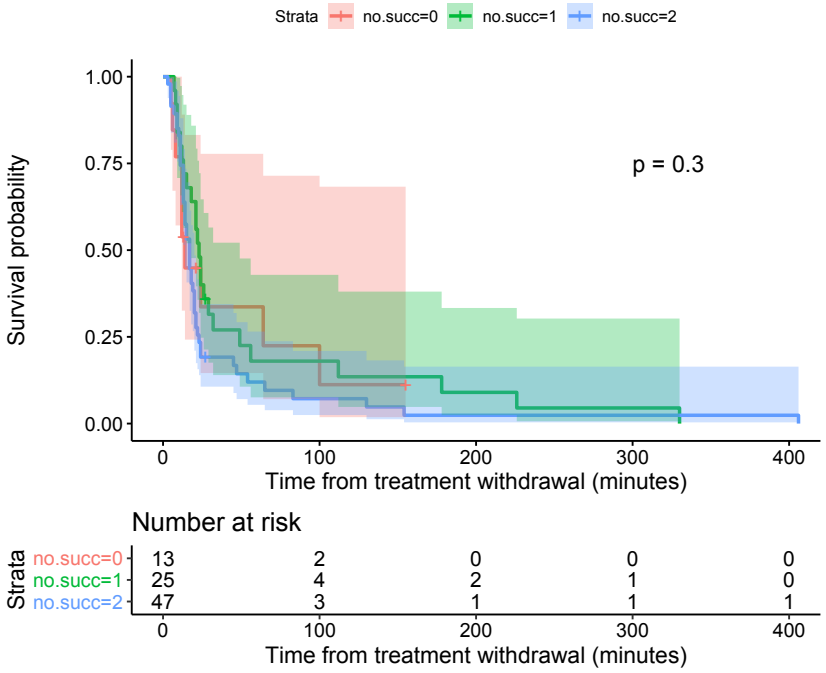


Figure 5.2: *Donor survival probability throughout treatment withdrawal (time in minutes) with 95% confidence interval (shaded regions) and number at risk table for donors that donated both kidneys, stratified by the number of grafts corresponding to the recipients that immediately functioned (`no.succ`). The p-value corresponds to the log-rank test.*

We now investigate whether various characteristics of the physiological variables in the withdrawal period are associated with recipient transplant outcome. Figure 5.3 and 5.4 are trajectory plots that display the observed donor trajectories for each physiological variable. These are presented in the same spirit as the Kaplan-Meier curves just discussed, in that the former plot relates to donors that donated a single kidney and latter corresponds to those that donated both. The trajectories are distinguished by transplant outcome (red is DGF and green is an immediate function). The thin lines represent the observed trajectories, where each line corresponds to a different donor's physiological profile throughout the withdrawal phase. The thick lines correspond to the mean profiles modelled by a non-parametric local regression (LOESS) curve (Cleveland 1979).

According to Figure 5.3 there is very little deviation between the mean profiles of the two groups at the beginning of the withdrawal period for each physiological variable except respiration rate. In particular, the mean profiles for SBP barely deviate throughout the whole withdrawal period. The difference is somewhat more apparent for O2 and DBP, indicating the those that maintain a higher DBP and lower O2 may result in more favourable transplant outcomes. However, this deviation is small enough such that it could be due to random chance.

Trends are difficult to detect from Figure 5.4, but much larger fluctuations between the three groups are apparent compared to Figure 5.3. For this reason, Figure 5.5 is presented, which is the same as Figure 5.4 except the x-axis is cut at 60 minutes. Figure 5.5 shows that the larger fluctuations do not occur within the first hour, except for O2, where after half an hour the no successful recipients group dramatically drops. However, it can be seen from 5.4 that after one hour the no successful recipients group has a sharp rise.

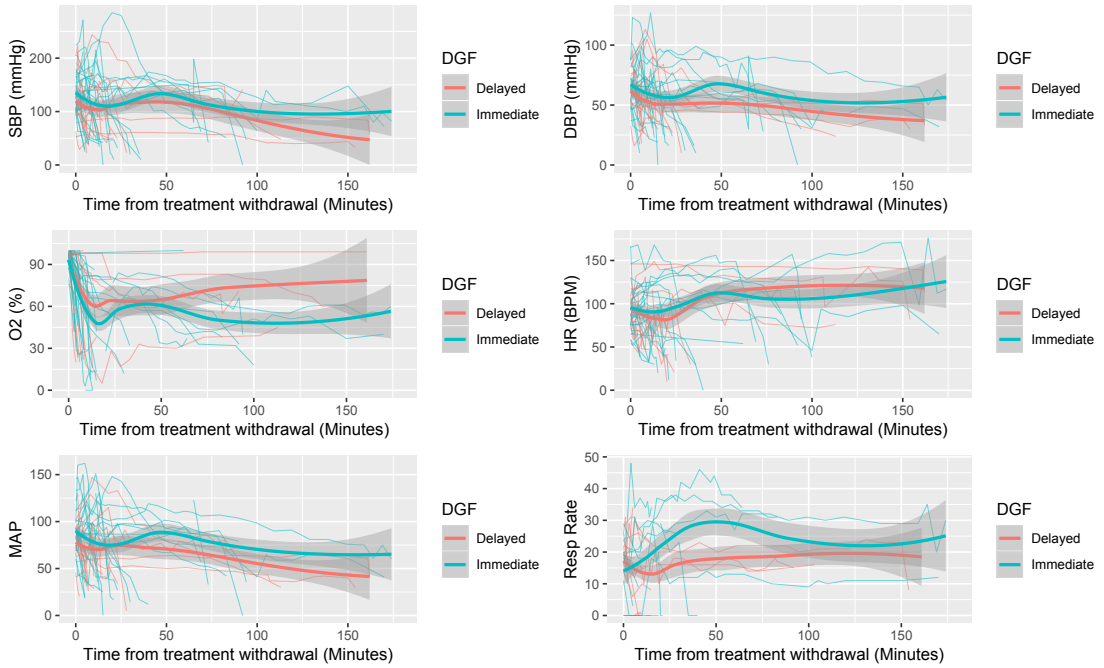


Figure 5.3: *Trajectory plots for DCD kidney donors that donated a single kidney, colour coded by the corresponding recipient transplant outcome DGF. The thick lines correspond to non-parametric local regression (LOESS) curves (Cleveland 1979) representing the conditional mean with 95% confidence intervals.*

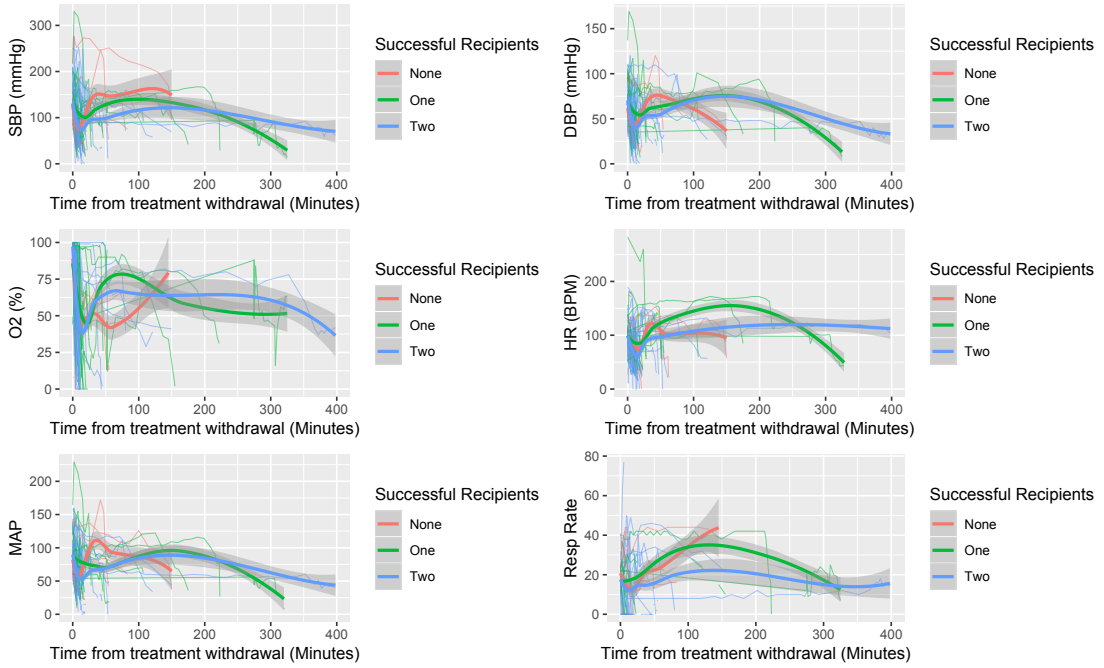


Figure 5.4: *Trajectory plots for DCD kidney donors that donated both kidneys, colour coded by the number of corresponding recipient immediate graft functions. The thick lines correspond to non-parametric local regression (LOESS) curves (Cleveland 1979) representing the conditional mean with 95% confidence intervals.*

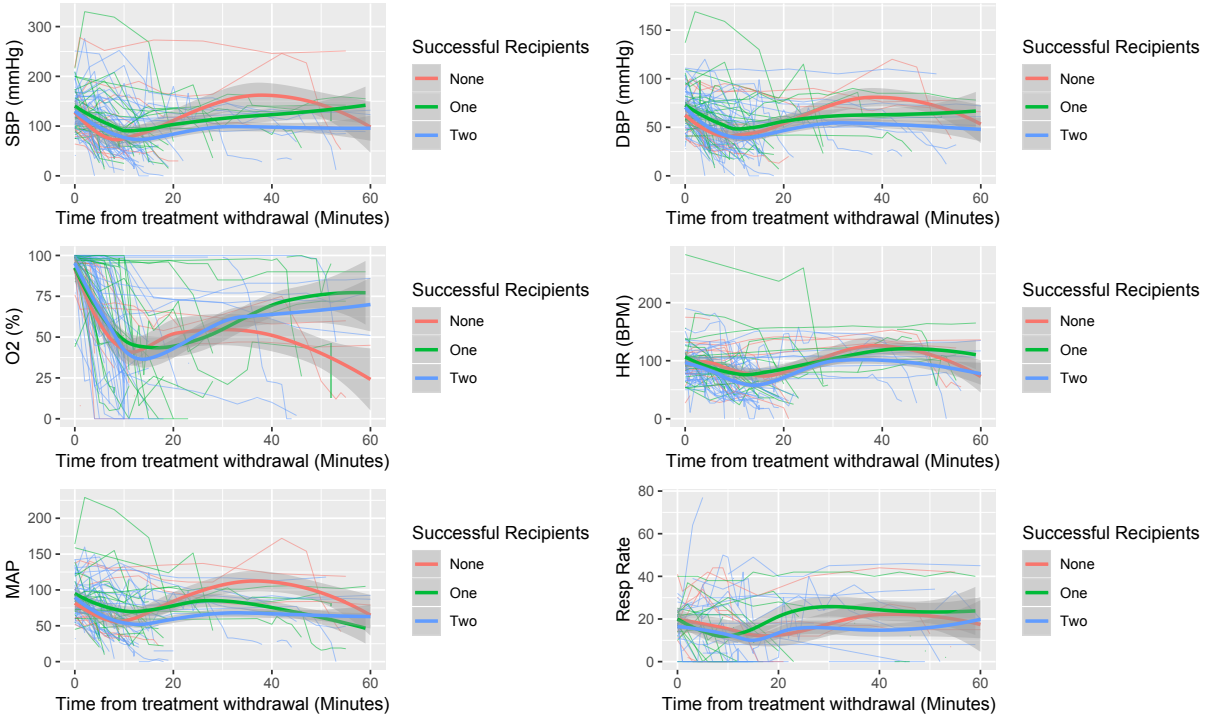


Figure 5.5: *A repeat of Figure 5.4 with the x-axis is cut at one hour, to make the beginning of the treatment withdrawal period more visually clear.*

5.4.2 Statistical Modelling

5.4.2.1 A Joint Modelling Approach

We now employ a more formal modelling approach that acknowledges the hierarchical structure of the data, in order to study characteristics of the mean profiles between groups of transplant outcome. In particular, we fit a JM whereas opposed to Chapter 4 interest is now in the longitudinal outcome (treating the survival outcome as a nuisance implicit outcome), rather than predicting event times.

Recall the standard JM discussed in Chapter 4 (Equations 5.11 and 5.12). A straightforward model selection procedure is carried out for each longitudinal outcome with the aim to derive a model that regresses the outcome group against the longitudinal outcome, while accounting for drop-out as a result of death or censoring. This selection procedure involves comparing four models per longitudinal outcome (choosing the model with the lowest deviance information criteria, DIC, value). For each model both the fixed and random components are flexibly modelled over time with a natural spline basis function that includes a single interior knot placed at the median of the follow-up times.

$$y_i = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \epsilon_i(t) = m_i(t) + \epsilon_i(t) \tag{5.11}$$

$$h_i(t) = h_0(t)\exp(\boldsymbol{\gamma}^\top w_i + \alpha m_i(t)) \tag{5.12}$$

The first model specified for each outcome includes an interaction in the fixed effects component between the flexibly modelled time effects and the transplant outcome group (`no.succ`). The survival component design matrix $\boldsymbol{w}$ is specified to include only an intercept term. The longitudinal component for this model can be written mathematically as follows:

$$\begin{aligned} m_i(t) = {} & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})\mathbf{B}(t, \kappa_1) + (\beta_2 + b_{2i})\mathbf{B}(t, \kappa_2) \\ & + \beta_3\{\mathbf{B}(t, \kappa_1) \times \texttt{no.succ}_i = 1\} + \beta_4\{\mathbf{B}(t, \kappa_2) \times \texttt{no.succ}_i = 1\} \\ & + \beta_5\{\mathbf{B}(t, \kappa_1) \times \texttt{no.succ}_i = 2\} + \beta_6\{\mathbf{B}(t, \kappa_2) \times \texttt{no.succ}_i = 2\} \\ & + \beta_7\{\texttt{no.succ}_i = 1\} + \beta_8\{\texttt{no.succ}_i = 2\}. \end{aligned}$$

We also specify for each longitudinal outcome a model that assumes no interaction between the transplant outcome groups and time, i.e.:

$$\begin{aligned} m_i(t) = {} & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})\mathbf{B}(t, \kappa_1) + (\beta_2 + b_{2i})\mathbf{B}(t, \kappa_2) \\ & + \beta_3\{\texttt{no.succ}_i = 1\} + \beta_3\{\texttt{no.succ}_i = 2\}. \end{aligned}$$

The third and fourth models are the same, except a log transformation is applied to the longitudinal response. The model with the lowest DIC for each outcome is compared against the same model that includes an extra interior knot in the natural spline basis function. The model with the lowest DIC is selected for each longitudinal outcome and the fitted values are plotted in Figure 5.6 (note that the time axis is cut at two hours).

The longitudinal variable names on the y-axis in Figure 5.6 begin with 'log' if the selected model included a log transformation. The models whose goodness-of-fit improved by including the time-dependent interaction effect correspond to the plots where the mean profile lines cross over.

It can be seen from Figures 5.6 and **??** that there is little evidence of a significant difference in mean profiles across the outcome groups (due to the cross-over of the credibility intervals), except for respiration rate for those that donated both kidneys. This sub-figure suggests that a low respiration rate is beneficial in terms of short-term transplant outcome DGF. Despite this, it is note worthy that respiration rate has a considerably higher missing data rate. Moreover, this method of analysis is used here only as an exploratory tool.

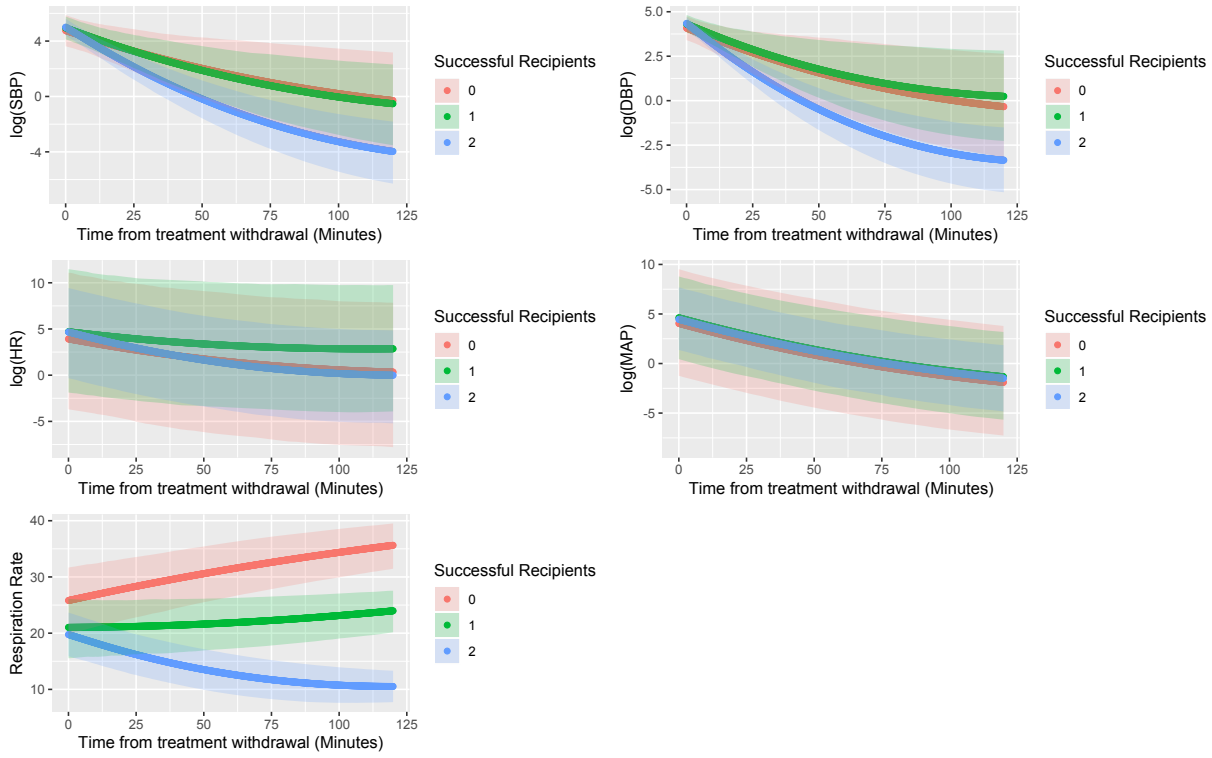


Figure 5.6: *Joint modelling DCD donor physiological variables throughout the treatment withdrawal period treating group of number of successful recipients as a fixed effect (including only donors that donated both kidneys).*

5.4.2.2 Random Intercept Logistic Regression

A univariate analysis was conducted, beginning with both characteristic variables and *observed* characteristics of the physiological variables (intercept, slope and AUC). The following step was to repeat the process using the intercept, slope and AUC estimated from the LMEM. Models were compared using the AIC and 5-fold cross-validation. Missing values were imputed using a method suitable for hierarchical data (Goldstein et al. 2009, Yucel 2011), which was implemented using the `jomoImpute` function in the `mitml` package in R.

Univariate RILR models were fitted with two random intercept terms. One corresponds to the recipient centre and the other relates to the donor unique identifiers, both are variables that we expect to have correlated observations within groups. All possible models were fitted that contained a single fixed effect.

Figures 5.7 and 5.8 display the AIC and the mean AUC ROC for each univariate model fitted. The former figure corresponds to an analysis where observed characteristics of the physiological variables are used and the latter corresponds to their estimated characteristics. The models are colour coded to highlight those that were ranked top six for each metric (six is chosen arbitrarily). The x-axis displays the formula of the univariate variable that is used in the model. Note that `bs(x , df)` and `ns(x, df)` represent a B-spline and natural spline basis function for variable `x` respectively with `df` degrees of freedom. The variable code name `scale(x, scale = TRUE)` subtracts the mean of the variable from its current value and divides by the standard deviation. This normalisation is performed to improve numeric stability during parameter estimation (particularly for the AUC variables, whose large values are likely to cause numerical difficulties). Variable names ending in `_scale` are also normalised. Physiological variables names containing `auc`, `slope` and `inter` correspond to the AUC, slope and intercept.

For each model containing a continuous covariate, the amount of flexibility specified for the corresponding variable was tuned. This was done by fitting models with all combinations of natural and B-spline basis functions with degrees of freedom ranging from the minimum (two for natural splines and three for B-splines) to a maximum of eight. Models assuming linearity were also fitted and compared. This tuning procedure was performed by selecting the model with the lowest AIC. This was repeated for both the observed and the estimated physiological characteristic variables, and the best univariate models are presented in Figures 5.7 and 5.8 alongside the other characteristic variables.

There is a notable difference between Figures 5.7 and 5.8. In particular, many of the covariates were found to have different functional forms as a result of different imputation models (one containing observed physiological variables and the other containing those estimated from the LMEM). The AIC remains reasonably consistent (most variables being around 274) with the exception of a few variables, in particular the observed HR

and respiration rate slopes seem to perform somewhat better than the estimated. Little variation can be seen in the mean AUC ROC nor the error bars. We proceed with the dataset imputed with the estimated physiological variables based on the favourable results in our simulation study compared to the observed values.

As it is not computationally feasible to compare all combinations of multi-variable models in an exhaustive selection procedure, it is necessary to reduce the set of candidate variables that are compared for goodness-of-fit. This is done by considering results from both the exploratory and univariate analyses, and judgement is used where necessary (taking into account variables that have been found to be predictive of DGF with the analysis of the larger dataset in Chapter 3) to remove likely weak predictor variables. Moreover, variables that were found to be predictive in the observed physiological characteristics univariate analysis, that were not predictive in the estimated analysis are not removed from the set of candidate variables.

Figure 5.7: *Univariate RILR models with two random intercepts (relating to recipient centre and donor ID) for each characteristic variable and* ***observed****: intercept, slope and AUC of the physiological variables. The mean AUC ROC and negative AIC are displayed and the variables ranked in the top six are colour coded.*

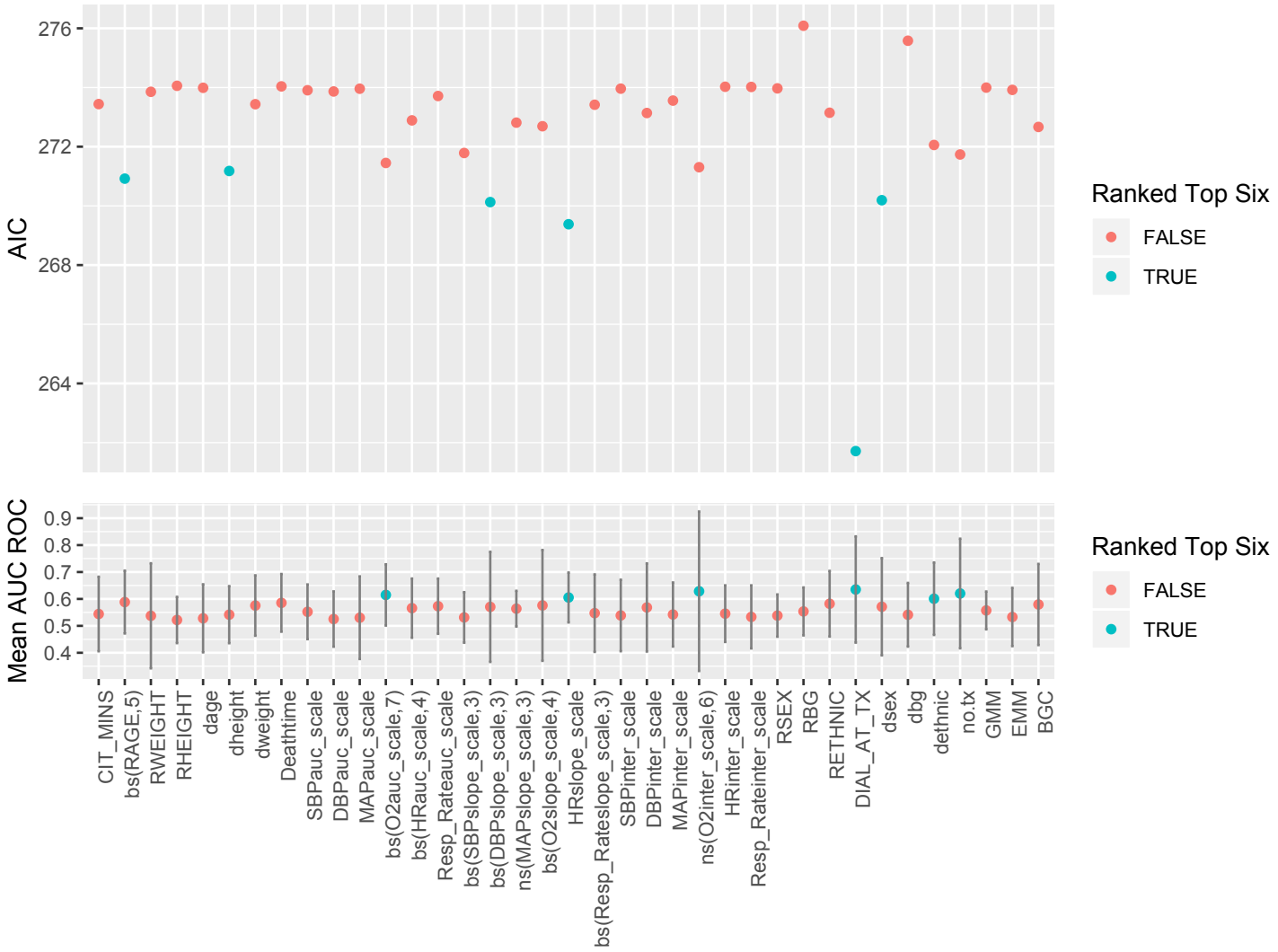


Figure 5.8: *Univariate RILR models with two random intercepts (relating to recipient centre and donor ID) for each characteristic variable and* ***estimated****: intercept, slope and AUC of the physiological variables. The mean AUC ROC and negative AIC are displayed and the variables ranked in the top six are colour coded.*

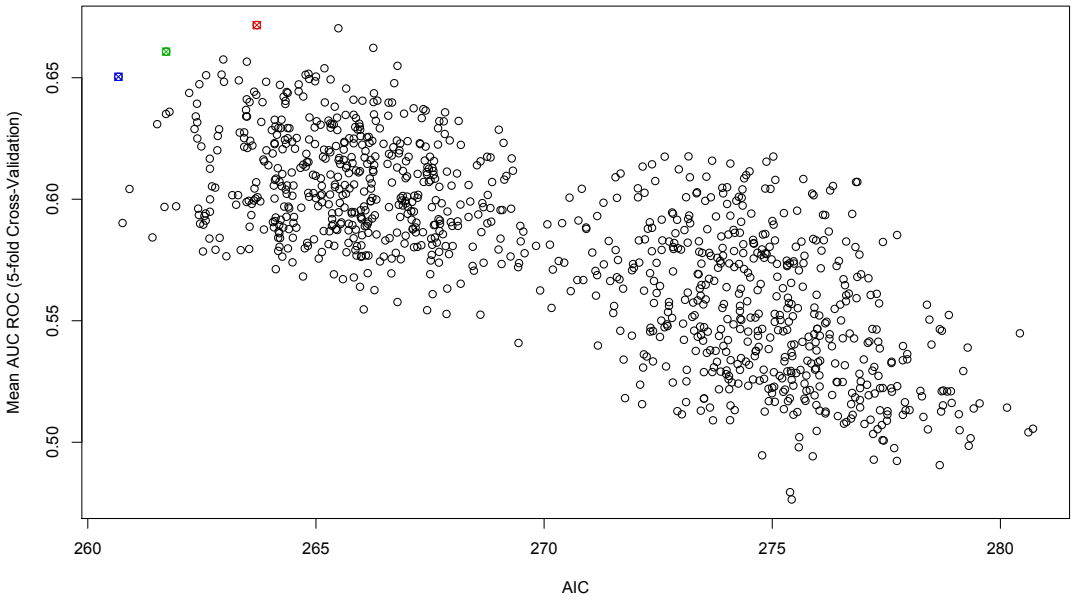


Figure 5.9: *Multi-variable model selection for choosing a baseline model (comparing models with the AIC and mean AUC ROC from 5-fold cross-validation). Each point represents a different fitted model.*

We proceed by deriving a predictive baseline model that is based solely on the characteristic variables (not including those relating to the physiological variables). Based on the criteria just discussed the following variables are retained for the multi-variable analysis: `DIAL_AT_TX`, `dsex`, `dheight`, `dethnic`, `deathtime`, `dage`, `rage`, `RETHNIC`, `RHEIGHT`, `CIT_MINS`. All combinations of these covariates in what was found to be their best functional form on a univariate basis (a total of 1,023 models) were tested and compared by the AIC and 5-fold cross-validation AUC ROC metrics. This model selection procedure was performed utilising parallel computing on the University of Southampton high performance computing facility Iridis 4. The results for the selection procedure can be seen in Figure 5.9. We restrict our interest to the three competing models highlighted in colour (blue: `DIAL_AT_TX + dsex`; green: `DIAL_AT_TX + dsex + dethnic`; red: `bs(RAGE,5) + DIAL_AT_TX + dage + RETHNIC`).

The final step for selecting a baseline predictive model was to test for an improvement in goodness-of-fit and predictive ability by including interaction effects. First, all combinations of models with a single interaction effect were fitted (not adjusting for other variables). The model with the best fitting interaction term was added to the three candidate baseline models. No improvement in fit or predictive ability was achieved by including interaction effects.

Judgement was used to select the model highlighted in blue as the baseline predictive model. It can be seen that penalising for complexity (with the AIC metric) resulted in

a small reduction in predictive ability. However, combining this model with non-linear effects from the physiological characteristic variables in the model highlighted in red is more likely to result in over-parametrisation and thus suffer from a loss of statistical power due to the very limited amount of data available.

The tuning selection procedure previously described (comparing all combinations of models with natural and B-spline basis functions with degrees of freedom ranging between the minimum and eight) was used for each estimated physiological characteristic variable appended to the selected baseline predictive model that includes `DIAL_AT_TX + dsex`. For each physiological characteristic that was tuned in terms of flexibility on the univariate basis can be found in Table 5.11 with the corresponding AIC, mean AUC ROC and standard deviation (SD) AUC ROC. Note that in order to avoid multicollinearity, combinations of models including more than a single physiological characteristic variable were not fitted. Three models were selected (highlighted in blue in Table 5.11).

Table 5.11: *Final model selection comparing models including physiological characteristic variables to the baseline model using goodness-of-fit and discriminatory ability metrics (AIC and mean AUC ROC).*

Baseline Including	AIC	Mean AUC ROC	SD AUC ROC
Baseline Only	260.67	0.65	0.09
Resp_Rateinter_scale	262.67	0.61	0.06
bs(SBPinter_scale,5)	262.10	0.62	0.08
DBPinter_scale	262.35	0.63	0.07
ns(O2slope_scale,4)	261.48	0.63	0.06
Resp_Rateslope_scale	262.31	0.63	0.09
Resp_Rateauc_scale	262.59	0.63	0.11
ns(HRauc_scale,3)	261.78	0.64	0.09
MAPinter_scale	262.29	0.64	0.07
bs(DBPslope_scale,3)	258.16	0.65	0.03
SBPauc_scale	262.60	0.65	0.07
HRinter_scale	262.48	0.65	0.06
bs(SBPslope_scale,3)	260.56	0.65	0.07
ns(MAPslope_scale,3)	261.80	0.65	0.10
DBPauc_scale	262.53	0.65	0.07
MAPauc_scale	262.62	0.66	0.07
bs(O2auc_scale,7)	260.05	0.66	0.04
HRslope_scale	258.06	0.67	0.10
ns(O2inter_scale,3)	259.54	0.68	0.04

It can be seen that only four of the physiological characteristic variable models resulted in an improved discriminatory ability compared to the baseline characteristic model. This improvement was marginal (an increase in mean AUC ROC between 1 and 2). Again compared to the baseline model, the AIC reduced for only three models (albeit by a very small amount). The model containing the estimated O2 intercept had the best discriminatory ability and the model containing the estimated slope of HR had the best goodness-of-fit.

As two of the three chosen models are semi-parametric due to the spline functions, their parameter estimates are not directly interpretable. These models are best interpreted by visualising the estimated probabilities over the unscaled range of the corresponding predictor variables. These plots are displayed for each chosen model in Figure 5.10, while stratifying for the two categorical baseline characteristic variables (`dsex` and `DIAL_AT_TX`). These plots display the marginal estimated probabilities (i.e., the average across recipient centres and donors), so that the random intercept terms can be incorporated into the interpretation. The estimated probabilities are presented with the 50% range that the predicted probabilities fell (marked by the shaded regions).

It can be seen that recipients that are not on dialysis at the time of transplant have a substantially improved chance of an immediately functioning graft compared to those that are. Moreover, there appears to be a less prominent gender effect, suggesting that males have an improved chance of success to females. The physiological characteristic variable that had the best discriminatory ability was the estimated oxygen saturation intercept, for which Figure 5.10 suggests an improved chance of success for those that have a larger oxygen saturation at the time of treatment withdrawal. In particular, a threshold effect appears around O2 intercept of 96, where patients with values greater than this have a much more improved chance of success. Figure 5.11 displays the density of the estimated O2 intercepts, indicating that there is credibility particularly between 90 and 100% O2 intercept. These results conform with intuition as a lack of oxygenation is known to impact graft quality.

The middle two plots in Figure 5.10 correspond to the chosen model including the estimated HR slope. It can be seen that a donor with a steep decrease in HR corresponds to an improved chance of success for the recipients. Conversely, those whose HR remains relatively constant or increases are less likely to be successful. Figure 5.11 shows that these probability estimates are credible between HR slopes values of approximately -7 and 1.

The bottom two plots in Figure 5.10 correspond the estimated AUC of O2. These results are less informative due to a lack of credibility across the majority of the range of the physiological characteristic predictor variable as shown by Figure 5.11. For the part range of the variable where there is credibility, there appears to be a negligible difference in the chances of success as a result of varying the estimated AUC. For this reason, the AUC O2 model is dropped from consideration.

As a final check we compare the competing models to the nested baseline model with the likelihood ratio test. These results are presented in Table 5.12. As O2 intercept and HR slope are not correlated variables, the model containing both of these variables is considered. It can be seen that including both O2 intercept and HR slope in the baseline model results in the lowest AIC (255.68) and is statistically significant at the 5% significance level.

As the variable time to death is of interest, we compare this to the current best model (including: O2 intercept, HR slope, dialysis status and gender). Assuming linearity for the death time variable resulted in AIC 257.28 (LRT p-value 0.53) and allowing flexibility by including a natural spline function with 3 degrees of freedom resulted in AIC 260.72 (LRT p-value 0.81). We proceed by choosing the model without time to death from treatment withdrawal.

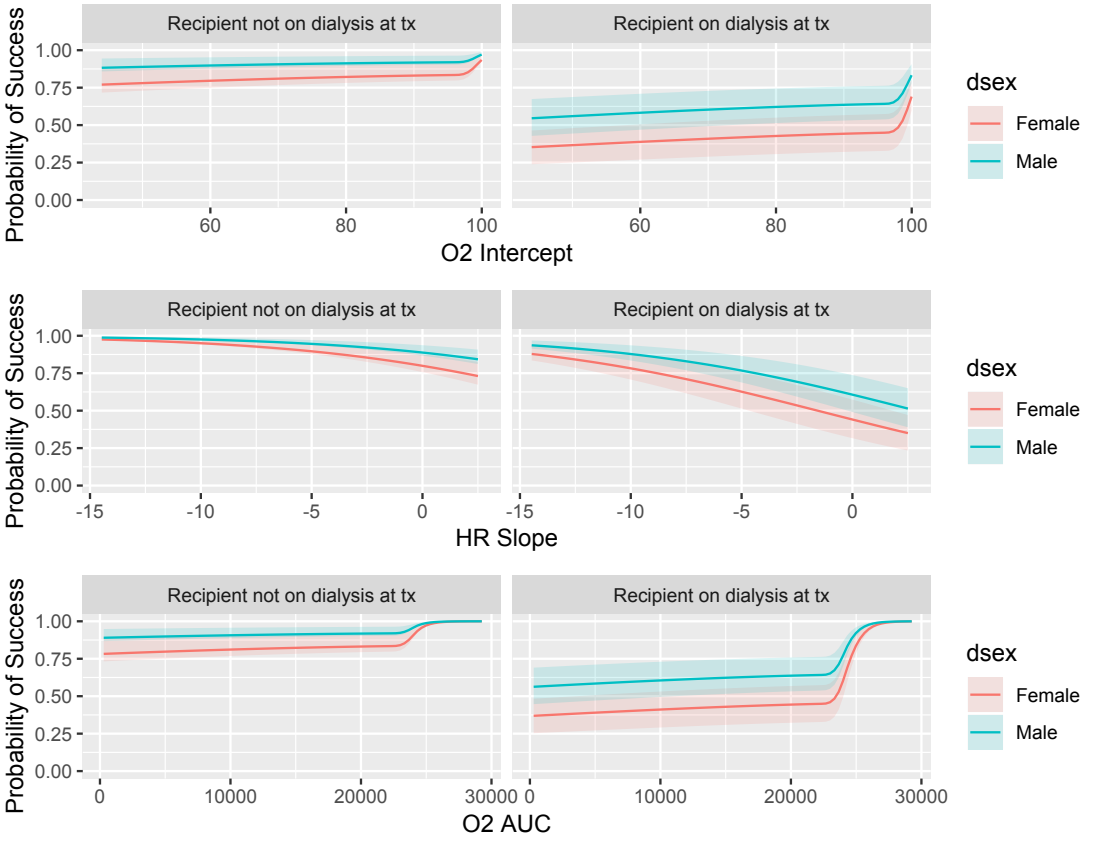


Figure 5.10: *Estimated marginal probability of an immediate graft function based on the three selected models with the range in which 50% of the predicted probabilities fell marked by the shaded regions.*

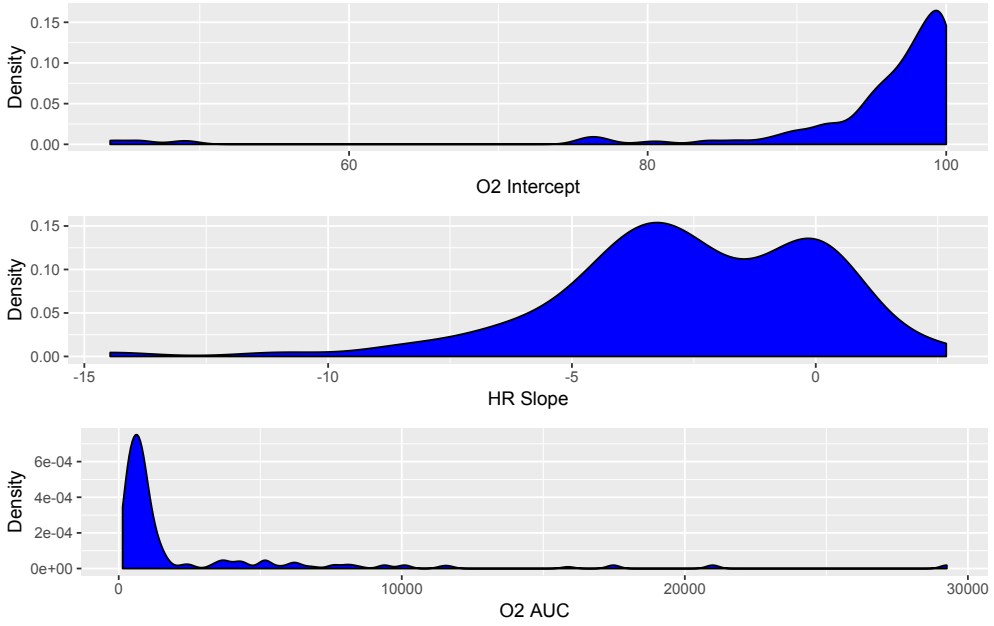


Figure 5.11: *Density plots for each estimated physiological variable characteristic that corresponds to the chosen models, to further determine credibility of probability estimates in Figure 5.10.*

Table 5.12: *The likelihood ratio test to investigate whether the additional parameters of interest are significantly different from 0, by comparing models to the nested baseline model.*

Model Including	AIC	Degree of Freedom	Deviance	P-value (LRT)
Baseline	260.67	5	250.67	-
O2 Intercept	259.54	8	243.54	0.07
HR Slope	258.06	6	246.06	0.03
O2 Intercept and HR Slope	255.68	9	237.68	0.01

We proceed by presenting the chosen model. The random intercept corresponding to the donor ID had a variance of 0.63 and that corresponding to the recipient centre had a variance of 0.41. The fixed effects estimates, standard errors, z values and p-values are displayed in Table 5.13.

Table 5.13: *Summary of chosen model, displaying parameter estimates, standard errors, z-values and p-values.*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.32	1.38	0.96	0.34
dsexMale	0.75	0.41	1.82	0.07
DIAL_AT_TXYes	-1.72	0.60	-2.88	< 0.001
ns(O2inter_scale, 3)1	-0.22	1.10	-0.20	0.84
ns(O2inter_scale, 3)2	13.16	7.15	1.84	0.07
ns(O2inter_scale, 3)3	22.16	10.56	2.10	0.04
HRslope_scale	-0.49	0.22	-2.24	0.03

The natural spline term in this model inhibits the interpretability of the parameter estimates. As an alternative interpretation, the marginal probabilities of recipient outcome (averaged over recipient centre and donor) that were estimated from this model are displayed in Figure 5.12. The 50% range of these values are represented by the shaded regions. Estimated probabilities are plotted over the range of values of the covariate O2 intercept. These probabilities are stratified by gender, the recipient's dialysis status at time of transplant, and the slope of HR (at the quantiles of the data).

Moving from the first quantile to the fourth of the covariate HR slope in Figure 5.12 results in a shift in probability towards a negative transplant outcome. Findings from Figure 5.10 still hold that those on dialysis at the time of transplant have a much larger risk of DGF (p-value < 0.001). This is somewhat worse for females (p-value 0.07). According to this model, donors with an oxygen saturation greater than approximately 95% at treatment withdrawal (that also have a steep decline in HR) have the best chance of a favourable transplant outcome.

The caterpillar plot in Figure 5.13 displays the conditional modes of the random intercept term corresponding to transplant centre with error bars. These estimated random intercept values represent the inherent risk of incurring DGF for recipients attending a given recipient centre. Those whose values are close to zero on the x-axis correspond

to an average risk. Centres with a large positive estimated value performed better from past experience compared to those with a large negative value.

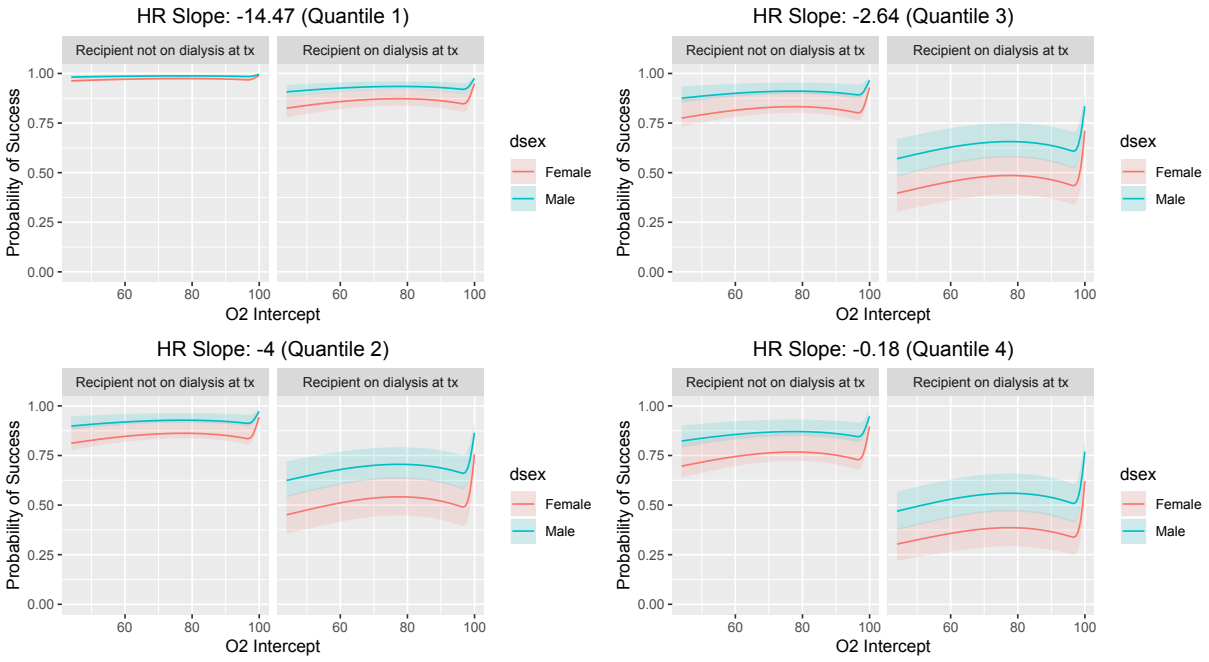


Figure 5.12: *Visualisation for the final selected model displaying the probability of success (immediate graft function) across the range of possible oxygen saturation intercept values, stratifying by gender, dialysis at time of transplant and the quartiles of the slope of heart rate.*

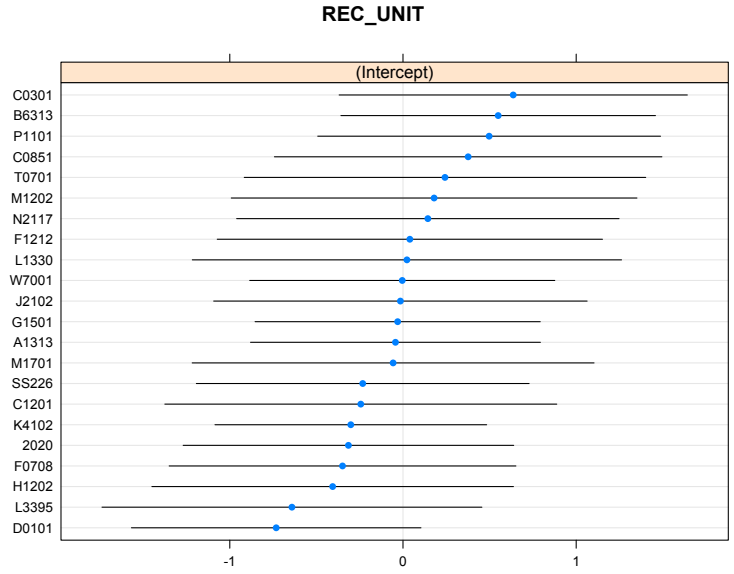


Figure 5.13: *The conditional modes (blue dots) of the recipient transplant centre random intercept values with error bars.*

5.5 Discussion

The aim of this chapter was to analyse the novel dataset provided by the NHS Blood and Transplant, that differs from those in the literature due to the availability of recipient outcomes corresponding to donors with a prolonged treatment withdrawal, to improve our understanding of how characteristics relating to this pivotal phase are associated (if at all) with recipient short-term transplant outcome DGF, while adjusting for confounding variables.

Recall Section 2.2.1 where it was discussed that causation cannot be established from an explanatory modelling approach when not all associated variables are available and adjusted for in the statistical model. Moreover, observational data collected not according to a well designed experiment is likely to inherit complications that violate explanatory modelling assumptions. For this reason, this analysis was posed as a predictive modelling task so that less assumptions were made at the cost of some interpretability. However, we were still able to test whether associations were statistically significant and visually interpret the model by observing how the probability of an immediately functioning graft varies across the range of predictor variables while adjusting for confounding variables.

Although we acknowledge that explanatory models are not optimal for prediction, the impact of our choice of method on statistical inference was still of interest as we expect it to be associated with predictive ability. Moreover, this impact on inference may be informative for future studies that employ explanatory models. We therefore conducted a simulation study to compare the MSE of estimates corresponding to the AUC of the observed physiological trajectories to those estimated from a LMEM and also those estimated from a joint modelling approach. Our results suggest that a two stage approach that uses the physiological trajectories estimated from a LMEM as a covariate in a classification model has better inferential properties as opposed to using the observed trajectories. However, the trajectories estimated from a joint modelling approach performed almost equivalently (yet somewhat worse) than those estimated from the LMEM. We expect this is a result of the expense of having to estimate many more parameters and having to approximate integrals numerically in the joint modelling approach. This led to the decision to implement the LMEM method for extracting relevant information from the longitudinal trajectories in our analysis.

An exploratory analysis was performed with the aim to understand the demographics that the motivating dataset is composed of and to obtain preliminary results that are indicative as to whether characteristics of donor physiological variables (observed AUC, slope and intercept) in the withdrawal phase could be related to recipient transplant outcome DGF. Simple univariate statistical tests (Chi-square, Mann Whitney U and Kruskal-Wallis) found that recipient dialysis status at time of transplant was the only baseline variable related to the DGF outcome at the 20% significance level (p-value < 0.001). Although donor death time was borderline significant (p-value 0.08), various

characteristics of the treatment withdrawal were found to be significantly associated with the DGF outcome. In particular, the following variables (with p-values in parenthesis) were found to be potential surrogates: SBP AUC (0.07), DBP AUC (0.09), MAP AUC (0.04), SBP slope (0.02), DBP slope (0.01), MAP slope (0.03) and HR slope (0.01) and O2 intercept (0.07).

The log-rank test was performed to test whether the survival curves of the outcome groups differ significantly based on the observed data. No evidence of a significant difference was found suggesting withdrawal time is a poor indicator of DGF outcome. Moreover, both of the recipients corresponding to the donor that survived the longest had immediately functioning grafts. As this is the first dataset to the best of our knowledge to contain donors with such a long withdrawal, this is an important clinical finding that suggests a potential for increased conversion from potential to actual donors by allowing transplants to proceed when the withdrawal phase is prolonged. Despite this, further validation is required before this advice can be applied in practice.

Our analysis proceeded by investigating whether the mean profiles showed evidence of trends between groups of transplant outcome DGF. In particular, a non-parametric local regression was fitted to the observed trajectories as an exploratory method and a JM was fitted allowing an interaction between outcome group and time (incorporating the hierarchical structure of the data and drop-out mechanism in to the model). Trends were difficult to detect from these methods, particularly for the part of the withdrawal period with the most credibility. Despite this, the JM approach suggested that patients with low respiration rate have favourable outcomes compared to those that do not.

A rigorous model selection procedure was performed to derive a model for DGF outcome based on baseline characteristics and also characteristics of the physiological variables (AUC, slope and intercept). We found that the model containing HR slope, O2 intercept, dialysis status at the time of transplant and gender had the best goodness-of-fit and discriminatory ability. In particular, we found that males that were not on dialysis at the time of transplant with O2 intercept above 96 and a large negative slope for HR led to the most favourable transplant outcomes.

We note that although our model accords with intuition, there was not sufficient evidence in these data of an association between DGF outcome and donor age or cold ischaemic time. This does not accord with findings in the literature which may suggest a limited statistical power to detect these associations, which could be a result of the relatively small sample size. It is also important to note that the JM approach suffers from the same assumptions and limitations discussed in Section 4.5 and the RILR models assume normality of the random intercept terms.

Part IV

Discussion and Concluding Remarks

Chapter 6

Discussion and Future Work

6.1 General Discussion

Throughout this thesis we have studied a range of problems that are of interest to clinicians that wish to ensure that the organ transplantation process is as effective and efficient as possible. The increased use of organs from DCD donors in the last decade has largely narrowed the deficit in the number of kidneys available by expanding the pool of potential donors. However, due to the nature of the DCD donor treatment withdrawal to death period, kidney recipients whose donors are DCD donors have a higher incidence of DGF compared to those from DBD donors. Although this is thought to be due to ischaemic injury resulting from a lack of oxygenation in the withdrawal phase, our understanding of how this period impacts DGF is limited, which has provided scope for this research.

The currently limited understanding of how the withdrawal phase is associated with negative transplant outcomes has led to restrictive protocols that prohibit donors proceeding to transplant under certain conditions as a safeguard to minimise the chances of negative transplant complications such as DGF. In particular, protocols currently restrict donors proceeding to transplant when the withdrawal phase is prolonged, however, in recent years the relevance of the treatment withdrawal period duration has been questioned and studies have implied that the behaviour of the physiological profiles may be a better predictor of DGF (Bradley et al. 2013). If this is the case, there is potential for an improved conversion of potential to actual donors by allowing patients with a prolonged treatment withdrawal duration to proceed to transplantation.

Graft quality is not the only factor that prohibits donors proceeding to transplant. This can also occur as a result of logistics and the removal team's limited resources. Once withdrawn from life-sustaining treatment, donors can take any time between a few minutes and a few days to become deceased. As this usually happens relatively fast, the

removal team prepare for transplantation once the donor is withdrawn from life-support. However, when the duration is prolonged the removal team do not have the capacity to wait and are required to perform transplantation elsewhere. Being able to predict when the donor is going to be deceased accurately would be invaluable for clinicians that aim to optimise the efficiency of the transplantation process.

In this work we have aimed to address these issues by applying innovative statistical methodology to the various datasets provided by the NHSBT. Each dataset proved to be inherent of complications presenting novel and interesting statistical challenges. This has provided scope for novelty in this work with regards to the methodological side of the project.

In Chapters 1 and 2 the problem at hand was introduced and background context was provided, describing both the clinical and statistical aspects of this project that are important throughout this thesis. In particular, we described the kidney transplantation process, types of kidney donors and the treatment withdrawal to death phase (and its clinical relevance). We then proceeded by distinguishing between explanatory and predictive models and outlined relevant statistical theory.

In Chapter 3 various machine learning methods (random forests, conditional random forests, AdaBoost and XGBoost) were introduced, and their ability to predict whether a patient would experience DGF based on donor, recipient and treatment withdrawal phase characteristics were compared to the standard statistical method RILR. We noted a range in predictive performance across the considered methods, but the overall best performers were XGBoost and RILR. Based on the AUC ROC and BER metrics, XGBoost was the best performing method. This suggests that machine learning methods can be a useful set of tools for the NHSBT when the aim is prediction, as minimal assumptions are required and a subjective model selection procedure is not required at the expense of having to tune the hyper-parameters of the learning algorithm and having a limited ability to quantify certainty. Although the standard inferential tools used for explanatory modelling are not available for non-parametric methods, visualisation tools offer a means of interpretation. In the case of semi-parametric models (such as a RILR model that contains spline terms), the significance of associations can be tested.

Although prediction in the setting of Chapter 3 is useful, it was also of interest to know which of the covariates used to train the algorithms were predictive of DGF (and to rank these variables' importance). For this reason, we ran a series of simulations to assess the ability of the machine learning algorithms to rank the importance of the covariates when applied to data with the same inherent complications as the NHSBT dataset (such as a multilevel structure, variables with many categories, and the presence of non-linear associations). To the best of our knowledge, no study has assessed the robustness of importance statistics from these machine learning algorithms when such complications are present in the dataset. Our findings were consistent with Strobl et al. (2007) in that

the Gini index is biased towards variables with many categories, and that the permuted importance index should be used. Although Strobl et al. (2008) claim that the permuted importance index is biased towards highly correlated variables, this was not evident in our simulation study. In particular, the variables representing the time it takes SBP to drop below 50, 60 and 70 mmHg are very highly correlated, yet were (correctly) found to be not important in this simulation study. The CRF was by far the best performer in terms of ranking variable importance. For the sample size of 1000 observations, the CRF was able to correctly rank the most important variable every time. As expected the ranking was less effective for variables with a weaker association with the response and performance also decreased as the sample size was reduced. This suggests that asymptotically the machine learning variable importance ranking methods perform as they should (even when faced with the complications inherent to the NHSBT dataset) and are therefore an appealing tool for exploratory analysis.

An interesting clinical finding occurred as a result of applying the importance ranking methods to the NHSBT dataset in Chapter 3. The CRF algorithm found that the duration of the treatment withdrawal period was not important for predicting DGF, but features relating to physiological profiles throughout this phase are (such as the time is takes the oxygen saturation to fall below 70%). Out of the three oxygen saturation variables (time it took to drop to 70, 80 and 90%), the time it took to drop to 70% was the most predictive, followed by 80 then 90%. This is consistent with the conjecture made by Bradley et al. (2013), whom claim that too much attention in practice is given to the duration rather than characteristics of the physiological profiles. They stated the requirement for further research to support this claim.

We recommend the CRF importance ranking method for the purpose of exploratory data analysis, but note that unlike standard statistical methods it is difficult to quantify the level of certainty of the findings. Despite this, variables found to be important for predicting DGF (in descending order of importance) include: recipient status of dialysis at the time of transplant, recipient transplant centre, CIT, donor age, recipient ethnicity, donor weight, ethnicity mismatch, the time it takes oxygen saturation to drop below 70%, the time it takes SBP to drop below 60 mmHg and the time it takes oxygen saturation to drop below 80%. Many other variables appear to improve predictive performance by a small amount as can be seen in Figure 3.4.

In Chapter 3 we proposed the use of a method of visualisation helpful for interpreting the ranks of variable importance when multiple imputation has been performed. We proposed the use of a heat-map (see Figure 3.5) that displays how many times each variable received a particular rank across the number of imputed datasets. A higher confidence can be given to the ranking of variables that received the same rank many times. In our application the variables ranked as the top five were assigned the same rank for each dataset, indicating that these ranks are assigned with confidence and are robust to the noise resulting from the imputation process.

In Chapter 4 we proposed the use of the MBJM to dynamically predict when donors reach asystole throughout the treatment withdrawal to death period. We began by using an extension of the RF suitable for survival data to determine baseline covariates that are important for predicting the time of asystole (age and height were found to be important). We then performed a model selection procedure using 5-fold cross-validation to determine the discrimination and calibration performance measures that were used as a basis to compare models. We found that using functions of SBP and oxygen saturation as well as the fact that the patient is still alive at the time of prediction, resulted in an encouraging discriminatory ability for predicting asystole at one hour from 15 minutes. We tried including the baseline covariates found to be important from the RSF in the MBJM, but this information appeared to have been already captured by the longitudinal covariates, and having to estimate these additional parameters somewhat reduced the predictive performance overall. The encouraging predictive ability of the chosen model is attributed to employing an extension recently proposed by Andrinopoulou et al. (2018) that allowed the association structure between oxygen saturation and the survival process to be modelled flexibly over time. In the remainder of Chapter 4, we conducted a further validation of the chosen model with a 15 repeated 5-fold cross-validation and assessed the convergence of the sampler, which were both satisfactory. We presented the chosen model and illustrated how dynamic prediction can be performed for a new subject that has just been withdrawn from their life-support machine.

In Chapter 5 we investigated a two-stage approach for regressing a specified function of a longitudinal covariate (that constitutes an endogenous time-dependent covariate in a survival model) against the odds of success for a binary outcome. In our case, stage one involved fitting a JM to the longitudinal covariate (such as SBP or oxygen saturation) as well as the survival outcome (the event time and censoring indicator of the donor). In stage two a specified function of the curve fitted by the JM (such as the AUC) was used as a covariate in a RILR to see how it is associated with the chances of a recipient experiencing DGF. We conducted a simulation study to compare the inferential properties of the proposed approach compared to the alternative which involved fitting a LMEM instead of the JM and using the fitted function as a covariate in the classification model. We found that the LMEM and JM approaches performed almost equivalently (as expected both better than the observed trajectory approach), indicating that the less computationally intensive method (LMEM) sufficed and was therefore employed in our analysis.

Our analysis in Chapter 5 revealed various clinical findings of interest. In particular, after a rigorous model selection procedure, the chosen model retained estimated characteristics of the physiological variables, but including the duration of the treatment withdrawal period added no value to the model (in terms of goodness-of-fit or discriminatory ability). This was consistent with findings in Chapter 3. It was found from this analysis that the risk of a recipient experiencing DGF depends on whether the recipient was on dialysis

at the time of transplant, gender, the oxygen saturation level at the time of withdrawal and the rate of decline of HR. Those with an oxygen saturation above 96% at the time of treatment withdrawal (that also had a rapid decline of HR) resulted in the best chances of an immediately functioning graft. Moreover, it was notable that donors with a treatment withdrawal duration longer than what has previously been studied resulted in both recipients experiencing an immediately functioning graft. This suggests that restrictive protocols prohibiting transplantation once the duration of treatment withdrawal is prolonged could be stopping transplants that would have otherwise been successful. However, we emphasise that further validation of these findings is required before kidney transplant policy can be amended and changes are implemented in clinical practice.

6.2 Assumptions and Limitations

Throughout this thesis as much effort as was possible was made to avoid making restrictive assumptions, which played a large role in the choice of methods and steered a large part of this work towards predictive modelling rather than explanatory modelling. However, it was not always possible to avoid making assumptions and various limitations apply as we now discuss.

As it was discussed in Section 4.5 implicit assumptions were made relating to the JM. This applies for all parts of this work where the JM was used. We assumed that the visiting process was not related to the event times, which we claim to be plausible in this application as no attempt was made to resuscitate the terminally ill patients. The JM also assumes that the random effects follow a multivariate normal distribution. Rizopoulos et al. (2008) show that the JM is fairly robust to this assumption, especially when the number of repeated measures is large, which is the case for most patients in our application.

We note an important limitation relating to Chapter 4, where we employed the MBJM to dynamically predict the time of DCD donor asystole once withdrawn from their life-support machine. This method can only be used to dynamically predict survival probabilities once a set of physiological variable measurements have been obtained. We acknowledge that in practice it is preferable to have a predicted time of asystole before the donor is withdrawn from their life support machine, however, we believe that this method still has scope to improve medical practice and can serve as an effective guide for clinicians and can assist with managing family expectations. If no longitudinal data is fed into the trained model for a prediction to be made, the MBJM defaults to a proportional hazards model. In our analysis, a limited number of baseline characteristic variables were available to make such predictions and the model was trained on only longitudinal measurements.

The physiological profiles in the datasets analysed in Chapters 4 and 5 were recorded at irregular intervals and the number of measurements between donors varied substantially. These data complications as well as the erratic nature of these longitudinal variables meant that various possible surrogates such as the time it takes the physiological variable to drop to a certain threshold would be highly biased and inaccurate when calculated from the data. For this reason, in this work we limited the potential surrogates to the observed and estimated intercept, slope and AUC. If these data complications were not present it would have been interesting to have calculated warm ischaemic time from the data available and to see how our findings change after including this variable in the regression model. We suggest that this is carried out as future work (as discussed in Section 6.4).

The analysis in Chapter 5 was limited by a relatively small sample size, and only a subset of variables were available that have been found to be predictive of DGF in the literature. We expect that the small sample size explains the lack of evidence of an association between the variables cold ischaemic time and donor age with transplant outcome DGF, which has been found many times in previous studies. This arguably makes our findings that relate to these variables questionable.

6.3 Conclusion

Taking into account the assumptions and limitations discussed in Section 6.2, we now discuss the final conclusions drawn from this work. We found that machine learning methods offer a variety of tools that proved to be beneficial in the analysis of the NHSBT dataset. These methods allow restrictive assumptions required by explanatory methods to be relaxed at the cost of interpretability. We encourage the use of these methods in future related analyses in conjunction with explanatory models. The predictive performance of these methods were comparable and for some methods (XGBoost) improved for certain metrics. These methods offer the ability to rank the importance of variables in the model. Alternative methods of interpretation through visualisation offer valuable insight into the predictive structure of the highly flexible model. We note that the choice of variable importance metric is important, as the Gini importance metric proved to be a poor measure of importance in our application (especially when variables with many categories are present), yet the permuted importance index was able to detect important variables successfully.

The MBJM achieved an encouraging predictive performance in our application, which was mainly attributed to the recent extensions to the JM that relax the assumption of a linear association between the longitudinal variable and the hazard as time progresses. Moreover, the ability to model functions of the longitudinal variable such as the current

gradient and also the ability to include multiple longitudinal covariates improved the model's predictive performance.

Although some variables that we suspect to be predictive of DGF were not available in the motivating dataset, we were able to draw important insight from the model fitted to improve our understanding of how characteristics of the treatment withdrawal phase are related to the chances of a recipient experiencing a DGF. We found that the dialysis status at the time of transplant, gender, the oxygen saturation at the time of withdrawal and the slope of HR were variables that impacted the chances of a successful transplant.

6.4 Future Work

Although a range of sophisticated methods have been employed throughout this thesis to address various clinical research questions, we have been limited by computational resources and available data (in terms of both quantity and quality) to fully address all questions of interest, which provides a wide scope for future work. In particular, we have restricted this work to focus on the outcome of DGF, due to such a low count of events, which would inevitably limit statistical power to detect associations between variables of interest with the recipient survival outcome should they exist. As time progresses more events (recipient deaths) will occur in the NHSBT data, allowing for a formal analysis of how characteristics of physiological variables in the treatment withdrawal period are associated with recipient survival outcome. As the data quality and quantity improve over time, the opportunity for insightful analysis increases.

To the best of our knowledge, Chapter 4 is the first analysis of its kind in transplantation medicine. More specifically, the application of joint modelling and its various extensions (multiple longitudinal covariates, flexible association structure, parametrisations that allow the hazard to depend on the current slope of the biomarker) for using physiological variables to dynamically predict survival probabilities of DCD donors throughout the withdrawal period. The encouraging results achieved in this analysis should spark interest from both researchers and stakeholders that may lead to more ideas for novel applications of these methods.

There is scope for future work by validating our findings externally. In addition, there is potential for improved results through the implementation of more JM extensions that have been provided in the literature, such as modelling competing risks (corresponding to various events in the withdrawal period, such as asystole and death). Future studies may have access to more data that has improved quality. More combinations of models (for example, including baseline covariates in both the survival and longitudinal components of the JM) may also lead to an improved predictive performance. Finally, our analysis could have been improved by a rigorous approach to imputing the missing data in the longitudinal covariates, which provides further scope for future work.

Our analysis in Chapter 5 provides a strong starting point for improving our understanding of how characteristics of physiological variables in the treatment withdrawal period impact transplant outcome, by considering the intercept, slope and AUC surrogates. The limitations discussed in Section 6.2 restrict the number of potential surrogates that could be used. A future study could include the time that it takes the physiological variables to drop to a certain threshold as surrogates, and determine the optimal threshold for predicting transplant outcome. Novel ideas for potential surrogates is valuable in this work. It is of interest to include warm ischaemic time in the model and to see if our findings still hold. Moreover, this analysis could be repeated for the survival outcome (replacing the RILR model with a Cox frailty model), when there is a plausible number of events incurred to estimate the parameters effectively.

6.5 Software

This work was implemented using R version 3.5.1. Code is available on demand. The University of Southampton High Performance Computing facility Iridis 4 was used for highly computational tasks, including both simulation studies, tuning the hyperparameters, cross-validation and rigorous model selection. Machine learning was performed using the `mlr` package and joint models were fitted using the `JMbayes` package. We selected the Hamiltonian Monte Carlo option for Bayesian estimation which calls the Bayesian sampling software `stan` (which is called using wrapper functions available in the `rstan` package).

Appendix A

Appendices

Table A.1: *Description of variables in the NHSBT dataset with corresponding code name and variable type.*

	Variable	Type	Description
1	dbg	factor	Donor blood group
2	DGF	factor	Function of kidney post transplant in first week
3	dage	integer	Donor age (years)
4	dcod_grp	factor	Donor cause of death grouping
5	dsex	factor	Donor gender
6	dethnic	factor	Donor ethnicity
7	dweight	numeric	Donor weight (kg)
8	dheight	integer	Donor height (cm)
9	sbp70time	integer	Time (in minutes) for SBP to reach 70mmHg
10	sbp60time	integer	Time (in minutes) for SBP to reach 60mmHg
11	sbp50time	integer	Time (in minutes) for SBP to reach 50mmHg
12	o2sat90time	integer	Time (in minutes) for oxygen saturation to reach 90
13	o2sat80time	integer	Time (in minutes) for oxygen saturation to reach 80
14	o2sat70time	integer	Time (in minutes) for oxygen saturation to reach 70
15	deathtime	integer	Time (in minutes) from treatment withdrawal to death
16	surgerytime	integer	Time (in minutes) from treatment withdrawal to surgery
17	CIT_MINS	integer	Cold ischaemic time (in minutes) of donor kidney
18	RWEIGHT	numeric	Recipient weight (kg)
19	no.tx.Freq	integer	Number of kidneys donated by donor
20	REC_UNIT	factor	Recipient transplant centre
21	RSEX	factor	Recipient gender
22	RBG	factor	Recipient blood group
23	RETHNIC	factor	Recipient ethnicity
24	DIAL_AT_TX	factor	Recipient on dialysis at time of transplant
25	recip_prd	factor	Recipient on dialysis at time of transplant
26	GMM	factor	Gender mismatch
27	EMM	factor	Ethnicity mismatch
28	BGC	factor	Blood group compatibility

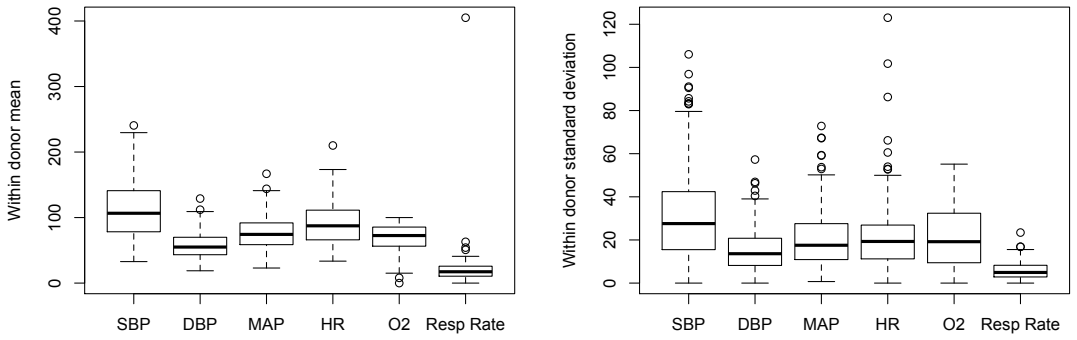


Figure A.1: *Box-plots displaying the summary statistics for the mean of the within donor profiles for each longitudinal covariate.*

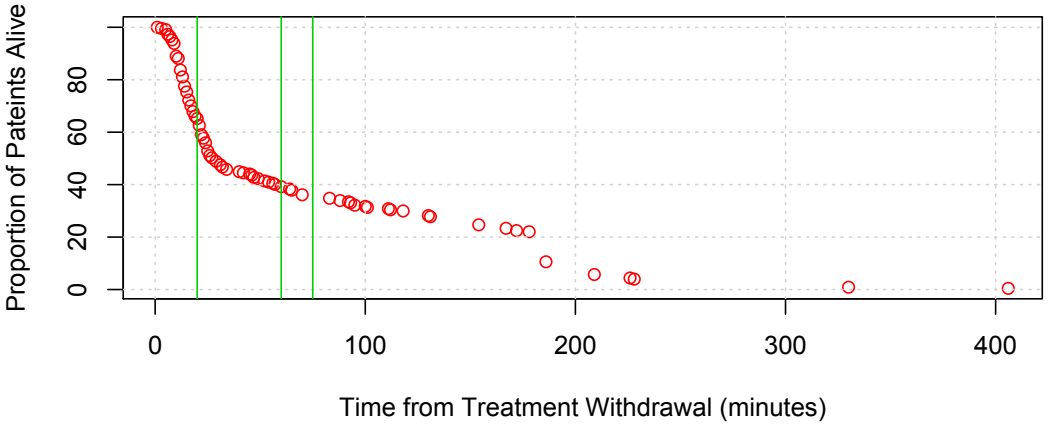


Figure A.2: *The proportion of donors remaining throughout the treatment withdrawal to death phase. The vertical green lines mark the 20, 60 and 75 minute marks, which are the times that predictions are made in Section 4.4.0.3.*

Bibliography

Akl, A., Ismail, A. M. & Ghoneim, M. (2008), 'Prediction of graft survival of living-donor kidney transplantation: nomograms or artificial neural networks?', *Transplantation* **86**(10), 1401–1406.

Albert, P. S. & Shih, J. H. (2010), 'On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure', *Biometrics* **66**(3), 983–987.

Allen, M., Billig, E., Reese, P., Shults, J., Hasz, R., West, S. & Abt, P. (2016), 'Donor hemodynamics as a predictor of outcomes after kidney transplantation from donors after cardiac death', *American Journal of Transplantation* **16**(1), 181–193.

Alvarez, I., Niemi, J. & Simpson, M. (2014), 'Bayesian inference for a covariance matrix', *arXiv preprint arXiv:1408.4050* .

Andrews, P., Burnapp, L. & Manas, D. (2014), 'Summary of the british transplantation society guidelines for transplantation from donors after deceased circulatory death', *Transplantation* **97**(3), 265–270.

Andrinopoulou, E.-R., Eilers, P. H., Takkenberg, J. J. & Rizopoulos, D. (2018), 'Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using p-splines', *Biometrics* **74**(2), 685–693.

Barnard, J., McCulloch, R. & Meng, X.-L. (2000), 'Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage', *Statistica Sinica* **10**, 1281–1311.

Bell, R. M., Koren, Y. & Volinsky, C. (2010), 'All together now: A perspective on the netflix prize', *Chance* **23**(1), 24–29.

Bellingham, J. M., Santhanakrishnan, C., Neidlinger, N., Wai, P., Kim, J., Niederhaus, S., Leverson, G. E., Fernandez, L. A., Foley, D. P., Mezrich, J. D. et al. (2011), 'Donation after cardiac death: a 29-year experience', *Surgery* **150**(4), 692–702.

Bergler, T. & Hutchinson, J. A. (2017), 'Tools for predicting kidney transplant outcomes', *Transplantation* **101**(9), 1958–1959.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. & Jones, Z. M. (2016), 'mlr: Machine learning in r', *The Journal of Machine Learning Research* **17**(1), 5938–5942.

Bradley, J., Pettigrew, G. & Watson, C. (2013), 'Time to death after withdrawal of treatment in donation after circulatory death (dcd) donors', *Current opinion in organ transplantation* **18**(2), 133–139.

Breiman, L. (1996), 'Bagging predictors', *Machine learning* **24**(2), 123–140.

Breiman, L. (2001*a*), 'Random forests', *Machine Learning* **45**(1), 5–32.
URL: *https://doi.org/10.1023/A:1010933404324*

Breiman, L. (2001*b*), 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)', *Statistical science* **16**(3), 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth.

Brier, M. E., Ray, P. C. & Klein, J. B. (2003), 'Prediction of delayed renal allograft function using an artificial neural network', *Nephrology Dialysis Transplantation* **18**(12), 2655–2659.

Brown, E. R. & Ibrahim, J. G. (2003), 'A bayesian semiparametric joint hierarchical model for longitudinal and survival data', *Biometrics* **59**(2), 221–228.

Brown, E. R., Ibrahim, J. G. & DeGruttola, V. (2005), 'A flexible b-spline model for multiple longitudinal biomarkers and survival', *Biometrics* **61**(1), 64–73.

Chatrchyan, S., Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Aguilo, E., Bergauer, T., Dragicevic, M., Erö, J., Fabjan, C. et al. (2012), 'Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc', *Physics Letters B* **716**(1), 30–61.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', ACM, pp. 785–794.

Chen, T. & He, T. (2015), Higgs boson discovery with boosted trees, *in* 'NIPS 2014 workshop on high-energy physics and machine learning', pp. 69–80.

Chen, T., He, T., Benesty, M., Khotilovich, V. & Tang, Y. (2015), 'Xgboost: extreme gradient boosting', *R package version 0.4-2* pp. 1–4.

Chi, Y.-Y. & Ibrahim, J. G. (2006), 'Joint models for multivariate longitudinal and multivariate survival data', *Biometrics* **62**(2), 432–445.

Clayton, P., McDonald, S., Snyder, J., Salkowski, N. & Chadban, S. (2014), 'External validation of the estimated posttransplant survival score for allocation of deceased donor kidneys in the united states', *American Journal of Transplantation* **14**(8), 1922–1926.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American statistical association* **74**(368), 829–836.

Cox, D. R. (1958), 'The regression analysis of binary sequences', *Journal of the Royal Statistical Society: Series B (Methodological)* **20**(2), 215–232.

Cox, D. R. (1972), 'Regression models and life-tables (with discussion)', *J Roy Statist Soc* **34**, 187–220.

D'Alessandro, A. M., Fernandez, L. A., Chin, L. T., Shames, B. D., Turgeon, N. A., Scott, D. L., Di, A. C., Becker, Y. T., Odorico, J. S., Knechtle, S. J. et al. (2004), 'Donation after cardiac death: the university of wisconsin experience.', *Annals of transplantation* **9**(1), 68–71.

Davila, D., Ciria, R., Jassem, W., Briceño, J., Littlejohn, W., Vilca-Meléndez, H., Srinivasan, P., Prachalias, A., O'grady, J., Rela, M. et al. (2012), 'Prediction models of donor arrest and graft utilization in liver transplantation from maastricht-3 donors after circulatory death', *American Journal of Transplantation* **12**(12), 3414–3424.

De Boor, C. (2001), *A practical guide to splines*, revised edn, New York: Springer-Verlag.

de Groot, Y. J., Lingsma, H. F., Bakker, J., Gommers, D. A., Steyerberg, E. & Kompanje, E. J. (2012), 'External validation of a prognostic model predicting time of death after withdrawal of life support in neurocritical patients', *Critical care medicine* **40**(1), 233–238.

Decruyenaere, A., Decruyenaere, P., Peeters, P., Vermassen, F., Dhaene, T. & Couckuyt, I. (2015), 'Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods', *BMC medical informatics and decision making* **15**(1), 83.

Diaconis, P. & Efron, B. (1983), 'Computer-intensive methods in statistics', *Scientific American* **248**(5), 116–131.

Díaz-Uriarte, R. & De Andres, S. A. (2006), 'Gene selection and classification of microarray data using random forest', *BMC bioinformatics* **7**(1), 3.

Diggle, P. & Kenward, M. G. (1994), 'Informative drop-out in longitudinal data analysis', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **43**(1), 49–73.

Dudoit, S., Fridlyand, J. & Speed, T. P. (2002), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *Journal of the American statistical association* **97**(457), 77–87.

Eilers, P. H. & Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical science* **11**(2), 89–102.

Evgeniou, T., Pontil, M. & Poggio, T. (2000), 'Statistical learning theory: A primer', *International Journal of Computer Vision* **38**(1), 9–13.

Fisher, A., Rudin, C. & Dominici, F. (2018), 'All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously'.

Foucher, Y., Daguin, P., Akl, A., Kessler, M., Ladrière, M., Legendre, C., Kreis, H., Rostaing, L., Kamar, N., Mourad, G. et al. (2010), 'A clinical scoring system highly predictive of long-term kidney graft survival', *Kidney international* **78**(12), 1288–1294.

Freund, Y. & Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer and system sciences* **55**(1), 119–139.

Friedman, J. H. (2002), 'Stochastic gradient boosting', *Computational statistics & data analysis* **38**(4), 367–378.

Friedman, J. H. et al. (2001), 'Greedy function approximation: A gradient boosting machine.', *The Annals of Statistics* **29**(5), 1189–1232.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2000), 'Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)', *The annals of statistics* **28**(2), 337–407.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', *Bayesian analysis* **1**(3), 515–534.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013), *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.
URL: *https://books.google.co.uk/books?id=ZXL6AQAAQBAJ*

Gelman, A., Rubin, D. B. et al. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* **7**(4), 457–472.

Geurts, P. (2002), Contributions to decision tree induction: bias/variance tradeoff and time series classification, PhD thesis, University of Liège Belgium.

Goldstein, H., Carpenter, J., Kenward, M. G. & Levin, K. A. (2009), 'Multilevel models with multivariate mixed response types', *Statistical Modelling* **9**(3), 173–197.

Greco, R., Papalia, T., Lofaro, D., Maestripieri, S., Mancuso, D. & Bonofiglio, R. (2010), Decisional trees in renal transplant follow-up, *in* 'Transplantation proceedings', Vol. 42, Elsevier, pp. 1134–1136.

Harrell, F. E. (2015), *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*, 2 edn, New York: Springer.

Hastie, T. & Tibshirani, R. (1987), 'Generalized additive models: some applications', *Journal of the American Statistical Association* **82**(398), 371–386.

Hastie, T., Tibshirani, R. & Friedman, J. (2008), *The elements of statistical learning: data mining, inference and prediction*, 2 edn, New York: Springer. URL: `https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf`. Last visited on 13/11/2018.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M. & Larochelle, H. (2017), 'Brain tumor segmentation with deep neural networks', *Medical image analysis* **35**, 18–31.

Henderson, R., Diggle, P. & Dobson, A. (2000), 'Joint modelling of longitudinal measurements and event time data', *Biostatistics* **1**(4), 465–480.

Henderson, R., Diggle, P. & Dobson, A. (2002), 'Identification and efficacy of longitudinal markers for survival', *Biostatistics* **3**(1), 33–50.

Hernández, D., Rufino, M., Bartolomei, S., Lorenzo, V., González-Rinne, A. & Torres, A. (2005), 'A novel prognostic index for mortality in renal transplant recipients after hospitalization', *Transplantation* **79**(3), 337–343.

Ho, K. J., Owens, C. D., Johnson, S. R., Khwaja, K., Curry, M. P., Pavlakis, M., Mandelbrot, D., Pomposelli, J. J., Shah, S. A., Saidi, R. F. et al. (2008), 'Donor postextubation hypotension and age correlate with outcome after donation after cardiac death transplantation', *Transplantation* **85**(11), 1588–1594.

Ho, T. K. (1998), Nearest neighbors in random subspaces, *in* 'Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)', Springer, pp. 640–648.

Hogan, J. W. & Laird, N. M. (1997), 'Model-based approaches to analysing incomplete longitudinal and failure time data', *Statistics in medicine* **16**(3), 259–272.

Hothorn, T., Hornik, K. & Zeileis, A. (2006), 'Unbiased recursive partitioning: A conditional inference framework', *Journal of Computational and Graphical statistics* **15**(3), 651–674.

Huang, X., Stefanski, L. A. & Davidian, M. (2009), 'Latent-model robustness in joint models for a primary endpoint and a longitudinal process', *Biometrics* **65**(3), 719–727.

Hwang, Y.-T., Wang, C.-C., Wang, C. H., Tseng, Y.-K. & Chang, Y.-J. (2015), 'Joint model of multiple longitudinal measures and a binary outcome: An application to predict orthostatic hypertension for subacute stroke patients', *Biometrical Journal* **57**(4), 661–675.

Ibrahim, J. G., Chu, H. & Chen, L. M. (2010), 'Basic concepts and methods for joint models of longitudinal and survival data', *Journal of Clinical Oncology* **28**(16), 2796–2801.

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

Irish, W. D., McCollum, D. A., Tesi, R. J., Owen, A. B., Brennan, D. C., Bailly, J. E. & Schnitzler, M. A. (2003), 'Nomogram for predicting the likelihood of delayed graft function in adult cadaveric renal transplant recipients', *Journal of the American Society of Nephrology* **14**(11), 2967–2974.

Irish, W., Ilsley, J., Schnitzler, M., Feng, S. & Brennan, D. (2010), 'A risk prediction model for delayed graft function in the current era of deceased donor renal transplantation', *American journal of transplantation* **10**(10), 2279–2286.

Ishwaran, H. & Kogalur, U. B. (2007), 'Random survival forests for r', *R news* **7**(2), 25–31.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. et al. (2008), 'Random survival forests', *The annals of applied statistics* **2**(3), 841–860.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An introduction to statistical learning*, New York: Springer. URL: `https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`. Last visited on 10/11/2018.

Johnson, R. J., Fuggle, S. V., Mumford, L., Bradley, J. A., Forsythe, J. L., Rudge, C. J., of NHS Blood, K. A. G., Transplant et al. (2010), 'A new uk 2006 national kidney allocation scheme for deceased heart-beating donor kidneys', *Transplantation* **89**(4), 387–394.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley.

Kearns, M. (1988), 'Learning boolean formulae or finite automata is as hard as factoring', *Technical Report TR-14-88 Harvard University Aikem Computation Laboratory* .

Kotsopoulos, A., Böing-Messing, F., Jansen, N., Vos, P. & Abdo, W. (2018), 'External validation of prediction models for time to death in potential donors after circulatory death', *American Journal of Transplantation* **18**(4), 890–896.

Krikov, S., Khan, A., Baird, B. C., Barenbaum, L. L., Leviatov, A., Koford, J. K. & Goldfarb-Rumyantzev, A. S. (2007), 'Predicting kidney transplant survival using tree-based modeling', *Asaio Journal* **53**(5), 592–600.

Laird, N. M. & Ware, J. H. (1982), 'Random effects models for longitudinal data', *Biometrics* **38**(4), 963–974.

Lang, S. & Brezger, A. (2004), 'Bayesian p-splines', *Journal of computational and graphical statistics* **13**(1), 183–212.

Lasserre, J., Arnold, S., Vingron, M., Reinke, P. & Hinrichs, C. (2011), 'Predicting the outcome of renal transplantation', *Journal of the American Medical Informatics Association* **19**(2), 255–262.

Lavalley, M. P. & DeGruttola, V. (1996), 'Models for empirical bayes estimators of longitudinal cd4 counts', *Statistics in Medicine* **15**(21), 2289–2305.

Lewandowski, D., Kurowicka, D. & Joe, H. (2009), 'Generating random correlation matrices based on vines and extended onion method', *Journal of multivariate analysis* **100**(9), 1989–2001.

Li, B., Cairns, J. A., Robb, M. L., Johnson, R. J., Watson, C. J., Forsythe, J. L., Oniscu, G. C., Ravanan, R., Dudley, C., Roderick, P. et al. (2016), 'Predicting patient survival after deceased donor kidney transplantation using flexible parametric modelling', *BMC nephrology* **17**(1), 51.

Lin, R. S., Horn, S. D., Hurdle, J. F. & Goldfarb-Rumyantzev, A. S. (2008), 'Single and multiple time-point prediction models in kidney transplant outcomes', *Journal of biomedical informatics* **41**(6), 944–952.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R. & Lipshultz, S. (2002), 'Parameter estimation in longitudinal studies with outcome-dependent follow-up', *Biometrics* **58**(3), 621–630.

Little, R. J. (1988), 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics* **6**(3), 287–296.

Little, R. J. (1993), 'Pattern-mixture models for multivariate incomplete data', *Journal of the American Statistical Association* **88**(421), 125–134.

Little, R. J. (1995), 'Modeling the drop-out mechanism in repeated-measures studies', *Journal of the american statistical association* **90**(431), 1112–1121.

Louppe, G. (2014), Understanding Random Forests from Theory to Practice, PhD thesis.

Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. (2013), Understanding variable importances in forests of randomized trees, *in* 'Advances in neural information processing systems', Springer, pp. 431–439.

Manara, A., Murphy, P. & O'Callaghan, G. (2012), 'Donation after circulatory death', *British Journal of Anaesthesia* **108**(S1), i108–i121.

Mantel, N. (1966), 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemother Rep* **50**, 163–170.

Mauff, K., Steyerberg, E., Kardys, I., Boersma, E. & Rizopoulos, D. (2018), 'Joint models with multiple longitudinal outcomes and a time-to-event outcome', *arXiv preprint arXiv:1808.07719* .

Michalak, M., Wouters, K., Fransen, E., Hellemans, R., Van Craenenbroeck, A. H., Couttenye, M. M., Bracke, B., Ysebaert, D. K., Hartman, V., De Greef, K. et al. (2017), 'Prediction of delayed graft function using different scoring algorithms: A single-center experience', *World journal of transplantation* **7**(5), 260.

Molnar, M. Z., Nguyen, D. V., Chen, Y., Ravel, V., Streja, E., Krishnan, M., Kovesdy, C. P., Mehrotra, R. & Kalantar-Zadeh, K. (2017), 'Predictive score for posttransplantation outcomes', *Transplantation* **101**(6), 1353.

Morgan, J. N. & Sonquist, J. A. (1963), 'Problems in the analysis of survey data, and a proposal', *Journal of the American statistical association* **58**(302), 415–434.

Murawska, M., Rizopoulos, D. & Lesaffre, E. (2012), 'A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies', *Journal of Probability and Statistics* **2012**.

Nelder, J. A. & Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384.

Nesterov, Y. (2009), 'Primal-dual subgradient methods for convex problems', *Mathematical programming* **120**(1), 221–259.

NHSBT (2018), 'Kidney activity report'.
URL: *https://nhsbtdbe.blob.core.windows.net/umbraco-assets-corp/12055/section-5-kidney-activity.pdf*

Ojo, J., Olatayo, T., Alabi, O., Akaike, H., Box, G., Jenkins, G., Furnival, G., Hannan, E., Parzen, E., Priestely, M. et al. (1973), 'A new look at the statistical model identification.', *Asian Journal of Scientific Research* **1**(5), 255–265.

Peters-Sengers, H., Houtzager, J., Heemskerk, M., Idu, M., Minnee, R., Klaasen, R., Joor, S., Hagenaars, J., Rebers, P., van der Heide, J. H. et al. (2018), 'Dcd donor hemodynamics as predictor of outcome after kidney transplantation', *American Journal of Transplantation* .

Pitman, E. J. G. (1936), Sufficient statistics and intrinsic accuracy, *in* 'Mathematical Proceedings of the cambridge Philosophical society', Vol. 32, Cambridge University Press, pp. 567–579.

Plate, T. A. (1999), 'Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using gaussian process models', *Behaviormetrika* **26**(1), 29–50.

Pugin, D., Hechinger, S., Mamjou, H., Arnaud, E., Brousoz, S., Flatres, S., Freitas, C., Rennesson, C., Simon, J., Moretti, D. et al. (2017), 'Donation after cardiac death (dcd), comparative of scores to predict death', *Transplantation* **101**, S54.

Rabinstein, A. A., Yee, A. H., Mandrekar, J., Fugate, J. E., de Groot, Y. J., Kompanje, E. J., Shutter, L. A., Freeman, W. D., Rubin, M. A. & Wijdicks, E. F. (2012), 'Prediction of potential for organ donation after cardiac death in patients in neurocritical state: a prospective observational study', *The Lancet Neurology* **11**(5), 414–419.

Rao, P. S., Schaubel, D. E., Guidinger, M. K., Andreoni, K. A., Wolfe, R. A., Merion, R. M., Port, F. K. & Sung, R. S. (2009), 'A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index', *Transplantation* **88**(2), 231–236.

Ratcliffe, S. J., Guo, W. & Ten Have, T. R. (2004), 'Joint modeling of longitudinal and survival data via a common frailty', *Biometrics* **60**(4), 892–899.

Reid, A., Harper, S., Jackson, C. H., Wells, A., Summers, D., Gjorgjimajkoska, O., Sharples, L., Bradley, J. & Pettigrew, G. (2011), 'Expansion of the kidney donor pool by using cardiac death donors with prolonged time to cardiorespiratory arrest', *American journal of transplantation* **11**(5), 995–1005.

Rice, J. A. & Wu, C. O. (2001), 'Nonparametric mixed effects models for unequally sampled noisy curves', *Biometrics* **57**(1), 253–259.

Rizopoulos, D. (2009), 'Package 'bootstepaic''.

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

Rizopoulos, D. (2012*a*), 'Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule', *Computational Statistics & Data Analysis* **56**(3), 491–501.

Rizopoulos, D. (2012*b*), *Joint models for longitudinal and time-to-event data: With applications in R*, CRC Press.

Rizopoulos, D. & Ghosh, P. (2011), 'A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event', *Statistics in medicine* **30**(12), 1366–1380.

Rizopoulos, D., Molenberghs, G. & Lesaffre, E. M. (2017), 'Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking', *Biometrical Journal* **59**(6), 1261–1276.

Rizopoulos, D., Verbeke, G. & Molenberghs, G. (2008), 'Shared parameter models under random effects misspecification', *Biometrika* **95**(1), 63–74.

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I. & McGregor, G. (2005), 'Boosted decision trees as an alternative to artificial neural networks for particle identification', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **543**(2-3), 577–584.

Royston, P. & Altman, D. G. (1994), 'Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling', *Applied statistics* pp. 429–467.

Royston, P., Altman, D. G. & Sauerbrei, W. (2006), 'Dichotomizing continuous predictors in multiple regression: a bad idea', *Statistics in medicine* **25**(1), 127–141.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Scalea, J., Redfield, R., Arpali, E., Leverson, G., Bennett, R., Anderson, M., Kaufman, D., Fernandez, L., D'Alessandro, A., Foley, D. et al. (2017), 'Does dcd donor time-to-death affect recipient outcomes? implications of time-to-death at a high-volume center in the united states', *American journal of transplantation* **17**(1), 191–200.

Schapire, R. E. (1990), 'The strength of weak learnability', *Machine learning* **5**(2), 197–227.

Schiffner, J., Bischl, B., Lang, M., Richter, J., Jones, Z. M., Probst, P., Pfisterer, F., Gallo, M., Kirchhoff, D., Kühn, T. et al. (2016), 'mlr tutorial', *arXiv preprint arXiv:1609.06146* .

Schröppel, B. & Legendre, C. (2014), 'Delayed kidney graft function: from mechanism to translation', *Kidney international* **86**(2), 251–258.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R. & Khovanova, N. (2017), 'Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation', *Biomedical Signal Processing and Control* .

Shmueli, G. et al. (2010), 'To explain or to predict?', *Statistical science* **25**(3), 289–310.

Shoskes, D., Ty, R., Barba, L. & Sender, M. (1998), Prediction of early graft function in renal transplantation using a computer neural network., *in* 'Transplantation proceedings', Vol. 30, pp. 1316–1317.

Soleimani, H., Hensman, J. & Saria, S. (2018), 'Scalable joint models for reliable uncertainty-aware event prediction', *IEEE transactions on pattern analysis and machine intelligence* **40**(8), 1948–1963.

Souter, M. & Van Norman, G. (2010), 'Ethical controversies at end of life after traumatic brain injury: defining death and organ donation', *Critical care medicine* **38**(9), S502–S509.

Stan Development Team (2018), 'RStan: the R interface to Stan'. R package version 2.18.2.
URL: *http://mc-stan.org/*

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008), 'Conditional variable importance for random forests', *BMC bioinformatics* **9**(1), 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007), 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *BMC bioinformatics* **8**(1), 25.

Strobl, C., Malley, J. & Tutz, G. (2009), 'An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.', *Psychological methods* **14**(4), 323.

Summers, D. M., Johnson, R. J., Hudson, A., Collett, D., Watson, C. J. & Bradley, J. A. (2013), 'Effect of donor age and cold storage time on outcome in recipients of kidneys donated after circulatory death in the uk: a cohort study', *The Lancet* **381**(9868), 727–734.

Summers, D. M., Watson, C. J., Pettigrew, G. J., Johnson, R. J., Collett, D., Neuberger, J. M. & Bradley, J. A. (2015), 'Kidney donation after circulatory death (dcd): state of the art', *Kidney international* **88**(2), 241–249.

Suntharalingam, C., Sharples, L., Dudley, C., Bradley, J. & Watson, C. (2009), 'Time to cardiac death after withdrawal of life-sustaining treatment in potential organ donors', *American Journal of Transplantation* **9**(9), 2157–2165.

Sweeting, M. J. & Thompson, S. G. (2011), 'Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture', *Biometrical Journal* **53**(5), 750–763.

Thorogood, J., Houwelingen, J., Persijn, G., Zantvoort, F., Schreuder, G. et al. (1991), 'Prognostic indices to predict survival of first and second renal allografts.', *Transplantation* **52**(5), 831–836.

Tikhonov, A. & Arsenin, V. Y. (1977), 'Solution of iii-posed problems', *VH Winston & Sons, Washington, DC* .

Tiong, H., Goldfarb, D., Kattan, M., Alster, J., Thuita, L., Yu, C., Wee, A. & Poggio, E. (2009), 'Nomograms for predicting graft function and survival in living donor kidney transplantation based on the unos registry', *The Journal of urology* **181**(3), 1248–1255.

Tsiatis, A., Degruttola, V. & Wulfsohn, M. (1995), 'Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids', *Journal of the American Statistical Association* **90**(429), 27–37.

Vapnik, V. N. (1999), 'An overview of statistical learning theory', *IEEE transactions on neural networks* **10**(5), 988–999.

Vijayarani, S. & Dhayanand, S. (2015), 'Data mining classification algorithms for kidney disease prediction', *International Journal on Cybernetics & Informatics (IJCI)* **4**(4), 13–25.

Wahba, G. (1975), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **24**(5), 383–393.

Wang, C., Wang, N. & Wang, S. (2000), 'Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements', *Biometrics* **56**(2), 487–495.

Watson, C. J., Johnson, R. J., Birch, R., Collett, D. & Bradley, J. A. (2012), 'A simplified donor risk index for predicting outcome after deceased donor kidney transplantation', *Transplantation* **93**(3), 314–318.

Wind, J., Snoeijs, M. G., Brugman, C. A., Vervelde, J., Zwaveling, J., van Mook, W. N. & van Heurn, E. L. (2012), 'Prediction of time of death after withdrawal of life-sustaining treatment in potential donors after cardiac death', *Critical care medicine* **40**(3), 766–769.

Wu, M. C. & Bailey, K. (1988), 'Analysing changes in the presence of informative right censoring caused by death and withdrawal', *Statistics in Medicine* **7**(1-2), 337–346.

Wu, M. C. & Bailey, K. R. (1989), 'Estimation and comparison of changes in the presence of informative right censoring: conditional linear model.', *Biometrics* **45**(3), 939–955.

Wu, M. C. & Carroll, R. J. (1988), 'Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process', *Biometrics* pp. 175–188.

Wulfsohn, M. S. & Tsiatis, A. A. (1997), 'A Joint Model for Survival and Longitudinal Data Measured with Error Author', *Biometrics* **53**(1), 330–339.

Wyatt, J. (1995), 'Nervous about artificial neural networks?', *The Lancet* **346**(8984), 1175–1177.

Yarlagadda, S. G., Coca, S. G., Garg, A. X., Doshi, M., Poggio, E., Marcus, R. J. & Parikh, C. R. (2008), 'Marked variation in the definition and diagnosis of delayed graft function: a systematic review', *Nephrology Dialysis Transplantation* **23**(9), 2995–3003.

Ye, W., Lin, X. & Taylor, J. M. (2008), 'Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach', *Biometrics* **64**(4), 1238–1246.

Yoo, K. D., Noh, J., Lee, H., Kim, D. K., Lim, C. S., Kim, Y. H., Lee, J. P., Kim, G. & Kim, Y. S. (2017), 'A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study', *Scientific reports* **7**(1), 8904.

Young, G. A., Smith, R. L. et al. (2005), *Essentials of statistical inference*, Vol. 16, Cambridge University Press.

Yucel, R. M. (2011), 'Random covariances and mixed-effects models for imputing multivariate multilevel continuous data', *Statistical modelling* **11**(4), 351–370.