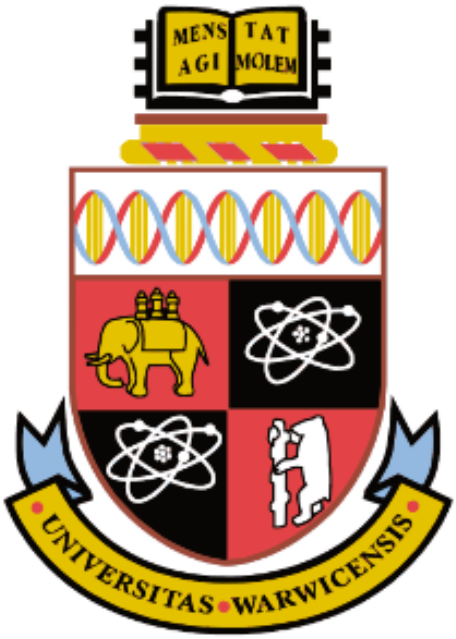


Dynamic Modelling of HLA-incompatible Renal Transplant Donor Specific Antibodies

by

Mason Reece Phillpott

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

SCHOOL OF ENGINEERING

April 2022

Table of Contents

List of Tables

List of Figures

Abbreviations

ABMR antibody mediated rejection.

AIC Akaike information criterion.

AiT antibody-incompatible transplantation.

AKI acute kidney injury.

ANOVA analysis of variance.

APC antigen presenting cell.

ARI adjusted rand index.

ARMA auto-regressive moving average.

BIC Bayesian information criterion.

BTS British Transplant Society.

C correct classification rate.

CDC complement dependent cytotoxic.

CI class I.

CI&II CI and CII.

CII class II.

CKD chronic kidney disease.

CV coefficient of variation.

DDT deceased donor transplant.

DFPP double filtration plasmapheresis.

DM distance matrix.

DSA donor specific antibody.

DTW dynamic time warping.

ECTS early classification of time series.

eGFR estimated Glomerular Filtration Rate.

ELISA enzyme-linked immuno sorbent assay.

ERF established renal failure.

ESRD end stage renal disease.

FC flow cytometry.

FN false negative.

FP false positive.

GF graft-failure.

GoF goodness of fit.

HLA human leukocyte antigen.

HLAi HLA-incompatible.

HR hazard ratio.

HSP highly sensitised patient.

iDSA immuno-dominant DSA.

Ig immunoglobulin.

IVIg intravenous immuno globulin.

LASSO least absolute shrinkage and selection operator.

LCM local cost matrix.

LCSS longest common sub-sequence.

LDT living donor transplant.

LOESS locally weighted scatter-plot smoothing.

LOOCV leave-one-out cross validation.

LQ lower quartile.

LS least squares.

MAP maximum-a-posteriori.

MAR missing at random.

MC Monte Carlo.

MCAR missing completely at random.

MCMC Markov chain Monte Carlo.

MFI mean fluorescent intensity.

ML machine learning.

ML maximum likelihood.

MNAR missing not at random.

MPL minimum predictive length.

NHS national health service.

NK natural killer.

no-GF no graft-failure.

NPV negative predictive value.

NRMSE normalised root mean squared error.

ODE ordinary differential equation.

PAA piece-wise aggregate approximation.

PDF probability density function.

PE phycoerythrin.

PL partial likelihood.

PPV positive predictive value.

ROC receiver operating characteristic.

RRT renal replacement therapy.

SAB single antigen bead.

SAX symbolic aggregate approximation.

SKB Sakoe-Chiba band.

Sn sensitivity.

Sp specificity.

SS summed square.

TCMR T-cell mediated rejection.

tDSA total DSA.

TN true negative.

TP true positive.

trx transplant.

UHCW University Hospitals of Coventry and Warwickshire.

UKLKSS UK living kidney sharing scheme.

UQ upper quartile.

VB variational Bayesian.

Acknowledgements

This thesis has been a challenging, however ultimately rewarding experience - with many people to thank along the way. Here I will name but a few who have made the largest impact and allowed me to produce my best possible work.

Firstly, my supervisors, Drs. Natasha Khovanova and Sunil Daga, for without whom this work would not have been possible. Over the years they have committed countless hours of their time and energy to discuss, revise and educate me towards achieving success in the biomedical engineering field. Perhaps most importantly, their positivity and excitement in my work was instrumental in me overcoming the most difficult junctures - for this I would like to express my gratitude to them both.

Secondly, to Dr. Manuel Eichenlaub, a colleague become friend who, despite graduating and moving abroad, was always available for discussion, advice and proofreading of my work - your time spent will not be forgotten.

Thirdly, I would like to extend my appreciation to the EPSRC, the Renal Transplant UHCW NHS Trust, and all those involved for the initiation and funding of this project. It has been a pleasure to collaborate with these people on such an excellent data set.

Special thanks also go out to all of my good friends that have made the PhD a far more enjoyable and memorable period of my life. This includes but is not limited to: Sean Perry, Ben Griffithsss, Peeps and Keeks. You all had a part to play and deserve your space on this list.

Last but not least, to my family and also my partner, Tarni, who put up with considerable periods of radio silence while I wrote this thesis - your patience and support throughout this process has been unwavering, and made a real difference throughout the years!

Declarations

This thesis is submitted to the University of Warwick in support of my application for degree of Doctor of Philosophy. The content is composed by me and has not been submitted in any previous application for any degree. From this work there have been five disseminations, two publications and a bursary award from three conferences.

Conference disseminations:

1. Plenary presentation: **M. Phillpott**, S. Daga, C. Imray, N. Krishnan, S. Fletcher, M. Leeson, R. Higgins, D. Lowe, D. Mitchell, D. Briggs, N. Khovanova. Monitoring patient DSA levels within six weeks post-HLAi kidney transplant can indicate the likelihood of failure. *British Transplant Society Annual Congress* (BTS19), Harrogate, 6^{th}-8^{th} March 2019.

2. Oral presentation: **M. Phillpott**, S. Daga, C. Imray, N. Krishnan, S. Fletcher, M. Leeson, R. Higgins, D. Lowe, D. Mitchell, D. Briggs, N. Khovanova. Identifying ideal monitoring periods post HLAi kidney transplantation to predict long term graft outcome. *5^{th} International Transplantation Conference*, Warwick University, 9^{th}-10^{th} May 2019.

3. Oral presentation: **M. Phillpott**, S. Daga, N. Krishnan, R. Higgins, D. Lowe, D. Briggs, N. Khovanova. Modelling early post-transplant antibody dynamics to predict transplant outcomes. *5^{th} International Transplantation Conference*, Warwick University, 9^{th}-10^{th} May 2019.

4. Oral presentation: **M. Phillpott**, S. Daga, C. Imray, N. Krishnan, S. Fletcher, M. Leeson, R. Higgins, D. Lowe, D. Mitchell, D. Briggs, N. Khovanova. Monitoring patient DSA levels within early stage post HLAi kidney transplant can indicate likelihood of graft failure. *European Society for Organ Transplantation*, Copenhagen, 15^{th}-18^{th} September 2019.

5. Oral presentation: **M. Phillpott**, M. Phillpott, S. Daga, C. Imray, N. Krishnan, S. Fletcher, M. Leeson, R. Higgins, D. Lowe, D. Mitchell, D. Briggs, N. Khovanova. Classification of post HLAi renal transplant antibody dynamics. *European Society for Organ Transplantation*, Copenhagen, 15^{th}-18^{th} September 2019.

6. Oral presentation: **M. Phillpott**, S. Daga, N. Krishnan, R. Higgins, D. Lowe, D.

Zehnder, D. Briggs, N. Khovanova. Unsupervised machine learning analysis of DSA levels after HLAi transplantation defines immunological activity and predicts outcomes. *One-lambda pre-EFI (European Federation for Immunogenetics) workshop*, 19th-20th April 2021.

Publications:

1. **M. Phillpott**, S. Daga, C. Imray, N. Krishnan, S. Fletcher, M. Leeson, R. Higgins, D. Lowe, D. Mitchell, D. Briggs, N. Khovanova. Dynamic time warping – an approach to grouping DSA responses in HLA-i kidney transplant patients. *Transplant International 32*, 179-179, 2019.

2. **M. Phillpott**, S. Daga, C. Imray, N. Krishnan, S. Fletcher, M. Leeson, R. Higgins, D. Lowe, D. Mitchell, D. Briggs, N. Khovanova. Monitoring patient DSA levels at specific days within early stage post HLA-I kidney transplant can indicate likelihood of graft failure. *Transplant International 32*, 119-119, 2019.

3. **M. Phillpott**, S. Daga, D. Lowe, N. Krishnan, D. Zehnder, R. Higgins, D. Briggs, N. Khovanova. Unsupervised machine learning identified fast and slow modulation of DSA dynamics following HLA-incompatible transplantation - 5 year outcomes. *Transplant International 34*, 124-124, 2021

4. **M. Phillpott**, S. Daga, R. Higgins, D. Lowe, N. Krishnan, D. Zehnder, D. Briggs, N. Khovanova. Dynamic behaviour of donor specific antibodies in the early period following HLA incompatible kidney transplantation. doi: 10.3389/ti.2022.10128, *Transplant International*. p. 11. Accepted. 03rd March 2022.

Awards:

1. Educational Bursary (£500) and dinner - Awarded by Kidney Research UK in collaboration with Chiesi Ltd. Best paper award is given to support the attendance to ESOT19 and was granted for the presentation listed above [4].

Abstract

End-stage renal disease (ESRD) is one of the leading causes of morbidity and mortality worldwide and its prevalence is only projected to increase as the average age and global population rises. By far the most effective treatment for ESRD is kidney transplantation with fully compatible donor and recipient human leukocyte antigen (HLA). Practically however, full compatibility is not always possible. Where a mismatch in HLA occurs an immune response may take place, leading to the formation of donor specific antibodies (DSA) which target the kidney. In this thesis analysis is performed on 110 HLA incompatible kidney transplants performed between 2003 and 2014 at the University Hospitals of Coventry and Warwickshire. The analysis aimed to address four main tasks: data processing, exploratory analysis, immune response classification, and dynamical modelling of DSA immune response. Under the exploratory analysis statistical techniques were utilised to investigate relationships between characteristic variables and daily post-transplant DSA levels with graft survival. Results of this work highlighted the significance of certain pre-transplant and post-transplant periods. For classification, a time series clustering technique was used to identify five early post-transplant DSA response types: no-response, fast-modulation, slow-modulation, rise-to-sustained and sustained. This discovery was a new contribution to literature and highlighted improved outcome for modulatory versus sustained outcome. Lastly, a set of parametric models were developed to describe the DSA post-transplant dynamic responses. Findings in this work reinforce the concept of higher DSA levels being associated with worse outcome, however, contribute further by suggesting that this relationship may be dependent on response type. Overall, these findings show that early period post-transplant monitoring can be a beneficial tool for clinicians in the monitoring of patient health. It has provided indicators towards the outcome of an incompatible kidney transplant. Future work should seek to validate these findings with a larger multi-centre cohort.

Chapter 1

Introduction

1.1 Motivation

End stage renal disease (ESRD) (or renal failure) is one of the significant causes of mortality and morbidity worldwide. It occurs due to a lack of kidney function which is primarily measured by the estimated Glomerular Filtration Rate (eGFR). Loss of function can occur quickly (days or weeks) in the case of acute kidney injury (AKI) or over an extended period of time (months or years) for chronic kidney disease (CKD), however, in all cases without medical intervention it will eventually lead to death. The scale of the problem is vast with an estimated 10% of the world population affected by CKD and millions dying each year due to lack of affordable treatment. CKD has also been shown by the 2010 global burden of disease study to be the fastest rising cause of death second only to that of HIV and AIDs [1].

Work by Liyanage et al. [2] revealed that in 2010 a total of 2.618 million people experienced renal replacement therapy (RRT) across the globe for treatment of ESRD. RRT is a collective term for a number of treatments such as dialysis, hemofiltration and hemodiafiltration [3]. However, of the available treatments, transplantation from a compatible donor is by far the most effective. Relative to dialysis-type treatments a systematic review of 110 studies [4] identified that transplantation offers superior survival[1] in addition to improved quality of life for ESRD patients.

From the perspective of the patient, receiving a donor organ is considered the ultimate step for treating ESRD [5]. Despite this there are major challenges to providing patients with donor organs - or even RRT treatment at all in developing countries. The UK Renal

[1] across lifetime period. Perioperative period shows increased hazard rate relative to dialysis treatment.

Registry report for 2010 shows, for example, that only 47% of RRT patients were transplant recipients with the remaining patients on various forms of dialysis [6]. The average wait time is 2.5 years (up to 3.5 for Black and Asian minorities) for a fully matched kidney [NHSBT 2018]. In developing countries the outlook is often considerably worse with fewer facilities and limited support from government for dialysis and transplant programs [7, 8].

One of the key barriers to transplantation is the lack of compatible donor organs. As a consequence, various countries have adopted different initiatives to improve the available transplant pool. The UK offers one of the most successful initiatives known as the UK living kidney sharing scheme (UKLKSS), put into action in April 2007. The scheme, developed at the University of Glasgow, implements a novel approach involving graph matching algorithms to link pairwise and 3-way exchanges of donor organs. In this system a willing donor may be incompatible for a certain recipient but will agree to be placed into a donor pool regardless. Their compatibility data is stored and an algorithm will attempt to form chains of compatible patients and recipients from the available pool. In a span between 2008 and 2017 UKLKSS algorithms led to 752 transplants taking place and saving the national health service (NHS) approximately £52M [9]. Additional schemes such as the 'Max and Kiera', or 'opt-out', law is also catching traction across the UK [2]. This law, which impacts all adults (except certain excluded groups), automatically agrees a person to donate organs when they die unless they specifically state otherwise [10]. In Wales this has seen consent rates for donation reach an all time high of 77%.

Despite the vast improvements in matching compatible donors and recipients seen in the last two decades there still exist a large number of ESRD patients for which identifying a compatible donor is a considerable challenge. In these circumstances patients may consider an antibody-incompatible transplantation (AiT). AiT outcomes are generally worse relative to compatible living donor transplant (LDT) transplantation - however are shown to be comparable to compatible deceased donor transplant (DDT) [11], and superior to waiting on RRT [12]. AiT is already greatly improving the lives for many ESRD patients, however it is only in its early stages of adoption and there is much left to be learnt about the underlying physiology determining short and long term outcomes. These gaps in knowledge offer a motivation for the work in this thesis, which ultimately aims to better understanding of risk in long term AiT outcomes.

[2] First Wales in 2015, England in 2020, Scotland 2021, Northern Ireland currently in legislative process.

1.2 Background

One of the main elements responsible for the success of a transplantation is the reaction of the immune system against the human leukocyte antigen (HLA) of the donor organ [13]. An ideal situation would be for donor and recipient to fully match HLA, ensuring that no immune response takes place against the transplanted organ. Indeed this was a requirement in the early days of transplantation [14–16]. Fortunately, developments in screening and availability of new drugs paved the way for looser compatibility rules [17–19]. One of the main challenges to overcome in HLA-incompatible (HLAi) transplantation is the immune response of antibodies produced by B-cells (white blood cell) targeting the donor organ. When directed at a donor organ, HLA specific antibodies are known as donor specific antibody (DSA) [20]. DSAs are usually formed due to previous exposure to the HLA such as through previous transplantation, blood transfusion, or pregnancy [21–23]. Their presence leads the immune system to be pre-sensitised to certain HLA - often presenting the risk of a faster and more aggressive immune response if encountered again.

Presence of DSA offers a substantial barrier to AiT, and is one that is carefully considered by clinicians in practice - even a small amount of DSA at the point of transplantation has been shown to compromise graft outcome [24]. High risk, or highly sensitised patient (HSP) - which have large amounts of circulating anti-human antibodies - must typically undergo immuno-suppressive treatment prior to transplantation, reducing antibody levels to below a pre-determined threshold [25]. This treatment is a challenging process and not without its risks. The main aim of immuno-suppressive treatment is to eliminate the pathogenic DSA targeting the donor graft. Current treatment types however are not sophisticated enough to target the specific DSA alone, often leaving the recipient vulnerable to opportunist infections and cancers. Current graft survival rates are notably lower for AiT cases due to these many complications. The British Transplant Society (BTS) notes that 3-year LDT graft survival rates are 94% for non-AiT and only 86% for HLAi- AiT cases [25]. As a consequence safe transplantation of patients with high levels of circulating DSA, i.e. HSP, is an ongoing problem and often results in increased waiting times for transplantation [13].

In recent years the value of screening for post-transplant DSA in addition to pre-transplant DSA has become more widely recognised [26, 27], and is subsequently gaining traction

as an important monitoring tool [25]. DSA can either form early following transplantation due to re-synthesis of preformed DSA, or develop some time (typically years) after transplantation towards the donor graft [28]. In the latter case DSA are referred to as *de novo* or 'new' if translated from Latin. Both preformed and de novo DSA are considered to have an important role in long term outcomes [29]. Preformed DSA following transplantation can become triggered towards the donor graft, stimulating large production of DSA that were previously removed. This highly active period is often associated with severe/early acute rejection towards the donor kidney, which has been shown to increase the risk of graft loss [25]. In contrast, presence of de novo DSA are associated with the longer term chronic rejection but similarly demonstrates worse longer term graft survival [28, 30, 31]. Despite these findings the underlying mechanisms governing DSA evolution are not fully understood and new research continues to generate unexpected findings. One of the most promising avenues for research is into early post-transplant DSA dynamic behaviour which has now seen a number of studies explore and demonstrate their link with clinical outcomes [25, 32–34]. Behaviour in this period can be highly dynamic, and varies greatly from case to case.

1.3 Problem statement

When considering AiT, the understanding as to why a given patient may experience rejection or graft failure is not truly appreciated. While it is recognised that presence of DSA both around the perioperative period and in the years following can be detrimental to graft outcome, there are still many notable examples where it is not the case. Higgins et al. [26] for example demonstrated a startling result that many cases which formed early high DSA levels post transplantation would regulate themselves following a dramatic increase, resulting in little long term effects on the donor organ. Large changes in DSA levels were previously understood to be highly detrimental to graft outcome and clinicians typically responded through administering plasmapheresis to remove the DSA, subjecting the patient to excessive and potentially harmful amounts of treatment. Revelations such as these are indeed valued insights, however do not yet address the true understanding of why these processes occur. True understanding may realistically be out of reach for current levels of technology to achieve as they would require validation of physiologically based models that analyse the underlying immune response. Current lab-

oratory assays do not present true values for antibody numbers but instead provide an indication of the number of antibodies reactive against a particular antigen. These values are derived from techniques such as enzyme-linked immuno sorbent assay (ELISA) and single antigen bead (SAB), which estimate antibody quantities based on a mean fluorescent intensity (MFI) indicator. In many cases the relationship between MFI and true antibody concentration can be difficult to quantify as the measurement process may result in nonlinear mapping between the two values (explained in later sections). Furthermore affinity of DSA may influence MFI values irrespective of concentration, which have each been shown to demonstrate very different in vivo kinetics [35]. Thus, relating MFI simply to antibody concentration is not a trivial task and leaves process driven models currently unobtainable in practice. Despite these limitations it is still widely considered that MFI data can provide valuable knowledge and should not be rejected as an informative resource when concerning transplant outcomes. In this thesis data-driven approaches are used as opposed to formulating and validating physiological models. Data driven analysis does not require a complete understanding of the underlying physiological system and can therefore be implemented where one is not present. There are two main types of data-driven analysis: parametric and non-parametric. In parametric analysis a parameter based model is developed to describe the system using a limited number of variables, such as in a multi-variable regression task or dynamic modelling using differential equations from physics. For non-parametric analysis the model structure is developed directly from the data itself, such as a nearest neighbour clustering task. Each approach satisfies a different role in analysis. Parametric approaches can greatly simplify the learning process by pre-defining the structure of the solution, but requires assumptions to be made about the model form. In non-parametric models there is less control over the solution, however no assumptions about model form need to be provided - potentially benefiting situations where less information about the solution structure is known. Parametric models have been used before such as in Zhang et al's work [34] on describing certain types of antibody decay rates following AiT. In this work daily DSA levels were considered as dependent events and mathematical models in the form of differential equations were developed to describe the DSA falls after transplantation and to find typical patterns associated with kidney rejection [34]. The aim of this work was to describe the regulatory mechanisms responsible for DSA evolution in various outcome scenarios. One of the biggest areas of weakness in AiT is the understanding of how antibodies contribute to

longer term outcome and more clinically concerned works look to find associations between the two which could be informative in clinical practice. Higgins et al. [26] have described the behaviour of DSA in the first month after transplantation by considering the rises and falls in DSA levels ultimately showing that higher levels associate with greater risk of acute rejection. Other more recent publications focus on use of DSA at specific time points both pre- and post-transplantation, either by applying classical statistical multivariable approaches [36–38] or more sophisticated machine learning algorithms [39]. Most studies were however centre specific and only limited to few samples [26, 40, 41]. Many of these studies are additionally difficult to compare to one another due to a wide variety in monitoring practices. A study of 66 histocompatibility laboratories which supported at least 25 transplantations each year showed substantial differences in defining sensitization status, method of reporting HLAi, and post-transplant DSA monitoring practices [42]. Even BTS guidelines remain relatively loose in defining key periods for monitoring [25]. It is clear then that best monitoring practices are still a developing area despite post-transplant associations with graft outcomes. Antibody monitoring is an expensive and time consuming practice, and defining a clear set of monitoring and interpretation protocols to maximise information gain from the fewest samples would be a a strong step towards improving transplant success rates and presenting opportunities for the development of multi-centre studies.

1.4 Thesis aim and objectives

In the recent two decades clinicians have begun to appreciate the value that post-transplant DSA monitoring can bring to assessing patient health, however there is not yet a complete understanding of how antibodies contribute to these outcomes. Modern DSA monitoring tools such as the SAB have offered substantial advantages over previous methods, however are still expensive to use in clinical practice. In this work a unique opportunity is presented to investigate short term post-transplant dynamic activity due to the high volume of DSA monitoring obtained (see Chapter 2) relative to other works in the field. In particular, the early post-transplant period (first two weeks) is of interest due to the presence of anamnestic memory B-cell dynamic behaviour and the occurrence of accelerated antibody mediated rejection (ABMR) episodes - which are thought to relate to long term graft outcomes. The advantages of this data set are sought to be fully utilised in this thesis

and the aim of the work can be summarised as follows:

Aim: To investigate the associations between early post-transplant DSA dynamics and long-term graft outcomes so as to develop new domain knowledge.

Emphasis is placed on the immediate two month period following transplantation which is hypothesised to contain valuable information pertinent to the acceptance of the kidney. To satisfy this aim, the work conducted in this thesis is guided by four main objectives:

Obj. 1: To acquire and prepare a set of data for use in satisfying the thesis aim.

Obj. 2: To determine the most valuable DSA observation for monitoring in practice.

Obj. 3: To identify and relate the types of DSA dynamic responses present in the AiT cohort to graft outcomes.

Obj. 4: To construct a generalised parametric based model of DSA dynamics for each of the previously identified response groups.

These objectives first seek to develop a solid foundation for analysis by acquiring, processing and defining a data set of post-transplant DSA dynamics before subsequently building upon this foundation through analysis of DSA data under different contexts. The purpose behind each of objective and how it relates to the thesis aim is briefly outlined below.

The first objective is satisfied in Chapter 2 where a HLAi kidney transplant cohort provided by the University Hospitals of Coventry and Warwickshire (UHCW) is investigated for its suitability in analysis of post-transplant DSA dynamics. This data set, once extracted, forms the basis of this thesis with all subsequent analysis performed on the same cohort. Most critical is an assessment of the DSA samples and a detailed breakdown of their meaning, measurement uncertainty, and limitations. This chapter additionally details relevant information relating to patient desensitisation, transplantation, and other treatment protocols as well as providing an overview of patient outcomes.

For the second objective a cross-sectional analysis of DSA observations is conducted in Chapter 3 whereby daily samples are investigated independently for their relevance towards long-term graft outcomes. This work builds upon the other studies in the area which, due to limited sampling rates, were only able to determine associations with graft outcome for specific time points pre- and post- transplantation [36–39]. Where DSA

measurements are expensive and time consuming to monitor, this work identifies the most informative period for a single sample to be taken in practice.

The third objective of this thesis, completed in Chapter 4, is to identify the different post-transplant DSA response types/groups which occur within the first two months. Multiple response types had already been shown in the literature before, for example Higgins showed that in some cases a marked rise in DSA levels in the first month was followed by a substantial fall back to the original levels, in others little to no activity was seen at all [26]. Despite this a conclusive study in the area had not been conducted on the cohort or otherwise. A deeper understanding of different DSA response types and their potential relationship with long and short term graft outcomes could be used to guide future treatment of patients.

The last objective of this work, completed in Chapter 5, is to develop a generalised parametric based model of the observed DSA dynamics as demonstrated in each of the previously identified response groups. Construction of these models allows for identification of the dynamic features present in the transient response of the underlying physiological system. From this analysis clinical insights can be identified from the use of model and the inferred parameters of the system.

1.5 Summary of thesis structure

Chapter 1 has began by exploring the history and current state of AiT before subsequently detailing the motivations, aims and objectives of the work contained within this thesis. Chapter 2 follows on by providing the necessary background information about the AiT transplant cohort which is used throughout all subsequent chapters. Chapter 3 details a cross-sectional investigation of DSA and concludes with demonstrating the most valuable day for monitoring a single DSA sample in the early post-transplant. Chapter 4 applies a novel grouping analysis to AiT kidney DSA response data - first demonstrating the types of response which occur and then relating these responses to graft outcomes. Lastly, Chapter 5 introduces a dynamic modelling methodology and applies this to each of the previously identified response groups allowing for clinical insights about DSA to be linked directly to dynamic features. Chapter 6 concludes the thesis by summarising its most important findings and exploring potential avenues for future works.

Chapter 2

HLAi kidney transplant cohort

2.1 Introduction

Extracting useful information from limited clinical data has been one of the great focuses of biomedical research. In the context of transplantation it could help guide or even develop new treatment protocols. When considering data there are two main types in clinical practice: static and dynamic. Static data do not change and typical examples may be patient characteristics such as age (at time of transplant) or gender. Dynamic properties change over time and are more frequently used in monitoring purposes. This thesis leans heavily on investigating the influence of DSA evolution on transplant outcomes however it is quite clear that DSA are not the only contributing factor. In this chapter the AiT cohort used throughout this thesis is introduced alongside both the static and dynamic variables relevant to future chapters.

2.2 Aim and objectives

The aim of the chapter is to introduce the background theory surrounding transplantation, treatment and monitoring - and then use it to select a suitable HLAi kidney transplant cohort for use in this thesis. Furthermore, DSA measurements, which form the foundation of this thesis are investigated for their underlying physiological meaning in addition to any limitations with the sampling approach. More specifically the chapter aims can be summarised by:

Aim 1: To select an AiT cohort which is suitable for subsequent analysis.

Aim 2: To detail patient treatment, transplantation and DSA monitoring protocols.

Aim 3: To extract a clean and usable DSA data set for use in subsequent analysis.

Aim 4: To describe short and long term patient outcomes for the cohort i.e., ABMR and graft survival.

To satisfy these aims an extensive literature search is completed which investigates the processes influencing an immune system response following a kidney AiT, the role of DSA, and potential outcomes for the patient (Section 2.3). This information is then utilised alongside key requirements for analysis, such as a minimum amount of DSA samples, to guide decision making on preparing, or pre-processing, the data set (Section 2.4). Section 2.5 summarises key decisions and findings which are used in later chapters.

2.3 Background

Presence of donor sensitised HLA antibodies play an important role in modern day decision making for renal transplantation. DSA are difficult to measure and modern day techniques still present challenges in interpretation. In this section an overview of the DSA monitoring tools used throughout this thesis are presented, however it is first required to detail the immunological background of DSA.

2.3.1 Immune response system

The immune response system consists of a set of mechanisms within the body that recognises and defends against bacteria, viruses and other substances which appear to be harmful. Initially when faced with a threat the body's innate immune system operates as the first line of defense in the form of natural killer (NK) cells - these function as some of the primary means of destroying infected cells or tumours. In cases where the innate immune system fails to protect the body from pathogens the active part of the immune system begins to respond (typically after 2-3 days). The functions of the active immune response are complex however can broadly be summarised into: recognition of specific non-self antigens, generating responses to eliminate pathogens, and development of immunological cells. These functions are carried out by the two types of active immunity mechanisms: humoral and cell mediated. The humoral system primarily consists of circulating antibodies which, once stimulated, initiate a rapid production of antigen-specific B lymphocytes (B-cells) that increases antibody titres and affinity to the inciting

pathogen. For cell mediated immunity there is no involvement of antibodies but instead activated production of NK cells, macrophages, and antigen-specific T lymphocytes (T-cells) alongside cytokines. A humoral type response is generally referred to as 'specific' as it directly detects and neutralises particular antigens operating on the surface of pathogenic substances. This is unlike cell mediated response which is less specific and can target intra-cellular pathogens which multiply within host cells.

The active immune system has been of primary concern for clinicians in transplantation due to a far stronger and more targeted response to the donor organ. Cellular response mechanisms in particular were one of the first areas of focus in the modern era of transplantation [25] where it was noted that up to a third of cases experienced graft loss in the first year from T-cell mediated cellular rejection. Over 70% of lymphocyte production is thought to be T-cell. Following this observation an intense period of research began, eventually culminating in the production of a therapeutic toolkit which involved inhibiting of T-cell production in addition to T-cell depletion routines. The available therapies eliminated almost all graft losses from this cause in adherent patients with one of the only limitations that treatments remain relatively ineffective against memory T-cell responses. Ultimately however, it is overall considered to be an effective set of treatment options [25]. Following the substantial improvements in reducing T-cell mediated rejection a greater deal of focus has more recently (in the last 20 years) been placed into the response of anti-donor antibodies, or DSAs, for which evidence had been accumulating that they play an important role in acute and chronic allograft rejection.

DSAs are part of the humoral response against the HLA of the donor organ [13]. HLAs are protein structures that exist on almost every cell in the human body. They function as a signal to the body's own immune system - informing that the cell belongs, or is 'self' [43]. The HLA system is controlled by the genes located on chromosome 6 and are split into two main classes: class I (CI) and class II (CII). CI and CII HLA each have three main variations: HLA-A, HLA-B, HLA-C, and HLA-DP, HLA-DQ and HLA-DR respectively. In total over 25,000 allele sequences already identified across the two HLA classes [44] with several studies showing that up to 40% of HLA allele have been identified in as little as only a single person [45, 46].

One of the major barriers to HLAi transplantation is the presence of pre-sensitised memory B-cells and antibodies specific to donor HLA. In Figure 2.1 the process of antibody

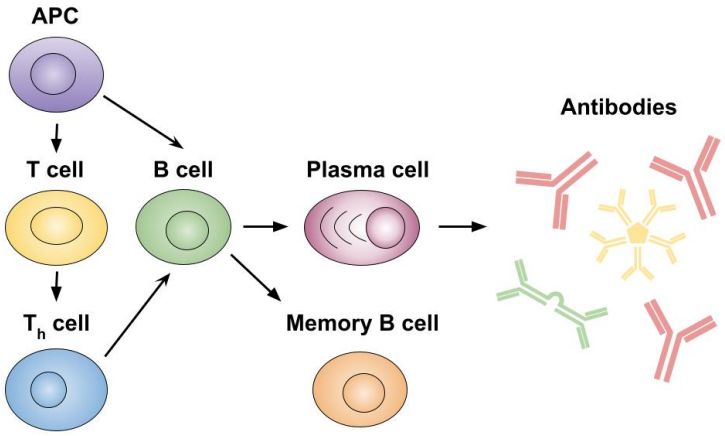


Figure 2.1: Diagram of humoral response upon first exposure to APC. APCs such as macrophages, dendritic cells, and B-cells develop upon contact with donor foreign antigens and stimulate humoral and cellular responses. Pathogen epitope-specific B-cell production is stimulated which subsequently either mature into plasma cells via assistance from helper T-cells (T_h) or develop into memory B-cells. Plasma cells ultimately facilitate the production of pathogen epitope-specific antibodies.

production and the development of memory B-cells is summarised following initial exposure to a foreign HLA - such as is often presented if the patient has experienced either pregnancy, blood transfusion, or previous transplantation. Initially, the signature of the foreign HLA is captured by antigen presenting cell (APC) which is then subsequently presented to B- and T- cells that can begin processing an immune response to the perceived pathogen. For the humoral response the process of antibody production, or synthesis, begins with B-cells which are produced in the bone marrow. Following a maturation phase the B-cell will differentiate into one of 2 types: memory cells, which allow for retaining of information relating to an antigen, or the plasma cell, which exist in the bone marrow and secrete antibodies directly into the blood. The vast majority of B-cells mature into plasma cells which function directly as a factory for production of antibodies to target foreign pathogens however it is the memory B-cells which remain present in the body (for up to decades) that allow for rapid response to previously exposed pathogens.

The structure of an antibody (also referred to as an immunoglobulin (Ig)) is made of two polypeptide sub units which together form a Y-shaped monomer, however some types of antibody can chain together to form dimers or polymers. Of the five main classes of antibody that are created in response to an infection IgD, IgE, and IgG are monomer types while IgA is a dimer and IgM is a polymer. The different types of Ig vary in function and location throughout the body. The most abundant anitbody type by far is the IgG isoptype and, in its four subclass forms (IgG1, IgG2, IgG3, and IgG4), it provides the majority of

antibody based immunity against invading pathogens [47].

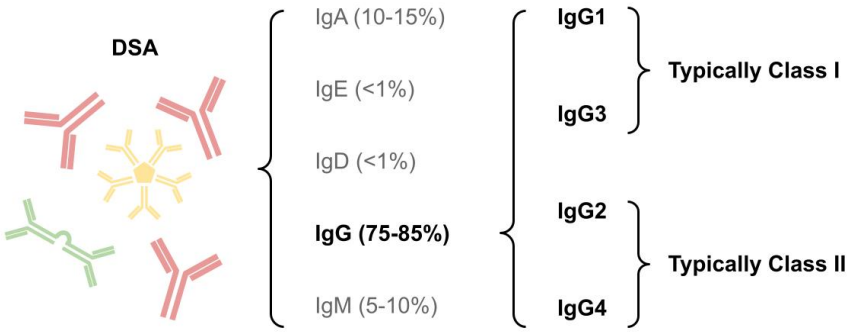


Figure 2.2: Diagram showing which antibody types contribute to the DSA immune response. IgG makes up the vast proportion of DSA volume and makes up the majority of the immunity against invading pathogens.

A common misconception is that antibodies themselves are responsible for the destruction of antigens however this is not the case. Antibodies are in fact not capable of penetrating cell walls and serve mainly to inform other cells of the presence of an antigen which then subsequently performs the destruction. Antibody antigen binding occurs through the F_{ab} region of the antibody (one of the forks). Once bound, the F_c region (trunk) is presented and cell destruction can either take place through one of two main mechanisms: phagocytosis or cell lysis. Phagocytosis occurs when effector cells (such as macrophages or NK cells) bind to the F_c region and proceed to ingest both the antibody and pathogen. For cell lysis the complement system is activated which enhances the ability of antibodies and phagocytes to clear pathogens. Complement binding begins with the C1q protein complex (component of innate immune system) which attaches to antibody F_c receptors. This process initiates the complement cascade that punctures the cell and neutralising the threat. Complement binding has been an area of considerable focus in transplantation as not all antibody types are capable of activating the complement cascade. Of the four subclass types IgG1 and IgG3 are often CI targeting antibodies and have strong complement binding properties. IgG2 and IgG4 are typically CII targeting antibodies with weak to no complement binding capability.

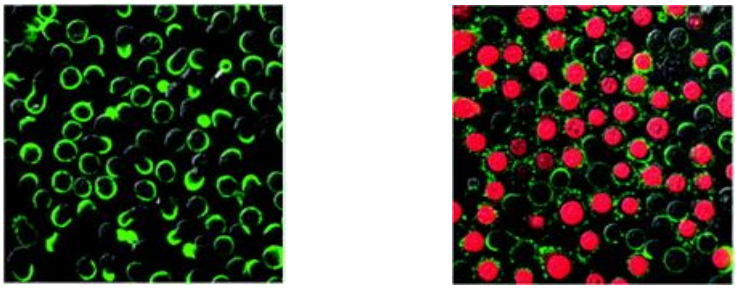
2.3.2 Monitoring of donor-specific antibodies

Investigative work on DSA in the context of transplantation can be dated at least as far back as 50 years, where Patel and Terrasaki showed that positive crossmatch tests were strongly correlated with delayed graft function [48]. Since then the importance of DSA in assessing donor and recipient compatibility is more fully appreciated and monitoring

of DSA both prior to and following transplantation is considered standard practice. Three main techniques are now used to monitor the presence of patient anti-donor HLA antibodies: complement dependent cytotoxic (CDC) crossmatch test, flow cytometry (FC) crossmatch test, and luminex assay tests. The roles of each in the context of transplantation are detailed as follows.

Complement dependent cytotoxic crossmatch

CDC is an assay used to determine if a potential organ donor organ and recipient are immunological compatibility [48] and is the standard test conducted prior to any organ transplantation used to detect complement-fixing antibodies. In brief, donor lymphocytes are isolated from peripheral blood, spleen, or lymph nodes before being further separated into T-cell and B-cell compartments. Recipient serum is then mixed into the compartments followed by rabbit serum containing a source of active complement. If DSA are present in the recipient serum they will bind to the donor cells causing activation of the complement - ultimately resulting in lysis of the lymphocytes. The level of DSA activity can be expressed as a percentage of cell lysis (ratio of live/dead cells) although is more generally either referred to as CDC+ or CDC- depending on if cell lysis is taking place or not (shown in Figure 2.3). Unlike later discussed DSA measurement techniques the CDC test is not able to distinguish specific types or classes of antibody and generally represents the summative result of all antibody types (HLA or non-HLA; Allo or Auto-antibodies; IgG or IgM) in the recipient serum. Despite this, CDC+ results are associated with worse transplant outcomes [33].



(a) CDC- result. **(b) CDC+ result.**

Figure 2.3: Assessment of CDC assay. (a) absence of lysis (shown in red) indicates non presence of DSA in recipient serum. (b) cell lysis present indicating binding of DSA to donor cells. Images adapted from [49].

Flow cytometry crossmatch

The FC crossmatch test is more sensitive than the CDC assay test as it can detect both complement fixing and non-complement fixing DSA [50]. This test is completed by initially mixing of recipient serum with donor lymphocytes in a similar process to that seen in CDC crossmatch test. Following an incubation period fluorescent labelled anti-human IgG monoclonal antibodies are added to the serum which bind specifically to recipient DSA that are bound to donor cells. The number of bound cells are then determined by using an FC machine which observes the scattered light bouncing off of cells as they flow past an emitted laser. Cut-off for a positive FC test is usually computed relative to a negative control serum which is without known HLA antibodies. If the fluorescence is two standard deviations greater than that of the negative control sample then it is considered FC+. The ability of the FC test to directly detect all bound donor specific IgG regardless of reliance on the complement secondary effector mechanism has been a great advantage of the test relative to CDC, however its reliance on statistical processes as opposed to biological relevancy has resulted in both tests still frequently employed in practice [51].

Luminex platform

The luminex bead is a relatively newly developed technology that offers a more practical approach to measure DSA on a daily basis. Unlike FC and CDC testing microbead analysis does not rely upon a steady supply of fresh donor lymphocytes to allow for daily monitoring. Instead, target HLA proteins are purified and subsequently attached to polystyrene beads which are each individually labelled with specific ratios of a fluorescent marker [52, 53]. Typically an assay consists of 100 beads which are used to measure different HLA specific antibodies (shown in Figure 2.4). Patient blood sera, which contains DSA, is then added to the assay and left to incubate allowing for binding of patient DSA to compatible microbead antigens. Following the incubation period a secondary anti-human IgG antibody is then added to the assay which contains a phycoerythrin (PE) agent that acts similar to the florochrome markers on individual beads. The PE anti-IgG antibodies bind to the anitbodies in the assay and - once non-bound antibodies have been removed from the assay - can be used to determine the concentration of antibodies in the patient sera. An estimate of antibody concentration on each bead is determined via an MFI measurement which assesses the fluorescent signal given by each bead using a laser. These values are then compared to an antibody negative control sample using a computer

algorithm.

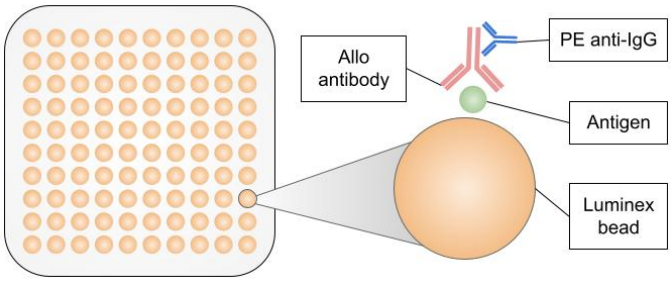
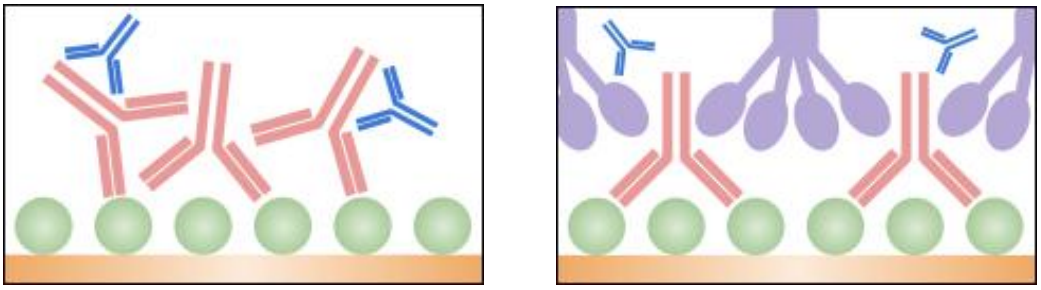


Figure 2.4: Luminex assay with example of single bead. Antigen specific antibodies will bind to antigen binding sites on luminex bead. Concentration of DSA is then estimated by fluorescent anti-IgG antibodies which bind to the antigen specific antibodies. The fluorescence is then measured using a targeting laser and computer algorithm while comparing emission to a negative control bead.

Earlier versions of the luminex platform utilise a technology known as phenotype beads which carry an entire HLA haplotype profile. This results in several antibodies potentially binding to the same bead some of which may or may not be donor specific. Developments upon this technology came about with the development of the SAB which is capable of targeting antibodies down to specific allele type (i.e. HLA-B7 and no others). This has substantially improved specificity of the assay. An additional advantage of luminex assays is that monitoring of non-DSA is also possible and may help distinguish between a general immune response - such as with inflammation - or a specific anti-graft response [26]. Positive reactivity of patient antibodies to a specific luminex bead HLA is not straightforward to determine. Luminex assays are highly sensitive and positive cut-off values between laboratories can vary substantially (from 500 to 3000 MFI) dependent on lab specific protocols and operator experience. It is noted that even with standardised protocols the coefficient of variation can be around 25% [54]. MFI levels can also be affected by artefacts [55] or presence of IgM HLA specific antibodies [56] or presence of complement proteins [57] and as such care should be taken not to directly equate with the true concentration of HLA-specific antibodies. Despite this SAB measurements are the best available option for determining DSA levels [58] and have improved the ability to identify and manage sensitised patients [59] by providing a semi-quantitative measurement of DSA with the MFI reading.

As previously mentioned low DSA level ranges provide large uncertainty for true concentration of DSA in patient blood serum as they approach the cut-off value for antibody positivity at around 500-3000 MFI. It is also necessary to be cautious of particularly large



(a) Close proximity to other DSA.

(b) Competition with complement c1.

Figure 2.5: Biological processes governing prozone effect at high DSA concentrations. (a) Close proximity of DSA on antigen binding sites restricts access for PE coated anti-IgG antibodies to bind to DSA causing a reduction of DSA relative to true concentration. (b) Concentration of DSA becomes high enough that complement c1 (purple) begins to compete for PE coated anti-IgG antibody binding sites. Complement c1 requires a divalent bond between two DSA binding sites and as concentration of DSA increases complement c1 becomes yet more competitive - resulting in a lower MFI level even as true DSA concentration increases.

DSA levels when using the luminex platforms. Two effects are documented in the literature which produce non-linear mapping of observed MFI mesasurements and true DSA concentration: the prozone (or hook) effect and the saturation effect. The prozone effect (Figure 2.5) occurs primarily due to competition for the PE coated anti-IgG binding sites. As concentration of DSA on the surface of the luminex beads increases they can become close enough that they overlap and subsequently block the anti-IgG antibodies from binding to the DSA - causing a decrease in the measured fluorescence. Additionally close proximity of complement activating DSA can allow for competition from complement c1 molecules which require a divalent bond between two DSA. As concentration of DSA increases more binding sites are available for complement c1 and therefore there are fewer opportunities for the anti-IgG antibodies to bind. Saturation effects occur as the concentration of DSA becomes so high that all antigen binding sites on the luminex bead are taken. Often in the design of a luminex bead saturation and prozone effects are weighed against one another in order to maximise the number of HLA present on the bead while minimising the effects of complement c1 and DSA overlapping. Work from the UHCW [60, 61] has shown that there is typically a 'Goldilocks' range where MFI and true concentration are linearly matched however the extent must be determined individually for specific HLA alleles. Figure 2.6 shows an example of MFI reading for HLA allele A2 and B7 before they experience saturation relative to true DSA concentration. To identify presence of prozone or saturated affected MFI measurements procedures typically dilute patient sera and check if observed MFI concentrationd react accordingly.

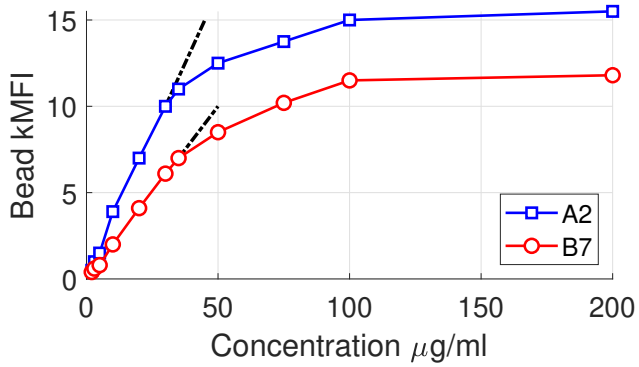


Figure 2.6: Example of linear range detected using luminex assay. Note linear range extends to different MFI for different types of anti-HLA antibodies. As true DSA concentration increases the linear relationship with MFI breaks down and a saturation effect is demonstrated (extracted from work by [60]).

DSA use in practice

In practice patients may often display more than one DSA targeting donor HLA and, where the goal is to relate these DSA levels to patient graft outcomes, has led to two main approaches for analysis. The first approach is to use the immuno-dominant DSA (iDSA) which is the DSA that demonstrates the largest MFI measurements. The second is to use the total DSA (tDSA) measurement which is the aggregate of all DSA for any given patient. No clear consensus is available on which approach is most appropriate for analysis and many studies will take the approach of assessing associations with both methods. In certain lung transplant cases, for example, tDSA has been associated with antibody rejection when iDSA was not [62] however numerous findings have been shown for only iDSA [63–65]. Figure 2.7 shows a case from the AiT cohort used in this thesis and demonstrates how the tDSA can be formed from a patient who is sensitised to three different donor HLA. Additionally, in the context of this patient the iDSA would be the B60 type DSA.

2.3.3 Allograft rejection

Allograft rejection is an important barrier to long term graft survival [66] and is a consequence of the interaction between patient immune system responding to HLAi expressing cells on the graft vascular endothelial cells. Both the innate and adaptive immune systems play a significant role in rejection however T-cells are the principal cells that identify the allograft [67]. Renal transplant rejections can be broadly classified into the following categories:

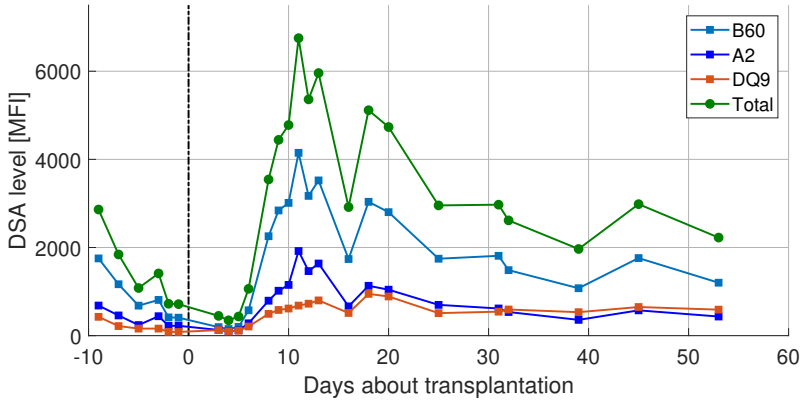


Figure 2.7: tDSA accumulation procedure for a patient with three different DSA.

1. **Hyper-acute rejection.** Occurs within minutes of graft reconnecting blood supply to organ (repurfusion) and occurs due to pre-existing antibodies in the recipient blood against donor HLA which are present at the time of transplantation.

2. **Acute rejection.** Occurs within a few days to weeks following transplantation and can occur either due to antibody or T-cell mediated response pathways.

3. **Chronic rejection.** Rejection is the result of a progressive loss of graft function occuring over months or even years following transplantation and can occur due to either antibody or T-cell mediated response pathways.

ABMR is caused by the binding of DSA to the graft endothelium and can result in classical pathway complement activation followed by cell death. This leads to restricted blood flow and oxygen (ischaemia) causing damage to the kidney. Additionally, as endothelium integrity is compromised as coagulation factors are able to pass through freely - ultimately leading to blood clotting (vascular thrombosis) and compromising of kidney function. It is usually defined with evidence of circulating DSA and immunological evidence of antibody-mediated injuries to the kidney - such as inflammation of glomeruli or peritubular capillaries, or even depositions of C4d (a degradation product of complement pathway) to the endothelial membrane of the donor kidney. T-cell mediated rejection (TCMR) occurs when patient lymphocytes become activated by recognition of foreign donor HLA by APC. This leads to activation and infiltration of T-cells to the donor graft causing cell lysis and is usually characterised by lymphocytic infiltration of the tubules, interstitum and arterial intima. If untreated TCMR causes irreversible nephron loss [68]. Chronic rejection is linked to both immune and non-immune mediated factors with the primary risk factor being non-compliance with immuno-suppression medication. Of the

three rejection types hyper-acute ABMR was the first to be recognised as early as in the 1960s [69] however it was largely solved in the same decade with the introduction of the CDC test which substantially reduced the rate of incidence [48]. Treatment for both acute and chronic rejection is similar. For ABMR plasma-exchange, intravenous immuno globulin (IVIg) and various cell inhibitor drugs such as Rituximab and Bortezomib are used to reduce antibody count. For TCMR T-cell, B-cell and macrophage levels are reduced via lymphodepleting treatments such as prednisone and ATG. In both cases clinicians may also vary dosage of immunosuppressive drugs to combat rejection onset.

2.3.4 End-stage renal disease

Kidneys are a vital part of the urinary tract whose purpose is to filter blood and remove waste products. Typically, a person is born with two kidneys which are situated under the lower ribs at the back, one on either side of the body where each kidney is shaped like a bean with approximate dimensions 12x6x3cm and weight 150g. Aside from the kidneys, there are three other main components of the urinary tract: ureters, bladder, and the urethra and their position relative to one another. Filtration begins with a kidney extracting waste products from the blood circulating from renal arteries. Common products to be removed at this stage are creatinine, urea, and water whilst products that are not removed are certain vitamins, amino acids, and hormones within the bloodstream. Once extracted, the waste products form urine are passed down the ureters to the bladder. A typical bladder can store 400 ml of urine before it is expelled from the body via the urethra.

In addition to its main function of removing waste products the kidney also provides several additional functions. Each kidney is a capable chemical factory when healthy and produces many hormones which help regulate the body. Production of prostaglandins and aldosterone for example influence the adsorption of sodium and water into the blood, which act to both balance the body's fluids and regulate blood pressure. It also produces vitamin D which has many functions however primarily aids in the maintenance of healthy bones and regulation of the body's immune response. Lastly of note, erythropoietin is also produced which stimulates the production of mature red blood cells and thus maintains healthy oxygen levels in tissues.

Kidney failure is the result of lack of kidney function. The length of time for which a kidney demonstrates loss of function dictates whether it is referred to as AKI (days or

weeks) or CKD (>3 months). Kidney function is measured via the eGFR which is based on a blood sample testing for creatinine levels amongst other baseline patient characteristics like gender and age. For CKD, where the loss of kidney function is more gradual, the effectiveness of the kidney to filter waste products is categorised based on its eGFR level. There are five stages with stage 1 corresponding to normal kidney function and stage 5 corresponding to kidney failure - otherwise known as ESRD. Once kidney function has deteriorated to the point of ESRD the kidneys' ability to regulate waste products is impaired substantially and without medical intervention will lead to illness and eventual death. Should patient kidney function deteriorate to the point of ESRD then the organ is considered to have failed, and, in the context of already having a previous transplantation, it is referred to as a graft-failure (GF) in this thesis.

2.4 AiT cohort

The data used in this thesis has been collected by University Hospitals of Coventry and Warwickshire who conducted 110 HLAi (and blood type compatible) kidney transplants between June 2003 and February 2014 with follow-up time available up until March 2019. One patient did not provide consent for their data to be used in the study and is excluded. Ethical approval for the study was obtained from the local ethics community (CREC-055/01/03 and 13/WM/0090) and written consent was obtained from all other patients included in the following studies. Of these 10 additional patients were excluded: 4 due to early death of patient (<10 days), 1 due to early graft failure, 3 due to missing/incomplete DSA data, and 2 cases were excluded due to saturation affected DSA samples. Following these exclusions 99 cases were available for all subsequent analysis in this thesis. Across the AiT cohort every single patient demonstrated at least one pre-sensitised HLA-specific antibody with their donor organ, known as a mismatch. The maximum number of mismatches was seven and the median was two. A total of 278 DSA were present across the whole cohort with 138 being CI, and 140 being CII. In transplantation it is also common to consider if a case consisted of solely CI ($N = 36$), solely CII ($N = 24$), or, a mixture of both CI and CII (CI&II) ($N = 49$) DSA.

2.4.1 Desensitization protocol

Approximately ten days prior to transplantation patient DSA levels were assessed via luminex assay MFI measurements. Sufficiently large DSA levels (tDSA >3,000 MFI) were to be met with desensitization treatments with a target of negative FC crossmatch or cumulative DSA MFI <3 000 for HLA -A, -B or -DR mismatch at the point of surgery. Two treatment options were considered which consisted mainly of double filtration plasmapheresis (DFPP) and IVIg only in cases where IgG levels were low. In total 73 out of 99 cases had received at least one DFPP session; the maximum number of sessions administered to a single patient was 13, and the median number of sessions received being four. In a third (33/99) of cases however the DFPP sessions could not achieve negative FC crossmatch and patients were transplanted in the presence of higher DSA levels due to clinical needs. There were 11 DDT patients which were among the 26 cases which did not receive pre-transplant DFPP treatments. Only a few cases received IVIg (N = 3).

Generally immunosuppressive treatments consisted of twice daily 1,000 mg mycophenolate mofetil, starting ten days before transplant with dosage reduced if white cell count dropped below 4.0 x 10^9 per litre. Daily administrations of tacrolimus were commenced four days before transplantation. Dosages were given at 0.15 mg/kg/day in increments with a target trough level of 10 to 15 μg/L in the first month. At surgery, a single 500 mg methylprednisolone dose was provided intravenously during the transplant operation; a 20 mg basiliximab induction was given twice on day zero and day four post-transplant.

2.4.2 Patient monitoring

HLA CI and CII specific antibodies were identified before transplantation through the use of One Lambda Inc.'s (Canoga Park, CA) SAB assay on the Luminex assay platform (XMap 200). For earlier cases, up to mid-2007, HLA phenotype beads were used (N=19) before being replaced by the SAB used to the current date. Some select cases were retrospectively retested with the SAB platform to confirm specificities. Protocol for monitoring DSA was built to collect a high number of samples early post transplant followed by fewer samples up until weeks 8-12 where protocol stopped.

Figure 2.8 indicates the stated protocol for each week following transplantation, with an example being to collect at least four samples in the first week. In practice this high intensity was difficult to achieve and success varied greatly from patient to patient. Relatively

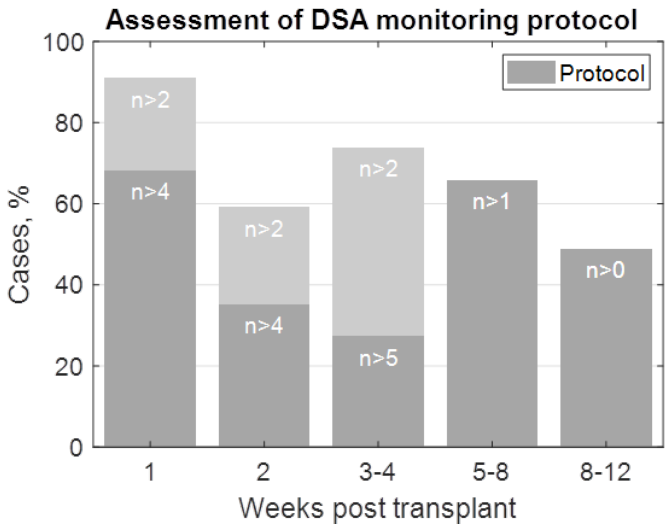


Figure 2.8: Assessment of DSA monitoring protocol. Following transplantation patients DSA would ideally be sampled at the stated protocol rate for a given period of time, i.e., week 1, at least 4 samples. In practice this was not always achieved - as shown by the % cases y-axis. For greater context some additional markers are included in light grey.

good rates are seen in weeks 1 and 2 however retention rates are notably at their poorest relative to the desired protocol in weeks 3-4. More typically then, protocol for post-transplant serum sample collection consisted of high-intensity testing over the first 15-20 days following transplantation followed by more sparse measurements as the weeks progressed, typically stopping at around sixty days. More limited testing is stated as occurring only when the patient's DSA levels and clinical measures were deemed stable. Figure 2.9 further adds to Figure 2.8 by displaying histograms comparing the number of DSA observations per case for a given period of time. The post-transplant period is of most interest in this thesis and it can be seen that the number of observations can vary greatly from case to case. At the extremes the lowest number of observations is six and the maximum is 52. The median number of observations is 21. Not all cases will present enough DSA observations necessary for analysis in this thesis - however this will be addressed in later chapters. Throughout this thesis tDSA levels will be used to represent the patient antibody response to donor HLA and the term will be used interchangeably DSA.

2.4.3 Handling early acute ABMR

Apart from post-perfusion biopsies taken in the operating theatre, biopsy was done for cause only, i.e. in cases of graft function deterioration or creatinine stuck. Acute rejection episodes occurring before 30 days post-transplant (N=45) were identified at incidence under the most recent BANFF guidelines [70]. In some instances, where the biopsy was not

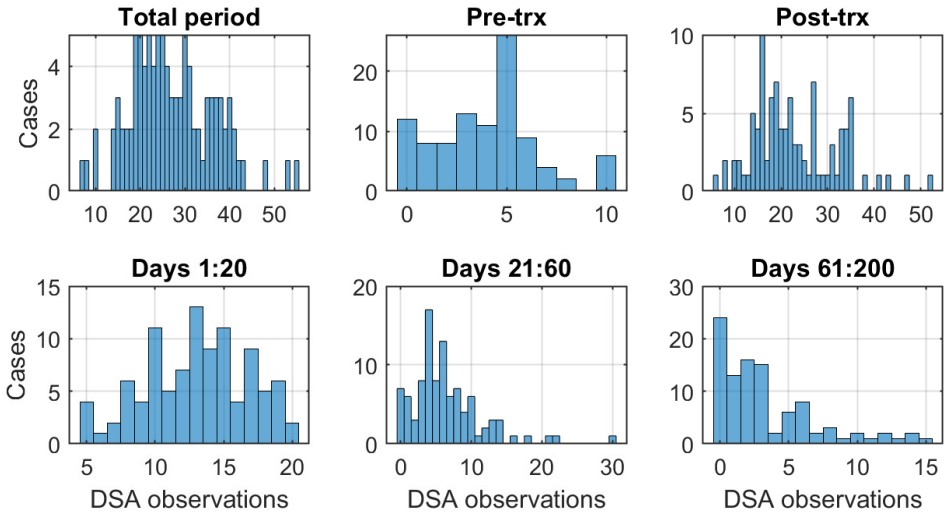


Figure 2.9: Histograms of observed number of data points per measurement period. Total period includes pre-transplant and post-transplant periods. trx=transplant. A general trend is that there are fewer observations as time progresses.

possible, for example, for patients on anti-coagulation therapy or during weekends, a rapid rise of DSA MFI values alongside a drop in urine output and increase in creatinine was defined as clinical rejection. All 45 rejection cases were treated with a course of pulse methylprednisolone 500 mg once a day for three days. Additionally, lympho-depleting agent (ATG or OKT3 or Campath) was administrated in N=31 cases, and DFPP treatment was performed in N=13 cases, of which 5 required IvIg replacement. One case additionally had ecluzimab as rescue therapy.

2.4.4 Graft failure

Within many biomedical studies the variable of interest can often be characterised as a "time to event" such as time to death or a failure event. In this study graft survival is of key interest. The first transplant was conducted in June 2003 with subsequent patients routinely referred to UHCW at an average rate of approximately ten per year since. The last patients included in our study were from February 2014. These transplantations have formed the basis of a longitudinal study whereby additional data on each patient is routinely collected following transplantation. Follow up time for each of these cases was available up until March 2019 providing vital information such as patient graft survival status. With over ten years separating the first and last cases in our study the available follow up time for each patient can vary considerably and many of the cases do not experience a graft failure event in this period. In total only 24 cases have experienced a

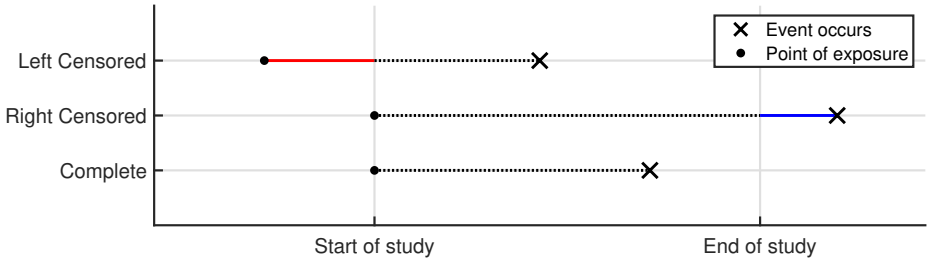


Figure 2.10: Comparison of censoring types. Left censoring - point of exposure occurs some time before start of study. Right censoring - event occurs sometime after end of study. Complete - both point of exposure and event occurrence are known.

graft failure event in the study time. By excluding those cases which do not experience a graft failure event a substantial number of data would not be used in analysis. To benefit the most from this survival data we therefore look to take advantage of censored data which allows for the inclusion of no graft-failure (no-GF) cases into survival analysis.

Censored data occurs when some information about survival time exists but the exact survival time is not known [71]. When considering patient graft survival as an outcome there are three reasons that case data may be censored: patient death, no-GF during follow up period, and contact with case lost during study period. In our cohort eight patients died during follow up period and are subsequently death censored, 61 experienced no-GF during follow up period and contact was not lost with any cases during study period. A total of 69 cases are censored. When censoring data the root cause of censoring must additionally be considered, and its subsequent impact on the confidence in survival time [72]. Consider the following censoring types shown in Figure 2.10. Left censored data occurs when the left side of the survival time becomes uncertain, such as in the case of a person being tested positive for an infection - the true point of exposure is not known. Right censored data is more commonly seen in transplantation literature where the point of exposure is clearly known. Cases in our cohort are either complete, or right censored, depending on the whether the time-to-GF occurs prior to the end of our study period.

When analysing patient survival data it is often convenient to compare in terms of a consistent reference point as opposed to analysing relative to a fixed starting date. One of the most powerful tools to achieve this analysis is through the Kaplan-Meier survival estimator [73] (detailed in Appendix section A.1). First publicised in 1958, the estimator determines the fraction of subjects which have not experienced an event for each point of time following a common starting point (i.e., transplantation). In doing so the estimator generates a non-parametric function describing patient survival over time. For a small

sample size a plot of the survival function will be represented by a series of large declining steps, however, as the sample size becomes large enough the plot will approach the true survival function of the population. In bio-medicine these types of analysis are often used to identify how different factors, such as gender, may influence long term outcome.

Through application of the product limit method the patient graft survival probability is estimated from the 99 HLAi kidney transplant cases at each time point following transplantation. The calculated Kaplan-Meier survival function is then plotted as shown in Figure 2.11. In total 25 cases have experienced a graft-failure event as of March 2019 and 74 cases are censored. Of the censored cases one case exceeds 15 years of follow up data and 5 cases have less than 5 years follow up data. The median (lower quartile (LQ) to upper quartile (UQ)) follow-up time for censored cases is 9.9 (7.4 to 12.0) years. The earliest graft failure event occurs at 71 days following transplantation and the last is at 12.8 years. The median (LQ to UQ) event time is 4.8 (2.3 to 6.3) years post-transplantation.

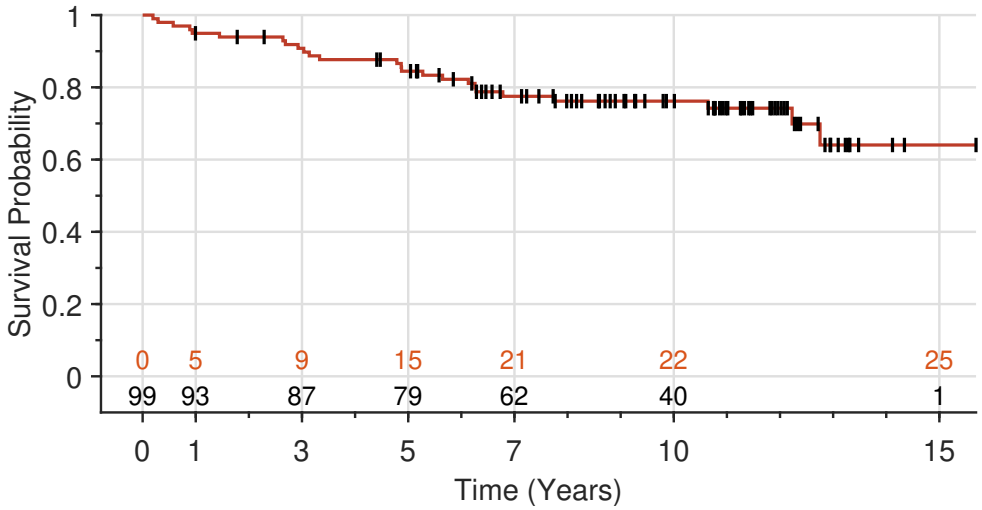


Figure 2.11: Kaplan Meier survival analysis for HLAi kidney transplant cohort Survival probability designates probability of graft-failure event for a given time. Cases (n=99) are either censored (n=74) or not-censored (n=25). Censored data include patient deaths or last record as non-event. Not-censored data are events. Black values indicate number of censored cases which have exceeded the indicated time point. Red values indicate the number of events which have occurred before the indicated time point.

2.5 Conclusions

Following a detailed literature review - which covered the relevant elements of the immune response system through to monitoring DSA and patient outcomes - a study of the AiT data set provided by the UHCW was conducted. The first aim of the chapter was

to identify a suitable cohort for use in subsequent chapters. This objective was achieved by accessing and assessing the UHCW HLAi (and blood type compatible) data set which was collected between June 2003 and February 2014, and with follow-up data extending up until March 2019. In total 110 cases were available for analysis however it was recognised that not all cases may be suitable, and a total of 11 cases were excluded for reasons including: early death (4), early graft failure (1), missing data (3), saturation effect (2), and lastly no-consent (1). This left in total $n = 99$ cases for subsequent analysis. The second aim of the chapter was also satisfied with a detailed account of the cohort, their treatment, transplantation and DSA monitoring protocols was produced in accordance with and records from the UHCW transplant teams. Thirdly, when considering developing a usable DSA data set it is noted that the number of possible DSA allele types is vast and substantial even for this cohort. Most cases have more than one DSA with the median being two and the maximum being seven. Additionally, nearly half of cases have DSA of mixed class ($n = 49$). Comparing like-for-like DSA allele, for example A1, would have substantially diluted the available cohort for analysis as few cases are likely to share that same feature. In light of this there are two main approaches used to analyse cases with multiple DSA in the literature, with the first assessing the patient iDSA and the second patient tDSA. With no clear consensus on the best approach all subsequent analysis utilises the latter approach - where each daily DSA sample consists of the aggregate of the individual allele. Analysis of the DSA sampling rate shows that a consistent monitoring pattern was not achieved amongst cases and in fact a wide variety of sampling rates is demonstrated. In the worst case only six data points are available in contrast to 52 for the case with the most observations. The median number of observations however was 21 with most samples present within the first 60 days. These limitations in sample rate will need be considered in later chapters for particular types of analysis. Lastly, DSA data were recognised as having a coefficient of variation of approximately 25% and limited reliability at both low and extremely high MFI values. Values at or below 500-3000 MFI may be subject to high levels of noise and careful consideration should be used for dynamic analysis. Additionally, high levels of DSA may induce a saturation effect whereby a nonlinear relationship is formed between the luminex measurement assay and the true DSA concentration. Two such cases in the UHCW AiT cohort were found by analysing diluted samples and have subsequently been removed from analysis. The last aim of the chapter was to describe the cohort in terms of patient outcomes, i.e., ABMR and graft

survival. For ABMR it was shown that several different definitions exist based on the timing and onset speed, however in this cohort information was only available for early-acute ABMR incidences. In total $n = 45$ cases experienced a rejection episode within the first 30 days. All incidences were treated with a course of pulse methylprednisolone however only $n = 31$ received lymphodepleting agent and fewer still, $n = 13$, receiving DFPP treatment. For graft failure, as considered by a patient reaching ESRD, a Kaplan-Meier survival analysis was produced to show how survival probability in the cohort changes over time. Notably, five cases are shown to fail within the first year and a total of $n = 25$ by the end of the study period.

Chapter 3

Identifying associations with survival outcome

3.1 Introduction

Throughout the 20th century many advances in biomedical knowledge were identified by statistical approaches with supporting evidence based on p-values and other accompanying methods [74]. This line of work contributes to what Varga et al. [75] describes as 'precision medicine' and is used in an explanatory or associative role but distinctively not towards detailed predictions for individuals - referred to as 'individualised medicine'. As the 21st century approached, various biomedical research communities sought progress towards individualised medicine which would ultimately allow for treatments and preventative actions to be tailored to specific persons. The goal of this analysis was simple, 'using previous knowledge, infer the future behaviour'. There is a critical difference between the two approaches: associative analysis find statistical strengths between variables which further develop understanding of a phenomena, and prediction studies focus on formulating predictions of future data based on previous examples. In this chapter associative analysis is performed with the objective of identifying the key variables that relate most closely with long term graft survival in the AiT data set. Variables here correspond to both the static characteristic variables and the dynamic DSA levels.

3.2 Aim and objectives

Kidney HLA AiT is an active area of research for which all contributing factors towards graft survival are not known. In this chapter an associative analysis is completed for the AiT kidney transplant cohort from UHCW as detailed in Chapter 2. Specifically, the influence of DSA on graft outcome are to be investigated as this cohort benefits from a high rate of DSA sampling - allowing for the influence of certain daily observations to be investigated. Due to the high costs associated with frequent DSA monitoring, this work aims to suggest an optimal measurement period which practices can utilise to gain the most information on long term survival. The aim of this chapter is therefore summarised by the following statement:

Aim 1: To identify the most informative DSA measurement period.

In order to identify the above aim the following objectives are established:

Obj. 1: Investigate and prepare a suitable set of independent variables for multivariable analysis.

Obj. 2: Investigate multivariable associations with long term graft survival.

Obj. 3: Investigate the association of daily DSA levels with long term graft survival.

The chapter is split into the following sections: survival model theory, data pre-processing, univariate analysis and multivariate analysis. Lastly, the results, reliability and implications of these analysis are discussed prior to chapter conclusions.

3.3 Survival model theory

In this chapter static and dynamic variables are considered in a univariate and multivariate analysis to identify associations with graft survival outcome. Given a continuous output variable, time to failure, one may typically look to a set of regression techniques to estimate the influence of a particular variable. For time-to-event data however, such as in graft survival, the task is additionally more complex due to the presence of censored data points which prohibit the use of standard regression models for parameter inference (or face model bias). In response, survival specific modelling approaches have been devel-

oped such as the Cox hazard models which can accommodate for censored data in their inference of variable coefficients. It should be noted that some alternative approaches are also present in the literature, depending upon the study objective. The most commonly seen alternative is to assess survival data at a fixed point in time e.g., at 5-years post start of study - ultimately reducing the task to a classification problem. Either approach has their advantages with the classification task lending itself towards easier interpretability and allowing for a wider range of modelling approaches to be implemented [76], while the survival task alternatively allows for a risk to be interpreted across time. Critically for this work however, by analysing at a particular time point the classification task is limited to use only the subset of the cohort which have either experienced an event or have been observed up until the end of that period. Given the relatively small data set (N=99) the effects of using yet smaller subsets can substantially reduce statistical power, the effect on case numbers can be seen in Table 3.1. Opting for a survival based approach here grants us the most cases for analysis while also allowing for consideration of time dependent risks, i.e., if a certain variable is more important in analysis of risk early as compared to later post-transplant.

	3-year	5-year	10-year	Survival
Cases	96	94	62	99
GF within period	9	15	22	-
Uncensored events	-	-	-	25

Table 3.1: Comparison of case numbers when considering graft survival as either a classification or survival task. Number of cases indicates the total number available for analysis for the given task. GF indicates the number of graft failures which have occurred up to the indicated period. Uncensored events are the number of known graft failures within the study period.

With a survival framework shown to be the best approach for this analysis we can consider a range of different models to perform the associative analysis from Cox based hazard models up to survival decision trees or random forests. For this chapter the use of Cox proportional hazards models are explored which are one of the most commonly used tools in survival analysis. In the model patient survival expectation can only be proportional to one another i.e., if twice as likely at a time point t then they will be twice as likely at all other time points too. For some cases this may be limiting if a model is expected to have non-linear terms with respect to time, however, in many cases this proportionality presents the main advantage of the model - eliminating the need to make assumptions about the type of baseline survival function that it follows: e.g., Weibull, exponential,

etc. With few data points many baseline functions could be considered plausible and therefore this model presents itself as a strong candidate for exploratory analysis.

3.3.1 Cox-proportional hazards model

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred to as a hazard rate. Predictor variables (or factors) are usually termed covariates. The Cox model is expressed by the hazard function, denoted by $h(t)$ which can be interpreted as the risk of dying at time t. The function can be estimated by,

$$h(t) = h_0(t)\, e^{\beta^T x} \tag{3.3.1.1}$$

where t represents the survival time, $h(t)$ the hazard function, and h_0 the baseline hazard. Given p covariates, $x = (x_1, x_2, \cdots, x_p)$, there will be p coefficients $\beta = (b_1, b_2, \cdots, b_p)$ denoting the corresponding effect size of each. The baseline hazard, h_0, corresponds to the value of the hazard if all the x_i are equal to zero $(e^0 = 1)$. Most frequently of interest are the quantities $e^{(b_i)}$, otherwise referred to as the hazard ratio (HR). HR can exist in a range between zero and infinity with a HR of one indicating that the particular covariate has no effect on survival outcome. Values lower than one indicate a reduced hazard ratio and therefore improved survival outcome, values higher than one vice versa.

Estimating covariate coefficients

As previously indicated one of the advantages of the Cox proportional hazards model is in its ability to determine a solution for coefficients β without making any assumptions for the baseline function, h_0. The solution, discovered by Cox in the early 21^{st} century, can be found through maximising the partial likelihood (PL) of the hazard function [77].

Defining the PL first requires each observation, i, to have a defined feature vector x_i, time T_i, and event indicator D_i. Following this the f unique failure times are then mapped into increasing order indices, $t_1 < \cdots < t_f$, where $j(i)$ is the index of the sample failing at time t_i. If the cohort can be described with at most one event occurring at any given time, then the PL for the Cox model can be expressed as,

$$\mathcal{L}(\beta) = \prod_{i=1}^{f} \frac{h_0(t_i)\, e^{\beta^T x_{j(i)}}}{\sum_{j \in R_i} h_0(t_i)\, e^{\beta^T x_j}}, \tag{3.3.1.2}$$

where h_0 can be eliminated from the equation,

$$\mathcal{L}(\beta) = \prod_{i=1}^{f} \frac{e^{\beta^T x_{j(i)}}}{\sum_{j \in R_i} e^{\beta^T x_j}}. \tag{3.3.1.3}$$

In the case where event times are tied the probability is less accurate (particularly so the more tied events that are in the cohort). The Breslow variation of the PL function should be used (not shown due to no tied events in the AiT cohort) [78].

Proportional hazards assumptions

All statistical models are based upon certain underlying assumptions and violation of these assumptions can result in less valid or even invalid model results [79]. Different model families each have their own specific assumptions which must be validated, in the case of Cox proportional hazards models the most critical assumption to validate is that of proportional hazards. The proportional hazards assumption implies that all individuals share the same baseline hazard function but only differ in scaling factor, i.e., for individual i with baseline hazard $h(t)$ the individual hazard is,

$$h_i(t) = a_i h(t), \tag{3.3.1.4}$$

where a_i is a scalar multiple that is invariant in time ($a_i(t) = a_i$). For any two individuals, i and j, a consistent hazard ratio should be present for all instances in time t,

$$\frac{h_i(t)}{h_j(t)} = \frac{a_i h(t)}{a_j h(t)} = \frac{a_i}{a_j}. \tag{3.3.1.5}$$

Where a is comprised of covariant coefficients $e^{\beta^T x}$ the proportional hazards assumption can thus be expressed as, $\hat{\beta} = \beta(t)$, where $\hat{\beta}$ denotes the PL estimate of the coefficients.

3.4 Pre-processing of data set

For biomedical applications, all predictors (or independent variables in associative works) in a data set are ideally clinically relevant, comprehensible, measured reliably, without missing data, and non-correlating - however this is seldom the case in practice [80]. In this section on pre-processing each of these areas will be explored and considered against the AiT data set. Additionally, more method specific pre-processing stages which are required to complete or improve analysis for the Cox proportional hazards models are considered, such as standardising and handling of categorical variables. Each variable used will be defined clearly and outlined how they are handled in the analysis.

3.4.1 Clinical relevance

Clinical relevance, comprehensibility and measurement reliability are all typically handled with the influence of background domain knowledge. An example of this could be in the case of observing that an increased risk of heart failure is linked to grey hair and baldness in a multi-variable analysis. From a clinical sense these associations are implausible and likely the relationship would not be demonstrated reliably under different sample populations. With adequate domain knowledge it is recognised that these findings are incorrect and more reasonably associated with age and male sex who also happen to demonstrate strong correlations with grey hair and baldness. In the AiT data set there are 146 characteristic variables and 211 DSA daily variables available. With only 99 cases this data set could be defined as wide and short with many more variables than samples for analysis. Regression type models are shown to perform poorly under short and wide conditions, yielding inconsistent modelling results with poor confidence on variable variance. Careful considerations along with consultation with clinical experts has allowed for the selection of just 15 characteristic variables for further analysis (highlighted in Table 3.3). Each of the DSA daily variables will be considered under their own merit.

3.4.2 Handling missing data

In many parametric based techniques, such as in the Cox proportional hazards model, missing data will result in complete omission of an observation from the analysis as a model output cannot be determined without complete data. For these cases imputation techniques are largely considered for predictive or associative analysis. Incomplete

data is commonplace in biomedical data sets and several mechanisms have been developed to handle missing values where one of the most popular techniques is single value imputation. In this technique missing values are predicted based upon other baseline characteristics either through substitution via global mean (or mode for categorical data), model-based imputation, or from the k-nearest neighbours. Each approach has its own drawbacks which should be carefully considered for a data set. The model-based approach for example may perform poorly given a weakly predictive model which is often the case in low data scenarios. Alternatively, substitution via mean does not consider correlations between other variables and if enough data are missing may introduce bias if imputed values are all at the mean. If a large enough proportion of data is missing from a sample it is recommended to use a multiple imputation technique which creates multiple versions of the data set with differently imputed values or otherwise resort to different strategies which can work around missing data [81]. Additional assumptions about the mechanism for which data is missing need also be considered and can be broken down into the following three categories:

- **missing completely at random (MCAR)** – probability of missing data is independent of all other variables, including the response being predicted. In this case there are no systematic differences between missing and observed values.
- **missing at random (MAR)** – probability of missing data is independent of only the response being predicted but may depend on other variables.
- **missing not at random (MNAR)** – probability of missing data is dependent on the values of unobserved data.

Current strategies for missing data are only able to operate under MCAR and MAR conditions as the probability of missing data is independent of the reason it is missing. In the AiT data set relatively few data are missing (<1% of samples) and are considered MCAR in nature, so a k-nearest neighbours approach is used to impute missing value. Within the literature a value of $k = 5$ is commonly used along with the Euclidean distance measure to identify sample similarity [82]. This measure is a special case of the Minkowski norm,

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \tag{3.4.2.1}$$

$p = 2$. Similarity is computed for each other set of observations and the k most similar

observations are utilised to form the imputed value via assuming the mean of those observations. In our case only continuous variables had missing data and so the case for developing a similarity measure given dichotomous, ordinal, and categorical variables is not considered however has been explored in literature.

3.4.3 Variable encoding

Data types can come in a wide variety of forms such as: temporal, spatial, image, continuous, ordinal, or categorical, however due to the numerical nature of statistical and machine learning (ML) techniques will typically be restructured to numeric format for analysis. For biomedical data sets it is common to see both ordinal and categorical data forms such as gender or integer counts. This data is often valuable and cannot simply be disregarded in models which only accept numerical input. As a consequence non-numerical data formats typically undergo a variable encoding pre-processing stage prior to analysis. For ordinal data it may be reasonable to assume the data as continuous in analysis, alternatively if there is suspected non-linearity, handling as categorical data may better represent the underlying structure. For categorical data it is not possible to directly convert categories to a numerical structure: e.g., category A $\rightarrow$1, category B $\rightarrow$2, etc due to a non-monotonic nature. Instead to address the challenges a categorical variable with n categories is replaced by n coded variables. We can represent this transformation in the following example where we look at how the patient hazard rate is effected by region in which treatment takes place. First we define the survival model with the regional categorical variable x_R and the associated coefficient b_R,

$$h(t) = h_0(t)e^{b_R x_R}, \tag{3.4.3.1}$$

where the regions may belong to any of the following four categories,

$$\text{Region} = x_R \begin{cases} \text{North} \\ \text{South} \\ \text{East} \\ \text{West} \end{cases} \tag{3.4.3.2}$$

Next, x_R is substituted for the four coded variables x_N, x_S, x_E, x_W each with their own

associated coefficients b_N, b_S, b_E, b_W,

$$h(t) = h_0(t)e^{b_N x_N + b_S x_S + b_E x_E + b_W x_W} \tag{3.4.3.3}$$

Several coding methods exist in the literature such as Dummy, Contrast and Helmert coding however when prioritising associative analysis Dummy coding is considered superior [83]. Each method changes the meaning of interpreting the coded coefficients. For Dummy coding, each coefficient, b_i corresponds to the difference in hazard ratio between category i and the reference category and an example of such coding can be seen in Table 3.2.

Region	x_N	x_S	x_E	x_W
North	1	0	0	0
South	0	1	0	0
East	0	0	1	0
West	0	0	0	0

Table 3.2: Example of dummy variable introduction for categorical variable. Categorical variable of region has $n = 4$ outcomes which is replaced by n dummy variables in the regression model. One category is considered as a reference, in this example West, which is effectively removed from the model by encoding of zero. The influences of the category West are considered in the baseline hazard function h_0.

The effective removal of one category as a reference category (incorporated into the baseline hazard function h_0) avoids over-parameterisation from occurring during model inference - otherwise known as the dummy variable trap [84].

3.4.4 Standardising variables

Standardisation is the process of adjusting variables to exist on the same scale as one another. This is often a critical process in regression based techniques which include interaction or polynomial terms as it reduces multi-collinearity between variables - improving consistency in estimation of coefficients. Additionally it allows for direct comparison of coefficients in order to estimate variable importance. In this case continuous variables are transformed via z-score scaling which re-scales the original variable to have a mean of zero and standard deviation of one. This is calculated mathematically using, $z = \frac{x-\bar{x}}{\sigma}$, where z is the z-score transformed variable, x is the original variable, $\bar{x}$ is the mean of x, and σ is the standard deviation of x.

3.4.5 Results following pre-processing

Following the pre-processing stages there are 17 variables (shown in Table 3.3) of which two correspond to output i.e., follow-up time and state of censorship, and 15 are potentially clinically relevant predictor variables. Of these predictor variables seven are categorical in nature and have undergone dummy coding with baseline category highlighted in bold font. Missing data, which makes up <1% of data is imputed using k-nearest neighbour (k=5) single value imputation and is indicated via values in the notation [*]. Factors were considered based on the available information present within the data set and can generally be related to one of the following seven categories: recipient details, recipient background, donor compatibility, pre-transplant (trx) status, pre-trx treatment, early post-trx health, and recipient outcome. Recipient details correspond generally to the demographic data available, such as age, weight, and gender. Generally speaking the cohort consisted of mostly middle aged persons with a median weight below the UK average of 76 kg. Notably, a larger population of females are present in the cohort, occupying approximately 60% of cases. Recipient background looks at the history of a patients' RRT treatments with the total established renal failure (ERF) time corresponding to the combination of time spent on dialysis and time spent with previous donor transplantations. The number of previous transplants, which ranges from 0-3 in this cohort, has been reduced from an ordinal to categorical variable for the purposes of analysis. Donor compatibility highlights factors which are typically considered by clinical teams when matching donor to recipient. DDT for example are shown to have worse long term outcomes [11]. For the purposes of the analysis patient DSA are investigated for the total number of mismatches between the donor and recipient, in addition to the class specific mismatches for CI and CII. Prior to transplantation clinicians were primarily concerned with ensuring that patients achieved a negative FC crossmatch status and would routinely administer DFPP treatment leading up to transplantation in order to achieve this. The number of cases which achieved this status, and the number of treatment sessions administered, are shown in the pre-trx status and pre-trx treatment sections, respectively. The number of DFPP sessions administered to patients to reduce DSA levels prior to transplantation is shown in the pre-trx treatment section. Lastly, two factors are used to represent the early post-transplant monitoring of patient health. Day-4 creatinine levels which were used to determine if a post-transplant biopsy was conducted, and occurrence of ABMR which is a factor commonly associated with poorer long term graft outcomes [65, 85].

Variables	Variable type	All cases (*N*=99)
Recipient details		
Age (years)	Continuous	43 (34-50)
Weight (kg)	Continuous	69 (61-77)[2]
Gender (**m**, f)	Categorical	39, 60
Recipient background		
Previous transplants (**0**, >0)	Categorical	40, 59
Total ERF time (years)	Continuous	11 (3-17)[1]
Dialysis time (years)	Continuous	5 (2-8)[1]
Transplant time (years)	Continuous	3 (0-10)[1]
Donor compatibility		
Donor status (**LDT**, DDT)	Categorical	88, 11
Number of DSA	Discrete	2 (1-7)
Number of CI DSA	Discrete	1 (0-4)
Number of CII DSA	Discrete	1 (0-5)
Pre-trx status		
Crossmatch (**FC-**, FC+, CDC+)	Categorical	27, 52, 20
Pre-trx treatment		
DFPP (**n**, y)	Categorical	26, 73
Post-trx health		
Day 4 creatinine (mg/dL)	Continuous	144 (104-234)
Early acute rejection episode (**n**, y)	Categorical	54, 45
Recipient outcome		
Follow-up time (years)	Continuous	8.7 (5.6-11.5)
Censored data (n, y)	Categorical	25, 74

Table 3.3: Characteristic table of AiT kidney cohort (*N*=99). [*] indicates number of values which have been imputed. Values for: continuous variables correspond to median (quartile range), categorical variables correspond to number per category. Baseline categories for predictor variables are shown in bold.

3.5 Univariate analysis of predictor variables

In most systems predictor variables are often not univariate nor linearly related to output. Despite this it is still often useful to explore univariate relationships between predictor and output variables. Univariate relationships can be indicative of where to explore for more complex nonlinear relations and offers broad context. Additionally, in cases which demonstrate a short and wide data set, i.e., with many predictor variables and few samples, narrowing down the number of predictor variables is a key process to extract useful information in the multivariable analysis. For these reasons a univariate analysis will be conducted first on the previously detailed variable list from Table 3.3. The univariate analysis is completed using the Cox proportional hazards model with a single variable inputted into the model at a time. For categorical models all encoded variables are included

in the same model. Each variable will yield a coefficient or hazard ratio which indicates the hazard relative to the baseline.

3.5.1 Testing model assumptions

Following the pre-processing stages, and prior to model regression, variables are first assessed for their adherence to the underlying assumptions of the Cox proportional hazards model which is primarily that of the proportional hazards assumption. As failure to comply with this assumption may impact the validity of a given model, violations will need to be addressed before analysis can be completed. There are primarily two approaches to testing the proportional hazards assumption in the literature: visual inspection via plotting survival curves or statistical analysis using methods such as the Schoenfeld residual test [86]. Frequently both are utilised where initially the Schoenfeld test screens potential violations before covariates are then visually inspected through survival plots to aid in decision making [80].

Approaches to testing proportional hazards assumption

Schoenfeld residuals are typically calculated prior to a multivariable analysis individually for each covariate. The coefficient β is calculated for each non-censored sample in an alternative model which allows β to vary with time, given as $\beta(t)$, these values are then compared to the maximum likelihood (ML) estimate $\hat{\beta}$,

$$\beta(t) = \hat{\beta} + \varepsilon_s(t), \tag{3.5.1.1}$$

where $\varepsilon_s(t)$ is the error term or Schoenfeld residual. A proportional hazard for non-censored observation at time t will yield $\varepsilon_s(t) = 0$ and validating of the proportional hazards assumption is achieved by plotting ε_s over time. An average or locally weighted scatter-plot smoothing (LOESS) [87] line is usually plotted to inspect the data, however a formal test (Schoenfeld F test [88]) has also been developed to determine if the data deviates significantly from a flat line about $\hat{\beta}$. An additional step typically normalises values about zero. Figure 3.1 shows an example of the scaled Schoenfeld residuals for a particular variable 'average calories consumed per meal' from a lung cancer data set (sourced from the R survival package [89], original data from [90]). The Schoenfeld residuals deviate significantly from a flat line (p=0.043, Schoenfeld F test) and therefore

proportionality is not valid.

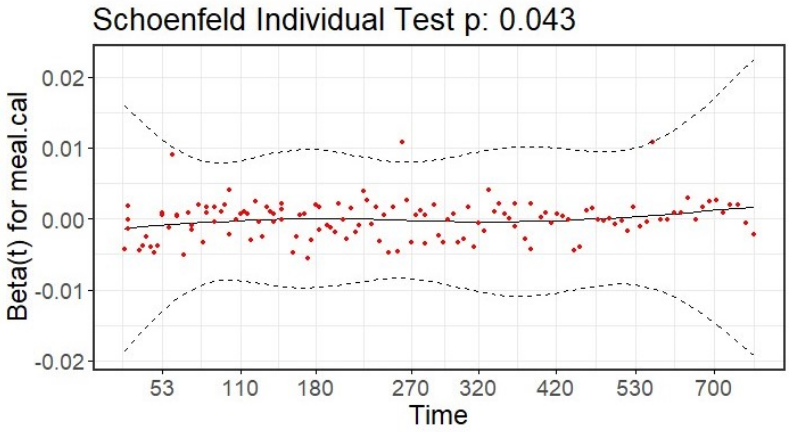


Figure 3.1: Schoenfeld residuals example. Schoenfeld residuals plotted against transformed time for a univariate Cox proportional hazards model investigating 'average calories consumed per meal' to lung cancer (data from [90]). Censored data (red) are smoothed with a LOESS curve (solid black line) overlaid with 95% confidence interval (dashed black line). Schoenfeld F test indicates non-proportional hazards at p<0.05.

Before assessing visually it is typically advantageous to first convert Kaplan-Meier survival function, $S(t)$, into the cumulative hazard function $H(t)$,

$$H(t) = -\log_e S(t), \tag{3.5.1.2}$$

which is then used to plot the $\log(-\log(S(t))$ vs. $\log(t)$ (referred to as the log-log plot) to determine hazard proportionality between two different data sets. This is advantageous over using survival function to check for proportionality as it can also be used to estimate the log hazard ratio by the vertical distances between two lines in the plot. Figure 3.2 shows three examples of how this plot may be presented for a two class data, i.e. patient sex, male and female. For the proportional case (A) the log cumulative hazard is parallel representing a constant hazard ratio across time. In (B) the hazard ratio is only partly proportional and in (C) not at all, thus violating the assumption.

Results of testing proportionality assumption

To begin with each variable is checked for their adherence to the Cox proportionality assumption via the Schoenfeld F-test as shown in Table 3.4. Significant variables (p-value<0.05) are those which show time dependent variation, and therefore violate the assumption. Only one variable, 'Cross match status', is shown to violate the assumption with a p-value = 0.046 (highlighted in bold). Several options are available to handle such

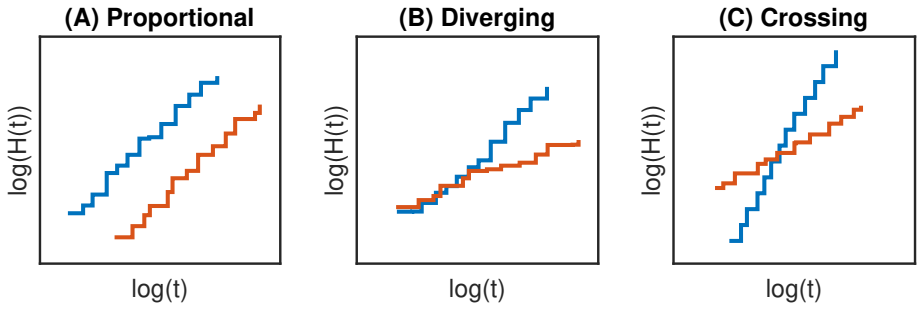
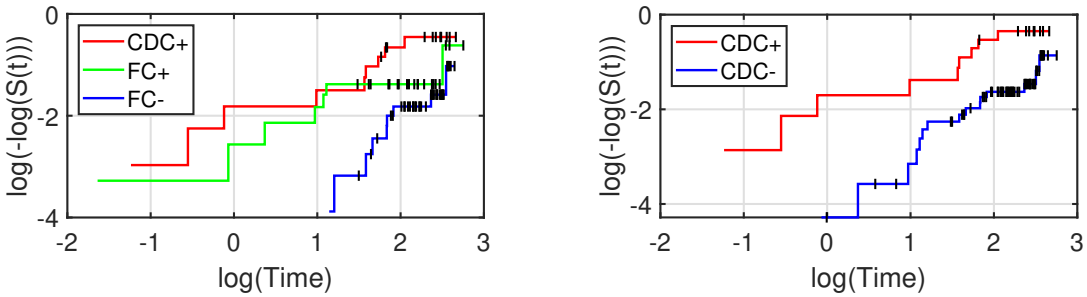


Figure 3.2: Example assessment of proportionality through log cumulative hazard function vs. log of time. Categorical variables can be assessed for adherence to proportionality via comparison of the log cumulative hazard function vs. log of time plot. (A) Proportionality satisfied when plots are parallel. Separation between curves indicates an estimate of log hazard ratio. (B) Proportionality is satisfied early on however hazard ratios diverge (or converge) at a later period thus invalidating the assumption. (C) Hazard ratios are not proportional as plots are not parallel thus invalidating the assumption.

a violation which have been explored in the literature, such as introducing a time dependent coefficient in its place [91], or combining groups if it is appropriate in categorical variables. In the case of crossmatch status it is reasonable to consider the combination of FC- and FC+ groups into a new CDC- group, as a CDC+ test is generally considered to be the worst outcome. This solution is then considered under the log cumulative hazard function vs. log of time plot to assess for its viability, shown in Figure 3.3. When observing the original grouping for crossmatch status the cause of the proportionality violation can be attributed to the crossing of FC+ group briefly over CDC+ group in the mid time period. This shows a time dependent proportional behaviour with similar characteristics early post transplant before diverging in later periods. The revised crossmatch groups solve for this violation and present an avenue for crossmatch status to be reliably included in the study.



(a) Log-log plot of crossmatch groups.

(b) Log-log plot of revised crossmatch groups.

Figure 3.3: Assessment of proportionality assumption among crossmatch status groups. The log-log plot compares the log of the cumulative hazard function, or $\log(-\log(S(t)))$, against the log of time, $\log(t)$. Proportionality between groups is satisfied when plots are close to parallel.

Variable	Data type	p-value
Patient age	Continuous	0.32
Patient weight	Continuous	0.31
Patient gender	Categorical	0.20
Previous transplants	Categorical	0.86
Time with ERF	Continuous	0.67
Time on dialysis	Continuous	0.56
Time transplanted	Continuous	0.95
Donor type	Categorical	0.14
Number of DSA	Continuous	0.80
Number of Class I DSA	Continuous	0.17
Number of Class II DSA	Continuous	0.45
Cross match status	Categorical	**0.05**
Pre-transplant DFPP	Categorical	0.56
Creatinine level*	Continuous	0.10
Early acute ABMR	Categorical	0.44

Table 3.4: Schoenfeld individual test. The Schoenfeld individual test is performed to identify any variables which violate the proportionality assumption, as given by a significant (i.e., p-value <0.05) test result. *Day four, post transplant.

3.5.2 Estimating univariate associations with graft outcome

Each variable has been shown to satisfy the Cox proportional hazard assumption of proportionality suggesting that model estimates should be reliable for the sample population. In this section each variable is considered in a Cox proportional hazards model independently for their associations with graft survival and an estimate of both the HR and its respective confidence interval is conducted. Estimation of variable confidence intervals is typically completed in one of two ways: parametrically, whereby a sample distribution type is assumed, or non-parametrically whereby no assumptions on the underlying distribution are made. In this case a non-parametric approach known as bootstrapping is used due to limited knowledge about the underlying distributions of the parameters under analysis. This section will first briefly cover bootstrap theory before its application in Cox proportional hazards analysis.

Bootstrap theory

Bootstrapping is a non-parametric statistical approach which, in this case, has the aim of creating an estimate for a population parameter. To do so, data is resampled multiple times from the original sample, creating many simulated data sets. Each simulated set is then used to independently form an estimate which, collectively, form a sampling distribution of the parameter. From this distribution statistical inferences about the parameter

can be made such as calculation of the standard errors, confidence intervals, and hypothesis testing. An underlying assumption for the application of bootstrapping is that the original sample is representative of the true sample distribution from which it is drawn - meaning that it is unable to account for any unforeseen bias in the sampled data set. Practically, bootstrapping serves to mimic the process of running multiple real experiments and then aggregating their results to form an improved estimate. This process is highlighted in Figure 3.4 for a simulated data set representing the height of a population with a true distribution from which samples are drawn being represented by a normal distribution $\theta \sim \mathcal{N}(\mu = 182, \sigma = 10)$. To estimate bootstrap confidence intervals a percentile bootstrap method is used [92]. This method observes the distribution of parameter estimates - in the previous section this would be $\bar{\theta}$ - and then orders them from highest to lowest. From this ordered set of bootstrap samples the confidence intervals are found by first identifying the integer positions of the lower bound (LB) and upper bound (UB),

$$\text{LB} = \lfloor \frac{\alpha \cdot N}{2} \rceil \tag{3.5.2.1}$$

$$\text{UB} = \lfloor N - \frac{\alpha \cdot N}{2} \rceil \tag{3.5.2.2}$$

where α and N are model specific parameters with α representing the desired level of significance (e.g., $\alpha = 0.05$ for 95% [93]) and N representing the total number of bootstrap samples. The lower and upper bounds of the confidence interval, $\bar{\theta}_{\text{LB}}$ and $\bar{\theta}_{\text{UB}}$, can then found by checking the corresponding θ values from the ordered set.

Results of univariate analysis

Results for univariate estimates HR have been conducted using a bootstrap analysis with $N = 2500$ simulated data sets. Each data set consists of resampled data up to the original size of the data set (i.e., 99 samples) with stratified sampling implemented to ensure fair representation of both censored and uncensored classes in each set (i.e. 25 uncensored and 74 censored samples, respectively). Figure 3.5 shows the results of this process for all 15 considered variables. Due to standardisation of the data occurring within preprocessing stages, interpretation of the continuous variable HR should be as follows - for every one unit increase in standard deviation of the variable, patient HR will experience a multiplier by the corresponding amount. Variables with the biggest mean difference

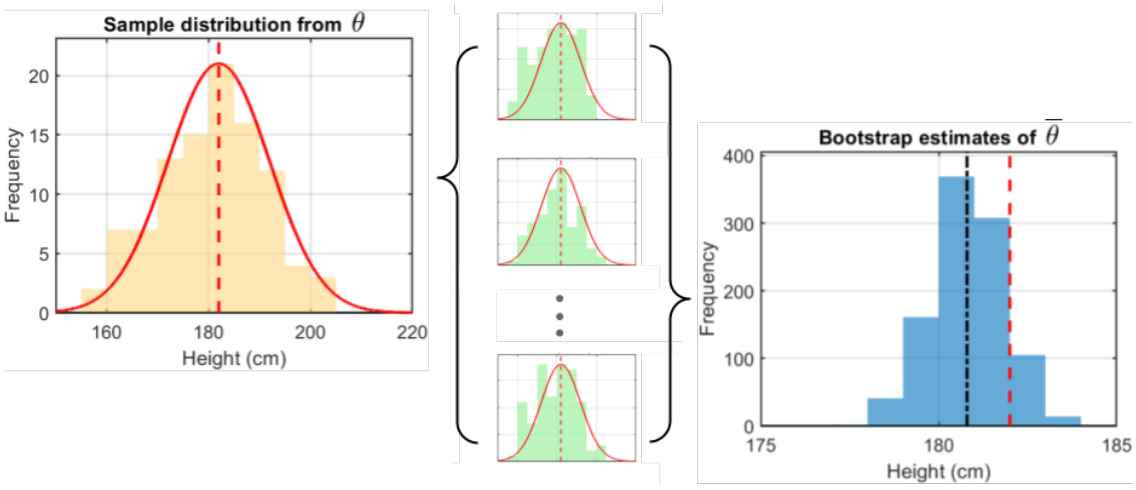


Figure 3.4: Example of bootstrapping approach to provide a parameter estimate. Left - A sample distribution of N samples (yellow histogram) is drawn from the true distribution (red outline, mean in red dashed line) representing the measurements of height, θ, from a study population. Middle - To estimate the mean height, $\bar{\theta}$, and confidence interval a bootstrap resampling approach is used to produce b simulated data sets. For each simulated set an estimate of the desired parameter, mean height θ, is computed. Right - resampling estimates are combined to form a distribution of the parameter estimate $\bar{\theta}$. Bootstrap mean estimate of $\bar{\theta}$ shown in black dashed line, true $\bar{\theta}$ in red dashed line.

from a HR=1 are ordered first in Figure 3.5 indicating that they are likely to be of most importance. Increased patient age, increased time on dialysis and negative CDC crossmatch status each demonstrate a reduced HR, by factors 0.66 (for every 12.0 years), 0.7 (for every 4.6 years), and 0.68, respectively. A high degree of confidence is shown in HR estimates due to tight bootstrap distributions. For less significant variables: time on ERF, no previous transplants, and early ABMR show slight associations with graft survival, however confidence in these estimates is relatively weak. The remaining 9 variables show little to no association with graft failure due to mean values centred close to HR=1.

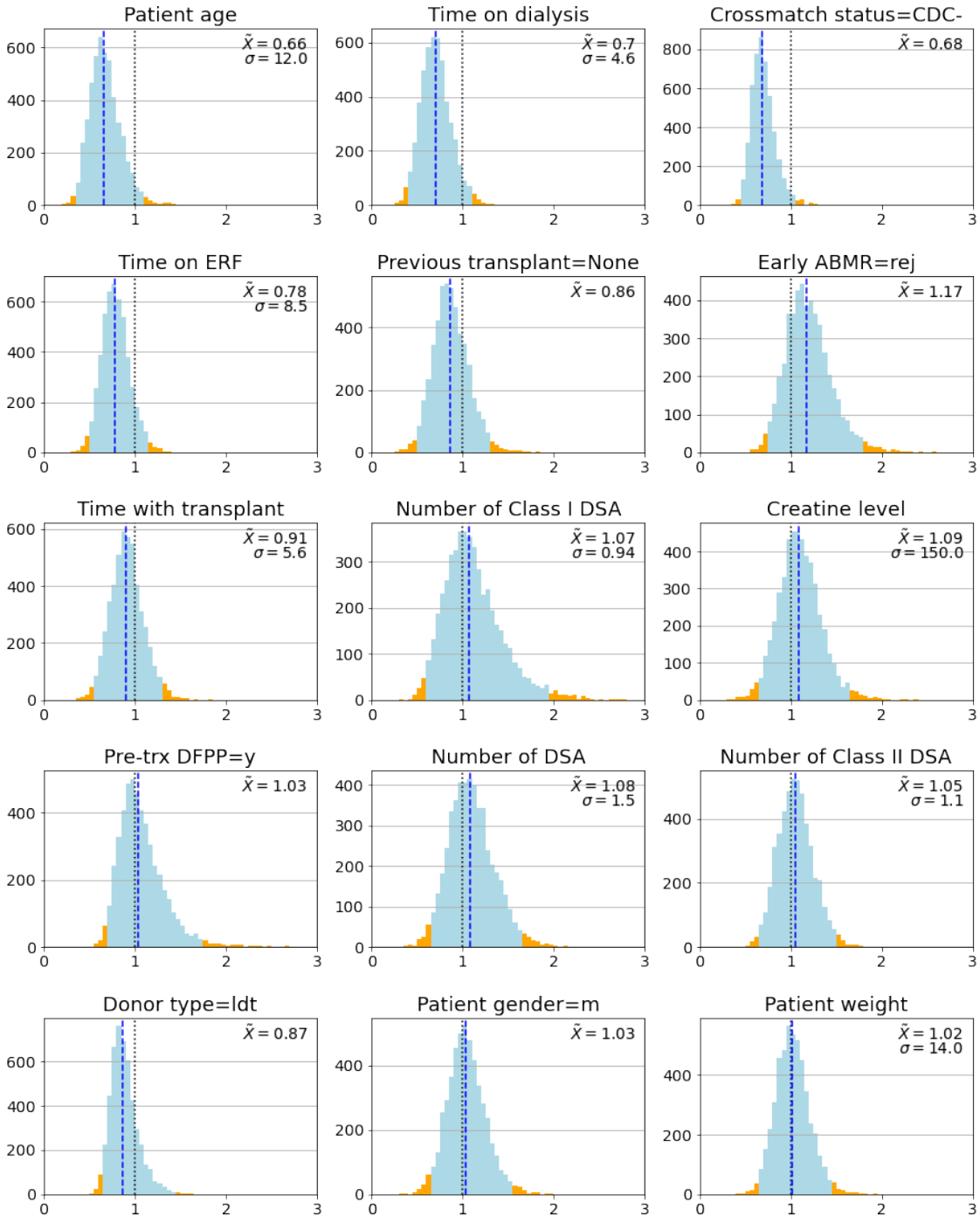


Figure 3.5: Histogram results of univariate Cox proportional hazards estimates for associations with graft survival. In total 15 variables were considered for associations analysis. Analysis was conducted via bootstrap resampling which simulated $N = 2500$ data sets for estimation of variable HR. Confidence intervals (95%) are highlighted in blue. Dark blue line corresponds to the mean estimate, the value of which is indicated in top right corner of each plot. A value of 1 indicates no impact on graft survival. σ indicates the amount corresponding to a one unit increase in the associated variable.

3.6 Multivariable analysis of predictor variables

From the univariate analysis it is clear that there are multiple factors that express associations with graft survival - indicating that their relationships should be considered under a multivariate environment. For an associative task, a multivariate analysis is used to better estimate variable coefficients under the condition of potentially confounding factors. A clear example could be in the association of a populations weight from a set of measurements of height or waist size. In either case the variable will be able to explain some variance, however an improvement would be possible when accounting for the influence of both variables in analysis. For biological applications there could be many variables considered in analysis due to the complexity of a system that has evolved over thousands of years. While it may be tempting to include large numbers of predictor variables this can often lead to dilution of the true associations with outcome; leading to large standard errors with wide confidence intervals in methods such as regression techniques [94]. In this section the 15 variables assessed under univariate conditions are further investigated for potential multivariate associations with graft survival. As the AiT data set could be considered both wide, meaning it has large numbers of variables, and short, meaning there are relatively few samples, it is practical to undergo a variable selection process prior to final assessment of multivariate associations. There are two stages of variable selection implemented, namely, an investigation of variable correlations and stability.

3.6.1 Variable correlations

An assumption of regression models is the lack of perfect multi-collinearity, or correlation, between variables due to the consequence of unidentifiable variable coefficients. For example, a model $y = A \cdot X_1 + B \cdot X_2$ where variables X_1 and X_2 are perfectly correlated, i.e., $X_1 = X_2 = X$, results in model simplification to $y = (A+B) \cdot X$ where neither values of variables A and B can be known. This leads to a case where the model regression coefficients cannot be trusted. In the case of estimating variable associations, high correlations should be removed from the data set - usually by removing one or the other variable in a pair. This differs from predictive analysis where the desired outcome is not an accurate estimate of variable coefficients. Variables in this case which are not perfectly correlated may still offer further predictive value to a model and therefore should still be included. To assess variable correlations a correllelogram is formed based on Pearson's correlation

coefficient calculated independently for each pair of variables in the data set. The theory and results of this process are covered next.

Assessing variable correlations

The Pearson's correlation coefficient [95] is an approach to measure the association between two variables on a continuous scale - lending itself useful in analysis of the 15 variables: 6 continuous, 6 binary (numerically encoded), and 3 (ordinal, handled as continuous). The Pearson's correlation coefficient, ρ, is defined by the covariance between two variables, X and Y, divided by the population standard deviations, σ_X and σ_Y, of those two variables,

$$\rho(X,Y) = \frac{COV(X,Y)}{\sigma_X, \sigma_Y}, \tag{3.6.1.1}$$

where for n samples in the data set, and population means μ_X, and μ_y,

$$COV(X,Y) = \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)/(n-1). \tag{3.6.1.2}$$

The values of ρ can occupy values between -1 and 1 which indicate the strength and direction of the correlation between two variables [96]. A value of 1 indicates two perfectly correlated variables, a value of -1 indicates two variables with a perfect negative correlation, and a value of 0 indicates that there is no correlation between variables at all.

Results of variable correlation analysis

Figure 3.6 illustrates the resulting correllelogram produced via computing the Pearson's population correlation coefficient for each pair of variables. The figure can be interpretted based on both marker size and colour. The size of the marker corresponds to the magnitude of the correlation coefficient and the colour to its direction, i.e., red being negative, and blue being positive correlation. Large correlations then, are found between certain variables which come as no surprise. Time on ERF (a), time with transplant (b), and time on dialysis (c) can, for example, be related to one another by the following expression $a = b + c$ - and show corresponding correlations. Additionally, a negative correlation is also formed between having no previous transplants and time with transplants/ERF which carry similar information. Other variables that have strong associations are the number of

DSA groups: total, CI, and CII - which similarly have an $a = b + c$ relationship. Several weak correlations are also present in the data, however are not strong enough to require action. Patient age, gender and crossmatch status for example show weak correlations with almost all variables. To handle multi-collinearity conflicts the results from the univariate analysis is used to assist in eliminating highly correlated variables. From this, only the most significant variables are sought to be included in further multivariate analysis. Time on dialysis, previous transplants (=none), and each of CI and CII DSA were selected to remain in the analysis based on this premise. Variables: time on ERF, time with transplant, and total number of DSA have been removed due to high correlations.

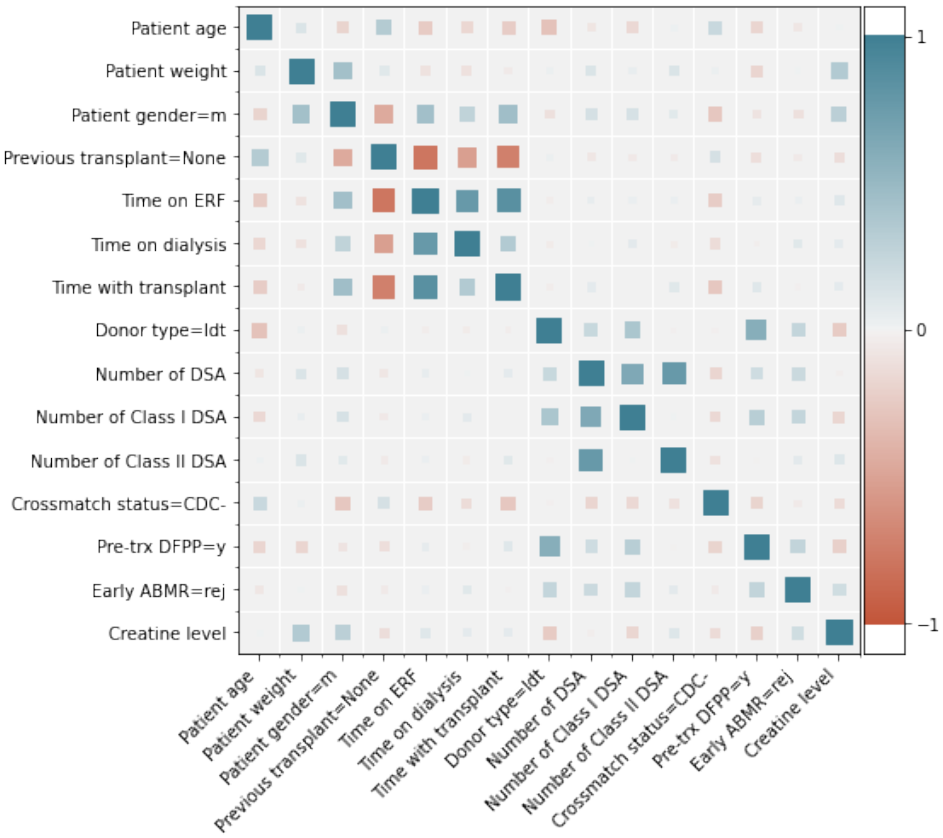


Figure 3.6: Correlation plot of univariate variables. Each variable is assessed for correlations via the Pearson's correlation coefficient. Variables are standardised and categorical variables are represented in dummy variable form. Each correlation is denoted by two features: size and colour. The size corresponds to the absolute correlation coefficient, and the colour indicates if the correlation is positive (blue) or negative (red).

3.6.2 Variable selection

After assessing and removing variables with high correlations a variable selection stage is conducted. Variable selection is a process whereby a larger number of candidate variables are reduced to a smaller subset which is considered for final inclusion in an associative

model. There are two primary reasons for doing so. Firstly, it aids in the selection of variables which are related to outcome, while simultaneously excluding those which are not - making the model more accurate. Secondly, it helps to select a model with the fewest number of variables by eliminating irrelevant variables that would otherwise decrease precision and increase model complexity [97]. Additionally, from a practical point of view, including a large number of predictor variables can often lead to dilutions of their true associations with outcome; leading to large standard errors with wide confidence intervals [94]. Simpler models are typically easier to generalise, interpret and apply in practice, however, one must be cautious to ensure that important variables are not excluded. In general there is no set rule for the number of variables to include in a model as it can depend on several factors. A number of 'traditional' rules commonly applied in clinical modelling strategy (e.g., for logistic regression and survival models) typically place this number between 10-15 events per variable [98], however in some cases this is higher or lower [97]. With $N = 99$ cases in the AiT cohort we could reasonably expect a multivariable model to support between 6-10 variables and maintain accuracy.

Variable selection techniques

There are many variable selection techniques available within the literature with two of the most common techniques being forwards and backwards selection. In these selection approaches variables first undergo a univariate analysis whereby each variable is fit independently of one another. Variables which show significant association with the output (typically $p < 0.25$) or variables which are already demonstrated/known to be clinically relevant are to be considered for further analysis. Optimising the variables within a model is then completed iteratively for both types of model. In forwards selection a model begins with no variables. Variables are then systematically tested to identify significant contribution to the model outcome. The most significant variable is added and the process repeated until variables no longer contribute a significant difference. For backwards selection this process is completed in reverse and begins with a full model including all variables. The least significant variables are then systematically removed until all remaining variables are considered to have significant contribution. The approaches have proved popular and effective in biomedical literature [99], with additional variations such as step-wise elimination (which includes both forwards and backwards steps), and all-possible subset selection (brute force analysis of all combinations) also used [97]. In

recent studies the advantages of using penalised regression techniques instead to perform variable selection have been investigated and are shown to be a worthy competitor or even superior to iterative selection techniques [100]. These approaches generally yield less variable and yet equally interpretable models, particularly in tasks with few samples as in the AiT cohort. For these reasons a regularised technique known as least absolute shrinkage and selection operator (LASSO) is used for variable selection in this chapter and is explained in the following section.

LASSO regularisation

The Cox proportional hazards model is often an effective choice for associative analysis due to its coefficients being interpreted in terms of a hazards ratio. With a high number of variables however the coefficients become unreliable. This is caused due to an attempt to invert a matrix which becomes singular due to correlations between features. Mathematically the problem can be avoided by introducing an l_1 penalty term on the coefficients, a process otherwise known as LASSO regularisation. LASSO regularisation differs from other forms of regularisation in that variable coefficients can be reduced to zero - effectively eliminating them from the model. This is in contrast to types such as ridge and elastic-net which may only shrink values towards zero. LASSO regularisation can be expressed by the optimisation problem for β,

$$\hat{\beta} = \arg\max_{\beta}(\log(\mathcal{L}(\beta)) - \alpha \sum_{j=1}^{p} |\beta_j|), \tag{3.6.2.1}$$

where $\hat{\beta}$ is the optimal set of β found by maximising the function, and α is a scalar value which implicitly determines the number of model features. The size of alpha influences the regression coefficients with a larger α corresponding to a larger penalty term and a value of $\alpha = 0$ corresponding to standard Cox regression analysis with no penalty. For LASSO variable selection approaches it is typical to optimise a model based on varying the α value. As α increases parameters are shrunk closer to zero, and the least influential parameters will be systematically eliminated from the model. Figure 3.7 shows this process for the investigated data set. A range of α's are considered from 10^{-3}, in which all 12 variables are still contained in the model, through to a value approaching 1 where by only a single variable remains in the model. From here selecting of an appropriate α is then dependent on maximising of a user specified quantity, in this case the Harrell's

C-index.

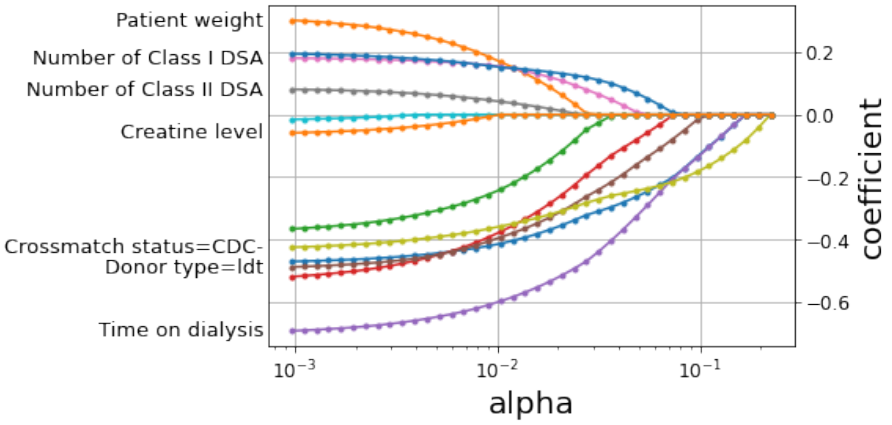


Figure 3.7: LASSO regularisation for variable selection. By varying α the magnitude of the l_1 penalty term is increased, systematically reducing variable coefficients to zero. *Only a subset (7/12) variables are labelled for clarity.

Harrell's C-index

Harrell's C-index (also known as the concordance index) is a goodness of fit measure for models which produce risk scores - such as the hazard ratio in Cox proportional hazards models. It is commonly used to evaluate risk in survival analysis where data may be censored and can be thought of as a generalisation of the receiver operating characteristic (ROC) for survival data. The intuition is as follows: for patient, i, assign risk score η_i from the risk model. If the risk model is good then patients who have shorter time to event, T, will have higher risk scores and vice versa. By considering any two patients it can then be inferred that the patient with the higher risk score should have had a shorter time to event. The C-index can then be computed by observing the risk scores and time-to-event for all patient pairs. The logic for Harrell's C-index can be expressed succinctly through the following equation for censored data (as described in [101]),

$$c = \frac{\sum_{i,j,i\neq j} I(\eta_i < \eta_j) \cdot I(T_i > T_j) \cdot \Delta_j}{\sum_{i,j,i\neq j} I(T_i > T_j) \cdot \Delta_j} \tag{3.6.2.2}$$

where Δ_i is the state of censoring for observation i, and $I(\cdot)$ is the indicator function mapping logic statements true = 1 and false = 0. The sum $\sum_{i,j,i\neq j}$ is carried out for all pairs i and j except $i = j$. The C-index can form values between zero and one with a value of 0.5 indicating random chance of determining which patient will live longer (i.e. no better than a coin flip) and a value of 1 being a perfect score. To increase confidence in the C-index estimation a stratified k-fold cross-validation scheme is implemented on the

data set. Cross-validation is an approach used to reduce over-fitting and serves to improve the generalisability of the model to new data. The cross-validation process, illustrated in Figure 3.8, requires partitioning of a data set into k approximately equal sized chunks. Chunks will then iteratively represent the testing set for which the desired quantities, i.e., concordance, are calculated. The rest of the data set take place of the training set for which model inference is conducted upon. Ultimately, k estimates of the calculated quantities are produced which can be used to determine a mean and confidence interval for further analysis. By lastly stratifying the partitioning process, chunks are ensured to have a fair representation of output classes (i.e, censored and not-censored data).

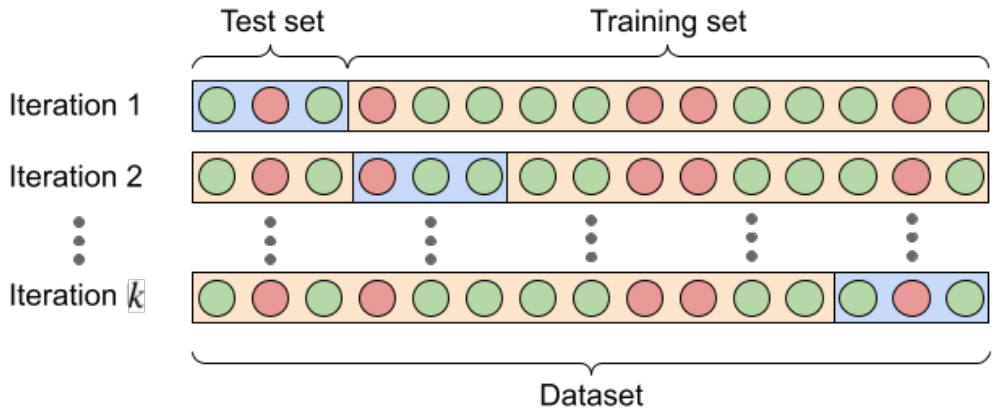


Figure 3.8: k-fold cross-validation scheme. For a k-fold partition scheme there will be k iterations through the data set. On each iteration a new partition of the data will represent the testing set, and all others will represent the training set. For each iteration the training set will be used to infer model parameters while the testing set will be used to calculate desired output quantities.

For each value of alpha investigated in LASSO regularisation analysis (Figure 3.7) the stratified 5-fold cross validation of Harrell's C-index is calculated (illustrated in Figure 3.9). The mean value of the C-index is shown in dark blue with the minimum and maximum values indicated by the highlighted region. In orange, the maximum C-index is found at a value of $\alpha = 0.154$, showing a distinctive peak with tighter confidence intervals than other values. From this it is found that crossmatch status, patient age, and time on dialysis are the variables which should be selected with all other variables discarded from the model.

3.6.3 Stability of variable selection solution

An important, and often ignored, element of data-driven variable selection techniques is that of model stability - this is to say the robustness of a given model to small sampling

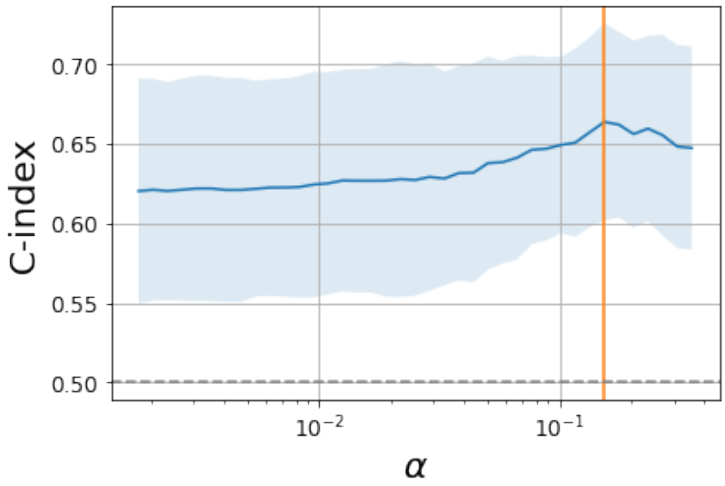


Figure 3.9: Cross-validated Harrell's C-index results. To select an optimal value of α for the LASSO regularised Cox proportional hazards model (Figure 3.7) the concordance index is maximised. Here the mean concordance-index is shown in dark blue, minimum and maximum boundaries are shown in blue, and α at the maximum concordance index represented by the orange line.

perturbations in the data set [102]. Irrespective of the given variable selection procedure, the goal of the associative analysis is to produce one specific model out of a large set of possible candidate models. In many cases however, it is likely for there to be several competing models which fit the data similarly well and may ultimately be chosen dependent on the characteristics of but a few of observations.This is to say that by only slightly altering the data a different model may be selected. The frequency at which variables may be selected then can be thought of probabilistically, with variables that demonstrate a stronger effect on outcome are selected at a higher rate, and those with weaker to no effect may be present only by a matter of chance [103, 104]. In this section a bootstrap analysis of the variable selection procedure (Section 3.6.2) is conducted where inclusion frequencies of specific variables, and particular models are investigated.

Bootstrap stability analysis methodology

This bootstrapping methodology builds upon the existing foundations established in Section 3.5.2, which used the approach to estimate parameter the coefficient mean and distribution based upon a defined univariate model. The basic idea for using bootstrap analysis to assess stability is to draw n bootstrap samples from the original data set and perform variable selection independently for each one [97]. The important quantities which are desired are (i) the bootstrap inclusion frequencies of variables, (ii) the bootstrap inclusion frequencies of particular model configurations. Selection of an appropriate model

can then be based off of a pre-defined cut-off value for inclusion frequencies. For the particular task of assessing stability in resampling based model selection approaches, it has been shown that a subsample of the data set without replacement performs superior to a regular bootstrap sample in the task of identifying uninformative variables (i.e., produces low inclusion rate) [105]. In this case a 90% subsampling procedure is performed without replacement $n = 5000$ times and the results are aggregated. The final model evaluation will be based on variables which appear in selected models at greater than 50% inclusion rate.

Results of multi-variable associations

The results of the bootstrap stability analysis are shown in Table 3.5, indicating both the variable inclusion frequency, rate (%), and the model inclusion frequency for the top three most commonly occurring models, model factors. In total there were 16 combinations of the variables which occurred among the 5000 bootstrap samples, however despite this the top tree models occupied 52% of the total occurrence rate, and the top model occurred in 35% of samples. Within these top three models exist combinations of only four variables: crossmatch status, time on dialysis, patient age, and early ABMR - with only early ABMR excluded of these for the top model. By considering the variable inclusion rate crossmatch status, time on dialysis, and patient age are demonstrated as by far the most frequently occurring factors ($\geq 87\%$). A second set of less frequently occurring, but potentially informative, variables includes: early ABMR, previous transplants, and donor type. All remaining variables occupy low to no inclusion rates across any of the considered bootstrap variations. Only crossmatch status occurred in all 100% of models. In line with the model selection criteria previously established, only variables with at least a 50% inclusion rate will be considered going forwards in analysis.

The final results of the multivariate associations can then be formulated based on a Cox proportional hazards model fitted from variables: crossmatch status, time on dialysis, patient age. Model coefficient estimates and confidence intervals are produced based on the bootstrap methodology outlined in the same approach as in the univariate case (section 3.5.2). As compared to the univariate results patient age and crossmatch status demonstrate similar mean HR values - with crossmatch status only slightly deviating from 0.68 to 0.67. Time on dialysis shows the largest deviation from the univariate model with mean HR changing from 0.70 to 0.61.

Variable	Rate (%)	Model factors		
Cross match status (CDC-)	100	✓	✓	✓
Time on dialysis	90	✓	✓	✓
Patient age	87	✓	✓	
Early ABMR (yes)	37		✓	✓
Previous trx (no)	33			
Donor type (living)	25			
Creatine level*	3			
Patient weight	0			
Patient gender (male)	0			
Number of Class I DSA	0			
Number of Class II DSA	0			
Pre-trx DFPP (yes)	0			
Occurrence (%)		35	16	11

Table 3.5: Results of multivariable stability analysis. Associations between factors and graft survival are assessed under multivariable conditions using a Cox proportional hazards model. To eliminate uninformative variables a variable selection approach known as LASSO is applied. This analysis assesses the stability of those variable selections by repeating the analysis under bootstrap conditions ($n = 5000$ simulations). The results of the analysis then shows (A) the frequency in which variables appear (Rate (%)), and (B) the three most commonly occurring models (Model factors). *Day four, post transplant.

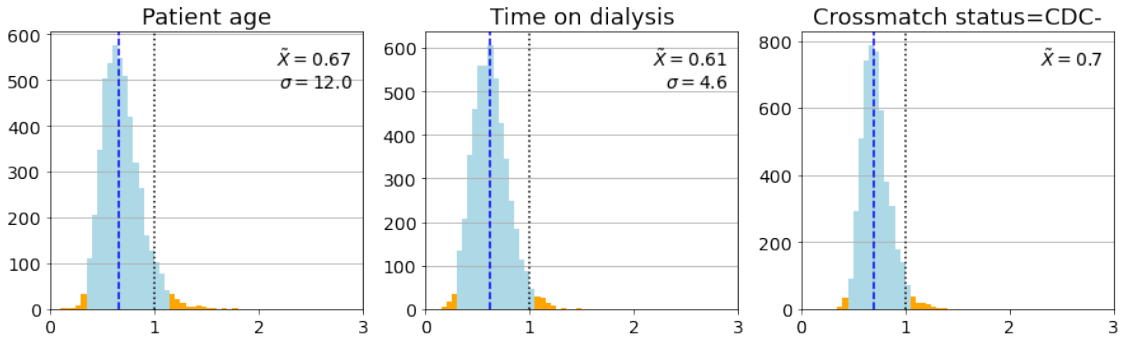


Figure 3.10: Histogram results of multivariate Cox proportional hazards estimates for associations with graft survival. Analysis was conducted via bootstrap resampling which simulated $N = 2500$ data sets for estimation of variable HR. Confidence intervals (95%) are highlighted in blue. Dark blue line corresponds to the mean estimate (indicated in top right corner of each plot). HR=1 indicates no impact on graft survival. σ indicates the amount corresponding to a one unit increase in the associated variable.

3.7 Exploratory cross-sectional analysis

The second aim of this chapter was to identify the most informative single day DSA observation for graft survival. In this section a multivariate analysis is performed independently for each DSA daily sample set surrounding transplantation - referred to collectively as a cross-sectional analysis. For each multivariate model the previously identified informative characteristics of: patient age, time on dialysis, and crossmatch status were introduced into a Cox proportional hazards model alongside the specific DSA sample set for a given day. DSA associations with graft survival were then assessed alongside the model concordance index performance. Prior to the modelling stage DSA responses were first subject to pre-processing to maximise on the available data for analysis.

Pre-processing of DSA for cross-sectional analysis

From Chapter 2 Section 2.4.2, it is shown that DSA monitoring protocols in the UHCW presented a high sampling frequency within the first few months. Despite this the specific monitoring periods varied substantially patient to patient resulting in relatively low sample numbers for each day. In many cases however, data is surrounded closely by samples that could be used to estimate non-measured daily data under the assumption that they closely represent the dynamic behaviour of the DSA response. To synthesise these non-measured DSA samples a linear interpolation procedure was implemented across the cohort - an example of which is shown in Figure 3.11. In each case data was never extrapolated and, due to the differing lengths of each observed response, would result in a data set which gradually decreases in size towards the extremes of the explored time period.

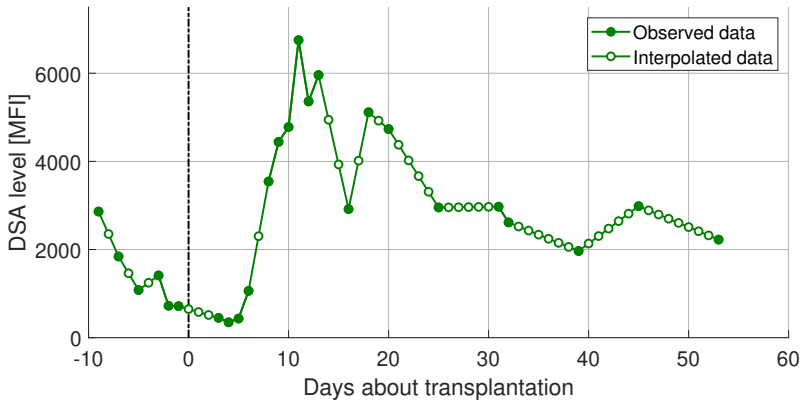


Figure 3.11: Example of DSA response with missing data filled by linear interpolation. Observed data can be seen to be most frequent in the earlier periods post transplantation and less frequent the further from transplantation.

Results of cross-sectional analysis

Each day surrounding transplantation was considered for analysis from ten days prior to transplantation up to 200 days post-transplant. From this data set only the cross-sectional models which presented at least 10 non-censored data would be further analysed - thus resulting in an explored period of 96 days from 7 days prior to transplantation up to 88 days post-transplantation. For each model confidence intervals on desired quantities such as parameter HR and C-index were calculated via bootstrap resampling approach (discussed in section 3.5.2, $n = 2500$). The results are highlighted in Figure 3.12 via three sub-figures: (A) which shows the DSA level (unit 1000 MFI) HR and confidence intervals (95%), (B) the HR significance (i.e., the proportion of bootstrap samples which satisfy HR$>$ 1), and (C) the available number of samples considered for each cross-sectional model in terms of total cases (black markers), censored cases (blue markers), and non-censored cases (red markers).

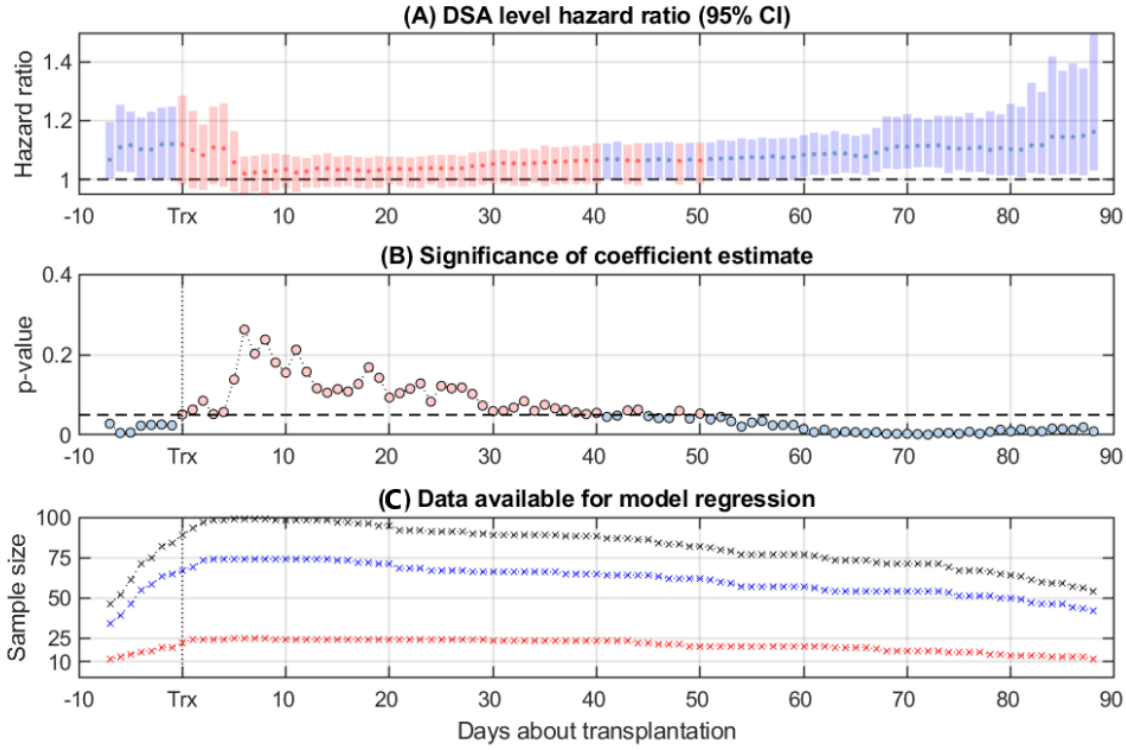


Figure 3.12: Cross sectional analysis of hazard ratios. Results for each given day correspond to an independent multivariate analysis of DSA levels using a Cox-proportional hazards model. (A) Shows the HR median and confidence intervals (95%) corresponding to an increase in DSA by 1000 MFI. (B) Coefficients are significant if null hypothesis is rejected, i.e., HR=1. Significance is deemed at a p-value < 0.05 and is shown in blue markers. Non-significance is shown in red. (C) Shows the number of samples available. In black the total sample size, in blue the censored data, in red the uncensored data. All variables satisfied Schoenfeld individual test.

Each of the sub-figures can be assessed independently:

(A) The median DSA HR (red and blue markers) demonstrates positive values across the explored time period. The highest median HR are found in the periods leading up to transplantation, immediately following, and in the later stages approximately >65 days post transplant. The highest recorded HR median is on day 88 at 1.16. In contrast the period which demonstrates the lowest HR median begins at 5 days post transplant and extends up until day 30 after which median HR begins to slowly rise. The lowest recorded HR median is found on day 6 post-transplant with a value of 1.01. Confidence in the HR estimates, visualised as coloured bands, is also shown to change throughout the analysis period. From the pre-transplant period through to 5 days post-transplant, DSA HR confidence interval size varies from 0.2-0.3 but generally shows an increasing trend up until day 5 post transplant. From day 6 until approximately day 30 confidence intervals are at their smallest around 0.1 before beginning another phase of gradually increasing interval size until the end of the study. Day 88 - the last day of analysis - has the widest confidence interval at 0.38.

(B) The significance of the DSA HR bootstrap confidence intervals (95%), visualised in (A) as bands, is shown relative to the significance threshold p-value of 0.05. A HR is significant if the null hypothesis is rejected, i.e., HR=1. Early DSA HR estimates are significant during the pre-transplant period where lower bound of confidence intervals are demonstrated to exceed HR of 1. Following transplantation significance in these estimates is lost with up to 25% of values shown below 1 on the least significant result - day 5. From day 6 onwards the significance of daily DSA HR estimates increases gradually, becoming significant again for the first instance on day 41 with the last instance of non-significance on day 50. Significance peaks at day 71 with a p-value of <0.001. For clarity, significant p-values (< 0.05) are colour coded blue and non-significant p-values in red (mapped similarly onto (A)).

(C) The sample size available for model regression is dependent on available DSA data for each patient on the day of analysis. In the early pre-transplant period the first observation for a given patient varied and therefore sample size began low at $N = 48$ before increasing and reaching maximum study size of $N = 99$ on day 5. The minimum number of samples reached by the end of the study was $N = 52$ on day 88. A minimum of 10 non-censored samples is present across all models.

3.8 Discussion

HLA-specific DSA monitoring following kidney transplantation can be a useful surrogate marker for immune response and associated outcomes and, as a result, most transplanting centres and national guidelines suggest such monitoring [25]. Specific monitoring protocol however is not defined and varies with uncertainty on the most useful time-points and duration of such monitoring around transplantation. In practice this currently translates to centre specific protocol due to little guidance available on specific and most informative monitoring periods [25]. Obtaining a DSA MFI sample is financially costly and imposes an additional burden on the patient. Therefore, knowledge of most informative monitoring period would be advantageous in clinical practice and allow for targeted application.

In this study, there was access to a high-frequency monitoring of HLA-specific DSA in the early post transplant period allowing for a detailed exploration of the association of daily DSA MFI values with long term graft outcome. This data set ($N = 99$ cases), provided by UHCW (Chapter 2), has presented a unique opportunity to observe these associations and how they may differ on a daily level and for a wide period from 7 days prior to transplantation up to 88 days following. Analysis was conducted in two primary stages. The first, an investigation of associations with long term graft survival, sought to identify relationships between graft survival and variables with historical and clinical significance. Associations between variables and graft survival were established based upon a set of multivariate Cox proportional hazard models which accounted for the influence of multiple confounding factors. A set of 15 variables were considered (Table 3.3) initially under univariate and subsequently multivariate conditions. The second stage of analysis looked to extract only those variables shown to demonstrate significant associations with graft survival and then use them as a basis for developing a set of multivariable models - one for each DSA day analysis.

Of the 15 variables, 6 were considered to demonstrate associations under a univariate analysis: patient age, time on dialysis, crossmatch status, time on ERF, occurrence of previous transplant, and occurrence of early ABMR - however confidence was weak in the latter three. Under multivariate conditions the most informative variables were found through a LASSO variable selection approach. Through this approach an l_1 penalty term on the model coefficients acts as a parameter to tune in order to maximise the cross-validated Harrell's C-index. As the penalty term increases less influential factors are

eliminated from the model. Furthermore modelling results were validated through a bootstrapping approach which sought to investigate the stability of variable selection. Data were resampled $N = 5000$ times and model selection repeated. Of these final models three variables: crossmatch status, time on dialysis, and patient age, were each present in most models (between 87-100%). They were additionally together the components of the most commonly forming model which occurred in 35% of resampled variable selection results. Long term graft survival following HLAi kidney transplantation is a previously explored multivariate problem, and individually each of these variables are known factors. Results in this chapter corroborate findings in literature, such as CDC+ crossmatch status at the time of transplantation [33, 48, 106], greater time on dialysis [107], and younger patient age [15, 38, 108], each being associated with an increased risk of long term graft failure. Certain other variables such as gender [15, 36, 38], DSA class type [26], and number of DSA [11] have demonstrated significance in other studies but not in this cohort. Early ABMR demonstrated some degree of association with long term graft failure in both the univariate and multivariate analysis - however not close to the extent of the top three selected variables. It ultimately only appeared in 37% of models during bootstrap analysis. Some studies have found strong associations with early ABMR, often associating it with recurrent rejection, chronic ABMR and poor graft survival [65, 85]. Such differences in results may be explained due to the smaller sample size of the UHCW data set and the elimination of less influential variables during regularisation. Additionally, the smaller sample size could have resulted in the under-representation of certain classes - such as in gender where male representation is lower at only 39%.

For the second phase of analysis a cross-sectional approach was used to identify associations between each set of daily DSA samples and graft outcome. On each day a separate multivariate Cox proportional hazards model was developed using the previously identified significant variables (crossmatch status, time on dialysis, and age) in combination with the DSA data representing the day in question. From this, four periods were identified with different characteristics: pre-transplant (up to 1 week prior), early post transplant (up to 5 days post), weeks 2 to 6, and weeks 6 to 12 (end of study). The first period, pre-transplant, shows strong association with graft outcome and demonstrates a median HR of 1.1 for every increase in DSA level. This finding is in line with other research which have also demonstrated that higher total DSA pre-transplant was associated with long term outcomes [109–112]. In the early period following transplantation

DSA median HR remain similar to pre-transplant levels, however confidence in these values decreases to the point of non-significance. Low confidence in this early period is an interesting observation. Higgins et al. for example [26], along with others [64, 113], have previously discussed the observation of subdued DSA dynamics in this period before and suggested that it may potentially arise due to adsorption of HLA antibodies onto the kidney allograft. The relative importance of measuring in DSA in this region may be influenced by the inability of the SAB tool to observe absorbed levels of DSA on the kidney itself - resulting in low MFI readings. The period between weeks 2 and 6 offers another striking result where DSA MFI observations here offer the weakest associations with long term graft outcome. Furthermore they offer the tightest HR confidence interval across the investigated period. Weak/ no association was unexpected and occurs in a period of known dynamic activity [26]. This may indicate that these early dynamics are not associated with long term outcome or may be a result of other factors such as dynamics occurring at different rates between patients. The penultimate analysis period at 6-12 weeks showed a trend of gradually increasing median DSA HR - implying that DSA levels were growing more strongly associated with graft outcome the further from time of transplant. Like our study, others reported these same longer term findings [110, 112, 114], however these were defined based on a single time-point at month-3 following transplantation. In all cases median DSA HR were positive, however significance was present only in the pre-transplant period and in weeks 6-12 up until the study period ended.

Throughout this analysis the dynamic complexity of DSA levels following transplantation is somewhat mitigated by assessment of each day in isolation and there is undoubtedly more complex activity not captured in this analysis. Despite this the objective of the chapter was to identify the most informative single DSA measurement for long term graft outcome. Recommendations of this work would place this day on either 2 days prior to transplantation - where it appears critical that DSA values are low for improved chance of long term survival - or, day 70, as a suitable monitoring point for patient follow up. Due to the aforementioned static nature of this analysis, yet stronger recommendations could come as a result of dynamic modelling, which may allow for more in depth understanding of time-dependent features in the DSA response: such as the period for which DSA levels are subdued following transplant, and time or presence of peak DSA. These may better yet equip clinicians in prognosis of graft outcome or aid in the understanding of

underlying physiological processes governing DSA response.

3.9 Conclusions

In this chapter a selection of potentially informative variables relating the UHCW AiT kidney data set ($N = 99$ cases) was formed and subsequently investigated for their relationship to graft survival. Final associations between variables and graft survival were established based upon a set of multivariate Cox proportional hazard models which accounted for the influence of multiple confounding factors. To achieve the primary aim of the chapter which was to identify the most informative DSA measurement period, the following three objectives were fulfilled:

1. **Investigate and prepare a suitable set of independent variables for multivariable analysis.** In multivariable analysis, biomedical data needs to be clinically relevant, without missing data and non-correlating [80]. For associations with long term graft outcome a set of 15 potentially informative variables were identified based on clinical literature. A small amount ($< 1\%$) of missing data were synthesized via k-nearest neighbour single value imputation, and for strongly correlating variables the weakest association with graft outcome were removed.

2. **Investigate multivariable associations with long term graft survival.** In total 12 variables were available for multivariable analysis following data pre-processing. For a data set with a relatively low number of observations, and large number of independent variables, a multivariate regression analysis may lead to a result with low precision. To address this limitation in the UHCW cohort a variable selection approach was used to remove variables of least importance. This was achieved via a bootstrapped LASSO regularisation technique which assessed consistency of the selected variables through a analysis of random perturbations in the data set.

3. **Investigate the association of daily DSA levels with long term graft survival.** Following multivariable analysis 3 variables were found to be significantly associated with long term graft outcome: patient age, crossmatch status, and time on dialysis. For daily DSA associations a new multivariate model was inferred for each day surrounding transplantation from 7 days prior to 88 days post (limited only by data available). Each model included the three significant variables in ad-

dition to the DSA data for a given day.

The results of DSA daily associations analysis indicate four distinct areas in the early period surrounding kidney AiT:

1. **Pre-transplant (>1 week prior).** Significant association with graft outcome. Higher DSA levels indicative of worse graft outcome.
2. **Immediately post transplant (<5 days post).** Not-significant with graft outcome. Low confidence in DSA HR interval.
3. **Early post transplant (weeks 2-6).** Not-significant with graft outcome. High confidence in DSA HR interval.
4. **Medium term (Weeks 6-12).** Significant association with graft outcome. Higher DSA levels indicative of worse graft outcome.

Two periods: pre-transplant, and medium term, demonstrate significant associations with graft outcome - with both also showing that higher DSA levels in their respective periods are indicative of worse outcome. Recommendations of this work would suggest that if few observations are to be made due to financial considerations, a monitoring protocol of two DSA samples on days -2 and 70 are likely to be most informative of long term graft outcome. Patients in this study were transplanted at a single University Hospital from the UK and the results in this study may therefore reflect specific immunosuppressive protocol, management, monitoring protocols, and case selections - potentially limiting generalisability to wider patient groups. For future work results should be corroborated via a multi-centre study, a larger number of cases may also present the opportunity to develop predictive models which could be used to assist in patient specific management.

Chapter 4

Grouping of post transplant DSA dynamic responses

4.1 Introduction

DSA monitoring is already prevalent in clinical practice as a monitoring tool for patient health and has had several studies demonstrate this link for both pre- and post- transplant periods [36–39]. Despite this however, there is not yet a complete understanding of how antibodies contribute to these outcomes. In this Chapter a novel analysis is conducted which utilises the high number of DSA samples available in the HLAi kidney transplant cohort to investigate for the different dynamic response types that have been demonstrated. It is hypothesised that this type of analysis may benefit clinical decision making while also providing further value in guiding future DSA monitoring protocol.

4.2 Aim and objectives

By observing the DSA data set it is apparent that several different response types are present in the cohort - however the number of response types and their distinctive features are not known. In this chapter the primary aim is to identify both the number of DSA response types and their respective features in addition to relating them to long term graft outcomes. The DSA responses in the cohort are complex and numerous providing a substantial challenge for manual grouping efforts. As such, an unsupervised machine learning technique known as clustering is to be utilised, reducing the influence of human bias in analysis and allowing for reproducible results. The aims of the chapter are then

outlined as follows:

Aim 1: To identify, using an unsupervised clustering technique, the different post-transplant DSA response types present in the cohort.

Aim 2: To determine the dynamic response type of each case in the cohort.

Aim 3: To identify the associations between post-transplant DSA responses type and long term graft outcome.

In most time series clustering approaches data must be uniformly distributed and evenly lengthened which presented some challenges for the DSA clustering task. Most notably, a compromise was made between the length of time series and the amount of cases which could be included in analysis. To address this limitation a secondary stage was introduced to label the shorter DSA responses through use of an early time series classification approach. Therefore, to satisfy the above aims the following objectives are established:

Obj. 1: Identify an ideal time series clustering approach for the DSA data set.

Obj. 2: Determine ideal DSA response length for time series cluster analysis.

Obj. 3: Implement clustering, assess the fitness and stability of model solution.

Obj. 4: Implement early time series classification on short time series in the cohort.

Obj. 5: Label DSA response type of each case in the cohort.

Obj. 6: Associate dynamic response type with long term graft outcome.

In Section 4.3 background information on similarity measures, clustering methods and cluster validation are discussed which provide the relevant framework to approach a time series clustering task. In Section 4.4 the clustering methodology for the DSA response is developed in addition to discussing pre-processing of the cohort for analysis. Section 4.5 provides the results of the clustering analysis on the DSA responses data set in addition to validating the solution via fitness and stability assessments. Section 4.6 next classifies the remaining shorter DSA responses into the determined response groups. Penultimately, Section 4.7 relates the now labelled DSA responses with graft survival before results are discussed are further discussed in section 4.8 and then summarised in section 4.9.

4.3 Background

In many fields of research it is often useful to structure unlabelled data. This can be achieved by organising data into homogeneous groups where both within-group similarity and between-group dissimilarity are maximised. The action of doing so, labelled as 'clustering', is necessary regardless of the object data type. Some of earliest instances of clustering research began in the mid 1960's [115] with such works as the monograph 'Principles of numerical taxonomy' by Sokal and Sneath (1963) [116] stimulating a world wide effort to discover superior techniques. The timing of this research is non coincidental as the advent of the first desktop computer in 1964 - the Programma 101 - presented new opportunities to researchers, allowing for automation of large numerical tasks. Its first uses were in anthropological data however it has since branched out to be a truly interdisciplinary technique and is used regularly in many fields from social science to engineering and bio-medicine [117, p. 50]. Due to the vast number of applications that employ clustering analysis there are a considerable number of clustering techniques present within literature [118, 119]. Additionally, there is no established methodology for selecting an ideal technique for a particular clustering task. In this section an overview of the literature and the main concepts of clustering are presented. A framework for selecting a clustering technique used in this chapter is also presented.

4.3.1 Similarity measures

When considering a data set to be clustered, an individual observation may be referred to as an object. Each object can be multidimensional and represented by any combination of features with the most common examples being that of binary, spatial, categorical, or temporal [119]. Calculation of the similarity, or dissimilarity, between two data objects is typically achieved via measurement of the distance between their multi-dimensional feature vectors - whereby the larger the distance between two objects the less similar they are to one another. Quantifying similarity between two data can be a challenging task and many measures may perform better or worse relative to a particular data set. Measuring closeness of the data may be specific to the desired objective of the clustering task. Due to the various requirements of different clustering tasks a large number of similarity measures have been developed across literature. The various distance measure approaches can generally be placed into one of the following three categories [118]:

- **Shape-based** - compares the overall shape of the time series. Time series are typically scaled thus placing emphasis on the relative differences in evolution.
- **Feature-based** - clustering is performed on extracted features from each time series. This is often used to reduce the dimensionality of the task.
- **Model-based** - raw time series are converted into a parametric model and defined by a set of parameters. Parameters are then clustered by conventional techniques.

Shape-based similarity measures are the first to be used historically and are by far the most popular within literature. They can be further subdivided into two categories: lock-step and elastic. Lock-step type measures include methods such as the Minkowski and Pearson correlation distance measures. The category is so-called as these measures require time series of equal length and sample rate. Each data point is compared on a one-to-one basis directly to its counterpart. Elastic measures are more flexible - at the cost of higher computational time. They allow for data points to be compared on a one-to-many or one-to-none basis which is advantageous in the handling of outliers or similar time series that are distorted in the temporal axis. Two of the most popular elastic measures are dynamic time warping (DTW) and longest common sub-sequence (LCSS). Figure 4.1 highlights these key differences by comparing how Euclidean and DTW calculate the distance between two time series.

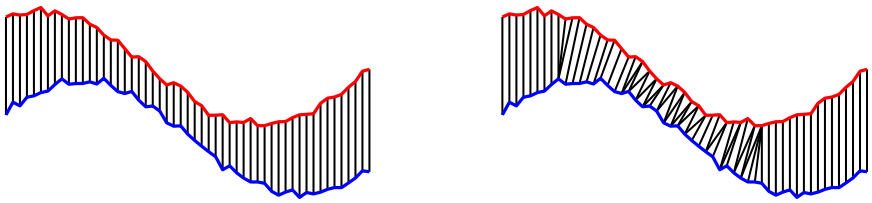


Figure 4.1: Examples of lock-step and elastic shape based distance measures. Left - Lock-step, Euclidean distance measure. Right - Elastic, DTW distance measure. Lock-step distance measures calculate distance on a one-to-one basis. Elastic distance measures can calculate distance using one-to-many - accommodating for distortion in time.

Time series are by their nature high dimensionality objects. In many situations this can present an obstacle to clustering algorithms due to the substantial amount of computational time required to complete the task. Feature-based similarity measures extract key features from the time series allowing for representation in a lower dimensionality form. The disadvantage of these sets of techniques is that some information is lost in this pro-

cess. One of the most common feature extraction techniques is piece-wise aggregate approximation (PAA) which segments the time series along the temporal axis. Each segment is then summarised with the mean of values in each segment before being clustered using conventional techniques. Building upon PAA, symbolic aggregate approximation (SAX) is a technique which first z-score normalises each time series before subsequently determining similarity based on bands which deviate from the time series mean [120]. This technique is powerful due to its ability to determine similarity across vastly different scales of the time series.

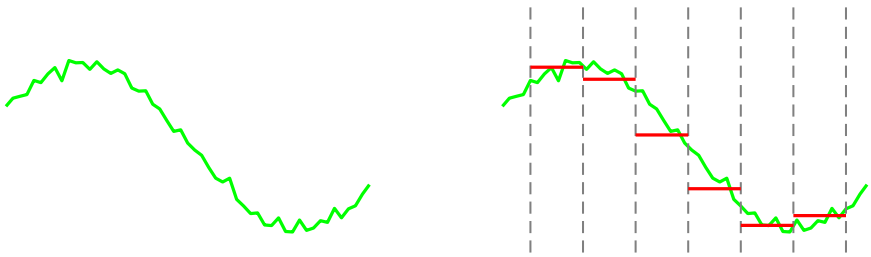


Figure 4.2: Examples of feature extraction used in PAA. Left - a time series in raw data form with high dimensionality. Right - time series features extracted by PAA. Time series is split into segments and each segment represented by its mean value.

Model based clustering is a statistical approach to clustering a data set. The premise of which is that the observed data is assumed to be generated from a finite set of component models [121]. The underlying models must be defined prior to clustering which may be a limitation depending on the task at hand. Model parameters are determined by a probability distribution. In addition to inferring the assignment of observations to a particular cluster most model based algorithms will estimate these parameter values. For example in the case of a parameter defined by Gaussian distribution the estimated values defining a cluster would be the mean and standard deviation. Some of the more popular approaches for time series clustering are auto-regressive moving average (ARMA), Time-Series Bitmaps, and Hidden Markov Models [122].

Many time series methods are proposed in the literature each with a set of advantages and disadvantages. The different trade offs in approaches should be carefully considered for a given task. In Figure 4.3 a flow chart is presented offering a route to identifying which particular set of techniques to explore for a given task. In general, if an underlying model is known, model-based techniques can offer a superior amount of control in a clustering task and should be considered first out of the clustering approaches. In many cases in

time series clustering an underlying model is not known prior to the task and so shape-based and feature-based measures should be considered. For extremely large data sets or for tasks which occur in real time, computational time may be of a concern. In these cases feature based methods may offer faster processing time due to their typical use of dimensionality reduction techniques. Shape-based techniques work with the raw data from observations, so no information is lost as in the case of feature based techniques and so may offer superior accuracy in clustering tasks where critical information is lost.

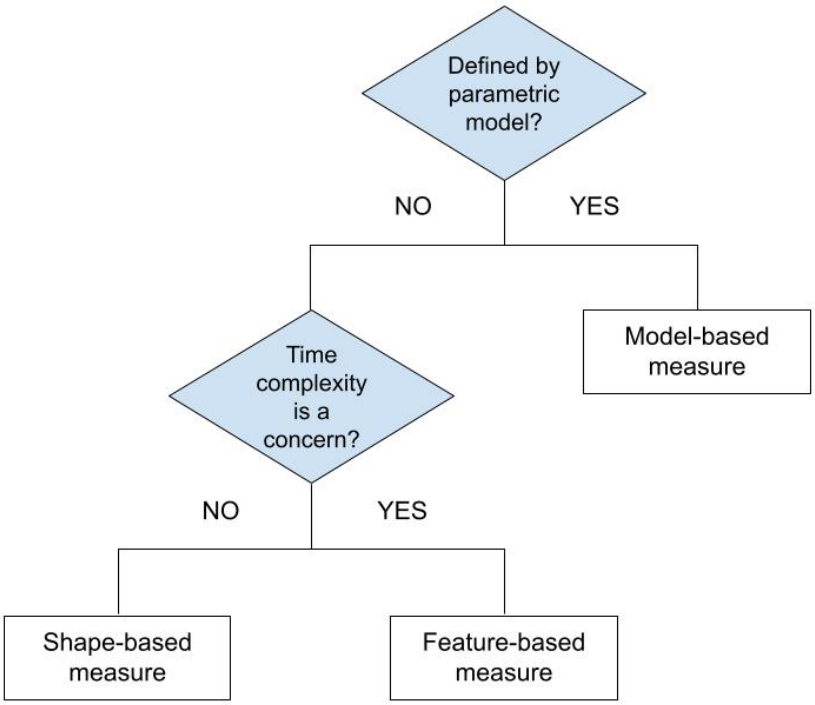


Figure 4.3: Time series distance measure selection flow chart. A flow chart details how one of three main groups of distance measure: shape-based, feature-based, and model-based, may be selected for a particular clustering task.

4.3.2 Clustering methods

Once a similarity measure is chosen a framework for clustering the time series needs to be decided. Many methods for clustering are available within the literature and are frequently classified into the following four categories: partitioning, hierarchical, density based and grid based [118, 122, 123]. Unlike the distance measure the choice of clustering method generally does not influence the solution a great deal [119]. As such only partitioning and hierarchical approaches will be explored here as they are more commonly

explored in the literature.

Partitioning methods are characterised by partitioning data objects into a predefined number of subsets k, where each subset is its own cluster. Objects may belong to only one cluster and clusters may not be empty. As stated by Sarda-Espinosa [124] partitional procedures are regarded as combinatorial optimisation problems. As such to find the global optimum set of clusters would require an exhaustive effort of trialling each combination – a feat not reasonably achievable even for small data sets. In response several heuristics for finding local optima have been developed, the most common of which being k-means [125] and k-medoids [126] which develop clusters around the mean and medoid of the data objects, respectively.

Hierarchical clustering [126] is an approach which establishes a hierarchy of most similar time series. Two main hierarchical approaches exist: agglomerative and divisive [118]. Agglomerative begins with each object assigned to its own cluster before an iterative process begins merging most similar clusters until ultimately only one cluster remains that describes the entire data set. A divisive approach is the complete opposite, in the sense that it begins with one cluster describing the data set before subsequently dividing each stage into two sub-clusters which are most dissimilar. The process is repeated until each object is assigned to its own cluster. Once complete the hierarchy can be represented by a dendrogram providing a clear visual tool to view the similarity structure of the data set. Several approaches exist for formulating cluster centres, or links, at each stage in the hierarchical structure. Some of the most used are single, complete, average linkage and ward's method.

4.3.3 Cluster validation

The objective of clustering is to identify the number of groups of objects within a given data set. Most clustering approaches either require pre-defining the number of groups before analysis, such as in partitioning methods, or extracting groups from a pre-determined structure, as in hierarchical methods. In either case identifying the correct number of groups is not always immediately clear. To identify the correct number of groups a cluster evaluation methodology should be implemented. Evaluation of cluster results in an objective manner is referred to as cluster validation [127], it can generally be broken down into three main sections:

1. Evaluating the fit of the clustered data.
2. Evaluating the stability of the clustered data.
3. Evaluating the performance of the solution relative to external sources of data.

Cluster fitness (1), and cluster stability (2), are both known as internal validation techniques which assess both the quality of the solution and robustness of the solution respectively. Evaluation relative to an external source (3), is known as an external validation technique and requires an additional data set with known solution - as this is not available for DSA response types we do not address external validation any further in this chapter. The following sections give a brief overview of the two internal validation approaches.

Cluster Fitness Validation

Assessing the fitness of a clustering solution is by far the more popular internal validation metric in literature. It assesses the quality of the solution, and is typically achieved via comparison of geometric properties of each identified cluster. The most commonly used geometric properties considered are compactness and separability which reward minimising distance of within-cluster (intra) objects and maximising distance of between-cluster (inter) objects respectively [128]. Figure 4.4 illustrates an example of how each of these geometric properties may be determined for a two dimensional data set in Euclidean space. Cluster distances are often represented by the mean difference between objects (inter) or cluster centres (intra). The figure shows how the correct number of clusters k=3 minimises the within-cluster distance between objects while also maximising the distance between cluster means. For too many clusters, k=4, the reduction in intra cluster distance is not enough to compensate for the much smaller inter cluster distance when compared relative to k=3 clusters. Aside from compactness and separability, other properties are also proposed in the literature and may offer different advantages in their use. One such measure is exclusiveness which can be beneficial in the detection of outlier data [128]. It is based upon the probability density function of the solution and implies that the closer objects tend to be to their cluster mean the more likely that it is a true cluster. Many fitness validation tests exist within the literature, with varying support for each metric [129]. Often it is the case that certain metrics will perform well for a given task as is the case for similarity measures. Some of the more popular measures are the Silhouette index [130], Calinski-Harabasz index [131] and the Gap statistic [132].

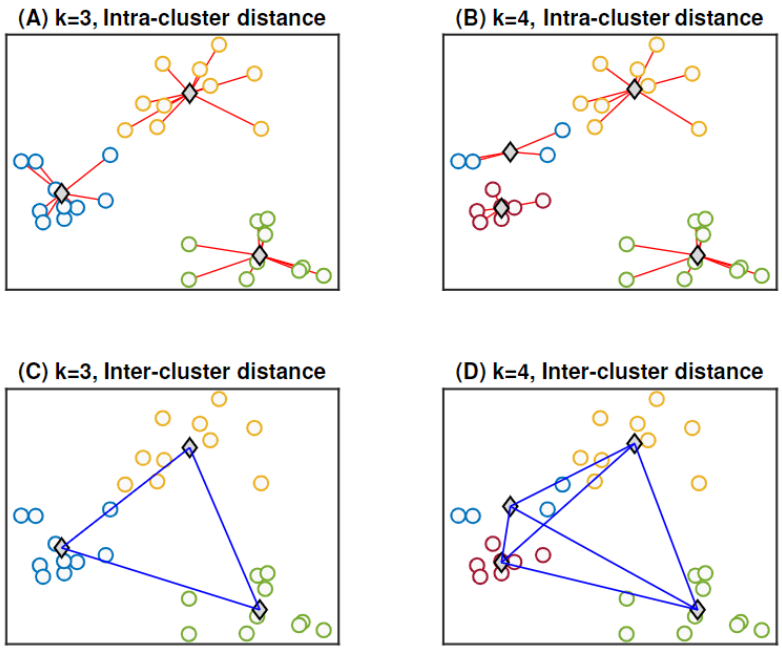


Figure 4.4: Compactness and separability concepts of cluster fitness. (A) and (B) represent compactness or intra-cluster distance for group size k=3 and k=4 respectively. (C) and (D) represent separability or inter-cluster distance for group size k=3 and k=4 respectively.

Cluster Stability Validation

In recent years a new approach to selecting the number of clusters has become increasingly popular based on cluster stability. The premise of which is that the clustering solution should not only be representative of the observed data set, but also the underlying structure from which it is sampled from. The philosophy of a stability assessment differs from that of fitness assessments by considering only if a solution can be constructed in a stable configuration and not through descriptive metrics which require defining what a good cluster is. Assessing stability refers to investigating the robustness of a clustering solution under perturbation or sub sampling of the original data [133, 134]. The concepts of using stability to select the optimal number of clusters is presented in Figure 4.5. Two samples are shown, both drawn from the same underlying distribution. The true solution, k=4, is represented by the black circles encasing individual observations belonging to each group. For each sample clusters are formed for k=2 and k=5 groups highlighting how instability can be formed from both too many and too few groups. In the case of too few groups, k=2, two separate solutions are formed from the differences in observed samples - a horizontal and vertical split - highlighting how small deviations leads to inconsistent grouping. In reality both the horizontal and vertical split are present in the true

solution. For too many groups, k=5, clustering solutions attempt to identify five groups where only four exist - requiring an existing cluster to be split, which, depending on how samples are drawn, could be any of the four.

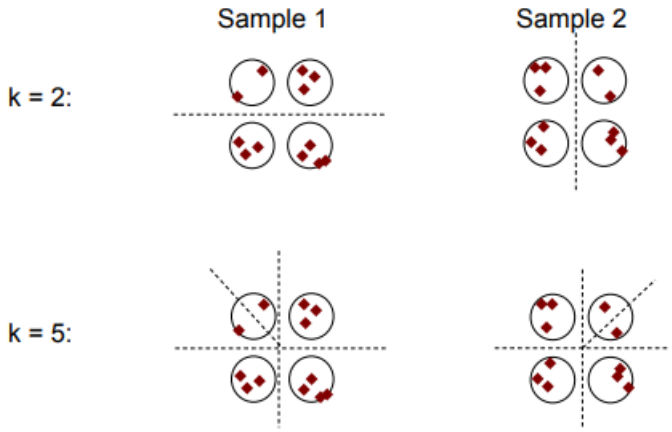


Figure 4.5: Concepts of cluster stability. Cluster solutions are shown to be unstable if number of clusters is either too small (k=2) or too large (k=5). [Image credit to U. Luxburg [135]]

4.4 Methodology

In the background section a wide range of clustering methodology was discussed highlighting how certain approaches may be more suitable for a particular task. In this section each element is revisited and discussed from the point of view of clustering the DSA time series data set. A framework for this clustering task is subsequently established. In order the methodology sections are: pre-processing of DSA data set, determining similarity between DSA responses, establishing a cluster solution, and identifying optimal number of cluster groups.

4.4.1 Pre-processing of DSA data set

Before beginning clustering analysis it is important to consider any pre-processing stages that should be completed. Pre-processing is an important step in machine learning and is required in order to prepare data for model development. In the case of time series clustering two main considerations are made at this stage: time-series dimensionality reduction and time series scaling. With relatively few cases in the cohort, dimensionality reduction techniques are not required to improve computational time and so this element

is not considered in analysis.

Scaling time series data

Depending on the desired goal of the clustering it is often beneficial to scale time series data. The DSA post-transplant responses in the AiT cohort have a wide range of maximum levels ranging from 34 kMFI to less than 1 kMFI. As the objective of the analysis is to identify different types of dynamic responses regardless of maximum levels it is advantageous in this case to perform scaling prior to clustering analysis. Different types of scaling options are available. As the most suitable scaling option is not known a priori the following three commonly implemented techniques are considered:

$$\text{Max scaling} = \frac{X}{X_{max}} \tag{4.4.1.1}$$

$$\text{Min-max scaling} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.4.1.2}$$

$$\text{Z-normalisation} = \frac{X - \mu}{\sigma} \tag{4.4.1.3}$$

where X is a vector representing a DSA time series, X_{max} the maximum DSA level, X_{min} is the minimum DSA level, μ is the mean DSA level and σ is the standard deviation of the DSA time series. The three scaling options in addition to their raw data observations are detailed in Figure 4.6. Four DSA responses are selected to highlight the advantages and disadvantages of each scaling option. The list below highlights some potential limitations identified in each scaling approach:

1. Clustering of similar responses is inhibited due to differences in peak DSA values.
2. Measurement uncertainty is exaggerated, adversely influencing the cluster solution.
3. Relatively little activity is exaggerated, introducing an undesired artifact.

Table 4.1 shows which limitations are present in which scaling option and cites examples of each through linking with Figure 4.6. It can be seen that none of the scaling options are immediately free of potential limitations. No scaling (or raw) has the most severe limitation which impacts almost every response in the DSA data set and so is disregarded as an option immediately. Min-max scaling and Z-normalisation both exaggerate the activity seen in the response for Case-95 which could otherwise be considered as exhibiting

relatively low changes in DSA level. Max scaling appears to offer the best balance of capturing the underlying DSA responses in the cohort whilst minimising the introduction of undesired artifacts. As no solution sufficiently addresses limitation 2, pre-emptive action is taken to isolate those impacted cases from influencing the clustering algorithm. Cases which exhibit a DSA MFI level consistently below 1,500 are excluded from the analysis. The DSA activity in these cases would at any rate have been considered less clinically relevant due to their low MFI level. Classifying DSA responses in this range has been observed before [26], these cases can generally be considered as 'no-response' due to their low activity. Following the removal of 'no-response cases', or otherwise now referred to as Group-0, of which there are N=20 in the cohort, the remaining N=73 cases are then scaled via max-scaling technique.

Scaling option	Limitation 1	Limitation 2	Limitation 3
Raw	C↔D		
Max scaling		E	
Min-max scaling		I	J
Z-score normalisation		M	N

Table 4.1: Comparison of scaling option limitations for DSA responses. Alphabetical code relates to time series responses seen in Figure 4.6.

Influence of DSA sample on cohort

Due to the detailed DSA collection protocols outlined in chapter 2 samples are rarely taken at consistent timing and frequency across any two cases. When considering lock-step clustering algorithms such as Euclidean distance measure for example, analysis would not be possible due to the requirement of one-to-one matching of measurements at each time point. Even the more flexible elastic measures such as DTW which can accommodate one-to-many and one-to-none matching may demonstrate worse performance with non-uniformly distributed samples. For these reasons missing data points in the data set are to be replaced by means of linear interpolation. An example of this process can be seen in Figure 3.11. Missing values are typically few in the first two weeks post-transplantation as DSA collection protocol aims to take measurements most frequently. The further from transplantation the less frequent DSA samples are observed. At the two week mark the median length of interpolated values is 0 days, at 4 weeks this extends to 2 days, at 50 days this is extends to 5 days, and at 100 days the median is considerably larger at 16 days. This same inconsistency in measurement sample rate also presents an

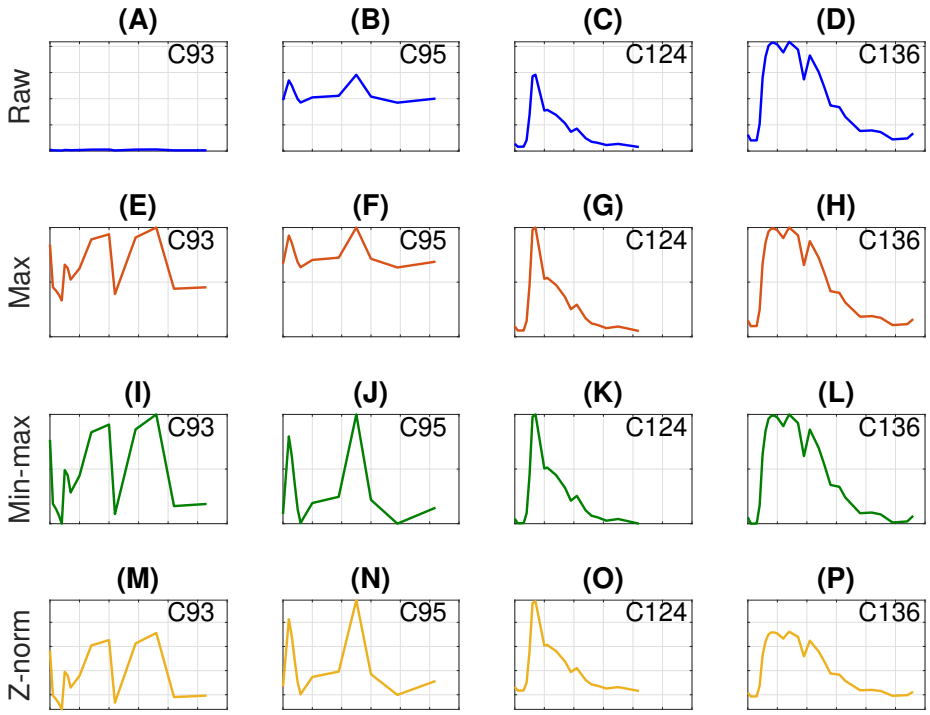


Figure 4.6: Comparison of scaling options for time series clustering. The influence of four different scaling techniques: no scaling (raw), divide through by maximum (max-scaling), min-max normalisation (min-max) and z-score normalisation (z-norm) are presented for four representative DSA time series responses: case-93, case-95, case-124 and case-136. When comparing cases, y-axis is consistent for a given scale.

additional trade-off consideration. For many samples the last available measurement is observed at different times post-transplantation. In Figure 4.7 the number of DSA responses is shown relative to the days from transplantation. This figure represents the number of cases whose DSA observation extend at least beyond a given day. By utilising all cases in the cohort at N=99 less than 10 days of data would be available for the clustering task - which is too small to capture even documented response types such as the modulation [26]. A reasonable balance was established at 50 days post transplant which extends the study period to 4 weeks past typically observed modulation response peaks. In total 17 cases are excluded from clustering analysis at this expense bringing the total to N=82.

4.4.2 Determining similarity between DSA responses

As discussed in the background section determining similarity between different time series responses is achieved through use of a distance measure. Many such distance mea-

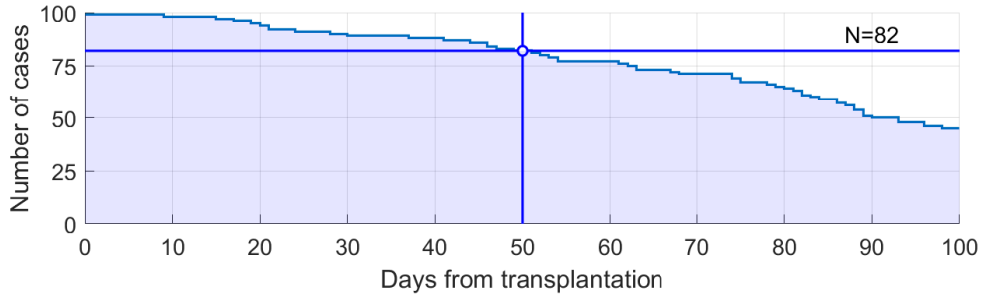


Figure 4.7: Cases vs. DSA response length. This graph assesses the trade-off between number of cases and length of DSA response in clustering analysis. The marked point at 50 days post transplant and N=82 cases indicates the selected trade-off used in subsequent analysis.

sures are available in the literature which often serve to benefit different clustering tasks. In Figure 4.3 a flow chart is presented to broadly narrow down the available set of techniques to either: model-based, feature-based, or shape-based depending on whether the data set can be defined para-metrically and whether time complexity is of great concern in analysis. In the case of the DSA response data set no parametric model is available to define the evolution of DSA data and time complexity is not a concern in analysis due to the relatively small sample size and due to no requirement of real-time output of clustering solution. Given this description of the data set a shape-based similarity measure would likely be the most suitable for the clustering task. Of these an elastic distance measure known as DTW is chosen due to its well documented advantages over lock-step distance measures such as Euclidean, and also other elastic measures such as LCSS [136].

DTW

DTW was developed to address the shift effect which was a major disadvantage of the Euclidean distance measure. It is so called as it effectively warps two sequences, x, and y, of data non-linearly in time to map data more effectively with one another when dealing with different speeds or shifts in the pattern. When calculating DTW between two time series first an $(n \times m)$ local cost matrix (LCM), referred to as d, is produced, whereby element (i, j) refers to the quadratic distance between elements x_i and y_j,

$$d\left(x_i, y_j\right) = \left(x_i - y_j\right)^2. \tag{4.4.2.1}$$

Following construction of the LCM a warping path, $W = w_1, w_2, \ldots, w_K$, is determined by forming a path through the LCM. The max length of the path meets the condition:

$\max(n,m) \leq K \leq m+n-1$, and paths are formed via three constraints:

- **Boundary condition**: A path must start and end at the diagonal corners of the LCM: $w_1 = (1,1)$ and $w_K = (n,m)$.
- **Continuity**: The path may only continue to adjacent elements within the LCM. Given $w_q = (i,j)$ then w_{q+1} is any of elements $(i,j+1),(i+1,j)$ or $(i+1,j+1)$ for $q = 1,\ldots,K-1$ and $i = 1,\ldots,n-1$ and $j = 1,\ldots,m-1$.
- **Monotonicity**: Subsequent steps on the path must be non-decreasing in time. Both indices i and j must either remain the same or be increasing with subsequent steps.

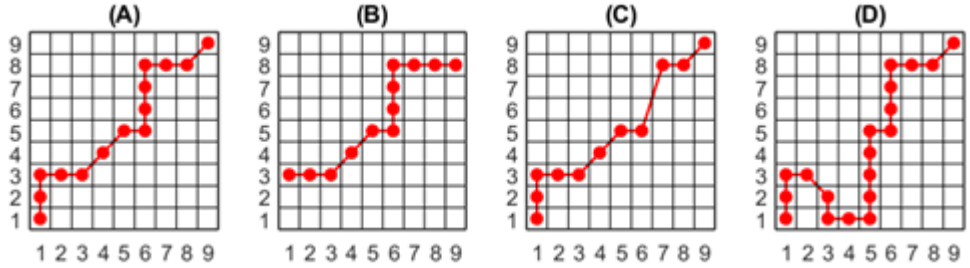


Figure 4.8: DTW constraint diagrams. (A) all constraints are met. (B) boundary condition constraint is broken. (C) Continuity constraint is broken. (D) Monotonicity constraint is broken.

The total distance for a warping path W is the sum of all individual distances of the LCM which the path traverses. To identify the DTW distance for any given pair of time series the minimum warping path is required. This is obtained algorithmically via exploring all possible warping paths in a brute force $O(nm)$ calculation using,

$$D_{i,j} = d\left(x_i, y_j\right) + \min(D_{i-1,j-1}, D_{i,j-1}, D_{i-1,j}). \tag{4.4.2.2}$$

Given the higher computational cost associated with this algorithm compared to many other lockstep or feature based approaches several techniques are developed to address this issue. The task can be reduced substantially by simply limiting the search area for identifying an appropriate warping path. In practice there are reasonable ways of achieving this as it is not likely, for example, that comparing first and last points is desirable. Additionally, of special note is that the lock-step method known as Euclidean distance measure can be considered a unique case of DTW whereby a diagonal warping path is formed from $d(1,1)$ to $d(n,m)$ $(n = m)$, hence satisfying one-to-one matching of all elements.

Global constraints

Imposing a global constraint on the potential warping path for DTW is a common procedure and is explored well in the literature for both classification and clustering tasks [137, 138]. Enacting of a constraint serves not only to reduce processing time for the algorithm – as fewer routes need be explored – but has also been shown to improve overall accuracy of said tasks [139–141]. While reducing the search space to identify a more accurate result may seem counter-intuitive, it acts to limit pathological alignments that may otherwise have been formed. There are two well known global constraints in the area: Sakoe-Chiba band (SKB) and Itakura parallelogram [141] (Figure 4.9). The SKB runs along the leading diagonal and has a fixed width R, where R is a positive integer. The Itakura parallelogram describes the region which constrains the warping path to a series of slopes. For example, with a fixed slope $S > 1$ the Itakura parallelogram consists of the cells that are bound by slopes with gradient S and $1/S$ intersecting at both origin and coordinates (N, N). Several more complex constraints also exist, such as those which generalise the band shape to the data set itself [138, 142], however they typically require large number of time series to establish constraint boundaries and so will not be explored in this work.

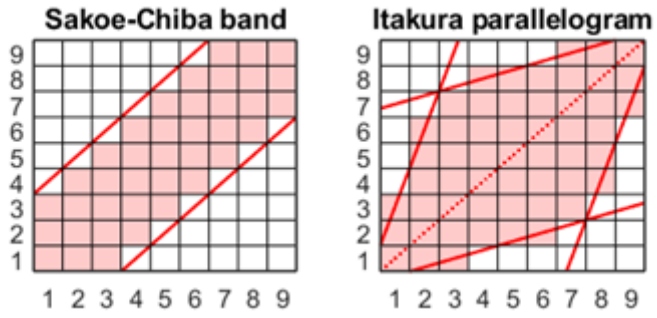


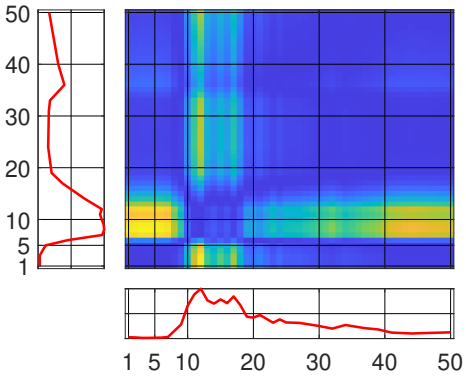
Figure 4.9: Global constraint types. Left: SKB. Right: Itakura parallelogram.

Identifying the optimal width R or slope S is often a task all in upon itself and many approaches to do so exist within the literature [143–145]. Much like the more complex constraint types, identifying optimal band width benefits from a large data set allowing for cross-validation of output. In the absence of large amounts of data this thesis opts to follow guidelines formed about studies in the area. A consensus suggests that an R width of 10% for the Sakoe-Chiba band is an optimal choice for maximising clustering accuracy - beating out the Itakura parallelogram and unconstrained band in most data sets [139, 146]. Use of these parameters has been standard practice for some time and will be

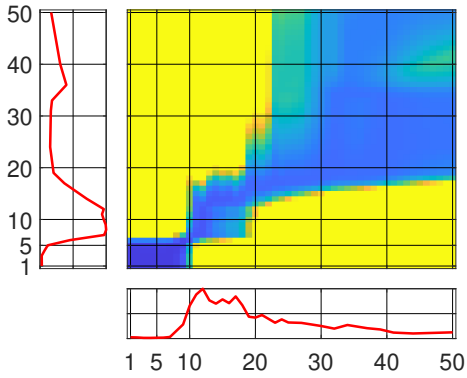
implemented in this thesis [147, 148].

Example of similarity between two antibody responses

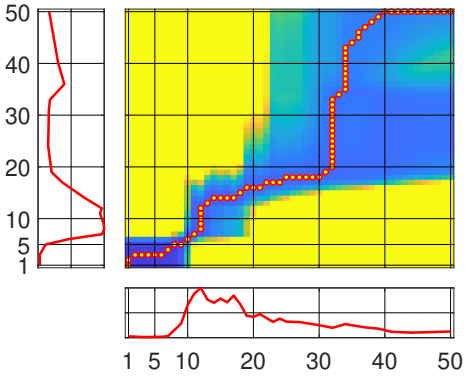
A complete example of the DTW distance measure and SKB constraint in action on two DSA responses from the AiT cohort can be found in Figure 4.10. The figure illustrates the DTW process starting with the distance matrix (DM) in Figure 4.10a and the LCM in Figure 4.10b. An example of the optimal warping path is then shown both without any constraints in Figure 4.10c and with the 10% SKB constraint utilised in this work in Figure 4.10d. Lastly, Figure 4.10e plots both DSA responses next to one another and illustrates how the optimal warping path (10% SKB) is mapped between the two. The aggregate distance of the mapped lines yields the value for DTW similarity between the two responses.



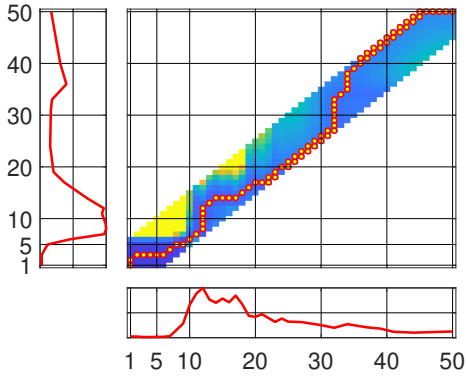
(a) DM



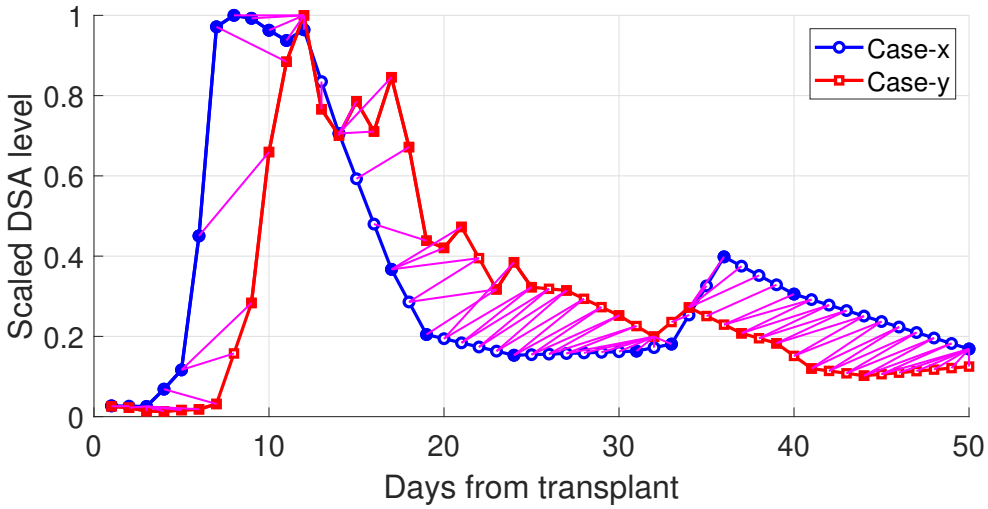
(b) LCM



(c) Optimal warping path



(d) Optimal warping path (SKB=10%)



(e) Optimal warping path (SKB=10%) as illustrated between Case-x and Case-y.

Figure 4.10: Example case demonstrating process of DTW distance measure. In this example, two time series for Case-x and Case-y are compared for similarity through the DTW measure in line with the methodology previously discussed. (a) More yellow regions indicate a larger distance between two time series observations. Distance is maximised when comparing distances further in time and in scaled DSA value (b) More yellow regions indicate a large 'cost' to reach from origin. (c) W is found by identifying the lowest cost path on LCM. (d) Similar to (c) however W is found by identifying the smallest route on the banded LCM. The path cannot exceed the boundaries and so W is constrained. (e) Each node on the warping path corresponds to a magenta line connecting each time series. The aggregate of all line distances forms the distance measure d. (b-d) Resolution of low cost paths are prioritised (hence yellow region saturation).

4.4.3 Establishing a cluster solution

By itself any similarity measure between two time series is fairly meaningless and only gains meaning when compared to additional pairings. In principle this is what a clustering algorithm adds to a similarity measure. Forming a cluster solution through any of the previously discussed clustering approaches (hierarchical, partitioning, etc.) takes the similarity information of all time series combinations within a cohort, and uses that to form a structure of most similar responses. As stated in the Background section there is little that separates clustering approaches in terms of the solution that they produce [119]. As such in this work an agglomerative hierarchical clustering approach will be used due to its clear presentation of cluster solution in structural format. Another advantage relative to other approaches such as partitioning approaches is that it does not require pre-defining of the number of clusters, k, prior to analysis. The membership of objects relative to a desired number of clusters is instead extracted from the hierarchical structure once the solution is complete. The structure being referred to is known as a dendrogram and an example is shown in Figure 4.11. The example shows a cluster solution with 10 objects, each represented at their intersection with the x-axis. The y-axis indicates the similarity between clusters (or objects at their lowest level). A clustering solution is made by establishing a threshold and then forming a cut in the dendrogram at that value. Figure 4.11 highlights how two different threshold values provide the membership solution for $k = 2$ and $k = 3$ clusters.

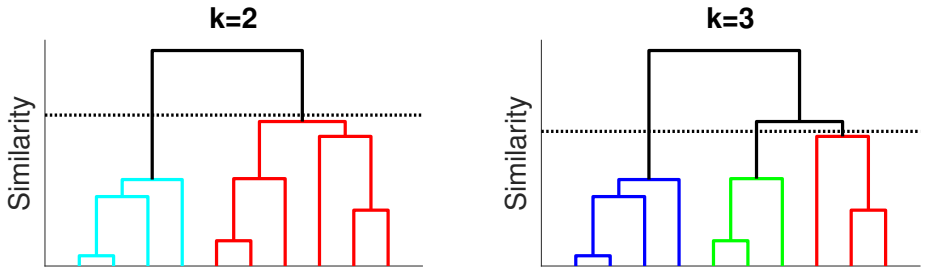


Figure 4.11: Example of dendrogram for hierarchical cluster solution. Both the left and right dendrograms represent the same solution but interpreted at different similarity threshold values. Left - a higher threshold yields a solution with two clusters. Right - a lower threshold yields a solution with three clusters.

The agglomerative hierarchical clustering method is a bottom up approach which starts with all objects in their own cluster before subsequently, at each step, merging the most similar pairs into groups. When complete only a single cluster remains which includes all objects. This process can be summarised by the following pseudo code:

Algorithm 1 Form agglomerative hierarchical cluster solution

1: Compute the distance matrix.
2: Initialise all N observations into their own cluster.
3: **while** there is more than 1 cluster **do**
4: Merge the two closest clusters.
5: Update the distance matrix to reflect previous step.
6: **end while**
7: **return** Set of nested clusters.

In step 1 a DM, D is formed which contains the distances between all pairs of objects in the data set. Step 2 establishes the bottom level of the hierarchical structure by placing each object into its own cluster. Steps 3-5 complete the iterative process of merging the most similar cluster pairs to one another. This is achieved through the merging process in step 4, where two clusters c_i and c_j are merged to form a new cluster $c_i \cup c_j$. Step 5 updates the DM by first removing the rows and columns of c_i and c_j and then subsequently adding an additional row and column for $c_i \cup c_j$ detailing the similarity of the new cluster relative to all other clusters in the DM. The final stage, step 7 returns the set of nested clusters which details the structure of the hierarchical cluster solution.

To calculate the distance measure for the new cluster the Lance-Williams dissimilarity update formula is used [149], where the distance between cluster $c_i \cup c_j$ and c_k is defined as,

$$D(i \cup j, k) = \alpha_i D(i,k) + \alpha_j D(j,k) + \beta D(i,j) + \lambda |D(i,k) - D(j,k)| \qquad (4.4.3.1)$$

where α_i, α_j, β, and λ are parameters used to determine how the distance function performs. Varying these parameters can yield results such as comparing the nearest neighbour between two clusters, known as single linkage ($\alpha_i = \alpha_j = 0.5$, $\beta = 0$, and $\lambda = -0.5$) or furthest neighbour, known as complete linkage ($\alpha_i = \alpha_j = 0.5$, $\beta = 0$, and $\lambda = 0.5$), however in this analysis the Ward's linkage type is used,

$$\alpha_i = \frac{n_i + n_k}{n_{ij} + n_k}, \quad \alpha_j = \frac{n_j + n_k}{n_{ij} + n_k}, \quad \beta = \frac{-n_k}{n_{ij} + n_k}, \quad \text{and} \quad \lambda = 0. \qquad (4.4.3.2)$$

where n_i and n_j represent the number of objects in clusters i and j, $n_{ij} = n_i + n_j$, and n_k represents the number of objects in cluster k. Ward's method is an analysis of variance (ANOVA) based approach that is based upon Ward's criterion [150]. It defines the dis-

tance between two clusters to be the increase in summed square (SS) that results from the merging of two clusters $d(i,j) = \mathrm{SS}_{i\cup j} - (\mathrm{SS}_i + \mathrm{SS}_j)$. This process is advantageous in that it minimises the total within-cluster variance - reducing the chance of outlier groups with few cases. Additionally it does not demonstrate some of the pitfalls of other techniques such as in centroid or median linkage whereby a non-monotonic structure can be formed.

4.4.4 Identifying optimal number of cluster groups

In the background section both fitness and stability internal validation measures were discussed as approaches to identify the optimal number of groups from a clustering method. Fitness measures are used to determine the quality of a solution while stability measures are used to determine how robust the solution is under random perturbations of the data set. Many such fitness and stability validation measures exist in the literature. For fitness validation the Gap statistic is utilised in this work due to its strong performances shown relative to other good measures such as the Calinski and Harabasz and Silhouette index [132]. The Gap statistic is also notably superior to many techniques as it evaluates clusters relative to a reference null distribution - providing the ability to distinguish whether data should be clustered at all. Stability validation refers to investigating the robustness of a clustering solution under perturbation or subsampling of the original data. A stable solution is one that is considered to have captured the underlying structure of the data set; any data set drawn from the same source would produce the same result. The assessment process for stability validation used in this work is based off of the techniques discussed in Luxburg's overview of clustering stability [135].

Fitness validation

The gap statistic [132] is a globally defined (across the entire data set) fitness measure which estimates the correct number of clusters by comparing the within cluster dispersion by that expected under an appropriate null distribution or reference set. For each data set the within cluster dispersion, W_k, is calculated for a given number of clusters, k, using the following equation,

$$W_k = \sum_{l=1}^{k} \frac{1}{2\,|n_l|} \sum_{q,r\in C_l} d\,(q,r)\,. \tag{4.4.4.1}$$

n_l is the number of observations in cluster C_l, and $d(q,r)$ is the distance between observations q and r. To avoid local minima and improve precision, use of multiple reference sets is recommended [151]. In this work 100 reference sets are generated from block bootstrapping the DSA data set. If the within cluster dispersion for a particular reference set is given as W_{kb} and the total number of reference sets is B then the gap statistic is calculated by,

$$Gap_k = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}) - log(W_k). \tag{4.4.4.2}$$

To identify the optimal number of clusters the gap statistic must be assessed across a range of likely group sizes. For N time series the maximum number of clusters is $K = N$ - where each observation has its own cluster - and the minimum number of clusters is $k = 1$ - where all time series belong to the same cluster. With $N = 50$ time series used in the DSA responses in analysis, the range of clusters assessed was from $k = 1$ to $k = 9$ which was satisfactory to find a solution. Once calculated, the gap statistic can be plotted for each number of clusters. For straightforward clustering tasks plotting the result may be sufficient to identify the solution which can be identified from an 'elbow' on the plot. This 'elbow' represents the optimal number of clusters k at which point the Gap statistic is maximised relative to the reference null sets. For many cases in clustering this point may not be clearly identifiable and so an additional rational step is provided to add more certainty to the estimation [132]. Details of the approach are given as follows.

By letting $\bar{l} = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb})$, compute the standard deviation of the B reference set $log(W_k)$,

$$sd_k = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(log(W_{kb}) - \bar{l})^2}. \tag{4.4.4.3}$$

Next, by defining $s_k = sd_k\sqrt{1+\frac{1}{B}}$, select the optimal number of clusters via the first instance meeting the following condition,

$$Gap_k \geq Gap_{k+1} - s_{k+1}. \tag{4.4.4.4}$$

This approach is known as the 'one-standard-error' rule and is used at least as early as

1984 in Breiman's work on classification [152] and has been shown to work well in both simulation studies and in real data examples for clustering tasks [132].

Stability validation

To assess the stability of a solution it is necessary to evaluate the performance of a fixed clustering algorithm upon several slightly varying data sets. There are several ways to approach the task such as: random sub-sampling without replacement [134, 153, 154], adding random noise to original data [155, 156], and dimensionality reduction [157]. Without clear indication of a superior methodology, in this work random sub-sampling without replacement is used to generate perturbed data sets. Additionally, evaluating the performance of the clustering solutions can either be completed relative to the original data set [153], or to compare clusterings of overlapping samples [154]. At this point the amount of sub sampling extracted from the DSA time series data sets, and the comparison approach needs to be carefully considered. By changing the data too much (sub sample is too small) there is a risk of destroying the underlying structure of the samples. By changing the data too little (sub sample is too large) changes in stability may be trivial. An overlapping clusters approach with sub sampling rate of 90% per data set was found to balance well the aforementioned considerations.

For each sub sampled data set clustering is performed using the methodology outlined in this chapter. Once clustering solutions are formulated the two are compared to one another for consistency. Comparison of consistency can be achieved by measuring the distance between clusterings [135]. For overlapping data this is straightforward to compute using a distance score such as the Rand index. The Rand index, R, identifies the agreement between partitions in data by calculating the percentage of object pairs that are similarly grouped in each clustering solution. It can be calculated simply by,

$$R = \frac{a+b}{\binom{N}{2}} \tag{4.4.4.5}$$

where a is the number of times a pair of objects belong to the same cluster across the two solutions, b is the number of times a pair of objects does not belong to the same cluster across the two solutions, and the binomial $\binom{N}{2}$ represents the total number of pairs of objects. The Rand index represents the frequency of occurrence of agreements relative to the total number of pairs. It can take values between 0 and 1, with 0 indicating the two

clustering methods show no agreement on any pair of objects and 1 showing complete agreement on all pairs of objects. In most cases it is more practical to use the adjusted rand index (ARI) which additionally corrects for random chance of agreement between two sets of clusterings. Such a correction is established off of the expected Rand index, $E(R)$, from a baseline random model. The ARI, is given by,

$$\text{ARI} = \frac{R - E(R)}{1 - E(R)}. \tag{4.4.4.6}$$

As a single ARI measurement may not be truly representative of the stability of the clustering solution the approach is completed numerous times on separately generated random sub samples of the cohort - allowing for sufficient statistics to be determined. In total the process is completed 100 times for each cluster size, providing each a distribution of ARI values. The simplest way to determine the optimal numbers of clusters is to calculate the mean or median ARI for each distribution and then identify where it is at its maximum [153].

4.5 Clustering analysis results

The aim of this section is to apply an unsupervised clustering technique to identify the DSA response types present in the AiT cohort thus addressing the primary objective of this chapter. The results in this section follow on from the development of the cluster structure using an agglomerative hierarchical clustering technique (with a DTW distance measure) as detailed in the methodology. The section first completes a cluster validation stage which applies both an assessment of cluster fitness (Gap statistic) and cluster stability (under perturbations of the data) to determine the optimal number of clusters before then subsequently discussing each identified cluster.

4.5.1 Identifying optimal number of clusters in DSA response data set

Fitness validation results

To determine the optimal number of clusters, k, present in the DSA response data set (with n responses) a fitness validation technique known as the Gap statistic is implemented [132] as is discussed in the methodology. The results of which are shown in Figure

4.12. The left figure shows results for the within cluster dispersion, w - a metric used to determine the quality of the cluster solution, where a lower value indicates a better fit. It is not advised, however, to simply select the number of clusters relative to lowest w as w naturally reduces with an increasing number of clusters approaching close to zero at $k = n$. To accommodate for this affect the within cluster dispersion is measured for both the true data set and also for reference data sets (N=100) which acts as a baseline for comparison. Identifying optimal k may now be clear and represented by an 'elbow' on the plot however for robustness a one-standard error test can be completed to provide a definitive result - as shown on the right of Figure 4.12. In this analysis the optimal k is given at the first incidence of a positive one-standard-error value which suggests that the benefit of adding clusters is no better than what is seen in the reference data. Under both the 'elbow' and one-standard-error analysis it can be seen that $k = 4$ clusters is optimal.

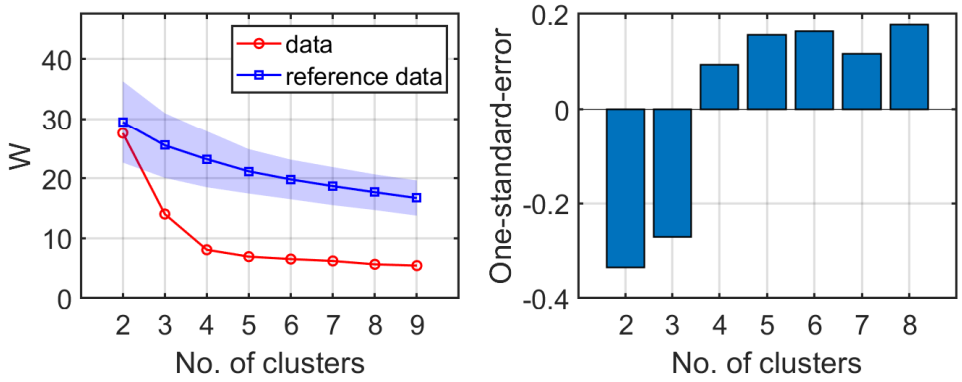


Figure 4.12: Gap statistic analysis to determine optimal cluster size. Results of the gap analysis for DSA time series. Left: within-cluster dispersion, the standard deviation is indicated for reference data via the shaded region. Right: one-standard-error measurements. The optimal number of clusters is indicated by the lowest k on the left in combination with positive one-standard-error measurement on the right; in this case k=4.

Stability validation results

Following fitness validation which assesses the quality of the clustering solution, the stability validation allows for a degree of confidence in arriving at a given solution consistently. For small data tasks as in this case with clustering only 65 responses it is important to assess whether the underlying structure of the data has truly been captured. In this analysis two random subsets of the DSA responses are extracted with each representing 80% of the total cohort (52 responses). Each subset forms a cluster solution before their degree of agreement is determine via the ARI. Figure 4.13 shows the ARI results across $N = 200$ runs of the stability analysis. On top a cumulative density function shows how

each cluster size, k, performs relative to one another and on the bottom the ARI results are displayed in terms of box-plots for easier interpretation of features. It is clear that for $k = 2$ and $k = 3$ a vast number of solutions perform poorly achieving less than 0.5 on the ARI. This result can generally be explained by observing that there are too few clusters to form a consistent solution, therefore different solutions are regularly formed and compared to one another (example shown in Figure 4.5). The most regularly forming solution is found at $k = 4$ which also demonstrates the highest ARI median value with the tightest inter-quantile range, producing a clearly superior result over 200 runs. For $k = 5$ and $k = 6$ the results demonstrate gradually worsening performance relative to $k = 4$.

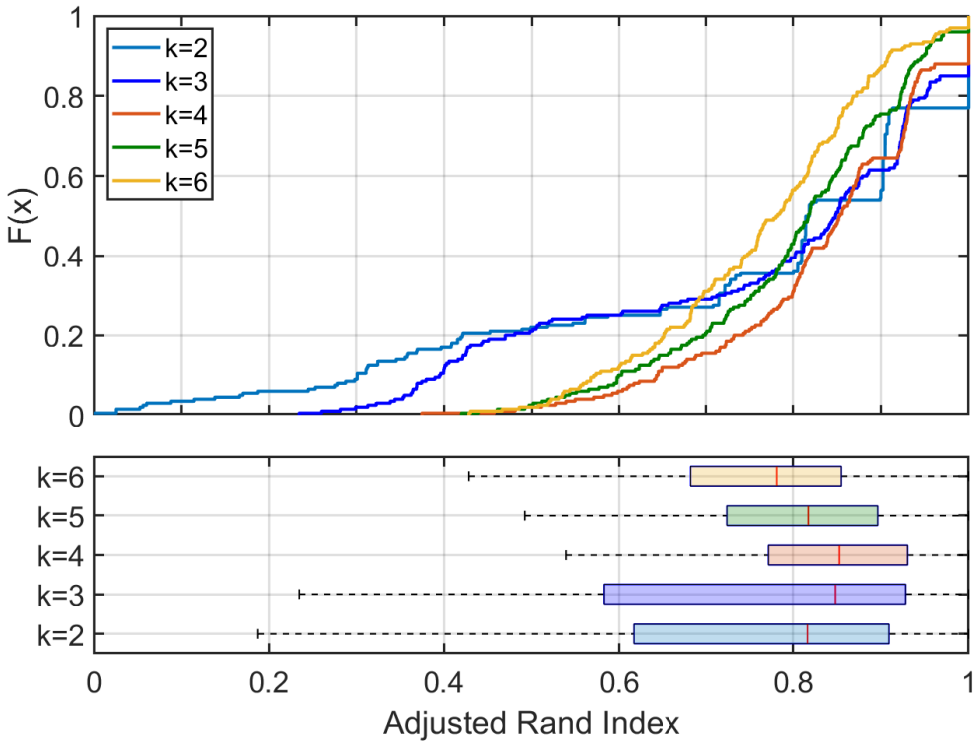


Figure 4.13: Stability analysis of cluster solutions for varying k. Top - accumulative density function of ARI results for each cluster size. ARI is based on comparing the clustering solutions of two 80% sub-samples of the DSA cohort i.e. a random 80% of 65 = 52 responses in each solution. Larger ARI indicates better agreement between two solutions.

4.5.2 Identified DSA response types

Following the results of both fitness and stability analysis on the cluster solution each of the DSA response types are now explored. In Figure 4.14 the cluster hierarchy is shown along with the four identified cluster groups (as colour coded). Each of the cluster

groups show good within cluster similarity relative to their between cluster similarity which is substantially larger - reinforcing previous validation efforts. It can be seen that any subsequent cuts would further split Group 4 while fewer cuts would result in the amalgamation of Groups 3 and 4 and then Groups 1 and 2 subsequently.

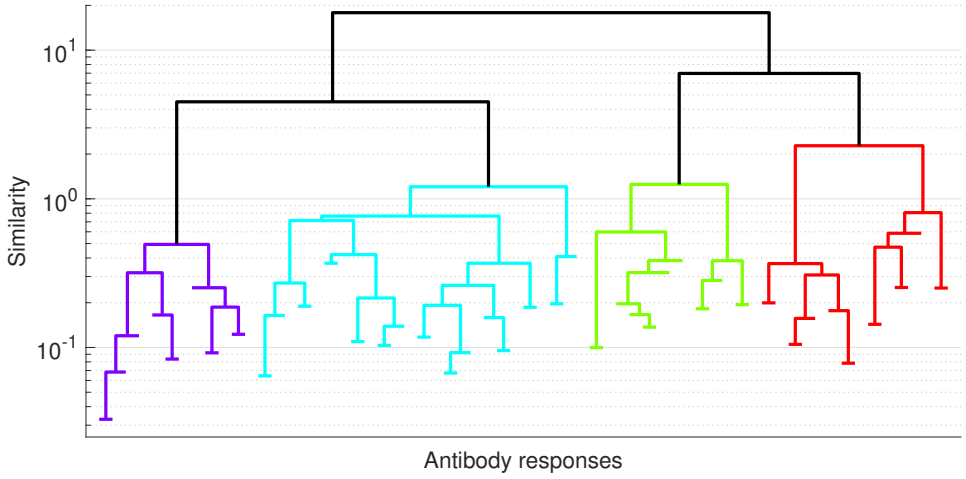


Figure 4.14: Agglomerative hierarchical structure of DSA cohort formed using DTW distance measure. The four identified clusters are highlighted by purple (Group-1), blue (Group-2), green (Group-3) and red (Group-4) colour code. Similarity is represented on a log-scale and the end of each branch designates a DSA response.

In Figure 4.15 each of the four groups are displayed relative to their cluster mean (dark blue line) and cluster standard deviation (blue highlighted band). The most populated cluster is Group-2 with 25 responses and the least populated is Group-1 with 12 responses. By additionally including the no-response group (N=19) which was excluded from analysis in the pre-processing stage there are a total of 5 observed response types in the AiT cohort. Each of the responses are detailed as follows:

No-response group (Group-0) - The no-response group is characterised by low DSA levels <1,500 MFI in post-transplantation follow-up. As the DSA levels are so low they are both considered clinically less relevant while at the same time DSA values are less reliable due to uncertainty in SAB measurement assay.

Fast modulation (Group-1) - Responses in the fast modulation group demonstrate a sharp rise followed by a sharp decrease is tDSA MFI values. The mean peak is day 13 post-transplant across the 12 cases. The dynamic behaviour can be summarised as having short peak duration, typically 3-4 days, before a sharp drop and settling at pre-peak DSA level. DSA are typically inactive for up to 5 days following transplant.

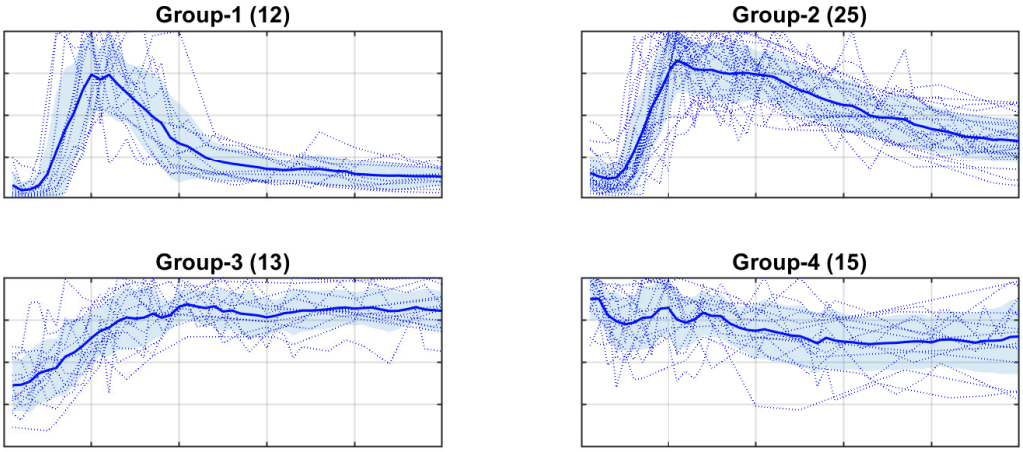


Figure 4.15: Identified DSA response types from cluster analysis. Each of the four identified response clusters are shown (Groups 1-4). Dark blue line represents cluster mean response and shaded region the cluster standard deviation. Dotted lines represent each individual response belonging to a given class.

Slow modulation (Group-2) - The slow modulation group shares similarities with the fast-modulation group, however dynamics occur over a longer period of time. They are characterised by a sharp rise followed by a gradual decline in DSA values. The mean peak day is 13 post-transplant however unlike the fast-modulation group DSA tend to remain at this level for some time before decay. In this group, peak duration is not easily defined and gradually reduces to approximately 30% of peak levels by the 50th day post-transplant. A period of inactivity following transplantation was also seen with DSA typically inactive for up to 5 days following transplant. In some cases large oscillations in DSA levels were present during the decay phase - not seen in group 1.

Rise to sustained (Group-3) - Cases in this group see a relatively lower increase in DSA levels from transplant levels when compared to either of the modulation groups (as indicated by the higher starting values). Additionally, whereas both modulation groups demonstrate a similar time to peak the rise-to-sustained group rise is much more gradual with the average response taking up to three weeks to achieve maximum value and little to no period of inactivity before the rise takes place. During the clustering period of 50 days there is no notable drop in DSA levels once the peak value has been reached indicating a relatively high settling level compared to modulation groups.

Sustained (Group-4) - The sustained group shows the least activity of the four response types (excluding Group-0) and has no substantial rise or fall in DSA from early post-

transplant levels. By observing the mean response a slight decay in DSA levels can be observed throughout the 50 day clustering period however individual responses do not necessarily reflect this. As indicated by the cluster hierarchical structure (Figure 4.14) Group-4 was the next consideration for an additional cut. Under further investigations this cut would produce two additional responses with 8 and 7 cases respectively and represent a sustained group and a steadily decaying group which reaches half its value by the end of the 50 day clustering period (shown in Appendix C.0.2). Although both fitness and stability analysis produced strong results and confidence in $k = 4$ this observation should be investigated further when a larger sample size is available for the clustering.

4.6 Classification of shorter time series

During the pre-processing stage for time series clustering outlined in this chapter it was described that the DTW distance measure would require two time series to be close to, or equal in length to determine similarity. This requirement has presented a challenge for clustering of the post-transplant DSA responses present in the AiT cohort. Monitoring of patient DSA levels on a consistent basis was rarely achieved at uniform sample rates and, more crucially, not always maintained for the same period following transplantation. As such selecting a time period to complete clustering analysis would need to come at a compromise of 'time analysed' to 'cases included'. Increasing the length of time period for analysis resulted in a decrease in the number of responses which could be included in analysis. Each patient record in the AiT cohort is valuable and the more cases which can be included in analysis the more confidence that can be obtained in the demonstrated results. In total there were 15 dynamic response type cases which did not present samples up to 50 days post transplantation and were subsequently excluded from clustering analysis.

Following this result two additional objectives are presented in order to extract the most useful information from the DSA data set as possible:

1. Classify as many of the remaining unclustered DSA responses into the pre-determined clusters as possible.

2. Maintain a high degree of confidence in classification results.

Time series classification is the act of grouping objects into previously defined classes

and is typically performed with a supervised learning technique that has labelled data. Functionally the technique shares many similarities with time series clustering as the objective is to identify similarity between two different temporal objects. In particular time series similarity measures are shared between the two techniques. For the following classification task a DTW distance measure with a 10% Sakoe-Chiba band is used to determine similarity between DSA responses, as was the case for the clustering task. The reader is subsequently referred back to previous sections in this chapter for the relevant methodology. The labelled classes for which DSA responses are to be classified into are the previously identified groups from cluster analysis and, more specifically, the similarity measure is calculated relative to the mean response of each cluster (shown in Figure 4.15).

4.6.1 Early classification of time series

In this early classification of time series (ECTS) analysis an approach is presented to classify the remaining responses into the now pre-defined classes from clustering analysis. The decision to classify each short response is based upon whether their length exceeds the minimum predictive length (MPL) - which is determined based on a predictive confidence threshold [158]. Those that exceed the MPL will then be classified and included in any subsequent analysis on DSA response groups. The concept behind MPL can be illustrated through Figure 4.16. In the figure a classification task is presented to place a time series into a pre-defined cluster. The two examples illustrate how classification can be dependent on the length of a time series. Without establishing some condition prior to classification, a time series which is too short can be easily placed into two or more clusters if their dynamics are similar in early periods. The result then is more down to chance as to which group it will ultimately be allocated - as is the case in Figure 4.16 (left). The MPL is a condition which establishes the minimum amount of data required to make confident predictions on the cluster allocation. Figure 4.16 (right) shows a time series with enough data to reliably distinguish between different clusters - allowing for confident allocation into the correct cluster group. Through use of MPL the left time series would be rejected and considered unsuitable for the classification task, the right time series would be accepted and allocated to the most similar cluster group.

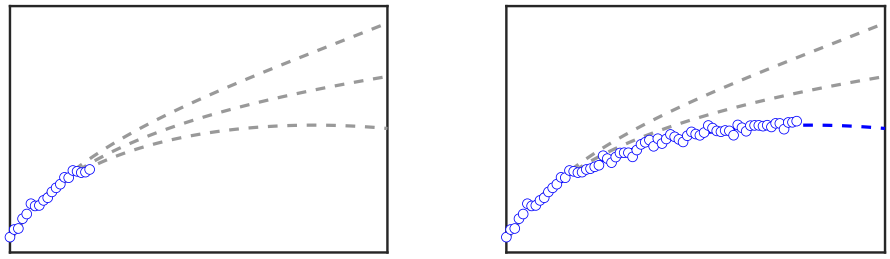


Figure 4.16: Example time series demonstrating the value of requiring a minimum number of samples in classification tasks. Left - too few samples do not provide adequate information to place the time series into the correct cluster. Right - enough samples are provided to correctly distinguish between clusters. Three separate clusters are represented by grey dashed lines. Initially they overlap one another before diverging at a later time point.

4.6.2 Minimum predictive length

Determining the MPL is achieved by systematic analysis of the classification accuracy at each given time series length. An overview of the process can be seen in Algorithm 2.

Algorithm 2 Determining minimum predictive length

Define L as the length of maximum response time.
Define C as number of cluster mean responses.
Define T as number of training set responses.
Define Y as list of true classes for training set responses.
Define A_{min} as the minimum classification accuracy to satisfy MPL.
for $i = 1$ to L **do**
    for $j = 1$ to T **do**
        Define t_T as the first i observations in the j^{th} training set response.
        Scale t_T according to its peak value.
        for $k = 1$ to C **do**
            Define t_C as the first i observations in the k^{th} cluster mean response.
            Determine the DTW distance measure $D(j,k)$ for t_T and t_C.
        end for
        Allocate the class prediction $P(i,j)$, determined from minimum $D(j,:)$.
    end for
    Determine predictive accuracy $A(i)$, from comparing $P(i,:)$ and Y.
    if $A(i) > A_{min}$ **then**
        return MPL i.
    end if
end for

The algorithm begins at a time series length of one and identifies the similarity between each training set response and each cluster mean response. Training set responses are scaled prior to analysis based on their peak value - reflecting the amount of information that is realistically available in practice. Once the similarity measures for a given training

response have been determined the training response is allocated to the most similar cluster. After all training responses are allocated to their respective clusters the classification accuracy can then be calculated by comparing cluster predictions to the true solution. The process is repeated for increasing units of time until the maximum response time is met or the pre-determined minimum classification accuracy is satisfied.

4.6.3 Classification matrix

For classification tasks a confusion matrix (as shown in Figure 4.17) is utilised to determine the overall classification accuracy known as the correct classification rate (C). For balanced (where representatives of each class are equal in number), or close to balanced class problems, C is one of the best measures used to determine the quality of the classification result. To determine C predicted values are compared to a known target class. For the binary class problem shown in Figure 4.17 there are two classes 'Positive' and 'Negative' there can be one of four possible outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) depending on the relationship between predicted response and target class. C is then given as the fraction of the total number of correct predictions TP+TN over the total number of predictions TP+TN+FP+FN. It can often be useful in analysis to observe the specific classification properties of any given class either relative to target or predicted values. For the binary class problem these are given specific names: specificity (Sp), sensitivity (Sn), negative predictive value (NPV), positive predictive value (PPV) - however the same concepts can be easily extended to multi class problems. Formula to calculate each is given in Figure 4.17.

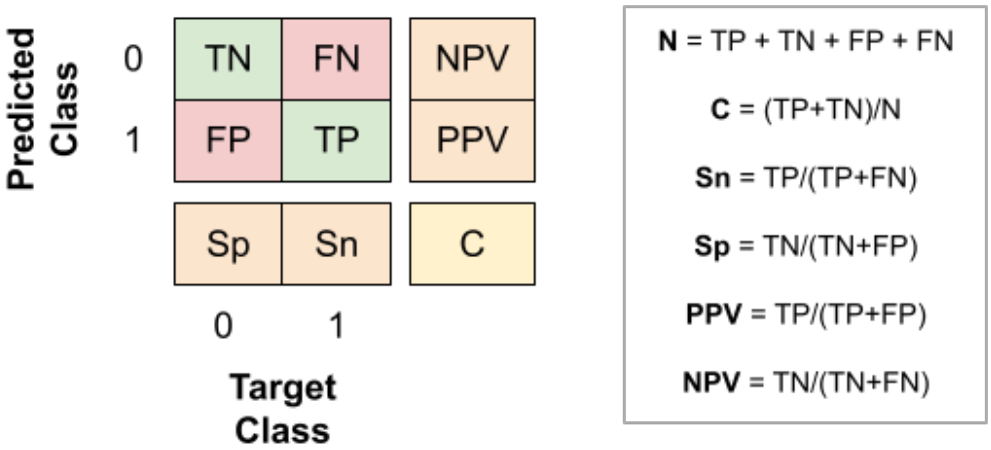


Figure 4.17: Confusion matrix notation and definitions for binary classifier.

4.6.4 Generating training data

Identifying the MPL requires analysing a set of training data which have complete observations up to the full length of 50 days. Due to the small sample size of the DSA responses there were few responses available to isolate for a training data set as they were all utilised in the development of the clustering structure. To address this limitation new data were generated through random perturbations of the original data. There are several approaches to doing so in the literature such as utilising bootstrapping techniques [159], model-based generation (if a model is available) [157], or adding random noise to the original data points [155, 156]. In this analysis 1000 new responses are generated through the latter approach and the methodology is detailed in Algorithm 3.

Algorithm 3 Generating new time series responses.

Define L as the length of maximum response time.
Define G as the number of generated training responses.
Define N as the number of clustered responses.
for i = 1 to G **do**
    Define idx as random integer from 1 to N.
    Set X as cluster response idx.
    for j = 1 to L **do**
        $M(i,j) = X(j) + \mathcal{N}(\mu = 0, \sigma = 0.05)$
    end for
end for
return M

In Figure 4.18 the results of the generation procedure are shown for a randomly selected subset of N=75 responses. The parameters values $\mu = 0, \sigma = 0.05$ were used to provide a representative sample of the DSA data set. Each generated response is shown relative to their allocated nearest neighbour cluster.

4.6.5 Results of MPL analysis

The results of the MPL analysis are shown in Figure 4.19. Classification accuracy, C, is shown relative to the length of data and steadily improves from the shortest measured length (2 days) which has an accuracy of approximately 25% up to the complete length (50 days) which has an accuracy rate of 100%. In general the accuracy trends upwards with an increasing amount of measurement samples available for classification analysis. The MPL is given at a pre-selected accuracy threshold specific to the needs of the classification task. Due to the high requirements of accuracy in this application a threshold

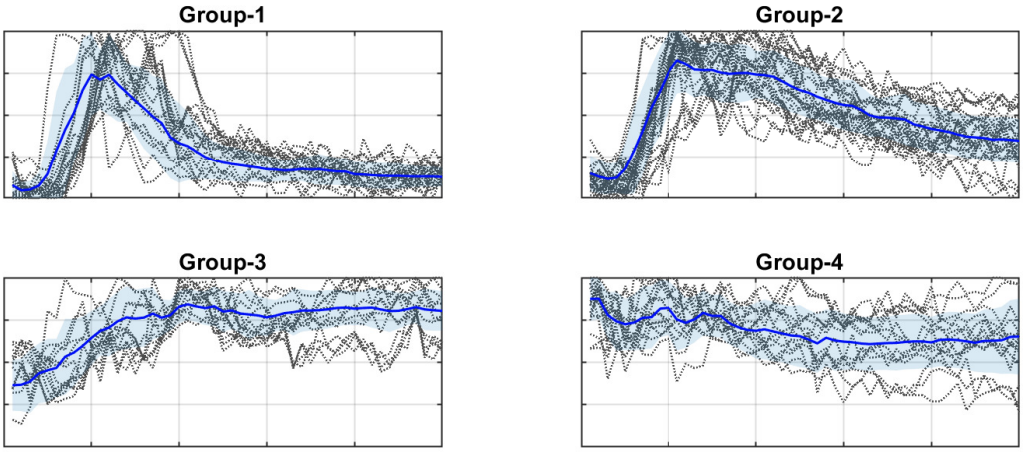


Figure 4.18: Example of generated DSA responses. G=1000 responses are generated using Algorithm 3. Following the generation procedure each response is classified into the pre-defined cluster groups. Here a random subset of N=75 responses are shown relative to the mean and standard deviation of their respectively allocated cluster groups.

of C=90% is used and yields an MPL of 27 days, indicating the minimum length that a DSA response must be observed to in order to be classified in this analysis. It is additionally useful to compare classification accuracy to a reference value. It may seem counter-intuitive for example to see that at 2 days an accuracy of 25% can be achieved when the lowest possible accuracy is a value of 0%. In fact 25% is the lowest accuracy that would be expected in this analysis as it represents that of a perfectly random classifier with 4 classes (as in this case). An accuracy of 0% is extremely unlikely through random allocation and would represent responses being deliberately allocated to incorrect classes. An additional exploratory MPL analysis was also conducted based on the assumption that the maximum post-transplant DSA value was known prior to the analysis and could form a basis for scaling each response (shown in grey, Figure 4.19). Notably a classification accuracy of nearly 75% can be achieved with only 1 to 2 early post transplant DSA measurements and, owing to the high cost of each SAB test, may represent an avenue worth exploring in future research.

By observing the confusion matrix at 27 days we can ascertain further information about the classification properties, as shown in Figure 4.20. In the figure the classification results of the 1000 generated time series can be observed. With 27 days of data it becomes clear that mis-classification occurs only between specific classes and not others. For examples, there are no cases of mis-classification between Group-1 to Group-3 or Group-4 however, it is present between groups 1 and 2, groups 2 and 3, and, to a much lower ex-

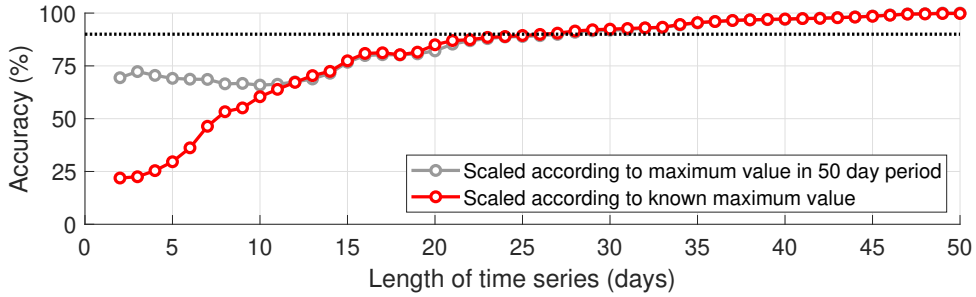


Figure 4.19: Results of MPL analysis applied to generated DSA responses. Y-axis is the correct classification rate, C, and a higher C indicates a better performance. A threshold at C=90% is established and represented by the black dotted line. As the length of time series increases, more data is available allowing for improved classification performance. Two results are shown in red and grey. In red is the results for which all subsequent analysis is to be based upon and follows Algorithm 2. In grey is a study of the classification performance excluding the scaling process (Algorithm 2, line 9) and may represent the potential ability of a classifier with few data points but a known peak value.

tent, between groups 3 and 4. Interestingly group-4 has extremely low mis-classification properties showing that at 27 days post transplant it can be identified far more consistently than other groups. In Xing's work on early classification of time series [158] a methodology for identifying the MPL individually for multi-class situations is also explored, and may potentially offer advantages over a single global MPL by presenting specific entry points for each class type. Although not considered here it does offer potential for certain time series to be classified at earlier points.

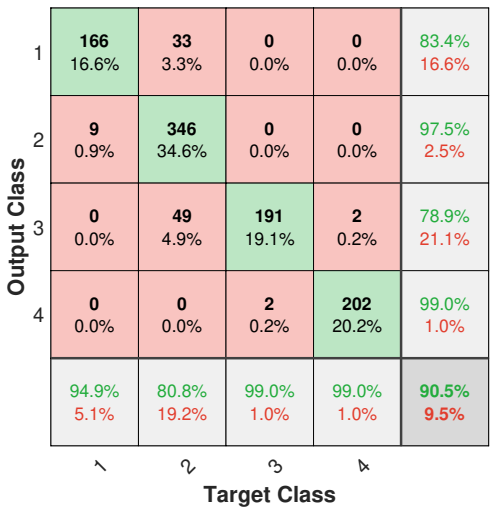


Figure 4.20: Confusion matrix showing the simulated classification results (N=1000) with post-transplant DSA data up to 27 days post transplant.

4.6.6 Classifying short responses

In total 15 responses were considered for early classification. By using the MPL of 28 days seven of these responses have now been classified with an estimated classification accuracy of 90%. The results of both the clustering and classification stages can be summarised as in Table 4.2.

Group	Response Type	Clustering	Classification	Total
0	No-response	19*		
1	Fast modulation	12	2	14
2	Slow modulation	25	2	27
3	Rise-to-sustained	13	2	15
4	Sustained	15	1	16

Table 4.2: Total number of cases grouped through classification and clustering processes. *No-response cases are completed through pre-processing stages prior to clustering analysis. Determined by maximum post-transplant DSA value <1500 MFI, shortest no-response time series length is 46 days. Out of 99 total cases 8 were too short to allocate to a response type.

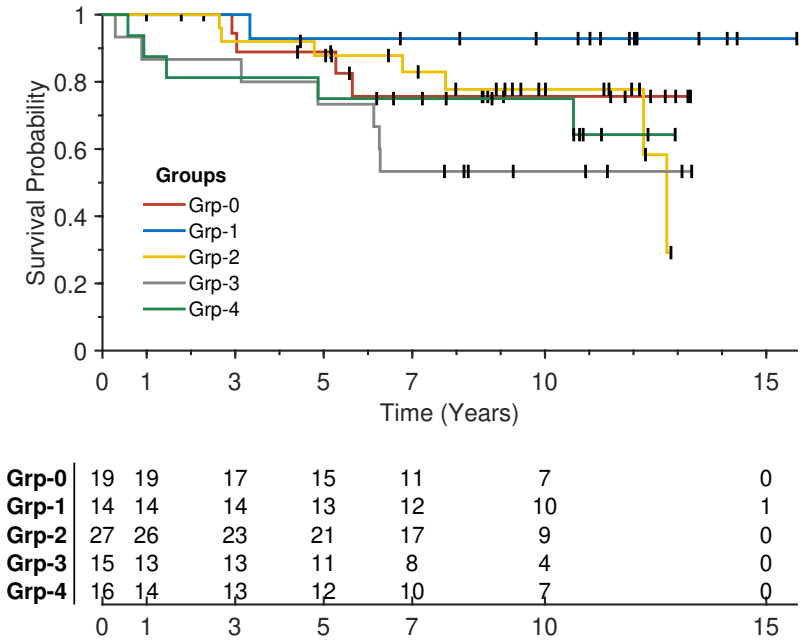
4.7 Analysis of DSA response types

By observing the total number of DSA responses allocated to each group as seen in Table 4.2, each of the identified DSA response types: groups 0 to 4 (no-response, fast-modulation, slow-modulation, rise-to-sustained, and sustained) are investigated for potential associations with graft outcomes.

4.7.1 Survival analysis

A Kaplan-Meier survival analysis is produced to investigate group associations with graft survival (Figure 4.21). Each of the 5 groups are shown in addition to censored data. Group membership can also be seen at each time point following transplantation. Survival rates of the groups differ considerably and do not in all cases function directly proportional to one another - indicating either influence of other confounding variables or presence of nonlinear hazard rates. This difference is most profound with Group 2 which appears to be a relatively good indicator of graft survival (compared to Groups 3 and 4) until 12 years post-transplantation when survival probability drops drastically and is ultimately presented as the worst group overall. Group 1 appears to be a good indicator of graft survival, it has only a single graft failure at 3 years post transplantation yet

demonstrates a median follow-up time of 11.6 years. Groups 3 and 4 share strong similarities with one another with the most striking difference being their poor 2 year survival rates at 87% and 81% respectively which functions in stark contrast to all other groups which have survival rates of 100% in this period. Group 0, interestingly, has worse long term survival than Group 1 despite the lack of DSA in the early post transplant period.



Grp-0	19	19	17	15	11	7	0
Grp-1	14	14	14	13	12	10	1
Grp-2	27	26	23	21	17	9	0
Grp-3	15	13	13	11	8	4	0
Grp-4	16	14	13	12	10	7	0
	0	1	3	5	7	10	15

Figure 4.21: Survival analysis of DSA response groups. Top: Kaplan-Meier survival plots for each group. Bottom: group membership following transplantation.

4.7.2 Early acute ABMR

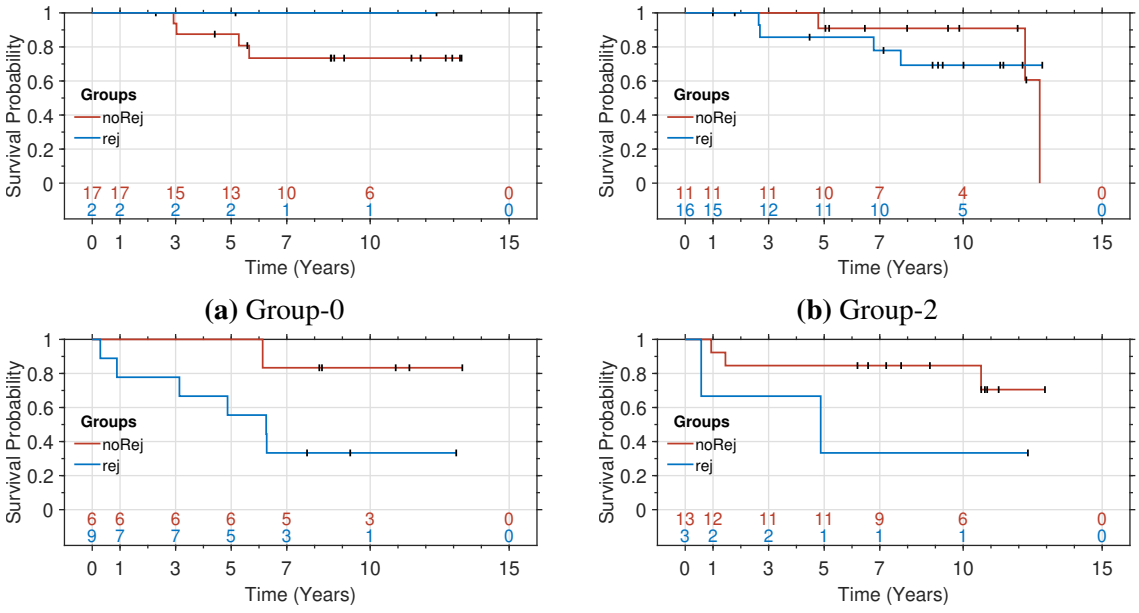
Presence of early acute ABMR is of key interest to clinical teams following AiT and is generally strongly associated with poorer long term graft outcomes. In Table 4.3 groups are shown relative to occurrence of early acute ABMR. Unsurprisingly, the no-response group demonstrated a lowest number of cases where acute ABMR was detected at only 10%, however this does highlight that ABMR can occur even at extremely low DSA levels. The next lowest group was Group 4 which demonstrates sustained or very slow decay of DSA levels over the analysis period. Groups 2 and 3 both share a relatively high ratio of cases which experience early acute ABMR at 60% however the most striking result is for Group-1 in which 100% of cases experienced early acute ABMR in this cohort. This analysis is extended in Figure 4.22 which compares survival rates for each group (except Group-1) when accounting for occurrence of early acute ABMR.

For Group-0 only 2 cases rejected and neither has thus far experienced graft failure, this is

Group	Response Type	Total	Early acute ABMR %
0	No-response	19	10
1	Fast modulation	14	100
2	Slow modulation	27	60
3	Rise-to-sustained	15	60
4	Sustained	16	19

Table 4.3: Analysis of early acute ABMR between identified DSA response groups

in contrast to the 17 cases which did not reject and have a lower overall survival probability after 10 years. This observation is intriguing however would require further samples in the rejection category to see if its superior survival rates are significant. There is little difference observed in Group 2 when assessing for influence of rejection on survival, in both cases the first instances of failures occur around the 3-5 year range. Both late failures occur in the non-rejection group. In Groups 3 and 4 a considerably larger risk is shown for rejection compared to non-rejection groups - suggesting that sustained DSA levels in combination with a rejection episode could be a concerning indicator for graft failure. Particularly for Group 3 where the earliest four failures occur in the rejection group, two of which in the first year following transplantation. For Group 4 both occurrence and non-occurrence of rejection can ultimately lead to graft failure, however it is a greater risk in rejection cases at 2/3 (66%) compared to non-rejection 3/13 (23%).



(a) Group-0

(b) Group-2

(c) Group-3

(d) Group-4

Figure 4.22: Influence of early acute ABMR on graft survival. Group-1 not shown as all cases were early acute ABMR. rej= early acute ABMR episode, nonRej= no early acute ABMR episode.

4.8 Discussion

In this Chapter an unsupervised machine learning technique known as 'clustering' has been used to extract the overall dynamic trends and patterns from the DSA responses data set. Furthermore, each dynamic trend was subsequently assessed for relationships with long term survival outcome. This analysis was possible due to the high-frequency monitoring of HLA-specific DSA made available from the UHCW which provided a uniquely detailed data set. DSA measurements were typically taken daily in the first two weeks, followed by three times a week for the first month before subsequently monitoring based on clinical progress of a given patient. This high rate of monitoring is unusual in practice where most transplant centres and guidelines typically suggest monitoring only a few points post transplant. Developing of the DSA response cluster structure took place using a DTW (10% SKB) distance measure with an agglomerative hierarchical clustering approach (Ward linkage) and the optimal number of clusters were identified both with fitness and stability validation techniques. In total 65 max-scaled DSA responses with length of 50 days were clustered using the algorithm. For both fitness and stability assessments the optimal number of clusters was independently identified at $k = 4$ showing that it had both the best quality and robustness of tested cluster size ($k = 2, \ldots, 9$). By accepting this strong solution, the four clusters would subsequently be referred at as Groups 1 to 4 (fast-modulation, slow-modulation, rise-to-sustained, and sustained, respectively), and, in addition to a previously outlined Group 0, indicate that five response types are found within the cohort. Group 0 is the no-response group ($N = 19$) and was isolated in the pre-processing stage of analysis due to there being consistently low DSA (< 1500 MFI) values which when scaled adversely affected the clustering solution. Additionally, clinically speaking, DSA in this range are also generally considered less relevant with low risk associated with long and short term transplant outcomes. Following clustering analysis 15 cases which had a limited amount of DSA data shorter than 50 days were classified based on an early time series classification approach. This approach used a set of simulated DSA responses (generated via random perturbations of true samples) to determine the MPL which a DSA response could be accurately classified ($> 90\%$). Ultimately 7 out of these 15 short responses displayed a length greater than the MPL (27 days) allowing for classification of the responses. When observing dynamic behaviours a rising trend is often considered more worrying and clinically useful than a steady state or

drop in the MFI values. Additionally it has been observed that dynamic behaviours can precede an episode of ABMR, both histologically and clinically, i.e. drop in urine output and a rise in creatinine. Here it has been shown that an aggressive rise in DSA correlates strongly with early acute ABMR, the fastest rising group is the fast-modulation (Group 1) which regularly experience their peaks within two weeks following transplantation and has a 100% early acute ABMR rate. The slow modulation and rise-to-sustained groups also demonstrate aggressive rises in DSA and have high rates of rejection at 60%. Most interestingly however is that these high rates of rejection do not always associate with a higher rate of graft failure. Group 1, with the highest rate of rejection, has the lowest rate of graft failure with survival rates comparable to those seen in non-AiT. In contrast 6 out of the 7 graft failures in Group 3 experienced an early acute ABMR episode suggesting that ABMR may only be a detrimental factor to long term graft health in certain types of DSA response. Certain features such as subdued DSA levels following transplantation, and sustained total DSA levels over extended periods of time, have been noted in several studies before - however it has not been achieved to this degree of resolution. Dynamic patterns for example were reported in individual donor-specific HLA antibodies using visual description [26]. The modulatory response case was also extracted and analysed under a more complex mathematical modelling of dynamic behaviour [34]. The delayed rise in DSA levels, a feature seen commonly in group-1 and group-2 cases, is thought to be caused due to adsorption of HLA antibodies onto the kidney allograft [26, 64, 113], and, like our study, others reported that sustained total DSA were associated with the worse outcome than resolved DSA [110, 114]. Early ABMR episodes are also typically associated with recurrent rejection, chronic ABMR and poor graft survival but a few studies have reported good graft outcome - highlighting that later episodes are typically more detrimental [65, 85]. One limitation in this work may be considered due to use of the One Lambda SAB assay. SAB are the best available option for determining DSA levels [58] and have substantially improved the ability to identify and manage allosensitised transplant patients [59]. Despite this they are subject to losses in measurement accuracy. They provide a semi-quantitative measurement of DSA in the form of a mean fluorescent intensity MFI reading which does not always have a linear relationship with true DSA levels. At extremely high DSA quantities the SAB assays can be subject to prozone and saturation effects [58, 160] which reduce accuracy in large DSA levels observations. Best efforts have been made to address these limitations during pre-processing stages and

two saturation affected responses were subsequently removed from the study (Chapter 2), however, it is still recognised that MFI levels cannot accurately represent true antibody strength. Despite these limitations, an in-depth and detailed description of DSA dynamic responses has been made and associated them with clinical end-points - satisfying the aim of the chapter. SAB measurements are reasonably attainable in clinical practice and the results may present an avenue for development of superior treatments types/ protocols to replicate responses associated with better outcomes. This approach improves understanding of dynamic response and stimulates future mechanistic studies, such as pre-transplant characterisation of immune status and post-transplantation role of idiotypic antibodies or soluble HLA or different stages of humoral immune maturation. Ideally, a larger multi-centre study is required to confirm the findings. Other limitations include cases discharged back to parent units or management protocol that may have influenced long term outcomes; additionally, protocol biopsies were not employed possibly resulting in missed sub-clinical ABMR. This work may help in future tailoring of treatment, so that lower risk HLAi patients are not subjected to over-immunosuppression even if they have had early acute rejection and that high-risk patients can be looked at more carefully even if they haven't had an early acute rejection. Lastly, despite strong fitness and stability validation results for optimal clustering size, the clustering process may still have missed other patterns due to sample size - including ones which may simply not be present in the cohort - and further studies with a larger sample may untie that.

4.9 Conclusions

In this chapter the primary aim was to identify through an unsupervised machine learning clustering technique the various types of post-transplant DSA responses that were present in the AiT cohort. Through application of a DTW based agglomerative hierarchical clustering algorithm the following result was achieved:

1. **Four active DSA response types have been identified.** These are referred to as: fast-modulation (Group-1), slow-modulation (Group-2), rise-to-sustained (Group-3), and sustained (Group-4) which each demonstrate distinctive characteristics from one another (validated via stability and fitness assessments).

2. **One inactive DSA response type identified.** Referred to as the no-response group (Group-0), it was isolated from the clustering set during data pre-processing stages

and is characterised by low DSA activity (< 1500 MFI).

Various elements of these response types have been identified before within the literature, such as observations of the modulation group [26], however no such analysis has been previously completed to characterise response types to such a high (up to daily) resolution. Building upon these analysis the second aim of the Chapter was to identify associations between response types and graft outcomes. These results can be summarise as:

1. **Group-1 has best long term survival outcomes.** Despite demonstrating comparably high maximum DSA levels to other groups Group-1 demonstrates excellent survival rates - comparable to that seen in not-AiT patients.

2. **Groups 0 and 2 experienced no early (<2 years) graft failure events.** Unlike other groups, Group-2 appeared to clearly demonstrate nonlinear hazard rates with a high drop in survival after 10 years.

3. **Groups 3 and 4 had worse early (<2 years) graft survival performance.** These response types demonstrate a superior risk of poor survival outcome, particularly in the first years following transplantation.

4. **Early acute ABMR rates are highest in Group-1 (100%) despite best graft-survival outcomes.** ABMR is thought to be a poor sign of long term graft outcome however this result serves to contradict such a finding. A survival analysis of each group highlights that almost all of Group-3 graft failures (6/7) are ABMR - suggesting that ABMR may only be detrimental in certain response types.

Chapter 5

Data driven dynamical modelling of post-transplant DSA responses

5.1 Introduction

Extracting useful information from limited clinical data has always been one of the main focuses of research in biomedicine. Data is frequently sparse and noisy, yet still often forms the basis of life changing decisions made by clinicians in practice [161]. The dynamic response of a physiological system following clinical intervention or treatment typically carries critical information about possible clinical outcomes, and is therefore often utilised in the stages of prevention, screening, diagnosis and prediction of the response to treatments or clinical intervention [162, 163]. In various physiological systems mathematical models have been developed to describe, interpret, or predict the dynamic behaviours in clinical data [164–166] which serve to either identify common groups or distinguish features at the individual case level. In the previous chapter four DSA dynamic response types were identified and shown to relate to different types of graft outcome. In this chapter a data-driven mathematical model is developed separately for each of these response types allowing for the representation of individual cases in terms of a parametric based model.

5.2 Aim and objectives

In the previous chapter, five types of post-transplant DSA response were identified: no-response (Group-0), fast modulation (Group-1), slow modulation (Group-2), rise-to- sus-

tained (Group-3), and sustained (Group-4). Of these, analysing the dynamic response groups (i.e, groups 1-4) present the greatest clinical interest as they are thought to be associated with both long and short term clinical outcomes. As it stands there is currently only a limited model available in the literature to analyse the personalised DSA dynamic response pattern which is capable of describing only the decay of certain modulation response types [34]. In the absence of a complete model the work in this chapter is undertaken with the goal of mathematically describing DSA evolution in each of the dynamic response groups. The chapter aims then can be summarised by:

Aim 1: To mathematically describe DSA evolution in each of the previously identified dynamic response groups.

Aim 2: To compare dynamic response characteristics between response groups.

Aim 3: To identify associations between dynamic response characteristics and long term graft outcome.

To describe DSA dynamics a data-driven modelling approach is used based upon differential equations - ultimately allowing for a model to be uniquely defined by a set of parameters. Each of the dynamic response groups is to be assessed independently and undergo a separate model selection process which consider best fit to data, physiological considerations and model stability. To achieve the chapter aims the following objectives are satisfied:

Obj. 1: To produce a suitable set of candidate models for analysis of each of the previously identified DSA dynamic groups.

Obj. 2: To develop and employ a data-driven model selection methodology on the DSA dynamic response data set.

Obj. 3: To extract and assess inferred model parameters for differences between groups and for associations with long term graft outcome.

The chapter is split into six main sections. Section 5.3 introduces the background theory surrounding the model selection and parameter estimation approaches. Section 5.4 introduces the model selection methodology. Section 5.5 is the results section which applies the model selection methodology to each dynamic response group. Section 5.6 extracts and then assesses inferred model parameters between groups and in relation to long term outcome. Lastly, the chapter conclusions are formed in Section 5.7.

5.3 Considerations for model development

5.3.1 Modelling approach

When developing mathematical models for the description of dynamics in physiological based systems there are two main approaches that are commonly used. The first approach, often referred to as physiological or mechanistic modelling, is based on the understanding of the underlying physiology which has generated the measured data. By utilising this approach the developed model can be used to interpret changes at the organ, tissue, cellular, or even molecular scale - all with parameters that have direct physiological interpretations [167]. Practically speaking this modelling process can be time-consuming or even impossible to implement in practice due to the detailed knowledge required about the structure of the system and its parameters. Secondly there is the data-driven approach. This approach differs in that it aims only to describe the data without accounting for the true nature of the underlying physiological system. In this system the internal structure is considered that of a black box with the only available information being presented as the measured inputs and outputs. A data-driven model need only be flexible enough to capture the various demonstrated dynamic characteristics in the data [168]. The ideal model to use depends upon the research aim, knowledge of the system, and availability of data. The mechanistic approach is advantageous in that it can present a deeper insight into the principles of the underlying system, however it is not always feasible in practice. In many situations where the system is less well understood, or access to data is limited, a data-driven modelling approach can still provide valuable information pertinent to the system and may even guide later establishment of mechanistic models [169].

When considering development of physiological systems - which typically demonstrate an extraordinary level of complexity due to their evolution over millions of years - there are two key challenges in establishing a model. (1) Systems are often highly complex and adaptable resulting in properties which are not able to be reduced to sub-units [170]. An example of this may be seen in the case of brain plasticity, where it is known that the nervous system is capable of reshaping itself following severe injury. These systems can present difficulties in model identifiability due to multiple feedback loops which can result in strong couplings between variables [167]. (2) Production of a physiological model, even approximately realistic, will require a large number of parameters [171]

which are not trivial to identify [172]. This is a formidable task in mechanistic modelling and, even in cases where a system can be defined, estimation of the parameters may again not be feasible - even with perfect data - due to limitations in structural identifiability.

For data-driven modelling techniques a complete understanding of the underlying physiological system is not required and therefore many of the difficulties presented for physiological models do not apply. Data-driven models may be considered as either parametric or non-parametric which differ dependent on whether the system is defined based upon a fixed number of characteristics - the parameters - or directly from the measurements themselves. For parametric models the information in the system is captured by a relatively small number of parameters as compared to non-parametric models which make no assumptions about an underlying system. Non-parametric models are therefore typically much more flexible, however this leads to less structural interpretability compared to parametric modelling approaches [168]. Additionally, parametric models typically require less training data due to the inclusion of modelling assumptions making them highly suited to many biomedical applications where data is limited.

Under consideration of the available modelling approaches in accordance with the available DSA data set, this thesis opts to consider only data-driven modelling approaches which are parametric in nature. Physiological models are not considered as the DSA data set is both noisy and sparse, and therefore does not provide sufficient resolution to model a highly complex system. Furthermore, as highlighted in Chapter 2 a DSA measurement can be influenced by several different molecules such as IgG and IgM which would render a true physiologically based system as structurally unidentifiable. A parametric based data-driven modelling approach is selected over a non-parametric approach due to the superior interpretability that it provides and the ability to enforce model constraints in line with physiological limitations. It additionally reduces the likelihood of overfitting to the DSA data given their noisy and sparse characteristics.

5.3.2 Modelling form

In this chapter data-driven models are developed based upon differential equations. Differential equations offer several advantages in the task of model development over alternatives such as difference models. (1) The estimated model is defined by a unique set of parameters, allowing for extrapolation of the model beyond sampled data. (2) Dif-

ferential models can easily accommodate for introduction of prior information through model initial conditions. (3) Irregularity in the sampled data is handled better compared to difference models. (4) Although at a molecular level a system may be discrete in nature, i.e., the creation/ destruction of a single antibody at a time, practically speaking the physiological system behaves as though it constantly evolves with time - resulting in a differential model being the more appropriate choice [34].

There are two main properties of a physiological system which aim to be captured. The first is that natural systems are inherently non-linear. Unlike linear systems - which can only lead to combinations of exponentially decaying/growing or damped periodic oscillating solutions - a nonlinear solution system can introduce irregular behaviour not necessarily attributed to an external input [173]. There is no general approach to developing a nonlinear parametric model found in the literature [168], meaning that a framework typically reduces to a trial-and-error process with a range of nonlinear solutions considered. In many cases, model non-linearity can be closely approximated by a linear model, which can be an attractive idea due to the simplicity and gain in interpretibility. Linear solutions for a non-linear system however may only approximate well for a given range. The second consideration is that of stochasticity. In a perfectly deterministic model the outcomes of a system can be reproduced given the same initial conditions, whereas for a stochastic system the model outcomes can only be estimated probabilistically. Stochasticity in a physiological model may originate from a number of external causes, such as unaccounted factors, environmental influences or even regulatory actions in response to the original disturbance [174] - however it is not practical to attempt to model these high dimensional dynamics. A small amount of stochastic noise can have a dramatic influence on a system response and should therefore not be ignored in the modelling process. By accounting for stochasticity a model can provide a more realistic representation of the underlying physiological system. To mimic the stochastic nature of a system without introducing additional factors (i.e., extra degrees of freedom) a deterministic model can be considered in conjunction with 'noise' which represents the uncertainty of the unconsidered variables. In this chapter a concept known as measurement noise is introduced which corresponds to the uncertainty in the value of an observation - typically manifesting itself as having a blurring effect on the solution.

5.3.3 Parameter inference

Two main schools of thought are often presented for parameter estimation. The first is based on classical (or frequentest) approaches where unknown parameters are considered deterministic and can be represented by a single value. The second are Bayesian based approaches where a probability density function is given to unknown parameters. In classical based approaches the parameters of each model need to be estimated before model selection. Model parameters are typically estimated via minimising of a cost function which represents the goodness of fit (GoF) of a model to the data. The GoF is evaluated by comparing the difference in observed values with expected values under the model in question and the differences between the two values are known as 'residuals'. Perhaps the most widely used method to minimise the cost function is the sum of squared residuals, named least squares (LS) estimation.

Other than constructing a cost function there are other approaches to where the most popular is known as the ML method. The ML method looks for the optimised parameters by maximising the likelihood of obtaining the measurement time series y given the parameters θ of model M. The likelihood is given by $P(y|\theta,M)$. In both cases LS and ML optimise the GoF; however, a model with a better GoF (smaller cost or higher likelihood) does not necessarily produce a better model. With a sufficiently complex model, parameters can be found to fit the observed data with a high precision but may ultimately have been overfit to the data through fitting of noise as well. Overfit models tend to be overly sensitive towards small fluctuations in measurements often leading to incorrect predictions. To compensate a penalty term is usually added to the model (such as Akaike information criterion (AIC) and Bayesian information criterion (BIC)) along with cross validation techniques to assess for model overfitting. Correct choice of the penalisation term is an essential task in model selection as one that is too large can also result in an underfit model. Underfitted models fail to capture important features in the experimental data, which introduce approximation errors into the model known as bias. A model is considered underfit if there are serial correlations between residuals. The model with the appropriate order and structure needs to balance both under and over fitting of data.

Compared with these classical approaches, Bayesian approaches have emerged as a more effective and informative alternative in the tasks of parameter estimation and model selection which can be done simultaneously without the need to select an appropriate penalty

term. Bayesian approaches interpret probability as a quantity that represents a state of knowledge instead of a frequency of an event happening. The states of knowledge are updated when more information is available in a process known as Bayes' rule. Applying Bayes' rule to model selection, a preference over several models $p(M)$ before accounting for any data is defined as the prior. Given the data y the conditional probability $p(M|y)$ of the model M being true is defined as the posterior distribution, and can be described mathematically as,

$$p(M|y) = \frac{P(y|M) \times p(M)}{P(y)} \tag{5.3.3.1}$$

where $p(\cdot)$ represents the probability density function (normalised function that integrates to one), $P(\cdot)$ represents a likelihood function which is not necessarily normalised. $P(y)$ can be referred to as the normalisation factor and represents the likelihood function of the data itself.

The model with the largest $p(M|y)$ is considered among the model candidates to be the most probable model. When there is no prior preference for any model, $p(M|y)$ is determined by the likelihood function $P(y|M)$, defined as the marginal likelihood. To calculate $P(y|M)$ the parameters of the model need to be estimated. There are three main approaches to estimating parameters: ML (classical method), maximum-a-posteriori (MAP) (bridge between classical and full Bayes' method), and full Bayes' method.

Maximum likelihood and Maximum-a-posteriori

As suggested, the ML looks for the most probable parameter values $\hat{\theta}_{ML}$ by maximising the likelihood function $P(y|\theta, M)$, which is the likelihood of obtaining y given the parameter θ and the model M as,

$$\hat{\theta}_{ML} = \arg\max_{\theta} P(y|\theta, M) \tag{5.3.3.2}$$

The ML parameter estimation method is widely used in practice due to its simplicity - however is not without some limitations. In some circumstances for example estimating the most probable parameter based only on data can be misleading. For example, assume model M includes a parameter which represents the probability of a car breaking down with no previous history of any breakdowns. The most probable value of the model would

be zero based on the record, and would underestimate the risk of a breakdown - especially if it was known that the car was old, had high mileage, and had been looked after poorly. Such information - known before taking the data into account - is known as the prior for the parameter. MAP methods incorporate this prior information into the ML model, which can be expressed as:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|M)P(y|\theta,M) \tag{5.3.3.3}$$

where $p(\theta|M)$ is the prior of the model parameters, and $P(y|\theta,M)$ is the likelihood of obtaining y given the parameter θ and the model M. The MAP method just as in the ML method only estimates the mode of the posterior distribution of the parameters. In situations where the confidence of the estimated parameter is of interest, both the ML and MAP methods are not sufficient and the full Bayes' method is required.

Full Bayes' method

The full Bayes' method provides the probability distribution of the parameters rather than the point estimations of parameters. Given the prior distribution of the parameters, $p(\theta|M)$, the posterior distribution of the parameters can be obtained by normalising the right hand side of Equation 5.3.3.3. This yields,

$$p(\theta|y,M) = \frac{p(\theta|M)P(y|\theta,M)}{P(y|M)}, \tag{5.3.3.4}$$

where the denominator is the marginal likelihood $P(y|M)$ and normalises the right hand side of the equation. It can be obtained by integrating over the parameter space

$$P(y|M) = \int P(y,\theta|M)d\theta = \int P(y|\theta,M)p(\theta|M)d\theta = \langle P(y|\theta,M)\rangle_{p(\theta|M)} \tag{5.3.3.5}$$

where $\langle\cdot\rangle_p$ denotes the expectation with respect to the probability density function p in the subscript. As the marginal likelihood is the normalisation constant of the posterior distribution of the parameters, it is obtained as a by-product of the parameter distribution estimation. As previously stated the model with the largest marginal likelihood is chosen to the most probable model; therefore the task of model selection is achieved simultane-

ously with the task of parameter estimation using the full Bayes' method.

5.3.4 Approximation schemes for Bayesian parameter estimation

For most models it is challenging to directly calculate the posterior distributions of unknown parameters and model evidence analytically. This has led to the development of approximation schemes for Bayesian parameter estimation. These can generally be categorised as either: stochastic or deterministic methods. Stochastic techniques, or sampling methods, construct the posterior density through iterative algorithms which are based on drawing random samples of parameters $p(\theta|y,M)$ (from Equation 5.3.3.4). The idea of these iterative sampling methods is to draw a set of samples $\theta^{(i)}$ (where i represents the i^{th} sample of θ, $i = 1,2,...,N$) independently from a sequence of distributions that converge, as iterations continue, to the posterior distribution $p(\theta|y,M)$. This process is known as Monte Carlo (MC) integration. One of the most prominent algorithms for this task is the Markov chain Monte Carlo (MCMC) approach which offers a transition probability for the next sample $\theta^{(i+1)}$ based on the previous sample $\theta^{(i)}$ improving convergence efficiency. Given enough computational power and time these processes will converge to the true posterior distribution. MCMC has been, and still is, one of the most prominent tools for Bayesian inference due to its flexibility in sampling from general distributions, however they are not without their shortcomings which manifest in the forms of: computational expense, uncertainty in convergence, and the inability to directly calculate the marginal likelihood. As such these approaches are typically reserved for parameter estimation tasks only. An alternative to the stochastic methods is the deterministic methods which do not need to take samples from the posterior probability distribution of parameters. They approximate the posterior distribution analytically by assuming it has certain properties such as taking a specific parametric form. In return they are computationally far less expensive and can approximate the model evidence along with the posterior distribution of unknown parameters, however at the cost of never truly representing the exact posterior distribution. Comparing deterministic and stochastic methods may then be viewed as a compromise between accuracy and time complexity with the former having the advantage of being faster and the latter of being non-parametric and asymptotically exact [175]. One particular deterministic sampling method, known as the variational Bayesian (VB) method, forms the structure of this work and a brief overview of the method is provided in the following section.

Variational Bayesian scheme

In the VB scheme the marginal likelihood is a key quantity used to compare different models in the task of model selection and is often referred to as the 'model evidence'. In this section the model evidence approximation is discussed. The goal of the VB algorithm is to approximate the posterior distribution $p(\theta|y,M)$ and to quantify the model evidence $P(y|M)$. For that, an approximation of the true posterior distribution, $q(\theta)$, is introduced and the Kullback-Leibler divergence, D_{KL} between the two distributions is formulated,

$$D_{KL}(q(\theta) \mid\mid p(\theta|y,M)) = \int q(\theta)\log\frac{q(\theta)}{p(\theta|y,M)}d\theta \tag{5.3.4.1}$$

$$= \mathbb{E}\left[\log\frac{q(\theta)}{p(\theta|y,M)}\right]_{q(\theta)} \tag{5.3.4.2}$$

providing a measure for the difference of the distributions. Through replacement of the posterior with the expression in Equation 5.3.3.4 and subsequent simplification procedures, the following expression for free energy $\mathcal{F}$ can be derived [176],

$$\underbrace{\mathbb{E}[\log P(y|\theta,M) + \log p(\theta|M) - \log q(\theta)]_{q(\theta)}}_{\mathcal{F}} =$$

$$\log P(y|M) - D_{KL}(q(\theta) \mid\mid p(\theta|y,M)) \tag{5.3.4.3}$$

D_{KL} is restricted to positive values, meaning that if $D_{KL} = 0$, then $q(\theta)$ and $p(\theta|y,M)$ are the same and the free energy $\mathcal{F}$ is equal to the log of the model evidence, $\log P(y|M)$. Maximising the free energy with respect to $q(\theta)$ is thus equivalent to minimising the D_{KL}, and simultaneously provides an approximation of the true posterior through $q(\theta)$ and of the lower bound of the model evidence through $\mathcal{F}$.

5.4 Model selection methodology

This section provides an overview of the methodology required for model selection as applied in this chapter. There are primarily three stages to the selection process - a dia-

gram of which is illustrated in Figure 5.1. Firstly, a series of candidate models is specified which are to be assessed for fit to the DSA response data. Each of these candidate models are then passed a through modelling criterion phase whereby models or modelling solutions which fail to meet the criterion are discarded. Within this criterion stage are three areas of consideration surrounding structural identifiability, practical identifiability and lastly physiological considerations. All successful models are then considered in the model selection phase whereby the best candidate model is identified based on fitness and stability of modelling solution. This final model is then to be used in subsequent analysis of the DSA dynamic response.

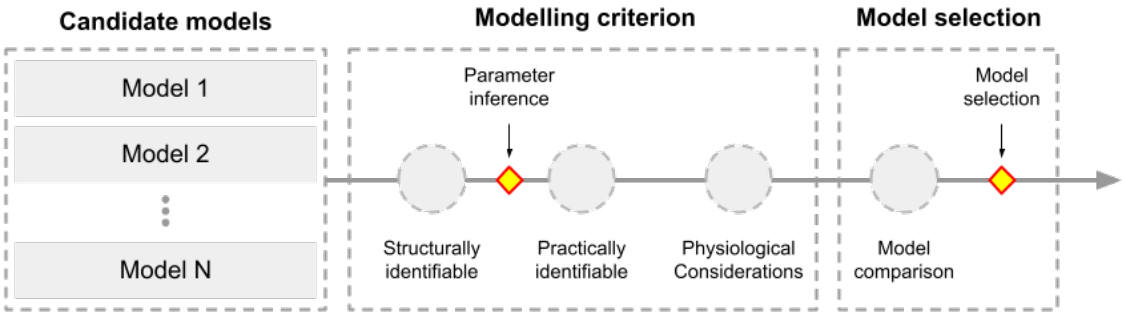


Figure 5.1: Diagram visualising model selection methodology. Initially a set of candidate models are proposed and considered in analysis. Each of these models must pass through a series of modelling criterion. Failure at any stage of the modelling criterion will result in the model not being considered in the modelling selection phase. Each of the successful candidate models which have passed through the modelling criterion are compared to one another for superior model fit and stability before a single model is selected and used in subsequent analysis of the DSA response.

5.4.1 Formulation of candidate models

The mathematical models explored in this chapter are those of ordinary differential equation (ODE) which describe how a system evolves through time in the absence of any external force. Several different system models are considered for the purpose of modeling DSA responses and each take the general form displayed in Equation 5.4.1.1,

$$\frac{d^n}{dt^n}x_t + \sum_{i=0}^{n-1} f_{i+1}\frac{d^i}{dt^i}x_t + f_0 = F(t), \tag{5.4.1.1}$$

where x_t and its derivatives correspond to the states of the system. The functions f_i allows for increased model complexity relative to the higher order ODE terms, and $F(t)$ allows for the introduction of external inputs. Additionally, each observation available in model

formulation is determined as a result of a measurement process subject to a random noise element ε_t. This measurement uncertainty can be mathematically represented by,

$$y_t = x_t + \varepsilon_t, \quad \text{where,} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{5.4.1.2}$$

where y_t is the unobserved true value or DSA level. The measurement noise is assumed to be additive and follows a normal distribution with mean zero and standard deviation, σ. Model formulation is operated systematically and independently for each of the investigated groups, beginning initially with the least complex linear first order model from Equation 5.4.1.1 before progressively increasing model order. Additionally, increased model complexity could also be achieved through the inclusion of nonlinear terms (i.e. polynomial based) or introduced via external inputs to the system. The increased model complexity - though useful in describing a wider range of system behaviours - may present an increased risk overfitting data and so the simplest model which describes the data is explored first in analysis with non-linear models only considered if necessary.

5.4.2 Structural identifiability analysis

Often in systems modelling the structure of the system is unknown and its parameters are determined in relation to the input data. For some systems multiple combinations of parameters may be possible with each yielding the same system output – this is often undesirable where the values of the parameters themselves are of most interest. Analysis of a systems' structural identifiability therefore asks whether a unique recovery of parameter values is possible. Equation 5.4.2.1 highlights an instance of an unidentifiable system, given that q_1 and q_2 are unknown. This is because an infinite number of parameter combinations can yield the same output. Furthermore, if q_1 was known then the system could be described as locally identifiable – as two possible solutions exist, i.e., $-q_2$ and q_2 both yielding the same output.

$$\dot{x} = q_1 q_2^2 x, \;\; x(0) = x_0, \;\; y(t) = x(t). \tag{5.4.2.1}$$

There are several approaches to testing structural identifiability of a system with some examples being: differential algebra methods, the Similarity Transform approach and the Taylor series approach. In this thesis we make use of the Taylor series approach and so

only this is described below. The Taylor series expansion of a model around a known time $t_0 \geq 0$, which is typically performed at t=0, is given by,

$$y(t,\theta) = y^{(0)}(t_0,\theta) + y^{(1)}(t_0,\theta)\frac{t-t_0}{1!} + \ldots + y^{(k)}(t_0,\theta)\frac{(t-t_0)^k}{k!} + \ldots \tag{5.4.2.2}$$

where,

$$y^{(k)}(t_0,\theta) = \frac{d^k}{dt^k}y(t_0,\theta) \tag{5.4.2.3}$$

For Taylor series expansion to term k, k+1 Taylor coefficients are produced alongside subsequent derivations of the model with respect to time. The coefficients of the Taylor series can theoretically be determined by the model output $y(t,\theta)$ and so to demonstrate whether a unique set of parameters θ exist for a given model the following k+1 simultaneous equations need be solved,

$$y^{(0)}(t_0,\theta) = y^{(0)}\left(t_0,\bar{\theta}\right) \tag{5.4.2.4}$$
$$y^{(1)}(t_0,\theta) = y^{(1)}\left(t_0,\bar{\theta}\right) \tag{5.4.2.5}$$
$$\vdots \tag{5.4.2.6}$$
$$y^{(k)}(t_0,\theta) = y^{(k)}\left(t_0,\bar{\theta}\right) \tag{5.4.2.7}$$

For a uniquely or globally identifiable system, parameters will have the solution $\theta = \bar{\theta}$ - whereby all Taylor coefficients have a unique solution. If a set of distinct and countable solutions exist for a particular parameter, then the system is only locally identifiable. Selection of the number of expansion terms, k, is an important factor to consider. Too few and a system may be incorrectly assigned unidentifiable. Work from Margaria et al. [177] has shown that for rational systems with a single model output, $n+m+1$ Taylor coefficients, where n is the number of state variables of the model and m the number of unknown parameters, is sufficient to determine the identifiability. To illustrate this approach an example is considered in Appendix B.1.

5.4.3 Parameter inference

In this chapter both the model and parameter identification were carried out using the VB scheme for the inversion of nonlinear state-space models. The approach itself is used to identify optimal states and parameters of a system given a set prior values and probability distributions for each. Specifically, this approach was conducted within a freely available MATLAB (The MathWorks, Inc., Natick, MA, USA) toolbox, referred to as the VBA toolbox, which implements an algorithm that iteratively optimises said states and parameters to maximise free energy – thus approaching the model log-evidence $\log(p(y_t|M))$. Its mathematical details are shown thoroughly in the literature [178–180], however for completeness, and a brief summary of the underlying theory is provided in Appendix B. For the differential models used in this work there are three types of unknown quantities, model parameters $\mathbf{p}$, model initial conditions $\mathbf{x_0}$, and the measurement uncertainty ε. Each of these unknown quantities - as previously mentioned - are characterised by probability density function (PDF) under the Bayesian approach. Within the VBA toolbox the quantities $\mathbf{p}$ and $\mathbf{x_0}$ are handled simultaneously and can therefore be described as a single variable $\vartheta = \{\mathbf{p}, x_0\}$. For ϑ, prior distributions in the toolbox must be modelled with a normal distribution, and are therefore specified with sufficient statistics μ_ϑ^0 and Σ_ϑ^0,

$$\vartheta \sim \mathcal{N}(\mu_\vartheta^0, \Sigma_\vartheta^0). \tag{5.4.3.1}$$

The result of this calculation is a Gaussian PDF over the unknown parameters that is specified by a mean and covariance matrix. For measurement uncertainty, ε, the prior distribution is limited to only positive values in the toolbox by defining the precision, $\kappa = \sigma^{-2}$, using a Gamma distribution,

$$\kappa_\varepsilon \sim \mathcal{G}a(a_\varepsilon^0, b_\varepsilon^0) \tag{5.4.3.2}$$

In order to maximise the free energy of the model across its multiple parameters, $\theta = \{\vartheta, \kappa\}$, it is required to assume that the approximate combined distribution $q(\theta)$ can be factorised into the separate densities of the two individual and unknown parameters -

known as the mean-field approximation. This can be summarised by,

$$q(\theta) = q_\vartheta(\vartheta) q_\kappa(\kappa). \tag{5.4.3.3}$$

With this assumption it is possible to maximise the free energy, $\mathcal{F}$, by optimising each parameter sequentially while keeping the other parameter fixed. The optimisation rules for this function are based off of variational calculus approaches determined by the Equation 5.3.4.3, hence the name of the approach. The optimisation process is iterative and is repeated until convergence, a visualisation of the approach is demonstrated in Figure 5.2.

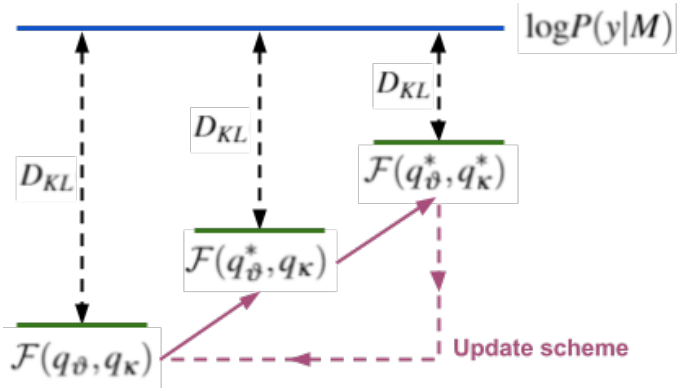


Figure 5.2: Schematic of the free energy maximisation used for VB parameter estimation. This approach seeks to minimise D_{KL} by maximising the free energy, $\mathcal{F}$, during the parameter optimisation scheme. Updated quantities are represented by *.

Input and output of the toolbox

In summary, to complete model inversion of a nonlinear state-space model within the VB toolbox it is crucial to specify both the model parameter and initial condition prior Gaussian distributions. This is introduced to the toolbox via two vectors containing their respective mean values, in addition to two covariance matrices. All priors in this chapter are loosely defined and with wide variances to allow for a large search space during model convergence. Non-diagonal elements of the covariance matrices are set to zero, indicating no prior covariance. Additionally, the measurement precision uncertainty element similarly needs defining prior to model inversion. In this case a Gamma distribution defined by shape and rate quantities is required. As an output the VB toolbox gives posterior distributions of all model parameters, initial conditions, and measurement uncertainty in the same forms as discussed for their respective prior inputs. These values allow for an approximation of the model sensitivity, discussed next in an assessment of practical iden-

tifiability. Lastly, the VB toolbox provides a single value calculation of the free energy $\mathcal{F}$ which is used in subsequent comparisons between model candidates of the same order.

5.4.4 Practical identifiability analysis

Throughout the structural identifiability phase a model is assessed for its ability to uniquely recover its parameters considering a continuous, error free data-set. In practice however model inference takes place using data which will likely not uphold these characteristics. There are certain scenarios then in which a model proven structurally identifiable does not necessarily equate to such under realistic modelling conditions - as in the case where parameter precision estimates are limited due to a large measurement error. Parameters of this nature are considered practically unidentifiable, and practical identifiability analysis is therefore performed following model inference to evaluate the precision of obtained parameters [181]. In the context of Bayesian parameter estimation as is the case in this chapter, unknown parameters are handled as random variables specified by a PDF - meaning that quantifying their uncertainty is already incorporated into the Bayesian methodology. Posterior estimates of parameter uncertainties therefore provide a direct method for assessing practical identifiability [182]. There are two ways considered for analysing the practical identifiability, otherwise referred to as the system sensitivity, within a given system through the VB toolbox. The first is a deterministic approach for which the model states at each time step are calculated directly from the covariance model parameters, the covariance of the initial conditions, and the model sensitivity trajectories (discussed in Appendix B) [183]. Though relatively easy to implement it is required to make an assumption that the model output is normally distributed - which may not necessarily be the case. An alternative stochastic approach instead makes use of Monte-Carlo sampling of individual parameter point estimates from their respective distributions. The approach then records many such samples which are used to plot a one-standard deviation range at each time point - thus representing the sensitivity of the system without the assumption of a distribution type. Though this approach comes at the expense of a larger simulation time it was deemed necessary due to the potential influences of system noise.

5.4.5 Physiological considerations analysis

In modelling of the DSA responses it must be recognised that the system represents the physical presence of antibodies and with it comes a set of physiological considerations

or constraints. Breaking of these constraints would be a strong indicator that the model is not representative of the underlying DSA dynamics.

The first constraint is that the modelled response is always positive as DSA measurements can only exist in the positive domain. Negative DSA estimates represent a clear indication of an incorrectly fitted model. In some cases this may occur due to few samples and correcting for this constraint was achieved through more strict prior distributions on parameters. 'Positivity' was assessed via observing if the stochastic sensitivity analysis demonstrated a 1-σ range exceeding the boundary of zero at any point in the response.

Lastly, each of the DSA responses have demonstrated a tendency to approach a settling value: groups 1 and 2 levels initially rise rapidly before subsequently falling and converging on a settling level, group 3 approaches one gradually following an initial rise, and group 4 cases are close to their settling levels immediately following transplantation. In each case an argument is presented that fitted solutions must be stable and converge upon a settling level. To validate system stability the Routh-Hurwitz stability criterion is used which requires assessment of both the necessary and the sufficient conditions. The necessary condition is that the coefficients of a models' characteristic polynomial should all be positive and nonzero - implying that the roots of the characteristic equation should have negative real parts. A solution is stable if all roots have negative real parts, however the inverse is not necessarily true. Satisfying the necessary condition is required for system stability however does not ensure it which is why a secondary sufficient condition is also used in analysis. The sufficient condition is satisfied when no sign changes are present in the first column of the Routh table. To illustrate these conditions the following general case is presented for linear and homogeneous differential equations,

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0, \tag{5.4.5.1}$$

where a_i is the i^{th} coefficient for term $y^{(i)}$. By assuming the solutions to this differential equation will be in the form $y(t) = e^{rt}$ Equation 5.4.5.1 is simplified to

$$e^{rt}(a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0) = 0. \tag{5.4.5.2}$$

By next simplifying with the removal of the exponential term e^{rt} Equation 5.4.5.2 is reduced to the characteristic equation,

$$a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0 = 0. \tag{5.4.5.3}$$

At this point the roots, $r_1, r_2, \cdots, r_n$, of the characteristic equation can be calculated to validate the necessary criterion. The general form of the Routh table, shown in Table B.3(a), is constructed first of the coefficients of the characteristic equation before subsequent rows are calculated sequentially via the determinant of the entries from the previous two rows (first column and next column following coefficient) divided by the first entry in the previous row. For example b_k could be calculated via,

$$b_k = -\frac{\det \begin{vmatrix} a_0 & a_{2k} \\ a_1 & a_{2k-1} \end{vmatrix}}{a_1}. \tag{5.4.5.4}$$

By observing that the first row of the Routh table cannot change signs (number of changed signs indicates number of negative real parts) else face instability a set of parameter constraints can be determined to ensure system stability. Table 5.1(b) shows an example of the solved Routh table for a third order linear ODE. With slight simplification it is possible to extract the following constraints for stability of the model: $A > 0$, $B > A/C$, and $C > 0$.

s^n	a_0	a_2	a_4	a_8	$\cdots$
s^{n-1}	a_1	a_3	a_5	a_7	$\cdots$
s^{n-2}	b_1	b_2	b_3	$\cdots$	
s^{n-3}	c_1	c_2	$\cdots$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
s^1	$\vdots$	$\vdots$			
s^0	a_n				

(a)

s^3	1	B
s^2	C	A
s	$B - A/C$	
1	A	

(b)

Table 5.1: Routh-Hurwitz stability criterion. (a) general form of the Routh table. (b) a solved form of the Routh table for the third order linear ODE, $\dddot{x} + C\ddot{x} + B\dot{x} + Ax = 0$.

5.4.6 Model selection criteria

Internally the VB toolbox optimises model fit through maximising the lower bound of the model evidence, otherwise known as the model free energy, F. This metric is therefore easily extractable and a useful indicator for model selection as it naturally penalises for

increased complexity in the model. To compare the free energy values of model candidates a difference in F was calculated between each pair. The superior model is then either the model with a greater value of F or the simpler model if the difference is not significant ($\Delta F < 3$). Free energy is effective as a model selection criteria however is limited in scenarios where the candidate models being compared are of a different order. This is because increasing the model order not only introduces additional degrees of freedom to the parameter space - but also adds dimensions to the system states, thus dramatically decreasing the free energy of the system and rendering comparisons less meaningful. For situations where different model orders are present a second modelling criterion is utilised - the normalised root mean squared error (NRMSE). NRMSE, is used to compare the quality of models which have different orders. First, mean values for the inferred parameters and initial conditions are fed back into the system equation to generate the deterministic solution for each DSA time series. Next, DSA measurements, y_t, and the inferred deterministic solution, $\widehat{y_t}$, are then used to calculate the NRMSE,

$$\text{NRMSE} = \frac{1}{y_{max}} \sqrt{\frac{\sum_{t=1}^{n} \left(\widehat{y_t} - y_t\right)^2}{n}} \tag{5.4.6.1}$$

where n is the number of data points in the measured DSA time series and y_{max} is the maximum value of y used as a normalisation factor. Normalising the DSA response adds further utility as opposed to comparing RMSE due to the vastly different DSA levels that are present amongst each case in the cohort – for example Case 057 has a maximum DSA level of 34,000 MFI in contrast to Case 064 with a maximum DSA level of 3,200 MFI. A lower NRMSE is considered optimal as it demonstrates the model is more capable of capturing the data accurately. NRMSE values have been used as a model selection criteria before in published works on DSA response with values < 0.15 considered satisfactory under the known inter-assay coefficient of variability for DSA measurements which is cited as between 10-30 [34]. As such this value will also be used in this work as a benchmark for model performance. Further to evaluating model fit an approach to model validation is also performed. Model validation is an important step in assessing whether a solution is stable given small perturbations to the original data set. This is particularly useful in the case of evaluating NRMSE where no penalty exists for increasing model order. Higher order models likely provide a better fit to data compared to lower order models due to the extra degree of freedom in the parameter space - at the cost of

being more likely to suffer from the problem of over-fitting. To assess for stability in the solution, a validation technique known as leave-one-out cross validation (LOOCV) is applied to compare between candidate models [184]. For a time series, y with length n the procedure outlined in Algorithm 4 is performed with each model M. By comparing NRMSE_{cv} for two models - one with greater fit but a higher model order - it can be assessed if the improvement is a consequence of over fitting.

Algorithm 4 Assessing model stability with LOOCV.

1: **for** $i = 1 : n$ **do**
2: Leave out data point i from the measurement time series.
3: Perform parameter inference on time series using VB method.
4: Obtain an estimate for missing data point $\hat{y}_i$.
5: Compute the estimation error term $e_i = y_i - \hat{y}_i$.
6: **end for**
7: Compute the NRMSE_{cv} from error terms $e_1, ...e_n$.

5.5 Results

For each of the found dynamic cluster groups (i.e., not Group-0) candidate models are presented and compared based on increasing complexity as outlined in the methodology. Table 5.2 highlights the first three models which were considered for each response group.

Model name	Model	Initial conditions
$M1$	$\dot{x} + x\theta_1 - \theta_1\theta_0 = 0$	$x_0 = x_{10}$
$M2$	$\ddot{x} + \dot{x}\theta_2 + x\theta_1 - \theta_1\theta_0 = 0$	$\dot{x}_0 = x_{20}, \quad x_0 = x_{10}$
$M3$	$\dddot{x} + \ddot{x}\theta_3 + \dot{x}\theta_2 + x\theta_1 - \theta_1\theta_0 = 0$	$\ddot{x}_0 = x_{30}, \quad \dot{x}_0 = x_{20}, \quad x_0 = x_{10}$

Table 5.2: A list of the five candidate models list used in Chapter 5 For each of the DSA model selection tasks the listed models were considered.

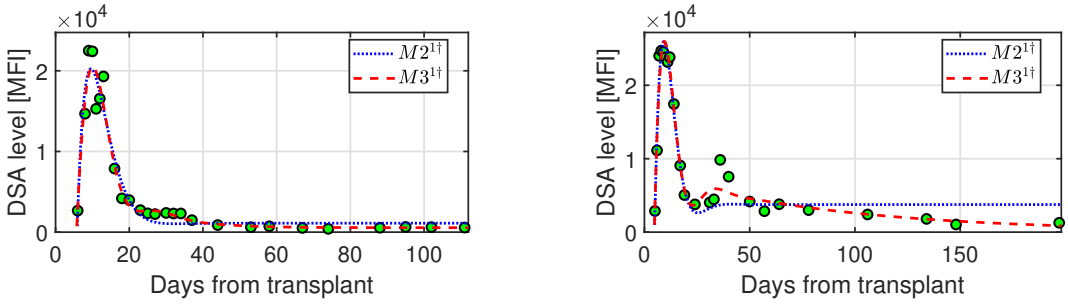
In each model the constant term is defined by $\theta_1\theta_0$, which benefits the modelling process due to simplifying to $x\theta_1 - \theta_1\theta_0 = 0$ at stable conditions (i.e., $\dot{x} = \ddot{x} = \dddot{x} = 0$). This can then be manipulated to show that when stable, $x = \theta_0$, otherwise referred to as the settling level in this analysis. Before model fit θ_0 prior distribution can be established based off of the last known DSA value. When applied to a particular group a model, M, is designated by a superscript, n, (model priors given in detailed in Appendix B.0.1). Each of the above models have been validated as structurally identifiable via Taylor series expansion implemented in Wolfram Mathematica, the number of terms required for each model is indicated in Table 5.3.

Model name	Taylor Series Expansion	n terms
M1	Globally identifiable	3
M2	Globally identifiable	5
M3	Globally identifiable	7

Table 5.3: Results of structural identifiability analysis for models $M1$, $M2$, and $M3$. Each of the explored candidate models were shown to be structurally globally identifiable via the Taylor series expansion approach, as implemented in Wolfram Mathematica software (example shown in Appendix B.1).

5.5.1 Early application of candidate models to Groups 1, 2, and 3

For Groups 1, 2, and 3 model candidates began at second order complexity ($M2$) due to notable rise and fall characteristics requiring at least two time constants. Early application of both models M2 and M3 provided generally poor results in large part due to the early adsorption period noted in almost every case for DSA responses. This period, which had an estimated range of 2-7 days, typically resulted in model fits compromising between accurate estimation of peak height or settling level while estimation of early post transplant period was poor in all cases. To overcome this challenge DSA were fit with models $M2^{\dagger}$ and $M3^{\dagger}$. These models were established as a compromise to the challenge and excluded early adsorption data points, ultimately beginning analysis at the start of the rise - a time point independently identified for each case which represented the first data point of the rising phase in the dynamic response. Although assessing from this period onwards excluded the adsorption period from each model, it ultimately allowed for far more accurate peak and settling values to be inferred by both models $M2$ and $M3$.



(a) Example Case COV028 (Group-1). **(b)** Example Case COV105 (Group-1).

Figure 5.3: Visual comparison of model fit for $M2^{1\dagger}$ and $M3^{1\dagger}$ from start-of-rise period. Models $M2^{1}$ and $M3^{1}$ provided poor fit to the data due to the adsorption period following transplantation. In response each subsequent model (†) is fit from the start-of-rise period allowing for improved fit of DSA dynamics such as peak and settling level.

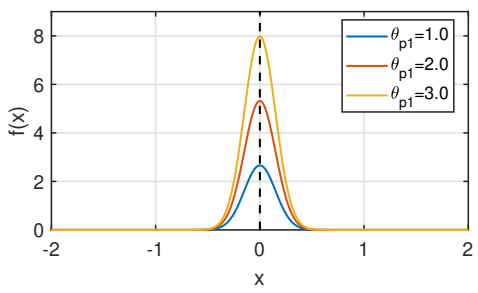
Figure 5.3 shows an example of the fit for two Group 1 cases modelled by $M2^{1\dagger}$ and $M3^{1\dagger}$ where parameters are inferred from start-of-rise data with early adsorption points excluded. In both cases the rise and fall dynamics are modelled similarly, capturing approximately similar and correct peak values according to the presented data. The rate of rise and fall in addition to the period spent at the peak can be inferred accurately. The differences in each model becomes apparent following the initial fall with model $M3^{1\dagger}$ being seen to capture a deeper level of dynamics within both of the shown cases. For Case 028, model $M3^{1\dagger}$ identifies two stages in reaching settling levels: the first, between days 20-35, is approximately twice that of the true settling level seen from days 35-110. In contrast model $M2^{1\dagger}$ shows only one settling value between the two. For Case 105, model $M3^{1\dagger}$ additionally captures a secondary oscillatory motion following the first rise and fall - before highlighting a much longer decay to a lower settling level than seen in $M2^{1\dagger}$. By comparison it can be seen that the true settling level is only approximately achieved by the end of the 200 day period in $M3^{1\dagger}$ in contrast to $M2^{1\dagger}$ which shows settling level achieved at day 35. Ultimately this analysis was effective at retrieving useful clinical data however is limited in that it requires manual selection of the starting day - a feature which can be subjective. To overcome this limitation an external input is considered known as a delayed impulse function.

The delayed impulse function, here on referred to as f_p, is typically defined as an infinitely high, infinitely narrow pulse with an area of unity - which is of course impossible to realise in a physical sense [185]. In this work a Gaussian impulse is used which produces a unit response with finite height and standard deviation designated σ^2, as shown,

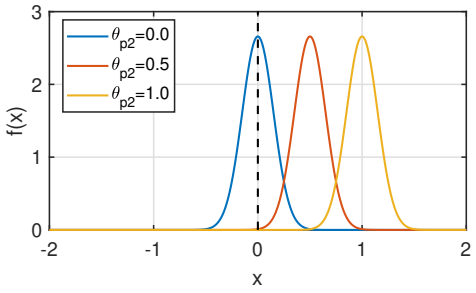
$$f_p(t) = \frac{\theta_{p1}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(t-\theta_{p2})^2}{2\sigma^2}\right). \tag{5.5.1.1}$$

With a small but finite standard deviation the Gaussian impulse will serve the same purpose as a unit impulse to initiate the model response. In this work $\sigma^2 = 0.2$ which is sufficiently smaller than the daily time frame which DSA observations may occupy. Given a pre-defined σ there are two additional parameters in the Gaussian impulse response, θ_{p1}, and θ_{p2} which define the magnitude of the impulse and time to mean respectively. The effects of these are illustrated in Figures 5.4a and 5.4b.

From here on the analysis conducted for Groups 1, 2, and 3 will utilise a delayed impulse function as shown in Table 5.4. Modelling will benefit from less bias due to θ_{p2} being



(a) Effects of increasing θ_{p1}.



(b) Effects of increasing θ_{p2}.

Figure 5.4: Demonstration of the Gaussian impulse function. The Gaussian impulse function, f_p, has three parameters σ, θ_{p1}, and, θ_{p2} of which $\sigma^2 = 0.2$ is fixed. The effects of magnitude parameter θ_{p1} and time delay parameter θ_{p2} are investigated.

optimised for in contrast to manually selected. Additionally, the model will be capable fully capturing the dynamic response, including the adsorption period.

Model name	Model
$M2_p$	$\ddot{x} + \dot{x}\theta_2 + x\theta_1 - \theta_1\theta_0 = f_p$
$M3_p$	$\dddot{x} + \ddot{x}\theta_3 + \dot{x}\theta_2 + x\theta_1 - \theta_1\theta_0 = f_p$

Table 5.4: Model list used in Chapter 5 for Groups 1, 2, and 3.

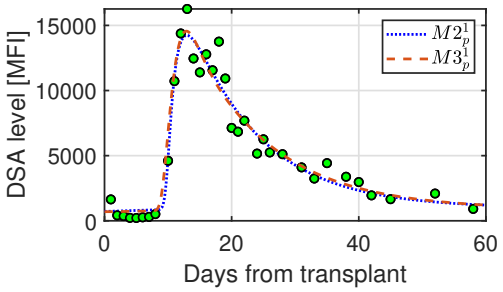
Following the introduction of models $M2_p$ and $M3_p$ each of the new models were assessed for their global structural identifiable via the Taylor series expansion approach as implemented in Wolfram Mathematica software however the approach was unable to ascertain a result. For these results a MATLAB (The MathWorks, Inc., Natick, MA, USA) software package known as STRIKE_GOLDD (version 3.0, created by Alejandro F. Villaverde, afvillaverde@uvigo.gal) was used that is capable of analysing nonlinear ODE models for parameter structural local identifiability (example shown in Appendix B.1). Both models were shown to be locally identifiable, as indicated in Table 5.5.

Model name	Taylor Series Expansion	Symbolic Engine
$M2_p$	Did not solve	Locally identifiable
$M3_p$	Did not solve	Locally identifiable

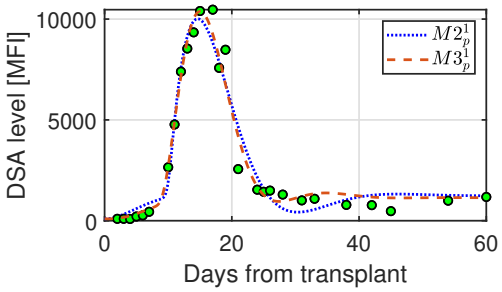
Table 5.5: Results of structural identifiability analysis for models $M2_p$, and $M3_p$. Each of the explored candidate models were shown to be structurally locally identifiable via the MATLAB software package STRIKE_GOLDD (example in Appendix B.1).

5.5.2 Group-1 model selection

By introducing the delayed impulse Group 1 cases could now be assessed for model fit to models $M2_p^1$ and $M3_p^1$, as indicated in Figure 5.5. Model fit is objectively improved upon those seen in $M2^{1\dagger}$ and $M3^{1\dagger}$ due to the ability to capture early adsorption period dynamics (seen within the first 2-7 days following transplantation). Cases COV010 and COV037 demonstrate this fact with little to no dynamic activity seen in the early period. By utilising a weak model prior for the delayed impulse mean time, θ_{p2}, set approximately at the start of rise for each case, the VB algorithm was able to optimise naturally without strong human bias. Comparing for fit in $M2_p^1$ and $M3_p^1$ little to no difference is seen in COV010 with both capturing the same peak and settling values in addition to demonstrating similar rise and fall rates. For COV037 model $M3_p^1$ demonstrates slight improvement over $M2_p^1$ at capturing the sudden change from fall to settling level dynamics, however as these deviations are within DSA expected coefficient of variation (CV) and should be assessed if this is a consequence of over fitting to the response data.



(a) Example case COV010 (Group-1).

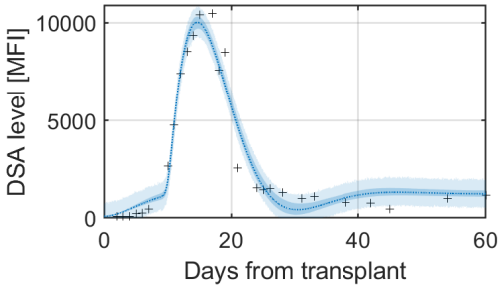


(b) Example case COV037 (Group-1).

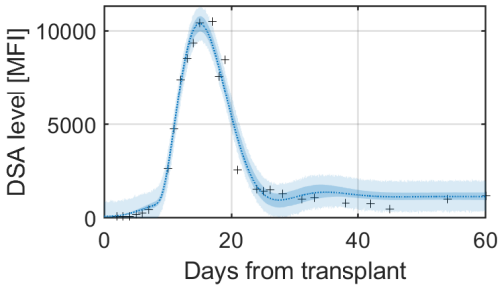
Figure 5.5: Example of model fit for Group-1 candidate models. Models $M2_p^1$ and $M3_p^1$ are compared for model fit. Objective improvements are seen over previous models $M2^1$ and $M3^1$ due to the ability to model the entire dynamics response including the early adsorption period following transplantation.

Following model inference parameter values are assessed for their practical identifiability, or sensitivity. For a given case, $n = 1000$ samples are taken from parameter posterior distributions and their response calculated. Figure 5.6 uses this analysis to visualise the model one-standard-error uncertainty at each time point (shown in dark blue). Additionally, a second analysis is performed including the measurement uncertainty and model uncertainty into the analysis (shown in light blue). Immediately it can be seen that confidence in model parameters are high in both $M2_p^1$ and $M3_p^1$ due to the tight bounds surrounding the mean parameter response with similarly strong performance for the in-

clusive measurement uncertainty analysis. Notably both models demonstrate high confidence in the start-of-rise period. Least confidence is shown in the model when capturing the transition from falling dynamics to settling level.



(a) Case COV036, model $M2_p^1$.



(b) Case COV036, model $M3_p^1$.

Figure 5.6: Example of model sensitivity for Group-1 case COV037. Models $M2_p^1$ and $M3_p^1$ are compared for model sensitivity. $n = 1000$ samples are taken from parameter posterior distributions and used to determine model and measurement noise uncertainty (dark blue and light blue respectively). One standard deviation bands are illustrated.

Assessing of the physiological considerations required analysis of model stability and confirmation that one-standard-error model uncertainty remained above the minimum established criteria of being positive. Model stability for a second and third order model were checked for by first calculating eigenvalues (Matlab solver) and confirming that all real parts were negative. Next parameter values were checked in accordance with Routh table criterion and in all cases satisfied the requirements. Similarly model uncertainty was also confirmed to meet positivity requirements meaning that both models $M2_p^1$ and $M3_p^1$ were able to successfully pass forward to the model selection phase.

As models $M2_p^1$ and $M3_p^1$ are of a differing model order the method of comparison will be that of NRMSE. NRMSE assesses model fit of the mean parameter value response compared to the observed data with a value of zero being a perfect fit. Figure 5.7 shows the results for all Group-1 cases comparing by both model type and showing standard and LOOCV stability assessments. When comparing for model stability a highly stable model will not deviate greatly under small perturbations of the data set - as experienced under LOOCV. By comparing LOOCV against the standard (all data) NRMSE an ideal result showing good stability will be indicated by little to no deviation. As seen in Figure 5.7 model $M3_p^1$ shows a large deviation for the LOOCV result with a substantial increase in NRMSE compared to the $M2_p^1$ model - despite the marginal difference seen in the standard NRMSE results. Due to the large difference it can be concluded that $M3_p^1$ model parameters are unstable under small perturbations to the data set and are likely over fit-

ting. In contrast $M2_p^1$ models perform substantially better - albeit still with some losses to NRMSE - leaving it a strong candidate for analysis of the Group 1 responses.

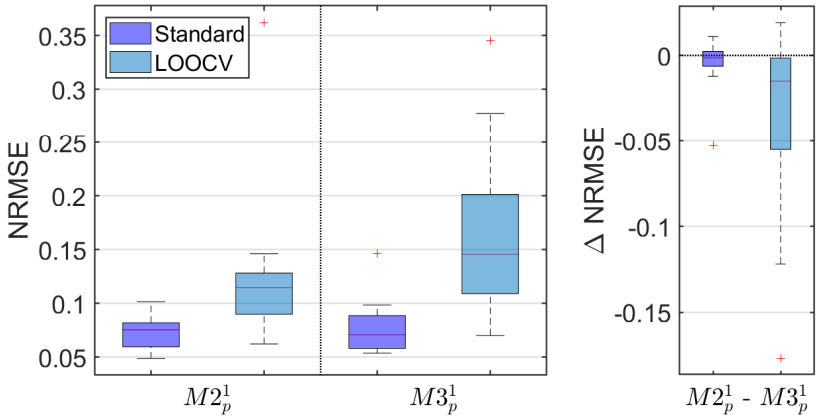
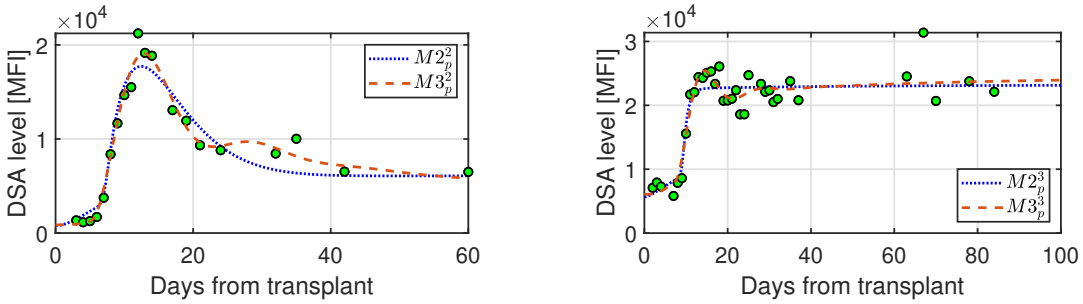


Figure 5.7: Comparison of NRMSE results for the Group 1 cases. Left - NRMSE box plots are shown comparing the fit for models $M2_p^1$ and $M3_p^1$ across Group 1. A standard NRMSE represents the fit of the mean parameter response to the observed data. LOOCV NRMSE is a measure of model stability. Large deviations from the standard NRMSE are a sign of unstable model inference. Right - a boxplot showing the difference between models $M2_p^1$ and $M3_p^1$ NRMSE.

5.5.3 Group-2 and Group-3 model selection

Despite undergoing separate model selection processes Groups 2 and 3 bare striking resemblance to the model selection for Group-1 cases and as such will be summarised more succinctly in this section. For Group-2 cases models $M2_p^2$ and $M3_p^2$ were used and for Group-3 cases models $M2_p^3$ and $M3_p^3$ were used (prior values detailed in Appendix B.0.1). Similar advantages were seen in these models over the implementation of $M2^\dagger$ and $M3^\dagger$ due to the ability to capture the adsorption period following transplantation as seen in the Group-1 cases. Two example cases shown in Figure 5.8, one from each group, show that model fits generally captured the data well. For the Group-2 case, COV036, there are large differences in the two model fits. While initially both models capture early DSA response similarly - including adsorption and rising period - model $M3_p^2$ notably fits peak DSA values more closely. Following from the peak $M3_p^2$ model offers a far superior fit by capturing the aggressive decay rate and identifying a secondary oscillatory motion. For the Group-3 case, COV071, little difference is seen between the two models with both seemingly capturing the adsorption period, rise and settling level dynamics.

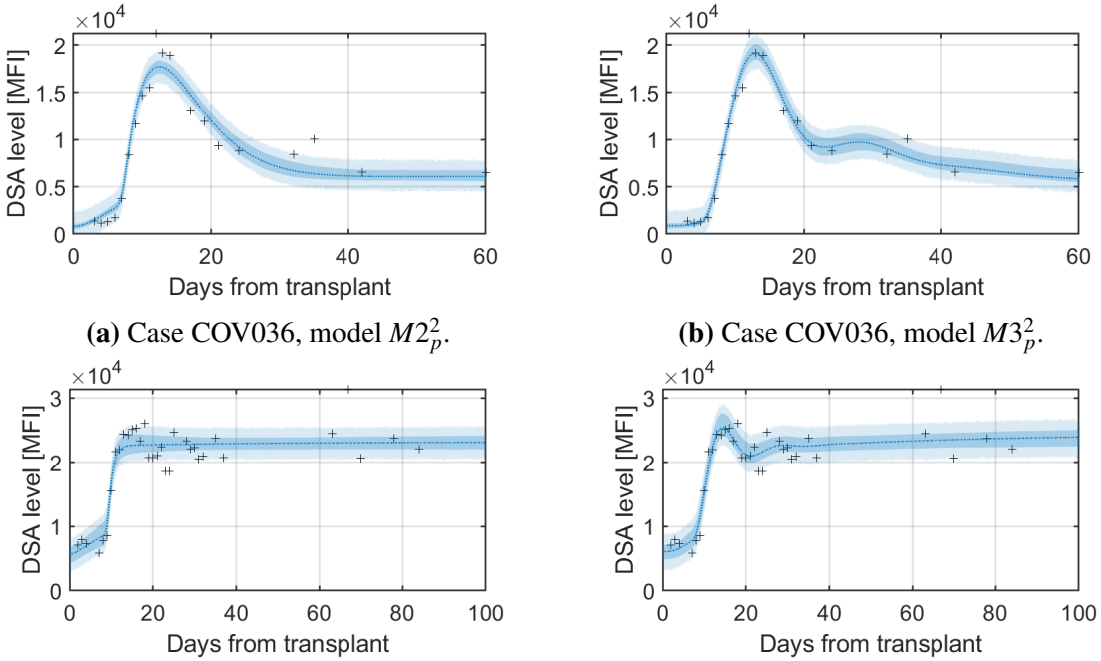
Next, Groups 2, and 3 are assessed for their practical identifiability by investigating the results of the Monte Carlo stochastic sensitivity analysis. Figure 5.9 displays these re-



(a) Example case COV036 (Group-2). **(b)** Example case COV071 (Group-3).

Figure 5.8: Example of model fit for Group-2 and Group-3 candidate models. Models $M2_p$ and $M3_p$ are compared for model fit. Objective improvements are seen over previous models $M2$ and $M3$ due to the ability to model the entire dynamics response including the early adsorption period following transplantation.

sults for two typical cases, COV036 from Group-2 and COV071 from Group-3. Model sensitivity among the four examples shows a strong level of confidence in the inferred parameters, owing to the tight dark blue band representing the model one-standard deviation uncertainty. When additionally considering the measurement uncertainty in the Monte Carlo analysis (shown in light blue) both models $M2_p$ and $M3_p$ offer convincing solutions and capture the observed data within the one-standard-error band. All cases in Groups 2, and 3 demonstrated similar such sensitivity characteristics and were considered to have passed the requirements for practical identifiability criterion.



(a) Case COV036, model $M2_p^2$.

(b) Case COV036, model $M3_p^2$.

(c) Case COV071, model $M2_p^3$.

(d) Case COV071, model $M3_p^3$.

Figure 5.9: Example of model sensitivity for Group-2 case COV036 and Group-3 case COV071. Models $M2_p$ and $M3_p$ are compared for model sensitivity. $n = 1000$ samples are taken from parameter posterior distributions and used to determine model and measurement noise uncertainty (dark blue and light blue respectively). One standard deviation bands are illustrated.

Comparison of models $M2_p^2$ and $M3_p^2$ for Group-2 cases and models $M2_p^3$ and $M3_p^3$ for Group-3 cases similarly required use of the NRMSE criteria due to the differing model orders between model candidates. The results of this analysis are shown in Figures 5.10 and 5.11, respectively. Notably for Group-2 cases there is little difference between NRMSE values in the two candidate models and the values are generally higher than seen in Group-1 cases. The subsequent LOOCV analysis also suggests that there is little gain in model fit by introducing an extra degree of freedom in model $M3_p^2$. For Group-3 models there may be some argument to be made about the use of model $M3_p^3$ over $M2_p^3$ due to an improvement in performance LOOCV NRMSE in the majority of cases. Despite this however, model $M2_p^3$ still shows strong performance with NRMSE well within CV criterion established at <0.15. The marginal improvement seen in $M3_p^3$ LOOCV NRMSE is therefore not considered significant enough to warrant its use over $M2_p^3$.

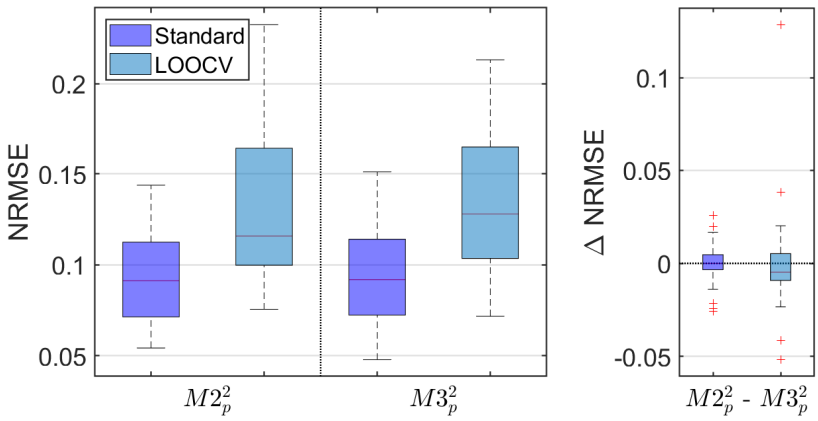


Figure 5.10: Comparison of NRMSE results for the Group 2 cases. Left - NRMSE box plots are shown comparing the fit for models $M2_p^2$ and $M3_p^2$ across Group 2. A standard NRMSE represents the fit of the mean parameter response to the observed data. LOOCV NRMSE is a measure of model stability. Large deviations from the standard NRMSE are a sign of unstable model inference. Right - a boxplot showing the difference between models $M2_p^2$ and $M3_p^2$ NRMSE.

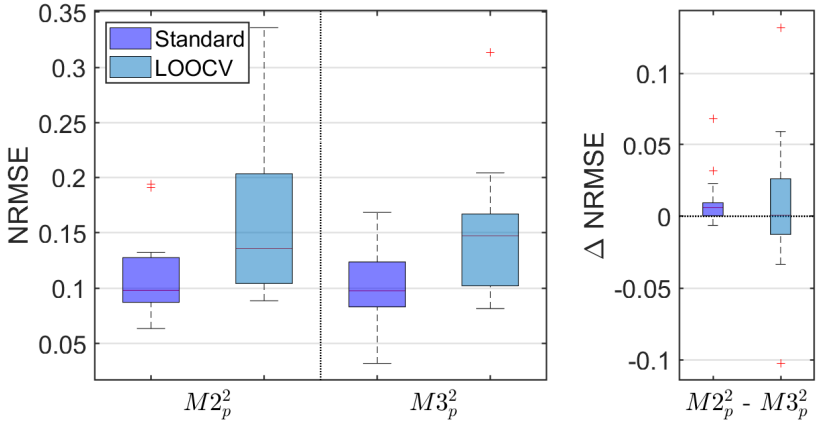
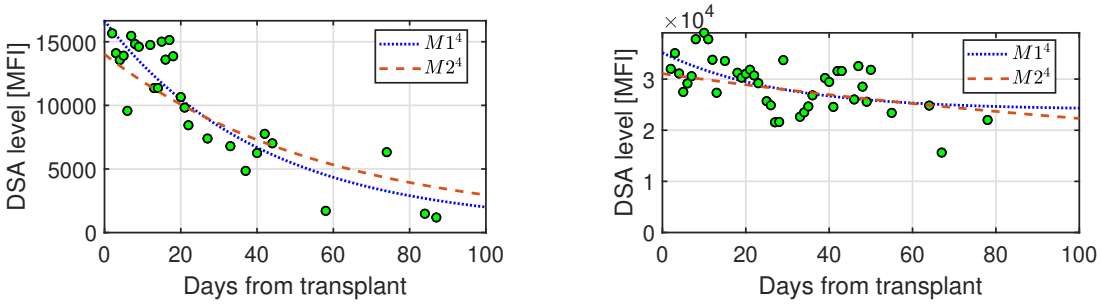


Figure 5.11: Comparison of NRMSE results for the Group 3 cases. Left - NRMSE box plots are shown comparing the fit for models $M2_p^3$ and $M3_p^3$ across Group 3. A standard NRMSE represents the fit of the mean parameter response to the observed data. LOOCV NRMSE is a measure of model stability. Large deviations from the standard NRMSE are a sign of unstable model inference. Right - a boxplot showing the difference between models $M2_p^3$ and $M3_p^3$ NRMSE.

5.5.4 Group-4 model selection

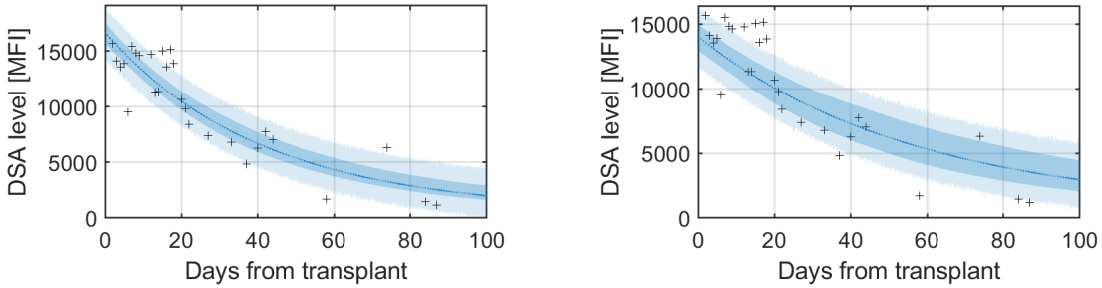
A lack of the rising DSA characteristic in Group-4 cases allowed for the consideration of the $M1^4$ candidate model for analysis as many cases appeared to demonstrate a slow decay over longer periods of time (exceeding that completed within the time series clustering task in Chapter 4). Additionally, without a time dependent rise the delayed Gaussian impulse served little to no purpose. As such, two candidate models were investigated for their possibility to fit the DSA responses, models $M1^4$ and $M2^4$ (prior values detailed in Appendix B.0.1). In almost all cases the two models showed little to no difference and generally followed the same decaying pattern - with $M2^4$ models tending only to fit to a marginally slower decaying pathway (shown in Figure 5.12). There were typically few characteristics to assess for model fit. Additionally, Group-4 cases appeared to demonstrate much greater levels of noise than other groups and many extreme value deviations can be seen in the data (such as Figure 5.12b data points $4 \rightarrow 5 \rightarrow 6$ showing wild changes in DSA levels from $14 \rightarrow 10 \rightarrow 14$ kMFI).



(a) Example Case COV063 (Group-4). **(b)** Example Case COV125 (Group-4).

Figure 5.12: Example of model fit for Group-4 candidate models. Models $M1^4$ and $M2^4$ are compared for model fit. Little difference is seen between candidate models with the main difference noted that $M2^4$ tends to take a shallower decay than $M1^4$ across the cohort.

The sensitivity of the two candidate models also illustrated a much lower degree of confidence in model fit across all cases. In both models a wider sensitivity (dark blue band) was demonstrated relative to model performance seen in Groups 1, 2, and 3. Of the two candidate models, $M1^4$ has the greater confidence - however the difference is slight.



(a) Case COV063, model $M1^4$. **(b)** Case COV063, model $M2^4$.

Figure 5.13: Example of model sensitivity for Group-4 case COV063. Models $M1^4$ and $M2^4$ are compared for model sensitivity. $n = 1000$ samples are taken from parameter posterior distributions and used to determine model and measurement noise uncertainty (dark blue and light blue respectively). One standard deviation bands are illustrated.

Lastly, to compare models $M1^4$ and $M2^4$ for the Group-4 cases the NRMSE measurement criteria was used due to the differing model orders between model candidates. The results of this analysis are shown in Figures 5.14. Model $M2^4$ notably exceeds the minimum criterion of NRMSE<0.15 on a number of cases in contrast to model $M1^4$ which stays within expectations. Additionally, $M2^4$ shows the most extreme drop in LOOCV NRMSE seen amongst any group and model showing large instability in the model results. This is potentially due to an increased number of extreme data points seen in the Group-4 DSA samples leading to overfitting from model $M2^4$. Final model selection for Group-4 is then to use model $M2^4$ for further analysis of the group.

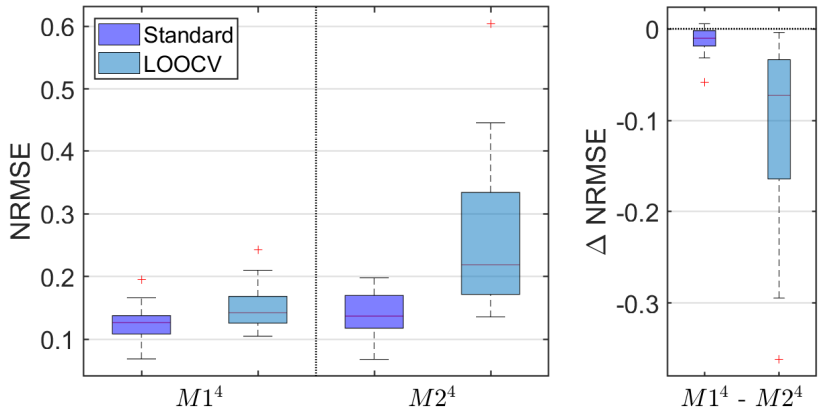


Figure 5.14: Comparison of NRMSE results for the Group 4 cases. Left - NRMSE box plots are shown comparing the fit for models $M1^4$ and $M2^4$ across Group 4. A standard NRMSE represents the fit of the mean parameter response to the observed data. LOOCV NRMSE is a measure of model stability. Large deviations from the standard NRMSE are a sign of unstable model inference. Right - a boxplot showing the difference between models $M1^4$ and $M2^4$ NRMSE.

5.6 Analysis of parameters

Groups 1, 2, and 3 were all found to be best modelled with a second order ODE with a delayed Gaussian impulse and can therefore be compared to one another to observe any significant differences between group responses. Differences between parameter and initial condition means are assessed via box plot analysis as shown in Figure 5.15. Significant difference between groups is determined by the non-parametric Wilcoxon rank sum test which assessed the null hypothesis that there is no difference between groups. Two distributions are considered significantly different when the test yields $p < 0.05$ for this work. From Figure 5.15 all five parameters in model $M2_p$ are compared alongside model initial condition x_{01}. A tight prior distribution was used to constrain initial condition x_{02} at zero for models $M2_p$ and is therefore not shown in further analysis. Within each box plot red and green markers represent the colour coded sample values in accordance with case censorship.

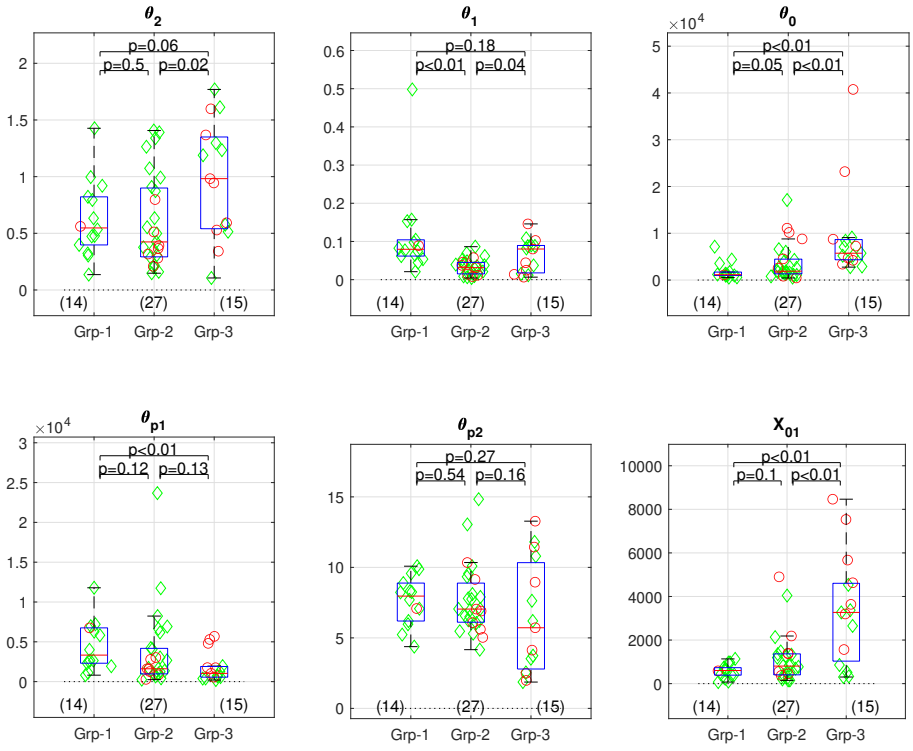


Figure 5.15: Parameter and initial condition mean values for model $M2_p$. Dynamic DSA groups 1, 2, and 3 are compared for differences in their response characteristics. Significant difference assessed via the non-parametric Wilcoxon rank sum test for differences between groups. In addition to the box plots sample values are indicated via green (censored) and red (not censored) markers, representing censorship.

The relative influence of a given variable on the response of the system when controlled independently can be seen in Figure 5.16 allowing for assigning intuition to the results in Figure 5.15. θ_{p2} for example correlates directly to the delayed impulse response time and indicates that no significant difference is seen between any of the groups. θ_{p1} has a large influence on the magnitude and steepness of the response, with higher values significantly associated with Group-1 response types. X_{01} and θ_0 correspond to the initial value and settling value of the DSA before and after the dynamic response has occurred. In both cases Group-3 is shown to have significantly higher DSA levels than Groups 1, and 2.

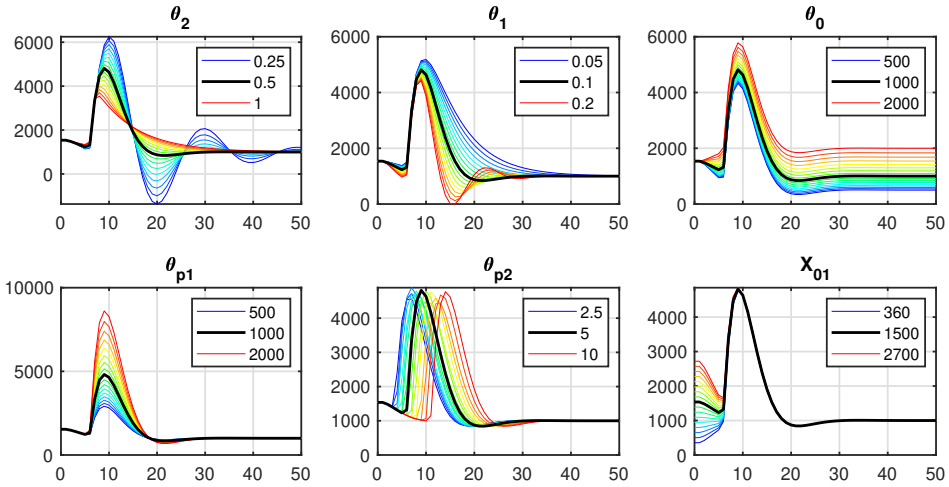


Figure 5.16: Univariate analysis of variable impact on the response for model $M2_p$. This set of subplots shows the impact of varying a single variable at a time for the response in an $M2_p$ model. An typical response taken from the Group-1 data set is shown in black consistently for each subplot. A colour spectrum is used to represent the shift between low variable values (blue) through to high variable values (red). Particular values for the low through to high are shown in the legend.

The inferred parameter and initial condition means were next analysed for relationship to graft survival in line with methodology outlined in Chapter 3 and will now be referred to collectively as variables. A short modelling pipeline was developed which included a standard scaling function applied individually to each of the variables. This stage allows for comparison between variables and also serves to limit the influence of outliers - such as the single occurrence of $\theta_1 = 0.5$ (Figure 5.15). Following this a univariate cox proportional hazards model was developed for each variable with bootstrapped sampling utilised to indicate confidence in the inferred output. Output in this case is HR. For every 1 unit increase in standard deviation for a given variable a corresponding increase in HR is expected. Bootstrapped confidence intervals are given at two standard deviations from the mean encompassing approximately 95% of observed results.

Due to Groups 1, 2, and 3 sharing the same model, $M2$, it was possible to assess their parameters under the same analysis, thus increasing the available cases and improving confidence in HR estimates. Figure 5.17 illustrates the histogram results of the bootstrap analysis for each of the model variables. From top left to bottom right variables are ordered according to their mean difference from zero - suggesting that X_{01}, θ_0, and θ_1 are likely to be the most informative variables for graft survival. In this case the 95% confidence intervals are indicated in the orange band. Confidence intervals are used to determine a significant finding by observing if $> 97.5\%$ of bootstrap samples are either side of the value one. In this case only two variables X_{01} and θ_0 demonstrate significance. In either case a higher value indicates a greater HR and therefore a lower chance of survival. For X_{01}, which corresponds to the initial value of DSA following transplantation, an increase in 1860 MFI is expected to increase the HR by 2.7. Similarly for θ_0, which corresponds to DSA settling level, an increase of 6470 MFI corresponds to an increase in HR by 1.95. Variables which correspond to the magnitude and time constants for rise and fall, such as θ_1, θ_2, and θ_{p1} are shown to have little influence on graft survival. Of all variables θ_{p2}, which determines the position of the impulse response in time, is shown to have the least impact on long term graft outcome.

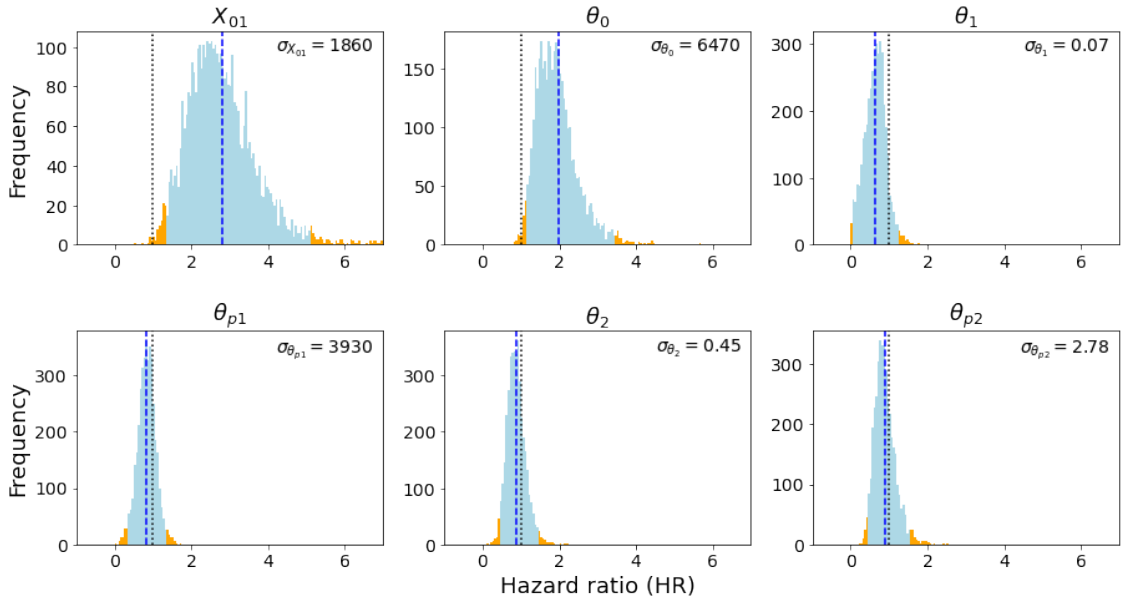


Figure 5.17: Univariate analysis of model $M2_p$ variables against graft survival. Groups 1, 2, and 3 model $M2_p$ variable means and combined in a univariate Cox proportional hazards survival analysis to determine associations with graft survival. Histograms represent the results of $n = 5000$ bootstrap samples, black dotted line indicates a HR=1, blue dashed line indicates the bootstrap mean. The 95% confidence interval is indicated by the orange bands. Significance is determined by $> 97.5\%$ of the bootstrap samples being on either side of HR=1.

With only a few cases ($n = 16$) in Group 4, bootstrap confidence intervals were far reaching and offered little value to explore. Where settling and initial DSA values had previously shown to be significant an alternative approach is used where Group 4 cases are now reassessed for these variables with all dynamic cases. The results of this are shown in Figure 5.18 alongside a reference distribution "Random" which is the results for an additional variable generated from a random number generator. Notably, the initial DSA variable which was previously significant when observed for cases in Groups 1, 2, and 3 is no longer demonstrated. This finding suggests that the value of observing the initial DSA value may be group specific and not necessarily a bad indicator for Group 4 cases though this is not possible to confirm without a larger sample size. Settling DSA level shares the same rounded standard deviation of 6470 MFI when including Group 4 cases (as compared to θ_0 in Figure 5.17) and is also similarly significant as in the previous study. A higher settling DSA level results in a larger HR and worse graft outcome for the patient. Observing a completely random variable is used here to check model performance is in line with expectations, where a tight confidence interval about one shows that the random variable has no association with graft survival.

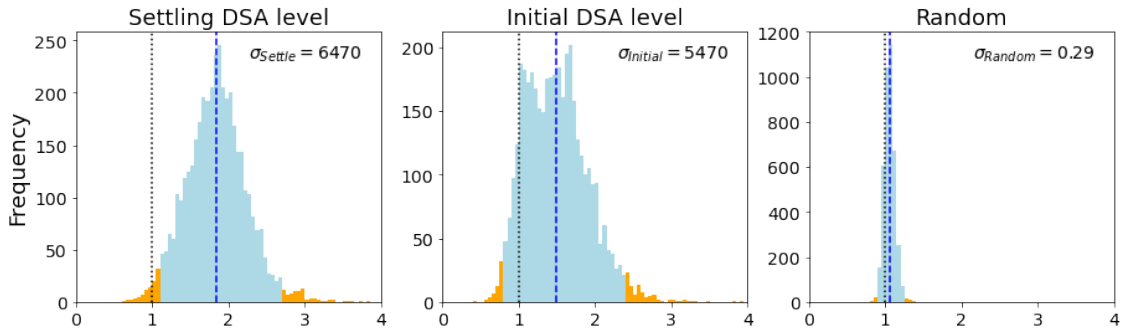


Figure 5.18: Univariate analysis of DSA initial and settling level variables against graft survival. Groups 1, 2, and 3 model $M2_p$ variables X_{01} and θ_0 are combined with Group-4 model $M1$ X_{01} and θ_0 to form initial and settling level DSA mean estimates for all dynamic cases. A univariate Cox proportional hazards survival analysis is then used to determine associations with graft survival. Histograms represent the results of $n = 5000$ bootstrap samples, black dotted line indicates a HR=1, blue dashed line indicates the bootstrap mean. The 95% confidence interval is indicated by the orange bands. Significance is determined by $> 97.5\%$ of the bootstrap samples being on either side of HR=1. A "Random" result is introduced to confirm the intuition of a baseline model.

5.7 Discussion

In this Chapter the primary aim was to mathematically describe the evolution in each of the previously identified DSA response groups (from Chapter 4), and, following a successful modelling approach, the secondary aims were to compare the dynamic responses between groups and then identify any response characteristic associations with long term graft outcome. The primary aim of this chapter was approached via three main stages: identification of a modelling framework, implementation within the VBA toolbox, and lastly, development of a model selection methodology. The secondary objectives of comparing group dynamics and observing associations with graft survival were assessed on inferred model parameters using both statistical techniques and univariate Cox proportional hazards models as outlined in Chapter 3.

In preparation of the primary aim, an appropriate framework for modelling physiological responses was considered, whereby decisions were made on modelling approach, model form, and parameter inference method. For the modelling approach a physiological based method was ultimately thought unfeasible due to the complexity of the underlying immune response system in addition to uncertainty limitations found within the current SAB DSA sampling methodology (Chapter 2). As a result a parametric data-driven modelling approach was utilised with a form based on an ODE. These choices allowed for great flexibility in selecting candidate models when seeking to capture the underlying physiological characteristics, and have been used before as the basis for forming such physiological studies [34, 176].

Inference of model parameters may commonly be performed within either a frequentest or Bayesian framework, where the former considers parameters as represented by a single value, and the latter considers them as a PDF. The advantages of utilising the Bayesian framework for this modelling task was two-fold as it served not only to give confidence distributions over model initial conditions, parameters, and uncertainties, but also allowed for establishing prior values to loosely nudge the response based on intuition. Both model and parameter identification were carried out using the MATLAB VBA toolbox which employs a VB scheme for the inversion of nonlinear state-space models. This is a deterministic approach and approximates parameter posterior distributions through the assumption of certain properties taking a parametric form.

Model selection as a process was split into three main sections: developing candidate models for each group, assessing model suitability via modelling criterion, and lastly the model selection stage which analysed successful models for final representation of the DSA group. Candidate models were considered sequentially with simplest first order ODE models considered first before introducing additional degrees of freedom with higher order models. External inputs were to be introduced as required. For a model to be suitable there were three main criteria it must satisfy: be structural identifiable, be practically identifiable, and satisfy the physiological considerations of demonstrating a stable (i.e., not diverging) and positive DSA output. In all considered cases in this work models were shown to be at least locally structurally identifiable and offered a high level of confidence in model output when considering their respective practical identifiability - presenting a strong degree of confidence in parameter estimates. Physiological considerations were also satisfied. Identifying a suitable model selection process would need to accommodate in particular for the large and inconsistent uncertainties presented in the sampling of DSA measurements. This process would manifest itself as an additional blurring effect on model confidence. Measurement uncertainty can be as high as 30 % in some DSA samples or considerably lower in others. Several factors can influence this uncertainty which may manifest during the SAB measurement process and result in the stated value itself being highly variable. To assess then that models were not being over fit to noise, a LOOCV scheme was considered during the model selection stages to calculate stability under perturbations of the data. Final model selections were based upon models which performed strongly in both quality and consistency of model fit relative to others.

For groups Groups 1, 2, and 3 which demonstrated similar rising characteristics, second and third order linear ODEs were considered in the form of models $M2$ and $M3$ - however resulting model fits were poor. The poor model fit was attributed to the early saturation phase in the first few days following transplantation whereby DSA levels demonstrate little dynamic activity. To accommodate for this process the introduction of an external input - the delayed Gaussian impulse - was considered. This input crucially introduced two extra degrees of freedom to the models (labelled $M2_p$ and $M3_p$) allowing for response to occur following a saturation phase. Only Group 4 differed in this modelling approach as there was little to no evidence of a delayed response amongst cases. In these cases the linear first and second order models $M1$ and $M2$ were considered. Final model selection

highlighted some important similarities between the previously identified cluster groups whereby model $M2_p$ was found to best represent Groups 1, 2, and 3 despite separate model selection processes. Amongst these groups, and for a minority of cases, model $M3_p$ was able to produced a superior fit compared to $M2_p$ - as compared via both model fit and stability - however improvement was always slight. For this work a generalisable model for each group was desired - due to the benefit it provided to subsequent stages of analysis, i.e., allowing for direct comparison of model parameters to one another. For the Group-4 cases, the first order model, $M1$ was shown to be superior both in terms of fitness and stability - indicating that the DSA dynamic response could be represented as just a single exponential decay from time of transplantation with a half life.

Results for the secondary objectives were conducted on the parameters and initial conditions of each system, with the mean value of said quantities taken as a reference point in analysis. Combined, Groups 1, 2, and 3 make up 78% of dynamic responses (56/72) and the characteristics of their respective DSA evolution following transplant are all shown to be well captured by the second order system with delayed Gaussian impulse. At only 22% of dynamic responses (16/72), Group-4 was the only group to be best described with an alternative model, $M1$.

By sharing the same model form it was possible to directly compare the parameters of each of these groups 1, 2, and 3 to one another. One area of interest in this work has surrounded the period of inactivity following transplantation. This period of subdued DSA levels following transplantation has been noted in several studies before [26, 64, 113], with potential causes arising due to adsorption of HLA antibodies onto the kidney allograft [26]. The median period of inactivity was between six and eight days however no significant difference was found between groups. More significant was analysis of the initial DSA and settling DSA values which highlighted how Group 3 cases had higher levels of DSA both immediately following transplantation and after dynamic periods were over. When considering that graft survival rates differ between Groups 1, 2, and 3 (Chapter 4), and that Group 3 cases have worse 1, 3, 5, and 10 year survival compared to Groups 1 and 2 it may be hypothesised that the rates of rise and fall dynamic are less important than the initial and settling DSA levels.

Following up on this analysis each model parameter was assessed for its influence on graft survival using an independent univariate Cox proportional hazards model (method-

ology as detailed in Chapter 3). This analysis echoed the previous sentiment on initial and settling DSA levels by highlighting a significant association between the factors and long term graft survival. For the initial DSA value an increase in only 1860 MFI would correspond to a HR factor of 2.7. Similarly, for the settling level DSA value an increase in DSA by 6470 MFI corresponded to an increase in HR by a factor 1.95. Like our study, others reported these same longer term findings [110, 112, 114], however these were defined based on a single time-point at month-3 following transplantation. For Group 4 cases it was not possible to analyse characteristics directly due to the limited sample size ($n = 16$), however a subsequent analysis included initial and settling level DSA values into an a larger study including all dynamic response types. In this case both of the previously significant characteristics saw reduced confidence in their HR and the estimate for initial DSA lost significance completely. These finding potentially indicate that settling and initial DSA levels may be important factors in the prediction of long term graft survival - but perhaps only for DSA responses demonstrating rise and fall dynamics.

5.8 Conclusions

In this chapter the primary aim was to mathematically describe the DSA evolution of each of the previously identified dynamic response groups (Chapter 4). To achieve this aim a model development pipeline was implemented to complete the task of model selection. For the task of modelling physiological based systems one can either look to physiological or data-driven approaches. A physiological based model can offer superior insight into the underlying immune response - however requires a deep understanding of the underlying dynamical systems. For the DSA immune response a complete understanding of the components and interactions of the physiological system is not known and this task instead focuses on the latter which can be implemented in its absence. The data-driven models in this work are formed using ODE's and are based within a Bayesian framework for parameter inference. For both the tasks of model and parameter identification a MATLAB VBA toolbox was used which employs the VB scheme for the inversion of nonlinear state-space models - allowing for a deterministic approximation of parameter posterior distributions via the assumption of certain parametric properties. For model development each of the DSA group responses would be investigated for suitability and fit versus a set of candidate models. Results of this process highlighted important similari-

ties and differences between dynamic groups:

- **Groups 1, 2, and 3.** For each of these DSA response groups it was independently found that the second order delayed Gaussian impulse model, $M2_p$, produced superior fit while considering model stability under LOOCV analysis.
- **Group 4.** For group 4 it was found that the linear first order model, $M1$, produced superior fit and stability to the alternate considered second order model, $M2$.

Via model $M2_p$, DSA responses in groups 1, 2, and 3 were shown to demonstrate a delayed response characteristic - previously noted by other groups [26, 64, 113] and hypothesised to arise due to adsorption of HLA antibodies onto the kidney allograft [26]. For the secondary objectives of the work the dynamic response characteristics were compared between groups for association with long term graft outcome. In this work the length of the delay period was captured for each of the group 1, 2, and 3 responses, however ultimately showing no association under Cox proportional hazards univariate analysis. Two characteristics were found to be significant for group 1, 2, and 3 responses: initial DSA level, i.e., in the immediate period of saturation following transplantation, and the settling DSA value which occurs following the period dynamic activity (length varies depending on response type). For both characteristics a higher DSA level was found to be indicative of worse long term outcomes. Group 4 by itself showed no significant outcomes - perhaps due to its relatively small group size ($n = 16$) - however its estimates for initial and settling levels could reliably be interpretted alongside other groups. Inclusion of these cases proved to weaken the association of those characteristics with respect to graft outcome - potentially indicating that settling and initial DSA levels may be important factors in the prediction of long term graft survival, however perhaps only for DSA responses demonstrating rise and fall dynamics. One potential avenue for future work in this area may be to include modelling in relation to more detailed characteristics of antibodies. An example of this may be to assess DSA individually as opposed to an aggregated sum - it is known for example that some patients have up to seven different DSA allele types. Alternatively this work may see application in the area of synthetic generation of response data - stimulating more detailed research into optimal monitoring protocols.

Chapter 6

Conclusions and future work

Kidney graft failure is a major problem following HLAi transplantation [25], and has resulted in considerable efforts by the research community to identify high risk patients. This information is likely to be highly valuable in clinical practice, potentially allowing for early preventative action to be taken [186]. Preferable approaches towards this task have been through the investigation of different minimally invasive biomarkers that can assessing risk [187], including the monitoring of DSA levels which has been shown to be a promising tool for clinicians. The role of post-transplant monitoring of DSA is not fully established however - even in cases that had pre-formed DSAs - and as such monitoring protocols are typically centre specific with wide variation amongst the different transplant centres [188]. The work conducted in this thesis sought to contribute new domain knowledge to the area of AiT kidney transplantation that may better equip clinicians for patient care in the future. More specifically the thesis aim is then structured to investigate the associations between early post-transplant DSA dynamics with long-term graft outcome. To fulfil this aim, four thesis objectives were devised (Chapter 1) and achieved - resulting in three main contributions of the work to the area of HLAi kidney transplantation:

1. Recommendations for optimal DSA monitoring periods around transplantation that associate most strongly with long term graft outcome.

2. Identification of the early period DSA dynamic response types following HLAi kidney transplantation.

3. Development of data driven dynamical models for parametric based description of DSA dynamic response types.

For each of these results analysis was based upon the HLA AiT kidney transplant cohort provided by the UHCW. In total 99 cases were assessed and collected between June 2003

and February 2014, with follow-up data available up until March 2019. Use of this cohort provided some of the first novelty in the work due to a high frequency of DSA monitoring in the early post-transplant period. Cases here had a median of 19 samples in the first 2 months following transplantation. Such high frequency is uncommon in clinical practice due to the costly DSA sampling process. Even in literature, studies typically either make use of single pre- and/or post-transplant DSA levels only, with just a select few studies found to be monitoring over the weeks or months following transplantation [27, 64]. Ultimately this has allowed for more detailed analysis of the early post-transplant period than others have achieved.

For the first contribution, work was conducted in Chapter 3 and looked to build upon the results of more recent publications that focused on use of DSA as a bio marker in associative or predictive models - such as those completed through classical statistical multivariable approaches [36–38, 189], or more sophisticated machine learning algorithms [39]. Motivation for this analysis formed due to the general recognition that presence of DSA both before and after transplantation correlates strongly with poorer graft outcome [187, 190–192] - however best practice in monitoring and use of DSA was still a developing area of research. A consistent and well defined protocol amongst different practices would benefit clinicians in decision making while also improving statistical power of future studies through an increased cohort size spread across multiple centres. To achieve the main aim of identifying the most informative DSA measurement period, a separate multivariate survival model was developed for each day surrounding transplantation, i.e., from 7 days prior up to 88 days post (day range based on minimum graft failure sample size of 10). Results of this work began prior to the model development where additional (i.e., non-DSA) factors were assessed under bootstrapped conditions for inclusion in multivariate models. Variables: patient age, crossmatch status at time of transplant, and time on dialysis, were found to be informative of long term graft outcome and subsequently included. These findings corroborated those of others [15, 33, 38, 48, 106–108]. Comparison of daily models revealed that monitoring in either the pre-transplant period or between 6-12 weeks post transplant offered strongest DSA associations with graft failure, with higher DSA values indicating worse outcome. Recommendations of this work would suggest that if few observations are to be made due to financial considerations, a monitoring protocol of two DSA samples on days -2 and 70 are likely to be most informative of long term graft outcome. This consistency would additionally allow for results

to be verified on a larger multi-centre cohort.

The second contribution of this thesis was the discovery and description of the early period DSA dynamic response types. Following kidney transplantation, allorecognition and memory response to the re-exposure of previously exposed HLA-protein produces immune response with an associated rise in DSA and rejection [193]. The response following can be variable [194] and may depend on the type of sensitisation events [195], age, baseline immunosuppressant, time since sensitisation and level of crossmatch before transplantation. Of these response types, a rising DSA trend is often considered more worrying and more clinically useful than a steady state or drop in the MFI values. Previously reported dynamic patterns have been seen in individual donor-specific HLA antibodies and third party HLA antibodies before, primarily identified via using visual description [26, 34] - however these studies did not investigate all outcomes. For the first time in this work (detailed in Chapter 4 and published in [196]), an unsupervised clustering approach was applied to DSA MFI time series - revealing five different types of response: fast-modulation (group-1), slow-modulation (group-2), rise-to-sustained (group-3), sustained (group-4), and no-response (group-0). Modulation of DSA post-transplant versus sustained DSA production has been observed by others [34, 110, 113, 114] with variable outcomes, however what is novel here is the ability to define a fast modulation group that has good outcomes - even though DSA levels may appear significant at 20-30 days post-transplantation. Clinically, these results additionally support observations that early acute ABMR is not generally deleterious to longer term transplant survival [65, 85, 195]. That is not to say that hyperacute ABMR and some cases of overwhelming acute ABMR may not cause graft failure, but that once acute ABMR is resolved the longer term outcome may be good. The results from this study have important implications for the practical management of HLAi transplants, and also for the understanding of the immunological dynamics post-transplantation. The role of unsupervised machine learning [118, 132] was important here, because clusters of DSA behaviour were not created in order to satisfy any pre-exisiting hypotheses; indeed, although the importance of the fast modulation group was expected, the critically important separation between the slow modulation group and the sustained groups was not anticipated in advance of the analysis. This is the first description of an objective, mathematical classification of transplant-driven HLA responses, and it is believed to be the first such high-resolution classifications of acute humoral responses in humans. The procedures that have been

used may therefore be generalisable to other specific responses, such as to pathogens and vaccines. Lastly, analysis was not able to indicate exactly what factors may determine modulation versus a sustained response, but they do provide an tight period for interrogation of immune responses. By not confusing slow modulation with a sustained response it will now be possible to search for the underlying physiological mechanisms.

For the third contribution a data-driven modelling task was completed whereby a parametric model was developed for each of the previously identified DSA post-transplant dynamic response types. Certain effects such as the subdued DSA levels following transplantation have been noted in studies before [26, 64, 113] - however mathematical descriptions of complete DSA response profiles are novel results. This work builds upon the results of a previous PhD project [197] which outlined a robust methodology, and ultimately published a mathematical model for the partial (decay phase) description of what is now recognised as the fast-response type [34]. Candidate models were proposed in the form of ODE and model parameters subsequently inferred via a VB parameter estimation task. The VB method makes use of a deterministic and efficient algorithm to approximate the true posterior distribution over unknown parameters, while simultaneously providing a lower bound on the model evidence. The work in this thesis (conducted in Chapter 5) focuses on the application of the VB method through use of a MATLAB toolbox - referred to as the VBA toolbox, with mathematical details found in [178, 179, 197]. A model selection framework was implemented whereby each group would be assessed independently. For Group 1, 2, and 3 cases it was found that a second order ODE with delayed response, fit best across their respective cases. This model allowed for representation of key features, such as the aforementioned subdued DSA levels dynamics and rise-and-fall dynamic phase. Group 4 cases produced different results from the others, where neither a subdued period nor rise-and-fall dynamics were experienced. It was instead found that a simple 1st order ODE produced best fit - indicating that these responses may be described by a simple decay rate when on a longer time frame. Lastly, parameters were assessed for their relationship with long term graft outcome - ultimately showing that for groups 1, 2, and 3, higher initial DSA and settling DSA levels were strongly associated with worse survival. However when including group 4 cases in analysis associations were less important or even non-significant. These findings may partly suggest why early period DSA levels in the daily DSA analysis (Chapter 3) had wide uncertainty - indicating that magnitude of DSA level may only be important for certain types of DSA response.

This study consisted of 99 cases transplanted at a single University Hospital with patients from multi-centre UK and Republic of Ireland and the results reflect specific immunosuppressive protocol, management and monitoring protocols, and case selections - limiting generalizability to wider patient groups. Ideally, a larger multi-centre study is required to confirm the findings. Other limitations for the study include cases discharged back to parent units, and management protocol that may have influenced long term outcomes; additionally, protocol biopsies were not employed and may have missed sub-clinical ABMR, which has been shown to influence long-term outcomes. Finally, SAB are the best available option for determining DSA levels [58] and have improved the ability to identify and manage allosensitised transplant patients [59]. They provide a semi-quantitative measurement of DSA in the form of a MFI reading however are not without limitations in measurement accuracy. It is recognised there can be a large coefficient of variation between assay runs [198] which increases with lower antibody levels and also these data were obtained without any form of serum treatments such as EDTA to resolve complement interference, which may mean MFI levels in some cases could be an underestimate. Despite these limitations an in-depth and detailed analysis of early post-transplant DSA dynamics and their associations with long term graft outcome has been conducted. These approaches improve the understanding and aids in the development of a DSA monitoring protocol by highlighting two main periods which demonstrate the most value to monitor in clinical practice, while also indicating key differences in early post-transplant DSA dynamics. This work may help in future tailoring of treatment, so that lower risk HLAi patients are not subjected to over-immunosuppression even if they have had early acute rejection and that high-risk patients can be looked at more carefully even if they haven't had an early acute rejection.

Future work would benefit most from a larger multi-centre cohort which could serve to validate this work beyond AiT patients at UHCW. This is not only due to the various treatment and monitoring protocols seen across transplant units, but also as the relatively small UHCW cohort did not have sufficient cases to perform predictive analysis which requires splitting of the data into test, train and validation subsets. Other avenues of work could focus on data that is already available, such as exploring the influence of different allele types on transplant outcome. Throughout this thesis DSA levels are formed of the aggregate of all DSA within a patient, however one could consider how results perform given use of the immuno-dominant DSA or even including multiple in a multivariable

analysis. Lastly, this work has shown that DSA response types play an important role in long term outcome - however it is not yet clear how DSA sampling can be specified in order to obtain class type. Development of group specific parametric models has now presented an opportunity to synthesise new data sets which may be used to conduct analysis into the effects of different sample rates.

Bibliography

[1] Vivekanand Jha et al. “Chronic Kidney Disease: Global Dimension and Perspectives”. In: *The Lancet* 382.9888 (July 20, 2013), pp. 260–272. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(13)60687-X.

[2] Thaminda Liyanage et al. “Worldwide Access to Treatment for End-Stage Kidney Disease: A Systematic Review”. In: *The Lancet* 385.9981 (May 16, 2015), pp. 1975–1982. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(14)61601-9.

[3] Sudha P. Jayaraman and Zara R. Cooper. “Renal Replacement Therapy (RRT)”. In: *Geriatric Trauma and Critical Care*. Ed. by Jay A. Yelon and Fred A. Luchette. New York, NY: Springer, 2014, pp. 323–333. ISBN: 978-1-4614-8501-8. DOI: 10.1007/978-1-4614-8501-8_32.

[4] M. Tonelli et al. “Systematic Review: Kidney Transplantation Compared With Dialysis in Clinically Relevant Outcomes”. In: *American Journal of Transplantation* 11.10 (2011), pp. 2093–2109. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2011.03686.x.

[5] Geoffrey M. Fleming. “Renal Replacement Therapy Review: Past, Present and Future”. In: *Organogenesis* 7.1 (Jan. 2011), pp. 2–12. ISSN: 1547-6278, 1555-8592. DOI: 10.4161/org.7.1.13997.

[6] Marion Kerr. “Chronic Kidney Disease in England:” in: (), p. 44.

[7] Ebun Ladipo Bamgboye. “The Challenges of ESRD Care in Developing Economies: Sub-Saharan African Opportunities for Significant Improvement”. In: *Clinical Nephrology* 86.S1 (Dec. 1, 2016), pp. 18–22. ISSN: 0301-0430. DOI: 10.5414/CNP86S128.

[8] Vijay Kher. “End-Stage Renal Disease in Developing Countries”. In: *Kidney International* 62.1 (July 1, 2002), pp. 350–362. ISSN: 0085-2538. DOI: 10.1046/j.1523-1755.2002.00426.x.

[9] David Manlove. "How Operational Research Helps Kidney Patients in the UK". In: *Impact* 2018.1 (Jan. 2, 2018), pp. 16–19. ISSN: 2058-802X. DOI: 10.1080/2058802X.2018.1435455.

[10] Thakshayene Mahenthran. "The New 'Opt-out' Organ Donation English Law: Is the NHS Ready for This?" In: *Clinical Medicine* 21.1 (Jan. 1, 2021), e92–e93. ISSN: 1470-2118, 1473-4893. DOI: 10.7861/clinmed.2020-0675.

[11] Laura Pankhurst et al. "The UK National Registry of ABO and HLA Antibody Incompatible Renal Transplantation: Pretransplant Factors Associated With Outcome in 879 Transplants". In: *Transplantation Direct* 3.7 (July 2017), e181. DOI: 10.1097/TXD.0000000000000695.

[12] Babak J. Orandi et al. "Survival Benefit with Kidney Transplants from HLA-Incompatible Live Donors". In: *New England Journal of Medicine* 374.10 (Mar. 10, 2016), pp. 940–950. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1508380.

[13] James Gloor and Mark D. Stegall. "Sensitized Renal Transplant Recipients: Current Protocols and Future Directions". In: *Nature Reviews Nephrology* 6.5 (5 May 2010), pp. 297–306. ISSN: 1759-507X. DOI: 10.1038/nrneph.2010.34.

[14] Daniel Brennan. *HLA Matching and Graft Survival in Kidney Transplantation*. Dec. 2006.

[15] Ilias IN Doxiadis et al. "Association between Specific HLA Combinations and Probability of Kidney Allograft Loss: The Taboo Concept". In: *The Lancet* 348.9031 (Sept. 28, 1996), pp. 850–853. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(96)02296-9.

[16] Anna V. Reisæter et al. "A STRONG IMPACT OF MATCHING FOR A LIMITED NUMBER OF HLA-DR ANTIGENS ON GRAFT SURVIVAL AND REJECTION EPISODES: A Single-Center Study of First Cadaveric Kidneys to Nonsensitized Recipients: 1". In: *Transplantation* 66.4 (Aug. 27, 1998), pp. 523–528. ISSN: 0041-1337.

[17] Steve Takemoto et al. "HLA Matching for Kidney Transplantation". In: *Human Immunology* 65.12 (Dec. 1, 2004), pp. 1489–1505. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2004.06.008.

[18] Steven K. Takemoto et al. "Twelve Years' Experience with National Sharing of HLA-Matched Cadaveric Kidneys for Transplantation". In: *New England Jour-*

nal of Medicine 343.15 (Oct. 12, 2000), pp. 1078–1084. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJM200010123431504.

[19] A. Sharif, N. Alachkar, and E. Kraus. "Incompatible Kidney Transplantation: A Brief Overview of the Past, Present and Future". In: *QJM* 105.12 (Dec. 1, 2012), pp. 1141–1150. ISSN: 1460-2725, 1460-2393. DOI: 10.1093/qjmed/hcs154.

[20] Jennifer A. McCaughan and Kathryn J. Tinckam. "Donor Specific HLA Antibodies & Allograft Injury: Mechanisms, Methods of Detection, Manifestations and Management". In: *Transplant International* 31.10 (2018), pp. 1059–1070. ISSN: 1432-2277. DOI: 10.1111/tri.13324.

[21] Rob Higgins et al. "New Choices for Patients Needing Kidney Transplantation Across Antibody Barriers". In: *Journal of Renal Care* 34.2 (2008), pp. 85–93. ISSN: 1755-6686. DOI: 10.1111/j.1755-6686.2008.00025.x.

[22] T. E. Starzl et al. "Renal Homografts in Patients with Major Donor-Recipient Blood Group Incompatibilities". In: *Surgery* 55.2 (Feb. 1, 1964), pp. 195–200. ISSN: 0039-6060, 1532-7361. DOI: 10.5555/uri:pii:0039606064902764.

[23] Andrea A. Zachary et al. "Specific and Durable Elimination of Antibody to Donor HLA Antigens in Renal-Transplant Patients". In: *Transplantation* 76.10 (Nov. 27, 2003), pp. 1519–1525. ISSN: 0041-1337. DOI: 10.1097/01.TP.0000090868.88895.E0.

[24] M. Willicombe et al. "Outcome of Patients with Preformed Donor-Specific Antibodies Following Alemtuzumab Induction and Tacrolimus Monotherapy". In: *American Journal of Transplantation* 11.3 (2011), pp. 470–477. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2010.03421.x.

[25] Michelle Willicombe et al. *Guidelines for Antibody Incompatible Transplantation*. Dec. 2015. URL: https://bts.org.uk/wp-content/uploads/2016/09/02_BTS_Antibody_Guidelines-1.pdf (visited on 03/29/2022).

[26] Rob Higgins et al. "Rises and Falls in Donor-Specific and Third-Party HLA Antibody Levels After Antibody Incompatible Transplantation". In: *Transplantation* 87.6 (Mar. 27, 2009), p. 882. ISSN: 0041-1337. DOI: 10.1097/TP.0b013e31819a6788.

[27] Patrícia Soares de Souza et al. "Dynamics of Anti-Human Leukocyte Antigen Antibodies after Renal Transplantation and Their Impact on Graft Outcome". In:

Clinical Transplantation 28.11 (2014), pp. 1234–1243. ISSN: 1399-0012. DOI: 10.1111/ctr.12451.

[28] Paul Terasaki and Kazuo Mizutani. "Antibody Mediated Rejection: Update 2006". In: *Clinical Journal of the American Society of Nephrology* 1.3 (May 2006), pp. 400–403. ISSN: 1555-9041, 1555-905X. DOI: 10.2215/CJN.02311205.

[29] Rubin Zhang. "Donor-Specific Antibodies in Kidney Transplant Recipients". In: *Clinical Journal of the American Society of Nephrology* 13.1 (Jan. 6, 2018), pp. 182–192. ISSN: 1555-9041, 1555-905X. DOI: 10.2215/CJN.00700117.

[30] L. G. Hidalgo et al. "De Novo Donor-Specific Antibody at the Time of Kidney Transplant Biopsy Associates with Microvascular Pathology and Late Graft Failure". In: *American Journal of Transplantation* 9.11 (2009), pp. 2532–2541. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2009.02800.x.

[31] C. Wiebe et al. "Evolution and Clinical Pathologic Correlations of De Novo Donor-Specific HLA Antibody Post Kidney Transplant". In: *American Journal of Transplantation* 12.5 (2012), pp. 1157–1167. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2012.04013.x.

[32] Rob Higgins et al. "Blood Levels of Donor-Specific Human Leukocyte Antigen Antibodies After Renal Transplantation: Resolution of Rejection in the Presence of Circulating Donor-Specific Antibody". In: *Transplantation* 84.7 (Oct. 15, 2007), p. 876. ISSN: 0041-1337. DOI: 10.1097/01.tp.0000284729.39137.6e.

[33] Robert Higgins et al. "Human Leukocyte Antigen Antibody-Incompatible Renal Transplantation: Excellent Medium-Term Outcomes With Negative Cytotoxic Crossmatch". In: *Transplantation* 92.8 (Oct. 27, 2011), pp. 900–906. ISSN: 0041-1337. DOI: 10.1097/TP.0b013e31822dc38d.

[34] Yan Zhang et al. "A New Data-Driven Model for Post-Transplant Antibody Dynamics in High Risk Kidney Transplantation". In: *Mathematical Biosciences*. Novel Models, Analysis and Methods in Medical Systems 284 (Feb. 1, 2017), pp. 3–11. ISSN: 0025-5564. DOI: 10.1016/j.mbs.2016.04.008.

[35] Sunil Daga et al. "Direct Quantitative Measurement of the Kinetics of HLA-specific Antibody Interactions with Isolated HLA Proteins". In: *Human Immunology* 79.2 (Feb. 2018), pp. 122–128. ISSN: 1879-1166. DOI: 10.1016/j.humimm.2017.10.012.

[36] Po-Da Chen et al. “Gender Differences in Renal Transplant Graft Survival”. In: *Journal of the Formosan Medical Association* 112.12 (Dec. 1, 2013), pp. 783–788. ISSN: 0929-6646. DOI: 10.1016/j.jfma.2013.10.011.

[37] Natasha Khovanova et al. “Subclass Analysis of Donor HLA-specific IgG in Antibody-incompatible Renal Transplantation Reveals a Significant Association of IgG4 with Rejection and Graft Failure”. In: *Transplant International* 28.12 (Dec. 2015), pp. 1405–1415. ISSN: 0934-0874. DOI: 10.1111/tri.12648.

[38] Cecilia Montgomery Øien et al. “Living Donor Kidney Transplantation: The Effects of Donor Age and Gender on Short- and Long-Term Outcomes”. In: *Transplantation* 83.5 (Mar. 15, 2007), pp. 600–606. ISSN: 0041-1337. DOI: 10.1097/01.tp.0000255583.34329.dd.

[39] Torgyn Shaikhina et al. “Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation”. In: *Biomedical Signal Processing and Control* 52 (July 1, 2019), pp. 456–462. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2017.01.012.

[40] Lucile Amrouche et al. “Long-Term Outcomes of Kidney Transplantation in Patients With High Levels of Preformed DSA: The Necker High-Risk Transplant Program”. In: *Transplantation* 101.10 (Oct. 2017), pp. 2440–2448. ISSN: 0041-1337. DOI: 10.1097/TP.0000000000001650.

[41] Ashley A. Vo et al. “Clinical Relevance of Posttransplant DSAs in Patients Receiving Desensitization for HLA-incompatible Kidney Transplantation”. In: *Transplantation* 103.12 (Dec. 2019), pp. 2666–2674. ISSN: 0041-1337. DOI: 10.1097/TP.0000000000002691.

[42] D. Dadhania et al. “P002 The Consistent Inconsistency in HLA Antibody Testing Practices: A Collaboration with the Banff Workgroup”. In: *Human Immunology*. The American Society for Histocompatibility and Immunogenetics: 42nd Annual Meeting Abstracts September 26 – 30, 2016 Hyatt Regency St. Louis at the Arch, St. Louis, Missouri 77 (Sept. 1, 2016), p. 41. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2016.07.067.

[43] Hung Do Nguyen et al. *The Evolution of HLA-Matching in Kidney Transplantation*. IntechOpen, Feb. 13, 2013. ISBN: 978-953-51-0985-3. DOI: 10.5772/54747. URL: https://www.intechopen.com/chapters/42879 (visited on 07/29/2021).

[44] James Robinson et al. “IPD-IMGT/HLA Database”. In: *Nucleic Acids Research* 48.D1 (Jan. 8, 2020), pp. D948–D955. ISSN: 0305-1048. DOI: 10.1093/nar/gkz950.

[45] F. F. Gonzalez-Galarza et al. “16th IHIW: Extending the Number of Resources and Bioinformatics Analysis for the Investigation of HLA Rare Alleles”. In: *International Journal of Immunogenetics* 40.1 (2013), pp. 60–65. ISSN: 1744-313X. DOI: 10.1111/iji.12030.

[46] D. Middleton et al. “A Bioinformatics Approach to Ascertaining the Rarity of HLA Alleles”. In: *Tissue Antigens* 74.6 (2009), pp. 480–485. ISSN: 1399-0039. DOI: 10.1111/j.1399-0039.2009.01361.x.

[47] Gerald Bryan Pier, Jeffrey B. Lyczak, and Lee M. Wetzler. *Immunology, Infection, and Immunity*. ASM Press, 2004. ISBN: 978-1-55581-246-1. URL: https://agris.fao.org/agris-search/search.do?recordID=US201300097172 (visited on 08/16/2021).

[48] Ramon Patel and Paul I. Terasaki. *Significance of the Positive Crossmatch Test in Kidney Transplantation*. http://dx.doi.org/10.1056/NEJM196904032801401. Jan. 14, 2010. DOI: 10.1056/NEJM196904032801401. URL: https://www.nejm.org/doi/10.1056/NEJM196904032801401 (visited on 07/21/2021).

[49] Yuji Mishima et al. “An Imaging-Based Rapid Evaluation Method for Complement-Dependent Cytotoxicity Discriminated Clinical Response to Rituximab-Containing Chemotherapy”. In: *Clinical Cancer Research* 15.10 (May 15, 2009), pp. 3624–3632. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1078-0432.CCR-08-1536.

[50] D. J. Cook et al. “The Flow Cytometry Crossmatch in Kidney Transplantation”. In: *Clinical transplants* (Jan. 1, 1987), pp. 409–414. ISSN: 0890-9016.

[51] Jennifer McCaughan, Qingyong Xu, and Kathryn Tinckam. “Detecting Donor-Specific Antibodies: The Importance of Sorting the Wheat from the Chaff”. In: *Hepatobiliary Surgery and Nutrition* 8.1 (1 Feb. 2019), pp. 372–352. ISSN: 2304-389X, 2304-3881.

[52] Rui Pei et al. “Flow Cytometric Detection of HLA Antibodies Using a Spectrum of Microbeads”. In: *Human Immunology* 60.12 (Dec. 1, 1999), pp. 1293–1302. ISSN: 0198-8859. DOI: 10.1016/S0198-8859(99)00121-4.

[53] Suchitra Sumitran-Karuppan. “THE CLINICAL IMPORTANCE OF CHOOSING THE RIGHT ASSAY FOR DETECTION OF HLA-SPECIFIC DONOR-

REACTIVE ANTIBODIES1". In: *Transplantation* 68.4 (Aug. 27, 1999), pp. 502–509. ISSN: 0041-1337.

[54] E. F. Reed et al. "Comprehensive Assessment and Standardization of Solid Phase Multiplex-Bead Arrays for the Detection of Antibodies to HLA". In: *American Journal of Transplantation* 13.7 (2013), pp. 1859–1870. ISSN: 1600-6143. DOI: 10.1111/ajt.12287.

[55] Andrea A. Zachary et al. "Naturally Occurring Interference in Luminex® Assays for HLA–Specific Antibodies: Characteristics and Resolution". In: *Human Immunology* 70.7 (July 1, 2009), pp. 496–501. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2009.04.001.

[56] Vasilis Kosmoliaptsis et al. "High-Resolution, Three-Dimensional Modeling of Human Leukocyte Antigen Class I Structure and Surface Electrostatic Potential Reveals the Molecular Basis for Alloantibody Binding Epitopes". In: *Human Immunology* 72.11 (Nov. 1, 2011), pp. 1049–1059. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2011.07.303.

[57] C. Weinstock and M. Schnaidt. "The Complement-Mediated Prozone Effect in the Luminex Single-Antigen Bead Assay and Its Impact on HLA Antibody Determination in Patient Sera". In: *International Journal of Immunogenetics* 40.3 (June 2013), pp. 171–177. ISSN: 17443121. DOI: 10.1111/j.1744-313X.2012.01147.x.

[58] Yue Gu et al. "Defining the Structural Basis for Human Leukocyte Antigen Reactivity in Clinical Transplantation". In: *Scientific Reports* 10.1 (Dec. 2020), p. 18397. ISSN: 2045-2322. DOI: 10.1038/s41598-020-75355-4.

[59] Thomas M. Ellis. "Interpretation of HLA Single Antigen Bead Assays". In: *Transplantation Reviews* 27.4 (Oct. 1, 2013), pp. 108–111. ISSN: 0955-470X. DOI: 10.1016/j.trre.2013.07.001.

[60] D. Lowe et al. "CLASS I HLA-SPECIFIC ANTIBODY QUANTIFICATION". In: 24 (Sept. 30, 2011), pp. 353–353.

[61] D LOWE. "The High-Dose Hook Effect in the Detection and Monitoring of HLA Specific Antibody by Luminex Assay". In: *Int J Immunogenetics* 34 (2007), p. 288.

[62] Antoine Roux et al. "Characteristics of Donor-Specific Antibodies Associated With Antibody-Mediated Rejection in Lung Transplantation". In: *Frontiers in*

Medicine 4 (Oct. 11, 2017), p. 155. ISSN: 2296-858X. DOI: 10.3389/fmed.2017.00155.

[63] Julie M. Yabu. "Should We Use Immunodominant DSA to Measure Response to Therapy in Antibody-mediated Rejection?" In: *Transplantation* 104.11 (Nov. 2020), pp. 2260–2261. ISSN: 0041-1337. DOI: 10.1097/TP.0000000000003146.

[64] Maria Cristina Ribeiro de Castro et al. "The Kinetics of Anti-HLA Antibodies in the First Year after Kidney Transplantation: In Whom and When Should They Be Monitored?" In: *Journal of Transplantation* 2018 (Apr. 23, 2018). ISSN: 2090-0007. DOI: 10.1155/2018/8316860.

[65] R. Carlin Walsh et al. "Early and Late Acute Antibody-Mediated Rejection Differ Immunologically and in Response to Proteasome Inhibition". In: *Transplantation* 91.11 (June 15, 2011), pp. 1218–1226. ISSN: 0041-1337. DOI: 10.1097/TP.0b013e318218e901.

[66] Scott Davis and James E. Cooper. "Acute Antibody-Mediated Rejection in Kidney Transplant Recipients". In: *Transplantation Reviews* 31.1 (Jan. 1, 2017), pp. 47–54. ISSN: 0955-470X. DOI: 10.1016/j.trre.2016.10.004.

[67] Ruchi H. Naik and Saed H. Shawar. "Renal Transplantation Rejection". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2021.

[68] P. F. Halloran. "T Cell-Mediated Rejection of Kidney Transplants: A Personal Viewpoint". In: *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 10.5 (May 2010), pp. 1126–1134. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2010.03053.x.

[69] G. Melville Williams et al. "Hyperacute Renal-Homograft Rejection in Man". In: *New England Journal of Medicine* 279.12 (Sept. 19, 1968), pp. 611–618. ISSN: 0028-4793. DOI: 10.1056/NEJM196809192791201.

[70] Kim Solez and Lorraine C. Racusen. "The Banff Classification Revisited". In: *Kidney International* 83.2 (Feb. 1, 2013), pp. 201–206. ISSN: 0085-2538. DOI: 10.1038/ki.2012.395.

[71] Ritesh Singh and Keshab Mukhopadhyay. "Survival Analysis in Clinical Trials: Basics and Must Know Areas". In: *Perspectives in Clinical Research* 2.4 (2011), pp. 145–148. ISSN: 2229-3485. DOI: 10.4103/2229-3485.86872.

[72] Shankar Prinja, Nidhi Gupta, and Ramesh Verma. "Censoring in Clinical Trials: Review of Survival Analysis Techniques". In: *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine* 35.2 (Apr. 2010), pp. 217–221. ISSN: 1998-3581. DOI: 10.4103/0970-0218.66859.

[73] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (June 1, 1958), pp. 457–481. ISSN: 0162-1459. DOI: 10.1080/01621459.1958.10501452.

[74] Danilo Bzdok, Denis Engemann, and Bertrand Thirion. "Inference and Prediction Diverge in Biomedicine". In: *Patterns* 1.8 (Nov. 13, 2020), p. 100119. ISSN: 2666-3899. DOI: 10.1016/j.patter.2020.100119.

[75] Tibor V. Varga et al. "Association Is Not Prediction: A Landscape of Confused Reporting in Diabetes – A Systematic Review". In: *Diabetes Research and Clinical Practice* 170 (Dec. 1, 2020), p. 108497. ISSN: 0168-8227. DOI: 10.1016/j.diabres.2020.108497.

[76] Shi Zhong and Joydeep Ghosh. "A Unified Framework for Model-based Clustering". In: *Journal of Machine Learning Research* (Nov. 2003), pp. 1001–1037.

[77] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. New York: Routledge, Jan. 31, 2019. 532 pp. ISBN: 978-0-203-75373-6. DOI: 10.1201/9780203753736.

[78] N. Breslow. "Covariance Analysis of Censored Survival Data". In: *Biometrics* 30.1 (1974), pp. 89–99. ISSN: 0006-341X. DOI: 10.2307/2529620.

[79] David A. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, Apr. 27, 2009. 459 pp. ISBN: 978-1-139-47731-4.

[80] James Warner. "Clinicians' Guide to Evaluating Diagnostic and Screening Tests in Psychiatry". In: *Advances in Psychiatric Treatment* 10.6 (Nov. 2004), pp. 446–454. ISSN: 1355-5146, 1472-1481. DOI: 10.1192/apt.10.6.446.

[81] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, June 9, 2004. 326 pp. ISBN: 978-0-471-65574-9.

[82] Lorenzo Beretta and Alessandro Santaniello. "Nearest Neighbor Imputation Algorithms: A Critical Evaluation". In: *BMC Medical Informatics and Decision Making* 16.3 (July 25, 2016), p. 74. ISSN: 1472-6947. DOI: 10.1186/s12911-016-0318-z.

[83] Ewout W. Steyerberg. "Coding of Categorical and Continuous Predictors". In: *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Ed. by Ewout W. Steyerberg. Statistics for Biology and Health. Cham: Springer International Publishing, 2019, pp. 175–190. ISBN: 978-3-030-16399-0. DOI: 10.1007/978-3-030-16399-0_9.

[84] Manoranjan Pal and Premananda Bharati. "The Regression Models with Dummy Explanatory Variables". In: *Applications of Regression Techniques*. Ed. by Manoranjan Pal and Premananda Bharati. Singapore: Springer, 2019, pp. 135–153. ISBN: 9789811393143. DOI: 10.1007/978-981-13-9314-3_8.

[85] Alicia B. Lichvar et al. "Reducing Donor-specific Antibody During Acute Rejection Diminishes Long-term Renal Allograft Loss: Comparison of Early and Late Rejection". In: *Transplantation* 104.11 (Nov. 2020), pp. 2403–2414. ISSN: 0041-1337. DOI: 10.1097/TP.0000000000003145.

[86] DAVID SCHOENFELD. "Partial Residuals for the Proportional Hazards Regression Model". In: *Biometrika* 69.1 (Apr. 1, 1982), pp. 239–241. ISSN: 0006-3444. DOI: 10.1093/biomet/69.1.239.

[87] William S. Cleveland. "Robust Locally Weighted Regression and Smoothing Scatterplots". In: *Journal of the American Statistical Association* 74.368 (Dec. 1, 1979), pp. 829–836. ISSN: 0162-1459. DOI: 10.1080/01621459.1979.10481038.

[88] PATRICIA M. GRAMBSCH and TERRY M. THERNEAU. "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals". In: *Biometrika* 81.3 (Sept. 1, 1994), pp. 515–526. ISSN: 0006-3444. DOI: 10.1093/biomet/81.3.515.

[89] Terry M. Therneau et al. *Survival: Survival Analysis*. Version 3.2-13. Aug. 24, 2021. URL: https://CRAN.R-project.org/package=survival (visited on 08/27/2021).

[90] C L Loprinzi et al. "Prospective Evaluation of Prognostic Variables from Patient-Completed Questionnaires. North Central Cancer Treatment Group." In: *Journal of Clinical Oncology* 12.3 (Mar. 1, 1994), pp. 601–607. ISSN: 0732-183X. DOI: 10.1200/JCO.1994.12.3.601.

[91] Kevin M. Veen et al. "A Clinician's Guide for Developing a Prediction Model: A Case Study Using Real-World Data of Patients with Castration-Resistant Prostate Cancer". In: *Journal of Cancer Research and Clinical Oncology* 146.8 (Aug.

2020), pp. 2067–2075. ISSN: 0171-5216, 1432-1335. DOI: 10.1007/s00432-020-03286-8.

[92] Kwanghee Jung et al. "Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation". In: *Frontiers in Psychology* 10 (2019). ISSN: 1664-1078.

[93] Bradley Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. 38. Philadelphia, Pa: Society for Industrial and Applied Mathematics, 1982. 92 pp. ISBN: 978-0-89871-179-0.

[94] Priya Ranganathan, C. S. Pramesh, and Rakesh Aggarwal. "Common Pitfalls in Statistical Analysis: Logistic Regression". In: *Perspectives in Clinical Research* 8.3 (2017), pp. 148–151. ISSN: 2229-3485. DOI: 10.4103/picr.PICR_87_17.

[95] Dominique Lord, Xiao Qin, and Srinivas R. Geedipally. "Chapter 5 - Exploratory Analyses of Safety Data". In: *Highway Safety Analytics and Modeling*. Ed. by Dominique Lord, Xiao Qin, and Srinivas R. Geedipally. Elsevier, Jan. 1, 2021, pp. 135–177. ISBN: 978-0-12-816818-9. DOI: 10.1016/B978-0-12-816818-9.00015-9.

[96] NJ Gogtay and UM Thatte. "Principles of Correlation Analysis". In: (), p. 4.

[97] Georg Heinze, Christine Wallisch, and Daniela Dunkler. "Variable Selection – A Review and Recommendations for the Practicing Statistician". In: *Biometrical Journal* 60.3 (2018), pp. 431–449. ISSN: 1521-4036. DOI: 10.1002/bimj.201700067.

[98] Frank E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer International Publishing, 2015. ISBN: 978-3-319-19424-0 978-3-319-19425-7. DOI: 10.1007/978-3-319-19425-7. URL: http://link.springer.com/10.1007/978-3-319-19425-7 (visited on 03/29/2022).

[99] Zoran Bursac et al. "Purposeful Selection of Variables in Logistic Regression". In: *Source Code for Biology and Medicine* 3 (Dec. 16, 2008), p. 17. ISSN: 1751-0473. DOI: 10.1186/1751-0473-3-17.

[100] Trevor Hastie, Robert Tibshirani, and Ryan J. Tibshirani. "Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso". July 29, 2017. URL: http://arxiv.org/abs/1707.08692 (visited on 03/29/2022).

[101] Frank E. Harrell Jr et al. “Evaluating the Yield of Medical Tests”. In: *JAMA* 247.18 (May 14, 1982), pp. 2543–2546. ISSN: 0098-7484. DOI: 10.1001/jama.1982.03320430047030.

[102] Willi Sauerbrei et al. “On Stability Issues in Deriving Multivariable Regression Models”. In: *Biometrical Journal* 57.4 (2015), pp. 531–555. ISSN: 1521-4036. DOI: 10.1002/bimj.201300222.

[103] Douglas G. Altman and Per Kragh Andersen. “Bootstrap Investigation of the Stability of a Cox Regression Model”. In: *Statistics in Medicine* 8.7 (1989), pp. 771–783. ISSN: 1097-0258. DOI: 10.1002/sim.4780080702.

[104] Chen-Hsin Chen and Stephen L. George. “The Bootstrap and Identification of Prognostic Factors via Cox’s Proportional Hazards Regression Model”. In: *Statistics in Medicine* 4.1 (1985), pp. 39–46. ISSN: 1097-0258. DOI: 10.1002/sim.4780040107.

[105] Riccardo De Bin et al. “Subsampling versus Bootstrapping in Resampling-Based Model Selection for Multivariable Regression”. In: *Biometrics* 72.1 (2016), pp. 272–280. ISSN: 1541-0420. DOI: 10.1111/biom.12381.

[106] J. M. Gloor et al. “Baseline Donor-Specific Antibody Levels and Outcomes in Positive Crossmatch Kidney Transplantation”. In: *American Journal of Transplantation* 10.3 (2010), pp. 582–589. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2009.02985.x.

[107] Herwig-Ulf Meier-Kriesche et al. “Effect of Waiting Time on Renal Transplant Outcome”. In: *Kidney International* 58.3 (Sept. 1, 2000), pp. 1311–1317. ISSN: 0085-2538. DOI: 10.1046/j.1523-1755.2000.00287.x.

[108] Tanya Pankhurst et al. “Young Adults Have Worse Kidney Transplant Outcomes than Other Age Groups”. In: *Nephrology Dialysis Transplantation* 35.6 (June 1, 2020), pp. 1043–1051. ISSN: 0931-0509, 1460-2385. DOI: 10.1093/ndt/gfaa059.

[109] Carmen Lefaucheur et al. “Preexisting Donor-Specific HLA Antibodies Predict Outcome in Kidney Transplantation”. In: *Journal of the American Society of Nephrology* 21.8 (Aug. 1, 2010), pp. 1398–1406. ISSN: 1046-6673, 1533-3450. DOI: 10.1681/ASN.2009101065.

[110] Aleksandar Senev et al. “Specificity, Strength, and Evolution of Pretransplant Donor-Specific HLA Antibodies Determine Outcome after Kidney Transplan-

tation". In: *American Journal of Transplantation* 19.11 (2019), pp. 3100–3113. ISSN: 1600-6143. DOI: 10.1111/ajt.15414.

[111] Igor Salvadé et al. "Clinically-Relevant Threshold of Preformed Donor-Specific Anti-HLA Antibodies in Kidney Transplantation". In: *Human Immunology* 77.6 (June 1, 2016), pp. 483–489. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2016.04.010.

[112] Malte Ziemann et al. "Preformed Donor-Specific HLA Antibodies in Living and Deceased Donor Transplantation: A Multicenter Study". In: *Clinical Journal of the American Society of Nephrology* 14.7 (July 5, 2019), pp. 1056–1066. ISSN: 1555-9041, 1555-905X. DOI: 10.2215/CJN.13401118.

[113] J. M. Burns et al. "Alloantibody Levels and Acute Humoral Rejection Early After Positive Crossmatch Kidney Transplantation". In: *American Journal of Transplantation* 8.12 (2008), pp. 2684–2694. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2008.02441.x.

[114] Andrea A. Zachary, Robert A. Montgomery, and Mary S. Leffell. "Factors Associated With and Predictive of Persistence of Donor-Specific Antibody After Treatment With Plasmapheresis and Intravenous Immunoglobulin". In: *Human Immunology* 66.4 (Apr. 1, 2005), pp. 364–370. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2005.01.032.

[115] Hans-Hermann Bock. "Clustering Methods: A History of k-Means Algorithms". In: *Selected Contributions in Data Analysis and Classification*. Ed. by Paula Brito et al. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer, 2007, pp. 161–172. ISBN: 978-3-540-73560-1. DOI: 10.1007/978-3-540-73560-1_15.

[116] R. R. Sokal and P. H. A. Sneath. "Principles of numerical taxonomy." In: *Principles of numerical taxonomy.* (1963).

[117] Anil K. Jain. *Data Clustering: 50 Years Beyond K-Means*. 2008.

[118] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-Series Clustering – A Decade Review". In: *Information Systems* 53 (Oct. 1, 2015), pp. 16–38. ISSN: 0306-4379. DOI: 10.1016/j.is.2015.04.007.

[119] T. Warren Liao. "Clustering of Time Series Data—a Survey". In: *Pattern Recognition* 38.11 (Nov. 1, 2005), pp. 1857–1874. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2005.01.025.

[120] Jessica Lin et al. “Experiencing SAX: A Novel Symbolic Representation of Time Series”. In: *Data Mining and Knowledge Discovery* 15.2 (Oct. 1, 2007), pp. 107–144. ISSN: 1573-756X. DOI: 10.1007/s10618-007-0064-z.

[121] Arindam Banerjee and Hanhuai Shan. “Model-Based Clustering”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 686–689. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_554.

[122] Esma Ergüner Özkoç. *Clustering of Time-Series Data*. IntechOpen, Feb. 3, 2020. ISBN: 978-1-83968-319-0. DOI: 10.5772/intechopen.84490. URL: https://www.intechopen.com/chapters/65772 (visited on 08/02/2021).

[123] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. “WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases”. In: (1998), p. 12.

[124] Alexis Sardá-Espinosa. “Comparing Time-Series Clustering Algorithms in R Using the Dtwclust Package”. In: (2017), p. 46.

[125] Lucien Marie Le Cam and Jerzy Neyman. “Some Methods for Classification and Analysis of Multivariate Observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather Modification*. University of California Press, 1967, pp. 281–297.

[126] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990. 369 pp. ISBN: 978-0-470-31748-8.

[127] Anil K Jain and Richard Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988. URL: https://dl.acm.org/doi/abs/10.5555/46712 (visited on 08/01/2021).

[128] L. Jegatha Deborah, R. Baskaran, and A. Kannan. “A.Kannan,“Survey on Internal Validity Measure for Cluster Validation”. In: *International Journal of Computer Science and Engineering Survey (IJCSES) Vol.1, No.2*. 2010.

[129] Mark Harman, Stephen Swift, and Kiarash Mahdavi. “An Empirical Study of the Robustness of Two Module Clustering Fitness Functions”. In: *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation - GECCO ’05*. The 2005 Conference. Washington DC, USA: ACM Press, 2005, p. 1029. ISBN: 978-

1-59593-010-1. DOI: 10.1145/1068009.1068184. URL: http://portal.acm.org/citation.cfm?doid=1068009.1068184 (visited on 03/29/2022).

[130] Peter J. Rousseeuw. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1, 1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

[131] T. Caliński and J Harabasz. "A Dendrite Method for Cluster Analysis". In: *Communications in Statistics* 3.1 (Jan. 1, 1974), pp. 1–27. ISSN: 0090-3272. DOI: 10.1080/03610927408827101.

[132] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (May 2001), pp. 411–423. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/1467-9868.00293.

[133] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. "A Stability Based Method for Discovering Structure in Clustered Data". In: *Biocomputing 2002*. WORLD SCIENTIFIC, Dec. 1, 2001, pp. 6–17. ISBN: 978-981-02-4777-5. DOI: 10.1142/9789812799623_0002.

[134] Tilman Lange et al. "Stability-Based Validation of Clustering Solutions". In: *Neural Computation* 16.6 (June 1, 2004), pp. 1299–1323. ISSN: 0899-7667. DOI: 10.1162/089976604773717621.

[135] Luxor von Ulrike. *Clustering Stability: An Overview*. Now Foundations and Trends. URL: https://ieeexplore.ieee.org/abstract/document/8187106/ (visited on 08/01/2021).

[136] Jason Lines and Anthony Bagnall. "Time Series Classification with Ensembles of Elastic Distance Measures". In: *Data Mining and Knowledge Discovery* 29.3 (May 2015), pp. 565–592. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-014-0361-2.

[137] Young-Seon Jeong, Myong K. Jeong, and Olufemi A. Omitaomu. "Weighted Dynamic Time Warping for Time Series Classification". In: *Pattern Recognition*. Computer Analysis of Images and Patterns 44.9 (Sept. 1, 2011), pp. 2231–2240. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2010.09.022.

[138] Chotirat Ratanamahatana and Eamonn J. Keogh. “Making Time-Series Classification More Accurate Using Learned Constraints.” In: Apr. 22, 2004. DOI: 10.1137/1.9781611972740.2.

[139] Vladimir Kurbalija et al. “The Influence of Global Constraints on DTW and LCS Similarity Measures for Time-Series Databases”. In: *Third International Conference on Software, Services and Semantic Technologies S3T 2011*. Ed. by Darina Dicheva, Zdravko Markov, and Eliza Stefanova. Advances in Intelligent and Soft Computing. Berlin, Heidelberg: Springer, 2011, pp. 67–74. ISBN: 978-3-642-23163-6. DOI: 10.1007/978-3-642-23163-6_10.

[140] “Dynamic Time Warping”. In: *Information Retrieval for Music and Motion*. Ed. by Meinard Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 69–84. ISBN: 978-3-540-74048-3. DOI: 10.1007/978-3-540-74048-3_4.

[141] Z. Geler et al. “Dynamic Time Warping: Itakura vs Sakoe-Chiba”. In: *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*. 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). July 2019, pp. 1–6. DOI: 10.1109/INISTA.2019.8778300.

[142] Vit Niennattrakul and Chotirat Ann Ratanamahatana. “Learning DTW Global Constraint for Time Series Classification”. Feb. 28, 2009. URL: http://arxiv.org/abs/0903.0041 (visited on 07/23/2021).

[143] Q. Chen et al. “Learning Optimal Warping Window Size of DTW for Time Series Classification”. In: *2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA)*. 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA). July 2012, pp. 1272–1277. DOI: 10.1109/ISSPA.2012.6310488.

[144] Martin Ester and Dino Pedreschi, eds. *Proceedings of the 2018 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics, May 7, 2018. ISBN: 978-1-61197-532-1. DOI: 10.1137/1.9781611975321. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611975321 (visited on 07/23/2021).

[145] Hoang Anh Dau et al. “Optimizing Dynamic Time Warping’s Window Width for Time Series Data Mining Applications”. In: *Data Mining and Knowledge*

Discovery 32.4 (July 2018), pp. 1074–1120. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-018-0565-y.

[146] C A Ratanamahatana and E Keogh. "Everything You Know about Dynamic Time Warping Is Wrong". In: *Third Workshop on Mining Temporal and Sequential Data*. Vol. 32. 2004.

[147] Chotirat Ann Ratanamahatana and Eamonn Keogh. "Three Myths about Dynamic Time Warping Data Mining". In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. Proceedings of the 2005 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Apr. 21, 2005, pp. 506–510. ISBN: 978-0-89871-593-4 978-1-61197-275-7. DOI: 10.1137/1.9781611972757.50. URL: https://epubs.siam.org/doi/10.1137/1.9781611972757.50 (visited on 07/23/2021).

[148] Alexis Sardá-Espinosa. "Comparing Time-Series Clustering Algorithms in R Using the Dtwclust Package". In: (), p. 45.

[149] W. T. Williams, H. T. Clifford, and G. N. Lance. "Group-Size Dependence: A Rationale for Choice between Numerical Classifications". In: *The Computer Journal* 14.2 (Jan. 1, 1971), pp. 157–162. ISSN: 0010-4620. DOI: 10.1093/comjnl/14.2.157.

[150] Joe H. Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (Mar. 1, 1963), pp. 236–244. ISSN: 0162-1459. DOI: 10.1080/01621459.1963.10500845.

[151] Martin Maechler et al. "Cluster: Cluster Analysis Basics and Extensions (2019)". In: *R package version 2.0.6* (2017).

[152] Leo Breiman et al. *Classification And Regression Trees*. Boca Raton: Routledge, Oct. 25, 2017. 368 pp. ISBN: 978-1-315-13947-0. DOI: 10.1201/9781315139470.

[153] Erel Levine and Eytan Domany. "Resampling Method for Unsupervised Estimation of Cluster Validity". In: *Neural Computation* 13.11 (Nov. 1, 2001), pp. 2573–2593. ISSN: 0899-7667. DOI: 10.1162/089976601753196030.

[154] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. "A Stability Based Method for Discovering Structure in Clustered Data". In: *Biocomputing 2002*. Proceedings of the Pacific Symposium. Kauai, Hawaii, USA: WORLD SCIENTIFIC, Dec. 2001, pp. 6–17. ISBN: 978-981-02-4777-5 978-981-279-962-3. DOI: 10.

1142/9789812799623_0002. URL: http://www.worldscientific.com/doi/abs/10.1142/9789812799623_0002 (visited on 03/24/2021).

[155] M. Bittner et al. "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling". In: *Nature* 406.6795 (6795 Aug. 2000), pp. 536–540. ISSN: 1476-4687. DOI: 10.1038/35020115.

[156] U. Moller and D. Radke. "A Cluster Validity Approach Based on Nearest-Neighbor Resampling". In: *18th International Conference on Pattern Recognition (ICPR'06)*. 18th International Conference on Pattern Recognition (ICPR'06). Vol. 1. Aug. 2006, pp. 892–895. DOI: 10.1109/ICPR.2006.42.

[157] M. Kathleen Kerr and Gary A. Churchill. "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments". In: *Proceedings of the National Academy of Sciences* 98.16 (July 31, 2001), pp. 8961–8965. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.161273698.

[158] Zheng zheng Xing, Jian Pei, and Philip Yu. "Early Classification on Time Series". In: *Knowledge and Information Systems - KAIS* 31 (Apr. 1, 2011). DOI: 10.1007/s10115-011-0400-x.

[159] G. S. Hongyi Li and Maddala. "Bootstrapping Time Series Models". In: *Econometric Reviews* 15.2 (Jan. 1, 1996), pp. 115–158. ISSN: 0747-4938. DOI: 10.1080/07474939608800344.

[160] Cynthia Garcia-Sanchez et al. "The Shared Epitope Phenomenon—A Potential Impediment to Virtual Crossmatch Accuracy". In: *Clinical Transplantation* 34.8 (2020), e13906. ISSN: 1399-0012. DOI: 10.1111/ctr.13906.

[161] Marzyeh Ghassemi et al. "A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1 (1 Feb. 10, 2015). ISSN: 2374-3468.

[162] Saeid Eslami et al. "Tight Glycemic Control and Computerized Decision-Support Systems: A Systematic Review". In: *Intensive Care Medicine* 35.9 (Sept. 1, 2009), pp. 1505–1517. ISSN: 1432-1238. DOI: 10.1007/s00134-009-1542-0.

[163] Kensaku Kawamoto et al. "Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success". In: *BMJ* 330.7494 (Mar. 31, 2005), p. 765. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.38398.500764.8F.

[164] David Collett. *Modelling Survival Data in Medical Research*. CRC Press, May 4, 2015. 538 pp. ISBN: 978-1-4987-3169-0.

[165] Lewis B. Sheiner, Barr Rosenberg, and Vinay V. Marathe. "Estimation of Population Characteristics of Pharmacokinetic Parameters from Routine Clinical Data". In: *Journal of Pharmacokinetics and Biopharmaceutics* 5.5 (Oct. 1, 1977), pp. 445–479. ISSN: 0090-466X. DOI: 10.1007/BF01061728.

[166] Nikolaos Tsamandouras, Amin Rostami-Hodjegan, and Leon Aarons. "Combining the 'Bottom up' and 'Top down' Approaches in Pharmacokinetic Modelling: Fitting PBPK Models to Observed Clinical Data". In: *British Journal of Clinical Pharmacology* 79.1 (2015), pp. 48–55. ISSN: 1365-2125. DOI: 10.1111/bcp.12234.

[167] Joachim Sturmberg and Carmel Martin. *Handbook of Systems and Complexity in Health*. New York: Springer, 2013.

[168] Johan Paduart et al. "Identification of Nonlinear Systems Using Polynomial Nonlinear State Space Models". In: *Automatica* 46.4 (Apr. 1, 2010), pp. 647–656. ISSN: 0005-1098. DOI: 10.1016/j.automatica.2010.01.001.

[169] J. Schoukens and R. Pintelon. *Identification of Linear Systems: A Practical Guideline to Accurate Modeling*. Elsevier, June 28, 2014. 355 pp. ISBN: 978-0-08-091256-1.

[170] Dean Rickles, Penelope Hawe, and Alan Shiell. "A Simple Guide to Chaos and Complexity". In: *Journal of Epidemiology & Community Health* 61.11 (Nov. 1, 2007), pp. 933–937. ISSN: 0143-005X, 1470-2738. DOI: 10.1136/jech.2006.054254.

[171] Andrew Gelman, Frederic Bois, and Jiming Jiang. "Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions". In: *Journal of the American Statistical Association* 91.436 (Dec. 1, 1996), pp. 1400–1412. ISSN: 0162-1459. DOI: 10.1080/01621459.1996.10476708.

[172] Jerry J. Batzel, Mostafa Bachar, and Franz Kappel. *Mathematical Modeling and Validation in Physiology: Applications to the Cardiovascular and Respiratory Systems*. Springer, Dec. 14, 2012. 270 pp. ISBN: 978-3-642-32882-4.

[173] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2004. 390 pp. ISBN: 978-0-521-52902-0.

[174] Ervin Sejdić and Lewis A. Lipsitz. “Necessity of Noise in Physiology and Medicine”. In: *Computer Methods and Programs in Biomedicine* 111.2 (Aug. 1, 2013), pp. 459–470. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2013.03.014.

[175] Tim Salimans, Diederik P. Kingma, and Max Welling. “Markov Chain Monte Carlo and Variational Inference: Bridging the Gap”. May 19, 2015. URL: http://arxiv.org/abs/1410.6460 (visited on 08/18/2021).

[176] Manuel M. Eichenlaub. “Mathematical Modelling of Blood Glucose Dynamics in Health and Type 2 Diabetes”. Warwick, May 2020. 200 pp.

[177] Gabriella Margaria et al. “Differential Algebra Methods for the Study of the Structural Identifiability of Rational Function State-Space Models in the Biosciences”. In: *Mathematical Biosciences* 174.1 (Nov. 1, 2001), pp. 1–26. ISSN: 0025-5564. DOI: 10.1016/S0025-5564(01)00079-7.

[178] J. Daunizeau, K. J. Friston, and S. J. Kiebel. “Variational Bayesian Identification and Prediction of Stochastic Nonlinear Dynamic Causal Models”. In: *Physica D: Nonlinear Phenomena* 238.21 (Nov. 1, 2009), pp. 2089–2118. ISSN: 0167-2789. DOI: 10.1016/j.physd.2009.08.002.

[179] Dimitris G. Tzikas, Aristidis C. Likas, and Nikolaos P. Galatsanos. “The Variational Approximation for Bayesian Inference”. In: *IEEE Signal Processing Magazine* 25.6 (Nov. 2008), pp. 131–146. ISSN: 1558-0792. DOI: 10.1109/MSP.2008.929620.

[180] Yongli Zhang and Yuhong Yang. “Cross-Validation for Selecting a Model Selection Procedure”. In: *Journal of Econometrics* 187.1 (July 2015), pp. 95–112. ISSN: 03044076. DOI: 10.1016/j.jeconom.2015.02.006.

[181] S. Hengl et al. “Data-Based Identifiability Analysis of Non-Linear Dynamical Models”. In: *Bioinformatics* 23.19 (Oct. 1, 2007), pp. 2612–2618. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btm382.

[182] Keegan E. Hines, Thomas R. Middendorf, and Richard W. Aldrich. “Determination of Parameter Identifiability in Nonlinear Biophysical Models: A Bayesian Approach”. In: *Journal of General Physiology* 143.3 (Mar. 1, 2014), pp. 401–416. ISSN: 1540-7748, 0022-1295. DOI: 10.1085/jgp.201311116.

[183] W. D. Penny. “Comparing Dynamic Causal Models Using AIC, BIC and Free Energy”. In: *NeuroImage*. Neuroergonomics: The Human Brain in Action and at

Work 59.1 (Jan. 2, 2012), pp. 319–330. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.07.039.

[184] G Cawley and N Talbot. *Fast Exact Leave-One-out Cross-Validation of Sparse Least-Squares Support Vector Machines - ScienceDirect*. URL: https://www.sciencedirect.com/science/article/pii/S0893608004001431 (visited on 12/31/2021).

[185] *Impulse Function - an Overview (Pdf) | ScienceDirect Topics*. URL: https://www.sciencedirect.com/topics/engineering/impulse-function/pdf (visited on 01/09/2022).

[186] L. L. Langan et al. "Post-Transplant HLA Class II Antibodies and High Soluble CD30 Levels Are Independently Associated with Poor Kidney Graft Survival". In: *American Journal of Transplantation* 7.4 (2007), pp. 847–856. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2006.01691.x.

[187] Roslyn B. Mannon. "Immune Monitoring and Biomarkers to Predict Chronic Allograft Dysfunction". In: *Kidney International*. Chronic Renal Allograft Dysfunction (CAD): New Frontiers 78 (Dec. 1, 2010), S59–S65. ISSN: 0085-2538. DOI: 10.1038/ki.2010.425.

[188] Brian D. Tait et al. "Consensus Guidelines on the Testing and Clinical Management Issues Associated With HLA and Non-HLA Antibodies in Transplantation". In: *Transplantation* 95.1 (Jan. 15, 2013), pp. 19–47. ISSN: 0041-1337. DOI: 10.1097/TP.0b013e31827a19cc.

[189] Denis Viglietti et al. "Value of Donor–Specific Anti–HLA Antibody Monitoring and Characterization for Risk Stratification of Kidney Allograft Loss". In: *Journal of the American Society of Nephrology* 28.2 (Feb. 1, 2017), pp. 702–715. ISSN: 1046-6673, 1533-3450. DOI: 10.1681/ASN.2016030368.

[190] Paolo Cravedi and Peter S. Heeger. "Immunologic Monitoring in Transplantation Revisited". In: *Current Opinion in Organ Transplantation* 17.1 (Feb. 2012), pp. 26–32. ISSN: 1087-2418. DOI: 10.1097/MOT.0b013e32834ee402.

[191] Alexandre Loupy, Gary S. Hill, and Stanley C. Jordan. "The Impact of Donor-Specific Anti-HLA Antibodies on Late Kidney Allograft Failure". In: *Nature Reviews Nephrology* 8.6 (6 June 2012), pp. 348–357. ISSN: 1759-507X. DOI: 10.1038/nrneph.2012.81.

[192] P. I. Terasaki, M. Ozawa, and R. Castro. "Four-Year Follow-up of a Prospective Trial of HLA and MICA Antibodies on Kidney Graft Survival". In: *American Journal of Transplantation* 7.2 (2007), pp. 408–415. ISSN: 1600-6143. DOI: 10.1111/j.1600-6143.2006.01644.x.

[193] Marc Lúcia et al. "Preformed Circulating HLA-specific Memory B Cells Predict High Risk of Humoral Rejection in Kidney Transplantation". In: *Kidney International* 88.4 (Oct. 1, 2015), pp. 874–887. ISSN: 0085-2538. DOI: 10.1038/ki.2015.205.

[194] Rafi Ahmed and David Gray. "Immunological Memory and Protective Immunity: Understanding Their Relation". In: *Science* 272.5258 (Apr. 5, 1996), pp. 54–60. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.272.5258.54.

[195] Rob Higgins et al. "Pregnancy-Induced HLA Antibodies Respond More Vigorously after Renal Transplantation than Antibodies Induced by Prior Transplantation". In: *Human Immunology* 76.8 (Aug. 1, 2015), pp. 546–552. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2015.06.013.

[196] Mason Phillpott et al. "Dynamic Behaviour of Donor Specific Antibodies in the Early Period Following HLA Incompatible Kidney Transplantation". In: *Transplant International* (Mar. 2022), p. 11. DOI: 10.3389/ti.2022.10128.

[197] Yan Zhang. "Variational Bayesian Data Driven Modelling for Biomedical Systems." PhD thesis. University of Warwick, Dec. 2016. URL: http://webcat.warwick.ac.uk/record=b3067335~S15 (visited on 10/18/2018).

[198] Chang Liu, Donna Phelan, and Thalachallour Mohanakumar. "Inter-Run Variation of Luminex Single-Antigen Assay for the Detection of Antibodies to HLA". In: *Human Immunology*. The American Society of Histocompatibility and Immunogenetics: 41st Annual Meeting Abstracts, September 28 - October 2, 2015, Savannah, Georgia 76 (Oct. 1, 2015), p. 45. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2015.07.064.

[199] Jean Daunizeau, Adam Vincent, and Rigoux Lionel. *VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data*. 2014. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003441 (visited on 10/30/2019).

[200] Manuel M. Eichenlaub. "On the Relationship between a Gamma Distributed Precision Parameter and the Associated Standard Deviation in the Context of Bayesian

Parameter Inference". Jan. 15, 2021. URL: http://arxiv.org/abs/2101.06289 (visited on 02/12/2021).

Appendix A

Additional theory

A.1 Kaplan-Meier survival function

A clear definition of the Kaplan-Meier survival curve can be given as "the probability of surviving in a given length of time while considering time in many small intervals". There are three assumptions to consider:

1. Patients who are censored have the same survival prospects as those who continue.
2. Survival probabilities are consistent for patients who have joined at any point in the study.
3. The event occurs at a specific time.

This may present a limitation when events are only determined at regular examinations. The survival probability for any point is calculated by the formula,

$$s_t = \frac{N_{\text{living at start}} - N_{\text{died}}}{N_{\text{living at start}}} \tag{A.1.0.1}$$

At each time interval the survival probability is calculated as the number of patients surviving divided by those at risk. At risk patients are those who are still in the study at the time interval of interest and do not consider those who have dropped out or whose follow-up time is shorter. Patients with lower follow-up time are considered as censored data from their last recorded update and not counted in the denominator. The product limit method (Kaplan and Meier 1958) is used to estimate S:

$$\widehat{S_t} = \prod_{t_i \leq t} \left[1 - \frac{d_i}{n_i} \right] \tag{A.1.0.2}$$

Where t_i is the duration of the study at point i, d_i is the number of deaths up to point i and n_i is the number of individuals at risk just prior to t_i. S is considered the probability of an individual surviving to the end of a time interval (given that the individual was present at the start of the time interval). The variance of S is estimated using the Greenwood (1926) method,

$$var\left(\widehat{S}_t\right) = \widehat{S}_t^{\,2} \sum_{t_i \leq t} \frac{d_i}{n_i\left(n_i - d_i\right)} \tag{A.1.0.3}$$

However, at the extremes of S this yields unrealistic confidence intervals (<0 and >1) so the confidence interval applies an asymptotic maximum likelihood solution by log transformation (Kalbfeisch and Prentice 1980).

Appendix B

Application of VBA toolbox

Model specification and ODE solution

When working in the context of the VB toolbox an ODE system will be represented in terms of its state space formula. The state space formula consists of a set of first order differential equations relating each of the state space variables which, when considered together, describe the various aspects of the system dynamics. The state space formula is denoted by,

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{F}(\mathbf{x}(t), \mathbf{p}, \mathbf{u}(t), t), \tag{B.0.0.1}$$

with system initial conditions,

$$\mathbf{x}(0) = \mathbf{x_0}. \tag{B.0.0.2}$$

In the context of Equations B.0.0.1 and B.0.0.2 bold letters represent vectors and matrices. Given n system states, $\mathbf{F} = \{F_1, ..., F_n\}$ is the set of formula which relate system state variables $\mathbf{x} = \{x_1, ..., x_n\}$ to $\dot{\mathbf{x}} = \{\dot{x}_1, ..., \dot{x}_n\}$ at time t. Parameter vectors $\mathbf{p}$ and $\mathbf{u}(t)$ correspond to the constant and time-dependent inputs o fthe system. For each system state an associated initial condition $\mathbf{x_0} = x_{01}, ..., x_{0n}$ exists at $\mathbf{x}(0)$.

During VB model inversion process, as previously described in Section 5.3.4, the ODEs defining the system have to be solved such that the current model estimation $y(t)$ can be compared to the observed data y. To achieve this the user must specify an update rule for calculating the next value of the state space vector x_{t+1}, from the current value x_t, within a fixed and pre-defined time step size Δt. This process allows for approximation of a

continuous state space model solution from discrete time steps and is performed using the Euler discretisation scheme [199],

$$x_{t+1} = x_t + \boldsymbol{F}(x_t, p, u_t)\Delta t \qquad x_{t=0} = x_0 \tag{B.0.0.3}$$

Within the toolbox only the mean of the model parameters, p, is passed to the user as an estimate of the associated PDF. A set of vector inputs, u_t, can also be specified at a given time t, allowing for the inclusion of time dependent model events. Time itself is considered a model input and is included in this manner. Through this approach an adequate approximation of model state trajectories can be computed if the time step Δt is at resolution smaller than the fastest decay of the system [199]. For all models explored in this chapter a step size of 0.1 days was selected and shown to be sufficiently small enough to approximate the true model solution while not compromising substantially in computational processing time.

In order to update the model parameters a Gauss-Newton optimisation scheme is implemented, as indicated in Section 5.3.4, whereby it is necessary to calculate the trajectory of the model sensitivity, i.e., the derivative of the model states relative to the unknown parameters, p, and initial conditions, x_0. Mathematical details on this procedure is provided below [176]. The equations for continuous time trajectories of the model sensitivity are,

$$S_p(t) = \frac{dx(t)}{dp} \quad \text{and} \quad S_{x_0}(t) = \frac{dx(t)}{dx_0}, \tag{B.0.0.4}$$

where $S_{x_0}(t)$ and $S_p(t)$ are the sensitivities of the model states relative to the initial conditions and parameters respectively. After application of the multi-variate chain rule to the state space representation of the system (shown in Equation B.0.0.1) the following is derived,

$$\frac{dS_p(t)}{dt} = \frac{d}{dt}\left(\frac{dx(t)}{dp}\right) = \frac{\partial \boldsymbol{F}}{\partial p} + \frac{\partial \boldsymbol{F}}{\partial x(t)}\frac{dx(t)}{dp}, \tag{B.0.0.5}$$

$$\frac{dS_{x_0}(t)}{dt} = \frac{d}{dt}\left(\frac{dx(t)}{dx_0}\right) = \frac{\partial \boldsymbol{F}}{\partial x(t)}\frac{dx(t)}{dx_0} \tag{B.0.0.6}$$

where $\frac{dx(t)}{dp}|_{t=0} = 0$ and $\frac{dx(t)}{dx_0}|_{t=0} = \boldsymbol{I}$. Symbols $\boldsymbol{F}$ and $\boldsymbol{I}$ are defined by the model equations and identity matrix respectively. At this stage a user defined input is provided to the toolbox for the partial derivatives of the model equations $\boldsymbol{F}$ relative to the model states $x(t)$ and parameters p,

$$\frac{\partial \boldsymbol{F}}{\partial x(t)} = \boldsymbol{J}(t) \qquad \text{and} \qquad \frac{\partial \boldsymbol{F}}{\partial p} = \boldsymbol{H}(t), \tag{B.0.0.7}$$

which, for the models investigated in this thesis, are analytically calculated by the computational software Wolfram Mathematica (Wolfram Research, Inc., Champaign, IL, USA) - an example is shown in Appendix B.1. Alternatively these solutions can also be approximated within the toolbox itself at the cost of additional computational time. Next by accounting for expressions B.0.0.4 and B.0.0.7, the equations B.0.0.5-B.0.0.6 can be reformulated as,

$$\frac{dS_p(t)}{dt} = \boldsymbol{H}(t) + S_p(t)\boldsymbol{J}(t) \tag{B.0.0.8}$$

$$\frac{dS_{x_0}(t)}{dt} = S_{x_0}(t)\boldsymbol{J}(t) \tag{B.0.0.9}$$

where $S_p(0) = 0$ and $S_{x_0}(0) = \boldsymbol{I}$. These equations are solved along with the ODE itself in discrete time using Euler's method,

$$S_{p_{t+1}} = [\boldsymbol{H} + S_{p_t}\boldsymbol{J}]\Delta t + S_{p_t} \tag{B.0.0.10}$$

$$S_{x_{0t+1}} = [S_{x_{0t}}\boldsymbol{J}]\Delta t + S_{x_{0t}} \tag{B.0.0.11}$$

where $S_{p_{t=0}} = 0$ and $S_{x_{0t=0}} = \boldsymbol{I}$. At each discrete time step the terms S_{p_t} and $S_{x_0 t}$ are $m \times n$ and $n \times n$ dimensional matrices respectively. Final implementation into the VB update cycle requires manipulation into the following format,

$$S_{p_{t+1}} = [\Delta t \boldsymbol{J}_t + \boldsymbol{I}] S_{p_t} + \Delta t \boldsymbol{H}_t \tag{B.0.0.12}$$

$$S_{x_{0t+1}} = [\Delta t \boldsymbol{J}_t + \boldsymbol{I}] S_{x_{0t}}. \tag{B.0.0.13}$$

B.0.1 Model prior values

Previously it was discussed that the VB method assumes certain parametric distributions over the unknown parameters - in particular the Gaussian and Gamma distributions. Practically it was beneficial to utilise different distribution types for model parameters prior to model estimation. Two such parametric transforms are applied: restriction to positive values and the transform of noise parameters.

Restriction to positive values

By assuming that a model parameter p has a PDF specified via Gaussian density f_p, with mean μ and standard deviation σ, a restriction to positive values can be achieved through the introduction of a substitute parameter s and then applying an exponential mapping whereby $s = exp(p)$. To calculate the density f_s the following theorem is used. If f_x is the PDF of a random variable x, and $y = h(x)$ is introduced as a mapping function, then the PDF over y is,

$$f_y(y) = f_x(h^{-1}(y)) \left| \frac{dh^{-1(y)}}{dy} \right|. \tag{B.0.1.1}$$

Which when considering the density f_s yields,

$$f_s(s|\mu,\sigma) = \frac{1}{s\sqrt{2\pi}\sigma} \exp\left(-\frac{(\log s - \mu)^2}{2\sigma^2} \right). \tag{B.0.1.2}$$

This expression is recognised as the log-normal distribution. This transform finds application in promoting stability of model solution. Further details pertaining to its implementation within the VB toolbox can be found in [176].

Transformation of noise parameters

The precision, $\kappa = 1/\sigma^2$, of noise parameter measurement uncertainties within the VB toolbox are defined via a Gamma PDF and are specified via shape and rate parameters a and b,

$$f_\kappa(\kappa|a,b) = \frac{b^a}{\Gamma(a)}\kappa^{a-1}\exp(-\kappa b) \qquad \text{for} \qquad \kappa,a,b,>0, \tag{B.0.1.3}$$

where $\Gamma(\cdot)$ is the Gamma function. The mean and variance of the precision PDF is,

$$\mathbb{E}[\kappa]_{f_\kappa} = \frac{a}{b} \qquad \text{and} \qquad \text{Var}[\kappa]_{f_\kappa} = \frac{a}{b^2}. \tag{B.0.1.4}$$

In this chapter the measurement process is modelled by the expression given in Equation B.0.0.1, where the measurement noise is characterised by the standard deviation λ. Defining measurement noise by λ instead of κ is beneficial in modelling. Firstly, it allows for characterising of uncertainty based off of the known inter-assay CV of luminex assays - which can be used to deduce λ. Secondly, it offers a more practical interpretation of κ. Transforming between $\mathcal{N}(\mu,\sigma^2)$ and $\Gamma(a,b)$ can be treated as a constrained numerical optimisation task, the details of which are included in [176, 200]. An official implementation of the function is included in the VB toolbox and validated under the following conditions:

- $0.005 < \mu_0 < 10^4$
- $0.005\mu_0 < \sigma_0 < 5\mu_0$,

where the values in this range are shown to be recoverable to within a 1% error. As such, this implementation will be used in this work to assist in defining noise uncertainty.

Model	$M1^{4}$	$M2^{1,2,3}$	$M2^{4}$	$M3^{1,2,3}$	$M2_{p}^{1,2,3}$	$M3_{p}^{1,2,3}$
X_{01}	$\mathcal{N}(\overline{y_{(1:3)}}, \overline{y_{(1:3)}}/10)$	$\mathcal{N}(\overline{y_{(1:3)}}, \overline{y_{(1:3)}}/10)$	$\mathcal{N}(\overline{y_{(1:7)}}, \overline{y_{(1:7)}}/10)$	$\mathcal{N}(\overline{y_{(1:3)}}, \overline{y_{(1:3)}}/10)$	$\mathcal{N}(\overline{y_{(1:3)}}, \overline{y_{(1:3)}}/10)$	$\mathcal{N}(\overline{y_{(1:3)}}, \overline{y_{(1:3)}}/10)$
X_{02}	-	$\mathcal{N}(0,0)$	$\mathcal{N}(0,0)$	$\mathcal{N}(0,0)$	$\mathcal{N}(0,0)$	$\mathcal{N}(0,0)$
X_{03}	-	-	-	$\mathcal{N}(0,0)$	-	$\mathcal{N}(0,0)$
θ_0	$\ln\mathcal{N}(y_{end}/2, 2)$	$\ln\mathcal{N}(y_{end}, 2)$	$\ln\mathcal{N}(y_{end}/2, 2)$	$\ln\mathcal{N}(y_{end}, 2)$	$\ln\mathcal{N}(y_{end}, 2)$	$\ln\mathcal{N}(y_{end}, 2)$
θ_1	$\ln\mathcal{N}(0.1, 2)$	$\ln\mathcal{N}(0.1, 2)$	$\ln\mathcal{N}(0.1, 2)$	$\ln\mathcal{N}(0.1, 2)$	$\ln\mathcal{N}(0.1, 2)$	$\ln\mathcal{N}(0.1, 2)$
θ_2	-	$\ln\mathcal{N}(0.5, 2)$	$\ln\mathcal{N}(10, 2)$	$\ln\mathcal{N}(0.1, 2)$	$\ln\mathcal{N}(0.5, 2)$	$\ln\mathcal{N}(0.1, 2)$
θ_3	-	-	-	$\ln\mathcal{N}(0.5, 2)$	-	$\ln\mathcal{N}(0.5, 2)$
θ_{f1}	-	-	-	-	$\ln\mathcal{N}(1e3, 2)$	$\ln\mathcal{N}(1e3, 2)$
θ_{f2}	-	-	-	-	$\ln\mathcal{N}(\phi, 2)$	$\ln\mathcal{N}(\phi, 2)$
ε	$\mathcal{G}(*,*)$	$\mathcal{G}(*,*)$	$\mathcal{G}(*,*)$	$\mathcal{G}(*,*)$	$\mathcal{G}(*,*)$	$\mathcal{G}(*,*)$

Table B.1: Chapter 4 model prior values and inputs for VB toolbox. Values are specified for each particular model and DSA group that the model is applied for. The impulse response delay variable, ϕ, represents a value between 2-8 selected independently for each response. Variables which are defined via the log-normal distribution, $\ln\mathcal{N}(\mu, \sigma^2)$, are initially defined as a normal distribution within the VB toolbox before undergoing the transform outlined in Section B.0.1. DSA coefficient of variation at 15% was used to establish a prior value for measurement uncertainty, this was defined via a normal distribution $\mathcal{N}(CV \cdot \bar{y}, CV \cdot \bar{y} \cdot 0.2)$, before conversion into the Gamma distribution, $\mathcal{G}(*,*)$, via methodology established in Section B.0.1.

B.1 Structural identifiability

An example of the Taylor series expansion to check for structural identifiability within a system i.e., to check if it is possible to determine the values of a systems' parameters from measurements of its model outputs. The approach is illustrated for the following nonlinear 2nd order system,

$$\ddot{x} = -(\theta_1 + \theta_n x)\dot{x} - \theta_2 x + \theta_2\theta_3 \tag{B.1.0.1}$$

with known initial conditions $x_0 = 0$ and $\dot{x}_0 = F$. Substitution of the initial conditions into Equation B.1.0.1 yields,

$$\ddot{x}_0 = -(\theta_1 + \theta_n x_0)\dot{x}_0 - \theta_2 x_0 + \theta_2\theta_3 = -\theta_1 F + \theta_2\theta_3 \tag{B.1.0.2}$$

Here we can compare the derivative terms determined through both the parameter vector θ and $\bar{\theta}$,

$$\theta_1 F + \theta_2\theta_3 = \bar{\theta}_1 F + \bar{\theta}_2\bar{\theta}_3 \tag{B.1.0.3}$$

Further equations can be identified through subsequent derivations of the system,

$$\dddot{x}_0 = -(\theta_1 + \theta_n x_0)\ddot{x}_0 - \theta_n\dot{x}_0^2 - \theta_2\dot{x}_0 = F\left(\theta_1^2 - \theta_2 - \theta_n F\right) - \theta_1\theta_2\theta_3 \tag{B.1.0.4}$$

Which yields,

$$F\left(\theta_1^2 - \theta_2 - \theta_n F\right) - \theta_1\theta_2\theta_3 = F\left(\bar{\theta}_1^2 - \bar{\theta}_2 - \bar{\theta}_n F\right) - \bar{\theta}_1\bar{\theta}_2\bar{\theta}_3 \tag{B.1.0.5}$$

As previously stated in Chapter 5 Section 5.4.2, $n+m+1$ defines the upper bound of Taylor coefficients terms required to determine the systems identifiability. In this case there are two states and four parameters thus 2+4+1 =7 Taylor coefficient terms are required and a further three derivations $x_0^{(4)}$, $x_0^{(5)}$ and $x_0^{(6)}$ should be calculated to solve the system. In practice solving these large sets of simultaneous equations can be time consuming and prone to human error and as such several computational solutions are used. In this thesis

the symbolic software package Mathematica (Wolfram Research, Inc., Champaign, IL, USA) is used, allowing for solutions to be extracted simply. Through these means the example system can be shown to be uniquely identifiable with the following solutions extracted from the simultaneous equations.

$$\theta_1 = \bar{\theta}_1, \quad \theta_2 = \bar{\theta}_2, \quad \theta_3 = \bar{\theta}_3, \quad \theta_n = \bar{\theta}_n, \tag{B.1.0.6}$$

B.1.1 Structural identifiability example task

To reiterate the example, Equation B.1.1.1 highlights an instance of a unidentifiable system, given that q_1 and q_2 are unknown. This is because an infinite number of parameter combinations can yield the same output.

$$\dot{x} = q_1 q_2^2 x, \;\; x(0) = x_0, \;\; y(t) = x(t). \tag{B.1.1.1}$$

To validate for structural identifiability in investigated candidate models the Taylor series approach was first trialled and implemented into Wolfram Mathematica software (Wolfram Research, Inc., Champaign, IL, USA). In some models however this approach was not able to determine a result, such as for models $M2_p$ and $M3_p$. For these results a MATLAB (The MathWorks, Inc., Natick, MA, USA) software package known as STRIKE_GOLDD (version 3.0, created by Alejandro F. Villaverde, afvillaverde@uvigo.gal) was used that is capable of analysing nonlinear ODE models for parameter structural local identifiability. Both solutions will be demonstrated on the example task.

B.1.2 Wolfram Mathematica solution

Structural identifiability analysis was first conducted analytically through use of the Taylor series implemented into Wolfram Mathematica software. A coded example is provided below following the example outlined in Chapter 5. The problem established in Mathematica by defining the model equation, initial conditions and output with,

```
(* 1st order sample problem *)
x'[t] = -(p1^2*p2)*x[t];
x[0]=x0;
y[t] = x[t];
```

Next, the Taylor series expansion is defined to expand around time `t0` up to `Nt` coefficients. Additionally, the vectors of unknown substitution parameters, `subst` is defined,

```
(* Define model conditions *)
Nt =5;
t0 = 0;
subst={p1->p1b, p2->p2b};
```

Lastly, this piece of code determines the Taylor coefficients before subsequently creating and solving for the system of simultaneous equations,

```
(* Calculates *)
Cof = {{y[t]/.t->t0}};
For[i=2,i<=Nt,i++,Cof=Append[Cof,{D[y[t],{t,i-1}]/.t->t0}]]
Cof
(* eqn = eqn with substitutes *)
Eqns = {Cof[[1]]==(Cof[[1]]/.subst)};
For[i=2,i<=Nt,i++,Eqns=Append[Eqns,Cof[[i]]==(Cof[[i]]/.subst)]];
Solve[Eqns ,{p1b,p2b}]
```

The result of this script is documented below showing that the parameters are not identifiable,

```
(* Output *)
Out[556]= {{x0},{-p1^2 p2 x0},{p1^4 p2^2 x0},{-p1^6 p2^3 x0},{p1^8
   p2^4 x0}}
Out[559]= {{p2b->(p1^2 p2)/p1b^2}}
```

Furthermore, if q_1 was known then the system could be described as locally identifiable – as two possible solutions exist, i.e., $-q_2$ and q_2 both yielding the same output.

B.1.3 STRIKE_GOLDD implementation

A similar approach is used when implementing the exmaple problem in the STRIKE_GOLDD MATLAB software package. First the system states, parameters and output are defined symbolically within MATLAB,

```
% 2 states
syms x1 x2; x = [x1];
% 1 output
h = x1;
% 2 parameters
```

```
syms p1 p2; p =[p1; p2;];
```

Next the initial conditions of the system are defined along with the model equation in state space form,

```
% initial conditions
syms x10; ics  = [x10;]; known_ics = [1];

% dynamic equations
f = [-((p1^2)*p2)*x1];
```

The script, once executed saves the system into a '.mat' file,

```
% save model
save('Example','x','p','h','f','ics');
```

Once saved the 'STRIKE_GOLDD.m' file is activated and produces the results summary for the above defined system (once linking to the appropriate 'Example.m' file in options). The results of the function are indicated below,

```
 ---------------------------------
 >>> STRIKE-GOLDD toolbox 3.0
 ---------------------------------
 Analyzing the Example model...
 >>> The model contains:
 1 states:
 x1
 1 outputs:
 x1
 0 known inputs:
 0 unknown inputs:
 2 parameters:
 [p1; p2]
 >>> Building the observability-identifiability matrix requires at
   least 2 Lie derivatives
     Calculating derivatives: 1 2
 >>> Observability-Identifiability matrix built with 2 Lie
   derivatives
     (calculated in 2.374540e-02 seconds)
 >>> Calculating rank...
     Rank = 2 (calculated in 6.087200e-03 seconds)
```

```
>>> Observability-Identifiability matrix built with 3 Lie
  derivatives
    (calculated in 3.746700e-02 seconds)
>>> Calculating rank...
    Rank = 2 (calculated in 4.522500e-03 seconds)
   The model is structurally unidentifiable as a whole
   => Parameter p1 is structurally unidentifiable
   => Parameter p2 is structurally unidentifiable
------------------------
>>> RESULTS SUMMARY:
------------------------
>>> The model is structurally unidentifiable.
>>> These parameters are identifiable:
>>> These parameters are unidentifiable:
     [p1, p2]
>>> These states are directly measured:
     x1
Total execution time: 2.224679e-01
```

Similarly to the Wolfram alpha approach this system has been defined as structurally unidentifiable along with parameters `p1` and `p2`.

B.2 Candidate models

In Chapter 5 there were five candidate models considered for analysis of the four DSA dynamic response groups. Each of these models are listed in Table B.2. In this section each model is assessed for its structural identifiability using

Model name	Model	Initial conditions
$M1$	$\dot{x}+x\theta_1-\theta_1\theta_0=0$	$x_0=x_{10}$
$M2$	$\ddot{x}+\dot{x}\theta_2+x\theta_1-\theta_1\theta_0=0$	$\dot{x}_0=x_{20},\quad x_0=x_{10}$
$M3$	$\dddot{x}+\ddot{x}\theta_3+\dot{x}\theta_2+x\theta_1-\theta_1\theta_0=0$	$\ddot{x}_0=x_{30},\quad \dot{x}_0=x_{20},\quad x_0=x_{10}$
$M2_p$	$\ddot{x}+\dot{x}\theta_2+x\theta_1-\theta_1\theta_0=f_p(t)$	$\dot{x}_0=x_{20},\quad x_0=x_{10}$
$M3_p$	$\dddot{x}+\ddot{x}\theta_3+\dot{x}\theta_2+x\theta_1-\theta_1\theta_0=f_p(t)$	$\ddot{x}_0=x_{30},\quad \dot{x}_0=x_{20},\quad x_0=x_{10}$

Table B.2: A list of the five candidate models list used in Chapter 5 For each of the DSA model selection tasks the listed models were considered. The external force $f_p(t)$ represents the delayed Gaussian impulse, $\frac{\theta_{p1}}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(t-\theta_{p2})^2}{2\sigma^2}\right)$, where $\sigma^2=0.2$ is fixed.

B.2.1 Model stability

To determine model stability, as described in Chapter 5 Section 5.4.5, there are two criteria: all of the real parts of the system eigenvalues must be negative, and there cannot be any sign changes in the first column of the Routh table.

Validating eigenvalues

Calculating eigenvalues can be achieved by first representing a higher order system in terms of a set of linear state space models shown here in matrix notation,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \tag{B.2.1.1}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \tag{B.2.1.2}$$

where $\mathbf{x} = \{x_1, \cdots, x_n\}$ are the states, $\mathbf{u} = \{u_1, \cdots, u_m\}$ the model inputs, and $\mathbf{y} = \{y_1, \cdots, y_r\}$ the model outputs. These may then be re-written in the Laplace domain as,

$$s\mathbf{X}(s) = \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s), \tag{B.2.1.3}$$

$$\mathbf{Y}(s) = \mathbf{C}\mathbf{X}(s) + \mathbf{D}\mathbf{U}(s). \tag{B.2.1.4}$$

The state equation may then be re-written to form,

$$s\mathbf{X}(s) - \mathbf{A}\mathbf{X}(s) = [s\mathbf{I} - \mathbf{A}]\,\mathbf{X}(s) = \mathbf{B}\mathbf{U}(s). \tag{B.2.1.5}$$

where $\mathbf{I}$ is the identity matrix. The eigenvalues of the system may then be found by taking the determinant of $[s\mathbf{I} - \mathbf{A}]$ and calculating the roots of the characteristic polynomial,

$$\det[s\mathbf{I} - \mathbf{A}] = 0. \tag{B.2.1.6}$$

Practically this is conducted within MATLAB using the eig() function. The results of this process are illustrated in Table B.1. For all instances models demonstrate stability.

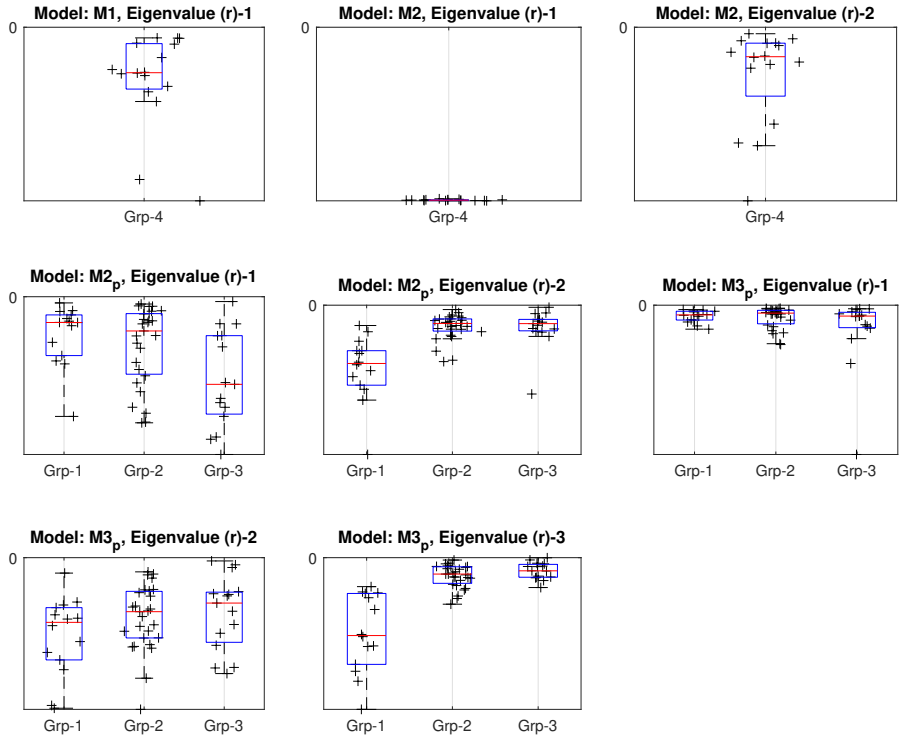


Figure B.1: Eigenvalues assessment of model stability. For each of the investigated candidate models $M1^4$, $M2^4$, $M2_p^{1,2,3}$, and $M3_p^{1,2,3}$ their eigenvalues are calculated and the real parts of those solutions shown. In all cases models are demonstrated as stable due to negative real parts to the eigenvalues.

Validating Routh criterion

Table B.3 shows the results of the Routh tables for the candidate models indicated in Table B.2. The criterion for stability is then deduced for model $M1$: $\theta_1 > 0$, for models $M2$, $M2_p$: $\theta_1 > 0$, $\theta_1\theta_2 > 0$, and models $M3$, $M3_p$: $\theta_1 > 0$, $\theta_2 > \theta_1/\theta_3$, and $\theta_3 > 0$. In all instances it can be seen that models satisfy the required conditions for model stability.

s	θ_1
1	1

(a) $M1$.

s^2	1	θ_2
s	θ_1	0
1	$\theta_1\theta_2$	

(b) $M2$, $M2_p$.

s^3	1	θ_2
s^2	θ_3	θ_1
s	$\theta_2 - \theta_1/\theta_3$	
1	θ_1	

(c) $M3$, $M3_p$.

Table B.3: Routh-Hurwitz stability criterion. Routh tables are displayed for the five candidate models presented in this work are detailed in Table B.2.

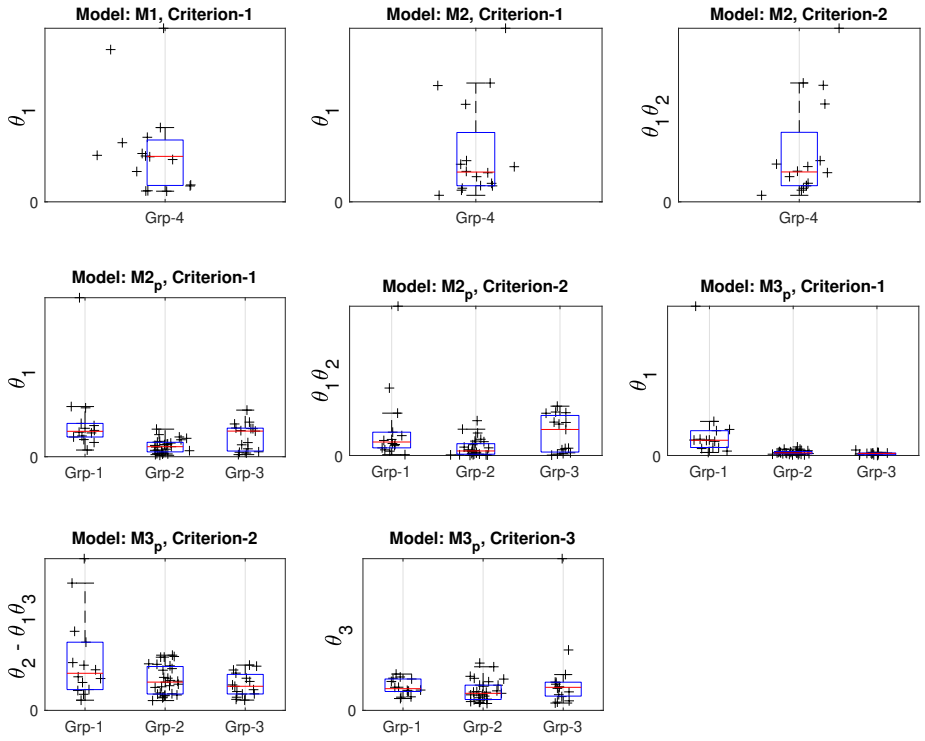


Figure B.2: Results for Routh table stability requirements. For each of the investigated candidate models $M1^4$, $M2^4$, $M2_p^{1,2,3}$, and $M3_p^{1,2,3}$ the results of the Routh stability requirements are displayed. Given negative real components to their respective eigenvalue solutions, the illustrated criterion must additionally be positive in order for solution to be stable. In all cases solutions are demonstrated to be stable.

Appendix C

Additional results

C.0.1 Results for three cluster groups

In Figure C.2 the three DSA response groups identified via the DTW agglomerative hierarchical clustering approach are shown. Figure C.2 shows the mean response and standard deviation for each group.

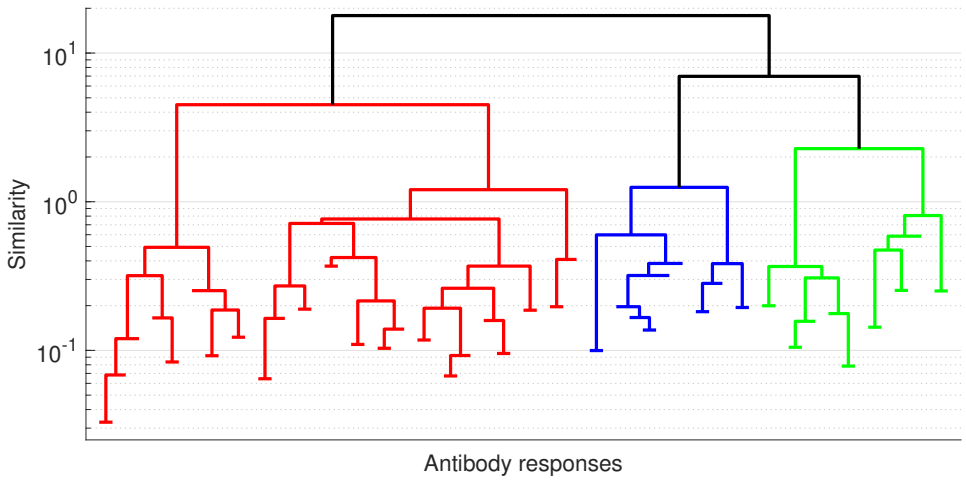


Figure C.1: Agglomerative hierarchical structure of DSA cohort formed using DTW distance measure ($k = 3$ groups). The three clusters are highlighted by red (Group-A), blue (Group-B), and green (Group-C) colour code. Similarity is shown on a log-scale. The end of each branch designates a DSA response.

C.0.2 Results for five cluster groups

In Figure C.3 the five DSA response groups identified via the DTW agglomerative hierarchical clustering approach are shown. Figure C.4 shows the mean response and standard deviation for each group.

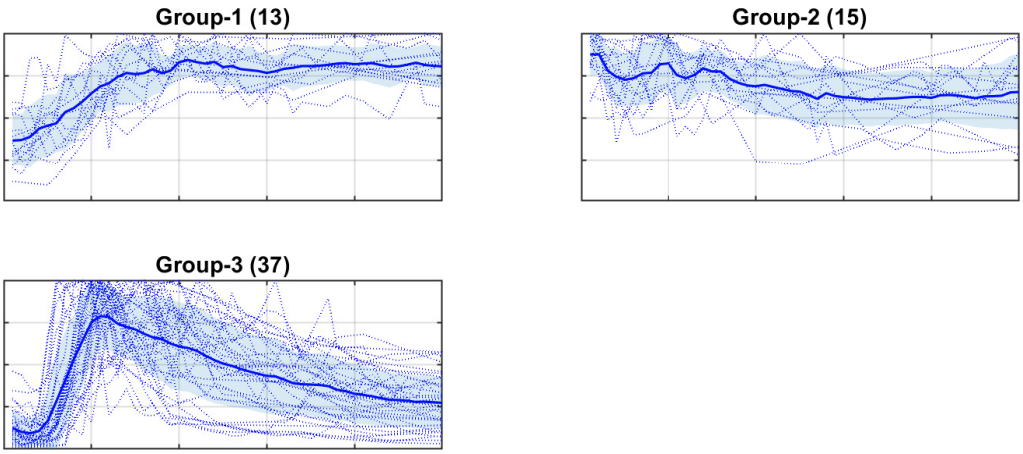


Figure C.2: Each of the three identified response clusters are shown (Groups A-C). Dark blue line represents cluster mean response and shaded region the cluster standard deviation. Dotted lines represent each individual response belonging to a given class.

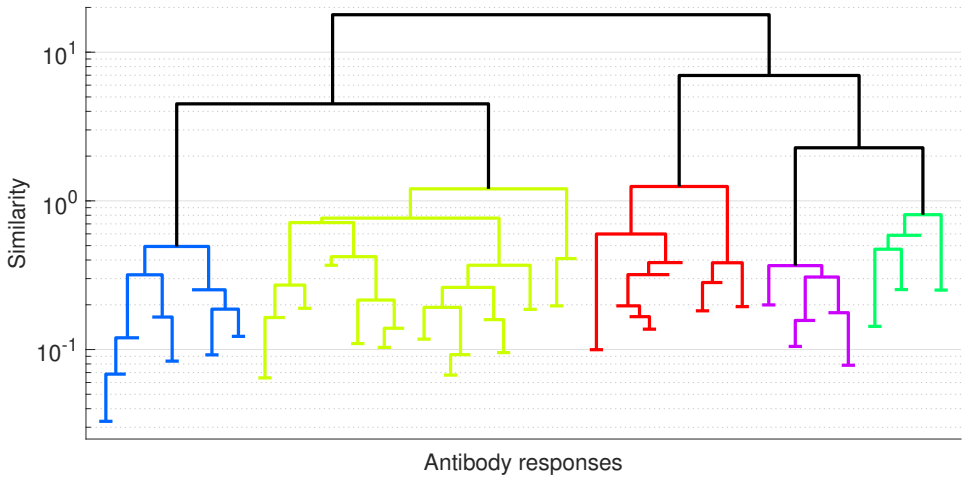


Figure C.3: Agglomerative hierarchical structure of DSA cohort formed using DTW distance measure (k=5 groups). The five clusters are highlighted by blue (Group-D), yellow (Group-E), red (Group-F), purple (Group-G), green (Group-H) colour code. Similarity is shown on a log-scale. The end of each branch designates a DSA response.

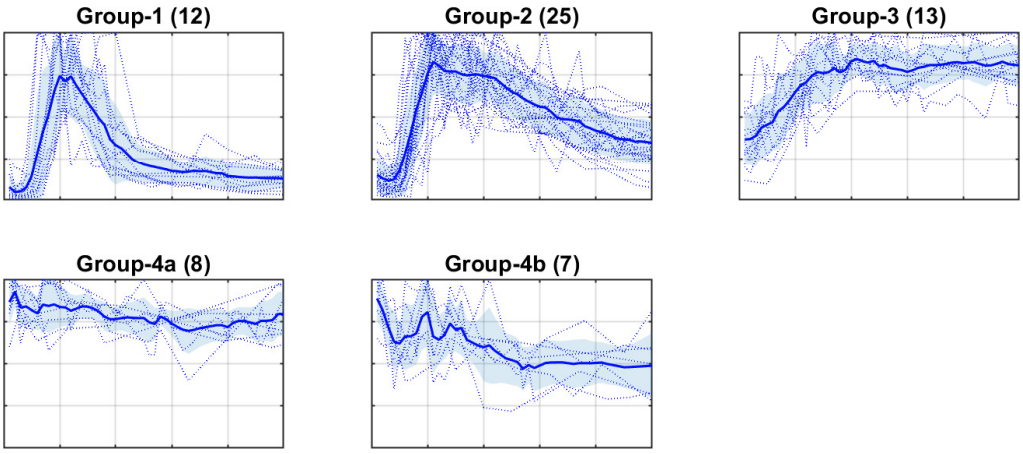


Figure C.4: Each of the five identified response clusters are shown (Groups D-H). Dark blue line represents cluster mean response and shaded region the cluster standard deviation. Dotted lines represent each individual response belonging to a given class.