Using Molecular Biomarkers to Identify ANCA-Associated Vasculitis Patients at Risk of Relapse

Barry Ryan MSc. Data Science

A Dissertation

Presented to the University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of

Masters of Computer Science (Data Science)

Supervised by Dr. Arthur White
Wednesday 18 August 2021

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Barry Ryan

Wednesday 18 August 2021

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Barry Ryan

Wednesday 18 August 2021

Acknowledgements

This project would not have been possible without the help and guidance of all those in the CDIG team. I would like to acknowledge not only the brilliant research being undertaken by the group but their support, encouragement and engagement with vasculitis patients in Ireland will no doubt lead to better understanding and treatment of the disease ANCA-associated vasculitis.

In particular I would like to thank Mark Little for welcoming me into the group and allowing me to explore the well curated RKD Registry database. I'd like to thank Jennifer Scott for her continuous support, guidance and expert knowledge on ANCA-associated vasculitis which she was always more than happy to share with me. Without Jennifer, my understanding of this disease would be severely limited and this project would not have been possible.

Finally, I cannot go without acknowledging Arthur White my supervisor. Arthur was always at hand with help and advice and his positivity and inquisition were very encouraging The time and support he allocated to me gave me confidence to explore this research topic for which I am very grateful.

Using Molecular Biomarkers to Identify ANCA-Associated Vasculitis Patients at Risk of Relapse

Barry Ryan, Master of Science in Computer Science
University of Dublin, Trinity College, 2021

Supervised by Dr. Arthur White

The primary challenge in the treatment of ANCA-associated vasculitis is the balancing of risk of relapse with the cumulative toxicities of immunosuppressive treatments. The goal of this research is to perform an exploratory analysis of the RKD Registry, a database containing patient biomarkers since its creation in 2012, and utilising these molecular biomarkers in this database to tackle this issue. Through the application of machine learning techniques it is hoped that the statistical information contained in these biomarkers can classify patients at risk of relapse from those who are off all treatment for a time period greater than one year. If successful, this research will be able to distinguish what biomarker or combination of biomarker values signal a safe termination of treatment.

Summary

ANCA-associated vasculitis is a rare auto-immune disease which if left untreated is fatal. That said, the treatment of AAV is also extremely challenging. The use of immunosuppressive therapies have reduced the mortality rate to about 2.3%, however it remains an issue that approximately 50% of people suffering from ANCA-associated vasculitis will relapse. While the disease of AAV leads to chronic symptoms, the immunosuppressive treatments, in particular their cumulative toxicity, have many equally harmful side effects. The immediate challenge of ANCA-associated vasculitis is therefore to balance this risk of relapse with the cumulative toxicities of the treatment at the maintenance therapy stage. The goal of this research was to tackle this problem through the application of machine learning techniques.

Patients were stratified into two groups, one year in remission off therapy and relapse, and biomarker samples were obtained from the RKD registry. These samples were taken from the maintenance therapy stage, i.e. before any patient has finished therapy fully and before any patient has relapsed. Three datasets were compared. A substantive sample dataset which contained raw biomarker values. A Principal Component Analysis dataset which grouped together biologically similar biomarkers into principal components. A delta analysis dataset which contained raw biomarker values as well as delta biomarkers which are differences in biomarkers values from diagnosis to the beginning of maintenance therapy. These datasets were then analysed using a Bayesian Logistic Regression and Lasso Logistic Regression algorithms.

A minimum accuracy of 72.5% was achieved for the substantive sample dataset using the Lasso Logistic Regression algorithm. A maximum accuracy of 95.7% was achieved for the delta analysis dataset using the Bayesian Logistic Regression algorithm. These accuracies significantly outperformed a baseline model confirming that the statistical information contained in the biomarker datasets can be utilised to distinguish between patients who have relapsed and those who are off all treatments for a time period greater than one year. Subsequent work was carried out to identify which biomarkers in particular are predictive of relapse. While it is clear that certain groups of biomarkers such as white cells and certain biomarkers such as ANCA titre are predictive of relapse, due to high levels of imputation and multicollinearity between variables it was difficult to conclusively identify individual important biomarkers.

Finally, a second Linear Regression analysis was performed to identify if the substantive sample dataset could be used to predict the number of relapses a patient will suffer throughout their disease course. This analysis was limited in dataset size and an optimal R^2 of 0.299 was achieved. This result is not strong enough to conclusively state that the statistical information in the substantive sample dataset can determine with confidence the Relapse Rate.

Glossary of Terms

Antineutrophil Cytoplasmic Antibody	**ANCA**
ANCA-associated Vasculitis	**AAV**
Granulomatosis with Polyangiitis	**GPA**
Microscopic Polyangiitis	**MPA**
Eosinophilic GPA	**EGPA**
Proteinase	**PR3**
Myeloperoxidase	**MPO**
Neutrophil Extracellular Traps	**NET's**
Cyclophosphamide	**CYC**
Glucocorticoids	**GC**
Rituximab	**RTX**
Rare Kidney Disease	**RKD**
Chronic Disease Informatics Group	**CDIG**
Birmingham Vasculitis Activity Score	**BVAS**
Mycophenolate Mofetil	**MMF**
C-Reactive Protein	**CRP**
End Stage Renal Failure	**ESRF**
Ear, Nose and Throat	**ENT**
Rheumatoid Arthritis	**RA**
Inflammatory Bowel Disease	**IBD**
Multiple Sclerosis	**MS**
General Practioners	**GP**
Trinity College Dublin	**TCD**
General Data Protection Regulation	**GDPR**
Comma-Separated Value	**CSV**
Immunosuppressive Therapy	**IS**

Long-Term Remission Off-Therapy	**LTROT**
C-Reactive Protein	**CRP**
Hemoglobin	**Hb**
Estimated Glomerular Filtration Rate	**eGFR**
Missing Completely At Random	**MCAR**
Missing At Random	**MAR**
Principal Component Analysis	**PCA**
Neutrophil Lymphocyte Ratio	**NLR**
Principal Component	**PC**
Markov Chain Monte Carlo	**MCMC**
Least Absolute Shrinkage and Selection Operator	**Lasso**
Area Under the Curve	**AUC**
Ordinary Least Squares	**OLS**
Akaike Information Criterion	**AIC**
Residual Sum of Squares	**RSS**
Total Sum of Squares	**TSS**

Table of Contents

List of Tables

Table of Figures

Chapter 1

Introduction

The objective of this dissertation is to improve personalisation of treatment plans for patients suffering from antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV). ANCA-associated vasculitis is a rare autoimmune disease with an incidence rate of 20 per million population (Karangizi & Harper, 2018). Untreated, ANCA-associated vasculitis is fatal and rapid diagnosis and treatment are essential to reduce organ damage and death caused by vasculitis (Karangizi & Harper, 2018). Despite treatment, AAV patient disease courses often follow a remitting-relapsing chronic course with 50% of patients experiencing a relapse within 5 years of diagnosis (Karangizi & Harper, 2018). By leveraging data analytical tools in a supervised manner, patient treatment plans may be personalised through identification of biomarkers which are indicative of relapse.

ANCA-associated vasculitis is characterised by inflammation of small- and medium-sized vessels (Berti & Specks, 2019; Geetha & Jefferson, 2019). The exact causes and pathogenesis of AAV is multifactorial, however it is believed that genetics, environmental factors and responses of the innate and adaptive immune system all influence the onset of ANCA-associated vasculitis (Geetha & Jefferson, 2019). AAV can result in inflammation and necrosis of capillaries, arterioles, venules as well as larger vessels such as the kidney. This results in systemic non-specific symptoms such as malaise, flu-like symptoms, fatigue and weight loss which make AAV difficult to diagnose (Karangizi & Harper, 2018).

ANCA-associated vasculitis can be stratified into 3 clinical diseases namely, granulomatosis with polyangiitis (GPA), microscopic polyangiitis (MPA) and eosinophilic GPA (EGPA). Each of these conditions is associated with inflammation of vessels which is caused by the presence of circulating ANCA directed against either proteinase (PR3) or myeloperoxidase (MPO) in the majority of patients (Berti & Specks, 2019). Vessel inflammation can occur due to ANCA being directed against both PR3 and MPO also. ANCA direction against PR3 and/or MPO are believed to be the key pathogenic triggers of neutrophil and monocyte activation resulting in small- or medium-vessel injury (Berti & Specks, 2019).

Neutrophils are the primary cause of vessel injury (Geetha & Jefferson, 2019). During an infection the body exposes neutrophils to inflammatory cytokines, lipopolysaccharide or complement C5a. The neutrophils become primed with movement of MPO and PR3 as a result. In this primed state, ANCA's may bind to these autoantigens on the cell surface resulting in cellular activation (Geetha & Jefferson, 2019). In patients with AAV, this immune response is incorrectly activated, therefore instead of the above process removing an infection it causes tissue damage which can lead to

organ destruction if left untreated. Tissue injury occurs as a result of neutrophil degranulation causing a release of reactive oxygen species and proteases. Activated neutrophils also undergo a form of cell death called NETosis in which neutrophil extracellular traps (NETs) are extruded from the cell containing entrapped MPO and PR3 resulting in tissue damage (Geetha & Jefferson, 2019). This response to infections or diseases causes flu-like and other aforementioned symptoms and if left untreated can result in organ damage and morbidity.

Treatment of ANCA-associated vasculitis is challenging. The introduction of immunosuppressive medications such as cyclophosphamide (CYC), glucocorticoids (GC) and novel medications such as Rituximab (RTX) have reduced patient mortality from in excess of 90% if untreated, into a manageable yet chronic disease with a relapsing nature (Berti & Specks, 2019). The difficulty in treatment of ANCA-associated vasculitis is that there is no one fit all approach and each patient disease course is unique. In general, the treatment plan is divided into two stages. The first is aimed at minimizing tissue damage by rapidly quelling the inflammatory process (Geetha & Jefferson, 2019). This is known as the induction phase and can last from 3 to 6 months. The following stage is referred to as maintenance therapy which takes place over the following 24 to 48 months. The maintenance stage is aimed at balancing the dangers of the diseases against those of treatment-related toxicity (Karangizi & Harper, 2018). The disease course and treatment plan is summarised below in Figure 1.

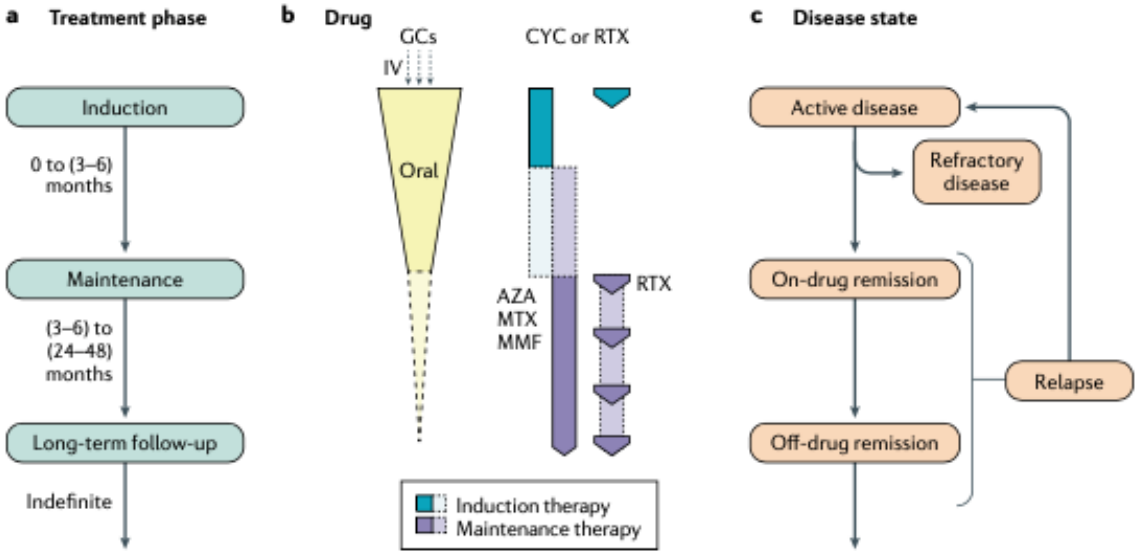

Figure 1 - ANCA-associated vasculitis disease and treatment course (Kitching, et al., 2020)

The reported Relapse Rate of patients suffering with AAV is about 50% within the first 5 years post diagnosis (Karangizi & Harper, 2018). Disease management must be tailored to the stage and severity of disease, however there is no precise disease course with most physicians recommending a minimum of 2 years of maintenance treatment (Karangizi & Harper, 2018). By

performing data analytics using the Rare Kidney Disease (RKD) registry, provided by the Chronic Disease Informatics Group (CDIG), it is hoped that novel biomarkers will be identified which can unobjectively tailor patient treatment plans to minimize the risk of relapse while subsequently optimizing the immunosuppressive treatment course.

Motivation for Research Topic

The majority of acute mortality and long-term morbidity associated with AAV is due to the use of long-term immunosuppressive agents (Berti & Specks, 2019). In fact, effective maintenance of remission with the least cumulative toxicity has been the focus of clinical trials in AAV since 1990 (Berti & Specks, 2019). Standard treatment of AAV during induction therapy utilises a combination of GC's with either CYC or RTX while azathioprine or more recently RTX are commonly used during maintenance therapy.

The use of GC's has been central in the management of AAV, however they are insufficient by themselves (Geetha & Jefferson, 2019). GC's are usually prescribed with an initial high dosage followed by a steroid taper. Despite its effectiveness in controlling disease activity there is an extensive side effect profile. Infection, bone disease, obesity, gastrointestinal bleeding, cataracts, adrenal suppression, dysglycemia and long-term risks for cardiovascular disease are all related toxicities with the maintained use of GC's (Geetha & Jefferson, 2019).

CYC's have been the most common therapy used in combination with GC's. This combination has proved to be effective in more than 90% of patients in inducing remission however it is limited by substantial toxicity associated with both CYC's and GC's (Geetha & Jefferson, 2019). CYC is associated with many adverse serious side effects. These side effects include but are not limited to haemorrhagic cystitis, bladder cancer, lymphoma, bone marrow suppression, infertility and teratogenicity (Karangizi & Harper, 2018). Once again all side effects are related to the cumulative dose of CYC (Karangizi & Harper, 2018).

RTX has been adopted as the primary treatment for ANCA-associated vasculitis in recent years given the toxicity of cumulative CYC's. RTX targets specific cellular and molecular pathways involved in the autoimmune response and has been shown to be as effective for remission induction (Geetha & Jefferson, 2019). RTX is therefore superior due to its decreased toxicity, however there are still adverse side effects associated with RTX. In clinical trials the rate of severe infection was not reduced in patients receiving RTX when compared to CYC however it was concluded that this is possibly due to the combined use of GC's. Progressive multifocal leukoencephalopathy is also a listed rare complication of RTX. RTX-induced hypogammaglobulinemia is a serious risk factor which occurs in 50% of patients treated with RTX and can lead to infection (Geetha & Jefferson, 2019).

Common to all treatments is that the cumulative dose of immunosuppressive therapies results in adverse side effects for the patient. Thus, the emphasis to minimize the use of said therapies by reducing the disease course. In essence, the clinical physician is searching for the optimal treatment time which minimises the cumulative dose of immunosuppressive therapy while ensuring that the patient is not at risk of relapse. Avoiding relapse is imperative as a patient who presents with a relapse is at a higher risk for subsequent relapses compared to newly diagnosed patients (Berti & Specks, 2019). Thus, a patient who relapses will have a longer disease course and higher cumulative dose of immunosuppressants.

Currently for ANCA vasculitis, there are some known indicators for patients who are at a higher risk of relapse. For example, with patients whose ANCA antibodies are directed at PR3, their risk of relapse is significantly higher than those whose are directed against MPO (Karangizi & Harper, 2018; Berti & Specks, 2019; Geetha & Jefferson, 2019). Other indicators of relapse include but are not limited to a rising ANCA titre, reappearance of haematuria in urine and increased serum creatinine (Karangizi & Harper, 2018). While these markers are useful indicators it is impractical to continuously monitor patients in this manner. Motivation for this research is therefore to leverage the power of data analytic techniques to identify patterns within these markers during maintenance therapy and to utilise this data to personalise patient treatment plans which minimise the cumulative dose of immunosuppressive therapies while ensuring a positive patient outcome.

Dissertation Overview

The structure of the dissertation is as follows. Chapter 2 reviews current literature and establishes the current state of the art research being undertaken in both the areas of AAV and supervised learning in a biomedical setting. Chapter 3 introduces the RKD Registry, the database used in this project, and outlines the work undertaken to clean and extract the final biomarker dataset. Chapter 4 describes the supervised learning algorithms applied to the dataset and presents the results of said algorithms. Chapter 5 concludes this research project. It discusses potential future areas of research as well as a reflection and conclusion on the findings of this project.

Chapter 2

Literature Review

This section provides a review of current literature as well as establishing research questions addressed throughout this dissertation. The first section presents open issues regarding ANCA-associated vasculitis as well as the state of the art research identifying known associated risks of relapse and their limitations in predicting patient relapse. The second section reviews supervised machine learning methods for identifying biomarkers in both ANCA-associated vasculitis and other autoimmune diseases and establish their use in improving patient treatment plans.

Related Work on ANCA-Associated Vasculitis

Current research undertaken to address the issue of relapse risk for patients diagnosed with ANCA-associated vasculitis suffers from conflicting results, small subgroup studies and varying inclusion criteria. The explanation for this is largely due to the rarity of the disease combined with several varying subgroups of the disease with many studies focusing on particular subgroups of AAV such as renal disease. As introduced in Chapter 1, there are three subtypes of ANCA-associated vasculitis, namely GPA, MPA and EGPA. Not only this but there are two phenotypes namely, PR3 and MPO at which the circulating ANCA is directed. The exact pathogenetic triggers of AAV are still, as of writing of this paper, unknown (Karangizi & Harper, 2018; Geetha & Jefferson, 2019; Kitching, et al., 2020). A widely held belief is that identifying biomarkers independent of these subtypes which captures disease activity is the key to not only understanding AAV but to tailoring patient treatment plans (Kitching, et al., 2020; Hogan, et al., 2019; Salama, 2020).

The Birmingham Vasculitis Activity Score (BVAS) is a measure of disease activity and was intended to be used as a measure to guide treatment decisions. A BVAS score of zero indicates there is no disease activity and a higher value indicates disease activity, however there are limitations to this system. Firstly is its inability to differentiate between remission and low grade vasculitis (Trejo, et al., 2019). For example, given two patients, one suffering a relapse and the other merely low grade disease activity, depending on the specific markers or symptoms of each, the BVAS score of both could be similar, however they should not be treated similarly. In addition to this, the BVAS system cannot differentiate active disease from organ damage (Salama, 2020). Using the BVAS scoring system, some symptoms such as persistent or mild haematuria, subtle elevations in creatinine and others may be related to scarring and active disease but may not be recorded (Salama, 2020). Thus, the requirement for accurate biomarkers and disease activity

identification methods to differentiate between disease remission, disease suppression (low activity) and disease activity.

In order to differentiate between different disease states, one must be able to obtain clinical markers which unobjectively identify that state. For ANCA-associated vasculitis the key differentiation between states that is of interest is identifying patients who are in remission and will remain in remission from those who will relapse. Various papers have successfully identified relapse risk factors of ANCA-associated vasculitis, however there are large differences in how researchers and clinicians have defined a relapse. Salama states that a relapse can only occur following a remission (Salama, 2020). Conversely, papers such as Hogan, et al., McClure, et al., and He, et al. loosely define relapse as active disease resulting in the change or increase in immunosuppressive therapy, although the exact criteria for relapse in all three vary. The disease and treatment course of each patient is unique. Defining a relapse based on variation of immunosuppressive therapy results in significantly higher reported rates of relapse. As a result given a definition of relapse which does not consider remission may result in overestimated relationships between risk factors and relapse. This is one possible explanation for conflicting results among papers.

Similarly, while the definition of remission for a patient with ANCA-associated vasculitis has generally been defined in papers as a BVAS equal to zero, there is no robust definition of remission (Salama, 2020). Hara, et al., have defined remission as a BVAS score of zero on two recurrent visits greater than one month apart while, Hogan, et al. have defined remission as the absence of dysmorphic red blood cells, no identifiable vasculitis lesions or symptoms in any organ and a BVAS equal to 0 (Hara, et al., 2018; Hogan, et al., 2019). Salama on the other hand, states that clinical remission is not straightforward, and evidence has been found where patients in remission have persistent inflammatory and immunological activity at levels above those of a healthy individual (Salama, 2020). For reasons outlined previously, the use of BVAS in identification of remission has known limitations. Thus, the variations in definitions of both remission and relapse emphasises the requirement for a strict definition of both of these quantities. A definition which is clinically validated is required in order to identify, unobjectively, biomarkers which are predictive and indicate an increased likelihood of relapse among all patients irrespective of their disease sub group.

Further evidence of this requirement is corroborated given the variation in results in inducing and maintaining remission across many immunosuppressive therapy trials. He, et al. performed a systematic review and meta-analysis on the risk factors of relapse which showed that intravenous CYC induction versus oral CYC was associated with a 1.74 increased chance of relapse and use of mycophenolate mofetil (MMF) during maintenance treatment resulted in a 2.33 increased chance of relapse when compared to AZA (He, et al., 2020). Hogan, et al. identified that a

combination of GC's and AZA for a time period greater than 2 years resulted in lower rates of relapse however this also resulted in a higher rate of infection and comorbid events (Hogan, et al., 2019). In two trials, namely the MAINRITSAN and RITAZAREM trials, RTX outperformed AZA in maintaining remission however the optimal duration for RTX maintenance therapy is still uncertain with the REMAIN trial supporting 3-4 years of treatment regardless of all other factors (Kitching, et al., 2020). Despite these many trials there is no known reason why one immunosuppressive therapy may reduce the risk of relapse. In addition to this, the problems of treatment length as well as the fact that patients do still relapse irrespective of their immunosuppressive treatment remain. If a biomarker or combination of biomarkers can be identified which are an accurate predictor of relapse and independent of treatment, not only can they be used to explain why different therapeutics perform better than others, they may also yield an explanation of the pathogenesis of ANCA-associated vasculitis and improve the overall understanding of the disease.

Currently no such biomarker has been discovered but there are various sub-group studies which have identified biomarkers which are associated with an increased risk of relapse. It has been concluded in most literature that patients who are PR3 positive are at a higher risk of relapse (Berti & Specks, 2019). However, there is a geographic incidence related to PR3 and MPO with the majority of PR3 positive patients having European ancestry whereas MPO is more common in those of Asian descent (Kitching, et al., 2020). Conversely, the study conducted by Hara, et al. of risk factors for relapse in AAV in Japan, where the majority of patients were MPO positive, did not find a significant difference in relapse rates between patients who were MPO positive and those who were PR3 positive (Hara, et al., 2018). These conflicting results could be attributed to the small studies performed, the geographic incidence of MPO and PR3 as well as differences in relapse definition. Nonetheless the literature highlights the requirement for further analysis. The use of ANCA titre is also commonly referred to in literature as being predictive of relapse but not sufficiently predictive to warrant a change to immunosuppressive therapy. Instead a change in ANCA titre is an indicator that closer monitoring of the maintenance therapy stage is required. While, it is commonly accepted that a change in ANCA titre alone is not predictive, Kemna, et al. hypothesize that a 'second hit' along with this is required for a patient to relapse (Kemna, et al., 2017). Such a hypothesis has been corroborated elsewhere in literature with Berti & Specks finding that a change in ANCA titre along with B cell reconstitution results in a higher incidence of relapse (Berti & Specks, 2019). Similarly, Kemna, et al. found that an ANCA rise in the autumn months were more frequently followed by a relapse (Kemna, et al., 2017). It was postulated in this study that the 'second hit' was due to low levels of 25(OH)D more commonly known as vitamin D. In a separate study, Kemna, et al. found that ANCA titre was a useful predictor in AAV with renal involvement but not in patients with nonrenal disease. Finally, Sanders, et al., Kitching, et al. and McClure, et al. all state that a persistently high ANCA titre is a good indicator for relapse however

it is possible that this is more indicative of a lack of response to treatment. Other biomarkers such as creatinine and complement C5a were both shown to have a relationship with relapse risk (Walsh, et al., 2012; Geetha & Jefferson, 2019). These findings however were not conclusive with the respective authors citing the requirement for further research.

The variation in findings in the literature mirror the difficulty of the challenge. Conflicting findings, small sub group analyses and the rarity of this disease has, in the past, meant that there are no conclusive biomarkers which are indicative of relapse. There is reason for optimism with more recent papers such as O'Reilly, et al. identifying a novel biomarker namely, CD163, which can accurately diagnose renal vasculitis in a non-invasive manner. CD163 is present in urine and is shed by monocytes during macrophage activation. Excessive macrophage activation is a biological phenomenon which occurs during active renal vasculitis thus a simple dip-stick urine test could be employed to diagnose a renal vasculitis relapse (O'Reilly, et al., 2016). It is hoped that through the application of data analytical methods similar novel biomarkers or patterns within biomarkers can be identified. Supervised learning has the ability to not only identify novel biomarkers but to combine known risk factors such as PR3 positivity or heightened ANCA titre. The ideal candidate biomarker or group of biomarkers is similar to CD163 whereby it is biologically associated with AAV, explains or aggregates findings in literature, is independent of the type of immunosuppressive therapy received and can unobjectively stratify all AAV patients who are at a risk of relapse during the maintenance therapy phase from those who are not.

Related Work in the Field of Supervised Learning in a Biomedical Setting

Supervised learning in the field of biomedicine is emerging as an area of increased utilization and importance. Machine learning can transform medicine into a data-driven, outcome-orientated discipline for disease detection, diagnosis and treatment due to its ability to collect and analyse large datasets (Goecks, et al., 2020). So far machine learning techniques have been used to diagnose breast cancer from X-rays, discover new antibiotics and predict the onset of gestational diabetes from electronic health records (Goecks, et al., 2020). ANCA-associated vasculitis is a complex disease. The exact pathogenesis and onset of ANCA-associated vasculitis is unknown although recent literature, as reviewed above, has begun to identify triggers and risk factors for the disease. Goecks, et al. state that currently clinical practices are limited to few markers which only reflect a narrow view of complex diseases such as AAV. The use of machine learning in this setting can leverage sophisticated algorithms operating on large multi-type datasets to uncover useful patterns that would be laborious or even unfeasible for well-trained individuals to identify (Goecks, et al., 2020). Applying supervised learning techniques to the challenge of optimizing maintenance treatment for patients with AAV is one that has the potential to unlock as of yet unknown triggers of disease relapse and activity.

There has been limited research utilising supervised learning techniques to predict the risk of relapse in patients with ANCA-associated vasculitis. Trejo, et al. performed a multivariable regression analysis which identified that one or more renal relapse was a strong predictor for End Stage Renal Failure (ESRF) however there were no factors found that were predictive of renal relapse (Trejo, et al., 2019). McClure, et al. performed research which aimed to create risk prediction models to help guide decision making during extended RTX maintenacne therapy (McClure, et al., 2020). Unfortunately this paper was unable to discriminate between individual patients at risk of relapse however they were able to categorise patientes into low and high risk of relapse. The main finding of this paper was that involvemnet of Ear, Nose and Throat (ENT) in AAV was predictive of relapse (McClure, et al., 2020). McClure, et al.'s paper is the closest study found to the proposed research in this paper. While the outcomes of the supervised learning models are identical, the inputs to the model differ significantly. McClure, et al. have focused on the use of RTX during maintenance therapy whereas this study is looking to identify biomarkers which can predict relapse irrespective of the immunosuppressive therapy for reasons outlined in the section Related Work on ANCA-Associated Vasculitis.

While there has been limited findings in the application of supervised learning to ANCA-associated vasculitis, other auto-immune diseases have benefited greatly from the application of machine learning techniques. In autoimmune diseases such as Rheumatoid Arthritis (RA), Inflammatory Bowel Disease (IBD) and Multiple Sclerosis (MS) supervised learning techniques have been used to diagnose diseases, classify disease subtypes and identify and asses autoimmune disease risk (Stafford, et al., 2020). One example is Edner, et al. utilising a gradient boosting model which was able to predict the response of patients with type 1 diabetes to immunosuppressive therapy using follicular helper T cells (Edner, et al., 2020).

Goecks et al. states that the application of machine learning to autoimmune diseases is a problem that requires disease-specific research to differentiate between indolent and fatal diseases, to avoid overtreatment and to identify disease subtypes and progress in order to guide the most effective treatment plan (Goecks, et al., 2020). The research undertaken by the CDIG and during this dissertation is closely aligned with Goecks statement, thus with the curated RKD database there is the opportunity to apply supervised learning methods to improve treatment plans for patients with ANCA-associated vasculitis.

Chapter 3

The RKD Database

Goecks, et al., state that in order for machine learning to play a transformational role in the area of biomedicine disease specific research with well-curated datasets will be imperative (Goecks, et al., 2020). The RKD Registry database is an example of one such dateset. Created by the CDIG and established in 2012, the RKD Registry is a nationwide biobank of patient encounters suffering from ANCA-associated vasculitis. This database was the sole source of the dataset used in completion of this thesis.

Security and Privacy Considerations

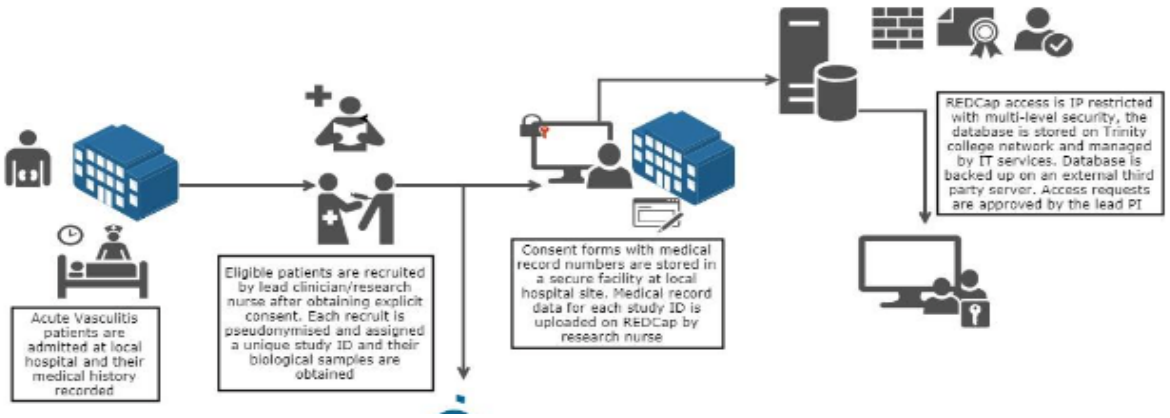

Figure 2 - RKD Patient Data Collection Flow (Little, 2019)

Figure 2 above shows the flow of data for an individual patient who has consented to inclusion in the RKD biobank. After a patient is admitted to hospital and diagnosed with ANCA vasculitis they are recruited by lead clinicians and researchers. De-identification is performed at this stage whereby each patient is identified with a unique RKD ID. For each subsequent encounter with a General Practitioner (GP) or clinician, patient data is uploaded to the RKD database using this RKD ID. User privileges allow for lead researchers to identify patients and high security measures are maintained through use of the REDCap database.

Data obtained includes all information about patients pertaining to the disease such as date of birth, gender, height, date of encounter, hospital name as well as disease specific information such as medication dosages and clinical biomarker samples. Patients are identified by a unique ID, however due to vasculitis being a rare disease factors such as clinical visit dates, unusual symptoms and others may allow for specific patient identity to be revealed if such data were known

to third parties. For this reason, all research participants, including myself, have agreed to and acknowledged a strict ethics declaration which can be found in the Appendix.

The RKD registry is hosted in its entirety using the REDCap data management system. REDCap was created in 2004 at Vanderbilt University and is compliant with 21 CFR Part 11, FISMA, HIPAA and GDPR (REDCap, 2021). The REDCap management system is hosted on the Trinity College Dublin (TCD) IT services which are compliant with the ISO27001/ISO27002 international best practice standards (Trinity Health Kidney Centre, 2021). CDIG personnel who have access to the non-anonymised data undergo additional training to ensure researchers adhere to the provisions set out by the General Data Protection Regulation (GDPR), Directive 2006/24/EC, Directive 2002/58/EC and Directive 95/46/EC. These provisions are intended to protect individuals with regard to the procession of their personal data and the freedom of movement of said data (Trinity Health Kidney Centre, 2021). The CDIG relies on consent as the mechanism for processing data. This is dictated by article 6(1)(e) of the Public Interest and Article 9(2)(j) Scientific Research which states - *Explicit consent is sought as an appropriate safeguard to rights of the data subject as mandated by the Health Research Regulations*. Only upon a patients consent can researchers use their data in a study. Through the aforementioned protocols and measures, the security and privacy of patient data is ensured.

Dataset Description

The entire usable dataset was pulled from the RKD registry in the form of a comma-separated value (CSV) file. This file contains 21060x699 data entries pertaining to patient disease information. There was a total of 868 patient ID's present in this dataset with an average of 24 entries per patient. The amount of data for each patient varied greatly. Depending on the disease stage and course the number of entries per patient could be very large. For example, a patient who relapsed multiple times could have as many as 40 observations. Conversely many patients were recruited during their remission stage, years after their disease course ended, thus they may only have one or two encounters. Large proportions of this data were unusable for the goals outlined in this project as discussed in the section Methodology and Data Preparation.

Biomarker Glossary

This section is a glossary of biomarkers available in the RKD database. Some of these biomarkers are biologically related to AAV. For example, Neutrophils are responsible for the inflammation of vessels as discussed in the Introduction. On the other hand, other biomarkers are not directly related to the disease but are good measures of bodily function. For example, Creatinine is a measure of Kidney function. Table 1 below is a summary of all biomarkers with sufficient data for

analysis in the RKD database. A full list of biomarkers available in the RKD database can be found in the Appendix in Table 23.

Biomarker Name	Description	% Missingness
Creatinine	Marker of Kidney Functionality	8.46
Estimated Glomerular Filtration Rate (eGFR)	Marker of Kidney Functionality	8.46
Hemoglobin (Hb)	Measure of Iron in blood	13.85
Total white cell count	Biological Measure	13.85
Neutrophil count	Biological Measure	20.00
Lymphocyte count	Biological Measure	20.77
NLR	Biological Measure	20.77
Anti-PR3/MPO level	Level of IgG ANCA directed against PR3/MPO	23.85
Weight (KG)	Weight of Patient	24.62
C-Reactive Protein (CRP)	Marker of Inflammation	28.46
Platelet count x10^9/L	Biological Measure	30.00
Eosinophil Count x10^9/L	Biological Measure	33.85
Urinalysis Blood	Quantity of Blood in Urine	43.85
Urinalysis Blood	Quantity of Blood in Urine	43.85
IgG g/dL	Immunoglobulin G - Antibody present in Blood	48.46
Age at Diagnosis	Age of Patient at Time of Diagnosis	0.00
Gender	Male/Female	Categorical
Smoking	Previous/Never/Unknown/Current	Categorical

Disease Subtype	MPA / GPA / EGPA	Categorical
ANCA Specificity	PR3 / MPO / Other	Categorical
Treatment Group	CYC / CYC & RTX / RTX / No Induction / Other	Categorical

Table 1 - Biomarkers with Sufficient Data for Analysis

In total there were 70 continuous biomarkers available for analysis in the RKD Registry. As can be observed in Figure 3 however there was a large proportion of biomarker missingness with some biomarkers such as Serum CD25 having 100% missingness. Consequently, any biomarker with a percentage missingness greater than 50% was excluded from the analysis. This left 16 continuous and 5 categorical biomarkers for analysis shown in Table 1. A biomarker or biological marker is a measure of biological state of a person. It can range from simple markers such as gender or height as well as biological measures such as those contained in Table 1. It is hoped that through the analysis of these biomarkers that a sub-clinical stratification of relapse will be attained. The RKD Registry in its current state is limited in the number of biomarkers available for analysis however it is expected as the database becomes more prevalent that missingness will be reduced and more biomarkers will be utilised in a similar analysis.

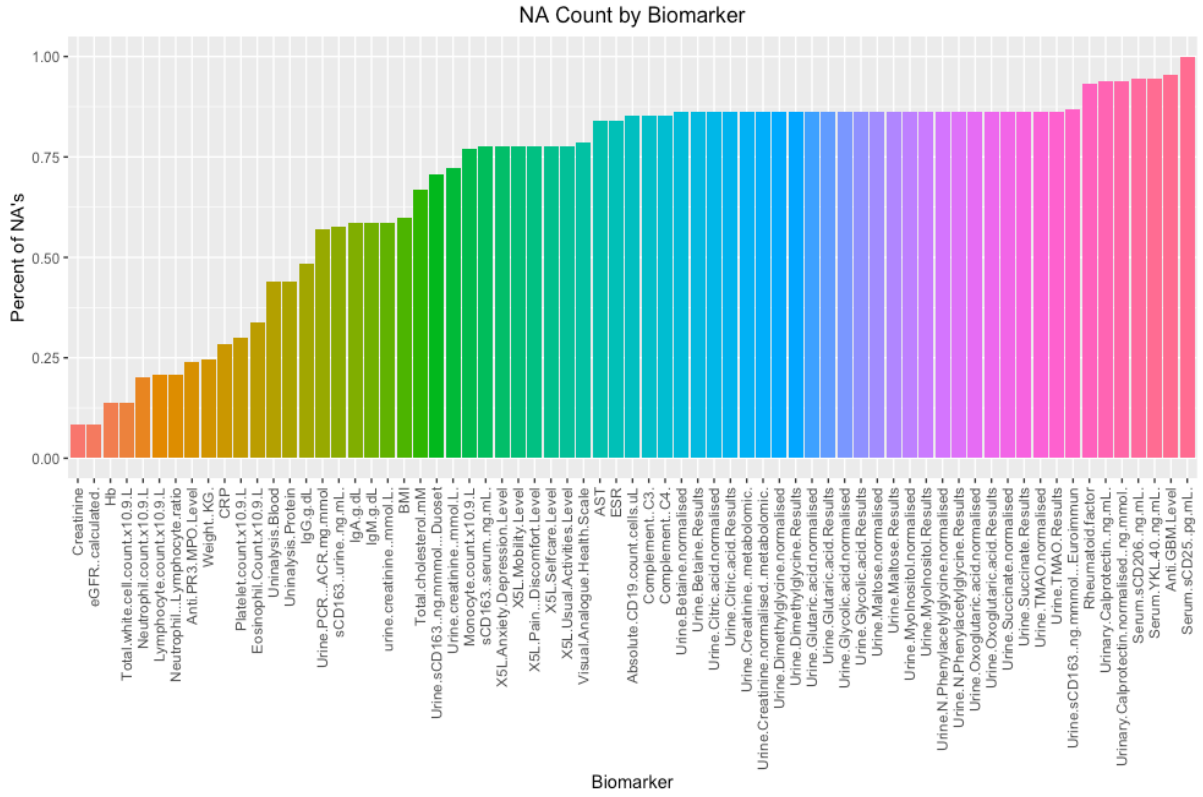

Figure 3 - Biomarker Availability

The use of the biomarkers in Table 1, to quantify risk factors and classify patients into target outcomes of relapse and long term remission, significantly differs from research found in the

literature review. Previous research tended to focus on the immunosuppressive therapy received or focused on a single marker such as ANCA titre. The focus of this research is to explore whether the biological factors can be used to identify unusual sub-clinical immunological activity in patients which distinguish those who relapse from those who don't using all available information in the RKD registry.

The following section provides a methodology outlining how the relevant biomarkers were extracted. As can be seen in Table 1, there is a reasonably high level of data missingness. This section also explains what measures were taken to maximise data use to reduce this percentage missingness.

Methodology and Data Preparation

The aim of this project is to stratify patients into those who are at risk of relapse from those who are in long term remission. The reason for this is to minimise the cumulative dose of the treatment stage due to the toxicity of immunosuppressive therapy received during the maintenance stage of therapy. The dataset provided contains patient biomarker data throughout the entire disease course however the methodology employed in this project considers patients who are in remission after the introduction of maintenance therapy. Application of data science techniques can then quantify sub-clinical immune activation which can aid the diagnoses of lingering ANCA-associated vasculitis in patients who are likely to relapse from those who will remain in remission long term off therapy. This section explains the work carried out to clean and prepare the data for analysis.

Clinical Definitions of Remission and Relapse

The first step in preparation was to define and stratify patients into the target outcomes of remission and relapse. As outlined in Chapter 2, there is no robust definition for relapse or remission. The definition of both changes throughout literature with various studies citing different definitions. A clinical definition of relapse was provided by the CDIG team to overcome this issue and ensure consistent stratification of patients. The definition provided is robust in that it considers patients who are in remission as well taking into account the patients BVAS score, objective data and both changes and responses to immunosuppressive treatment.

The occurrence of a relapse was defined by:

- New symptoms of AAV (BVAS > 0)
- In a patient with prior remission
- Requiring an increase in Immunosuppressive Therapy (IS)
- With a clinical response to IS
- Supported by objective data

Patient relapses are tracked in the dataset under the column ‘adjudicated probability of relapse’ with entries of ‘definite’ and ‘high probability’ being taken as a clinical confirmation of relapse. No patient data was used after they had relapsed as the aim of this work is to differentiate and ultimately prevent patients who will relapse from those who will not.

The clinical definition of remission is referred to as Long-Term Remission Off-Therapy (LTROT). This is a patient who has never had a clinically defined relapse and has stopped all IS for a period of time greater than one year. This is monitored in the RKD registry under the columns ‘adjudicated probability of relapse’ with either no entry or an entry of ‘No’ as well as an entry of ‘Remission’ under the column ‘disease activity since last return’.

Figure 4 on the right shows the distribution of labels in the dataset. The total number of patients was 868. Of those 22 have no useable data. These are patients for whom there is no data available during the maintenance therapy phase. There may be several reasons for this such as they were only recruited years after their disease course finished, or they may only have one encounter and they are missing key dates such as date of diagnosis.

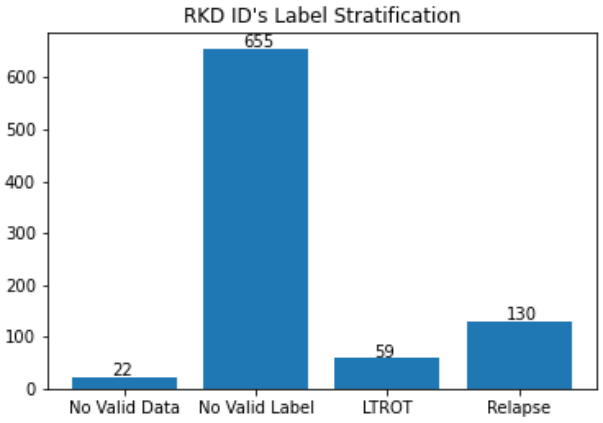

Figure 4 - RKD ID Remission and Relapse Stratification

656 patients do not have a valid label. These patients are either still undergoing induction or maintenance therapy or have active disease. The final 189 patients have successfully met the criteria for either LTROT or Relapse.

While only approximately 22% of the patients can be labelled as LTROT or relapse, this is to be expected for several reasons. First, the strict definition of LTROT and relapse results in many patients being excluded. The dataset is also still in its infancy despite being first curated in 2012. This was significantly evident as there are currently 626 patients who are currently undergoing some form of treatment and thus, while they have not relapsed, they also have not met the 1 year off treatment criteria. The incidence of AAV-associated vasculitis makes for very few cases. Given an incidence between 20-40 per million population it would be expected that there are 100-200 active cases of AAV in Ireland which corroborates the numbers seen.

Relapse Rate

Given the small number of labelled ID’s available it was hypothesised that by looking at the incidences of relapse, more data could be used. The Relapse Rates was defined as the number

of times a patient relapsed, given the above definition for relapse, during the disease course counted in years. The disease course for each patient was the time from their 'Date of Sample' (explained below) to their last data entry logged in the RKD Registry. For LTROT patients or patients who have not suffered a relapse, their Relapse Rate will be zero. Conversely a patient who has suffered a relapse will have a Relapse Rate greater than zero.

This method targets patients who have not suffered a relapse but who do not have sufficient time off therapy to be considered an LTROT. For these patients, given an accurate model, this method could help to distinguish LTROT patients from relapse patients while also identifying important features used to make this distinction.

Data Cleaning & Biomarker Extraction

Despite the RKD registry being a well curated dataset, much data cleaning and filtering was required before biomarker data was able to be extracted. The first step in this process was to stratify patients into LTROT and relapse as per the aforementioned definitions. The workflow for this process is shown in the Appendix.

Identifying key dates was integral to the data cleaning stage and identifying the window for each ID from which data could be sampled. Some dates such as date of diagnosis were included in the dataset however other dates such as the date of treatment stop and the date at which maintenance therapy began were not included. Finding the date at which maintenance therapy began and ended for each patient was imperative as this is the timeframe from which data was taken.

The start date from which data was sampled, referred to as date of sample, was found using the below workflow process in Figure 5. The end date from which data was sampled, referred to as the date of treatment stop, was found using the below workflow process in Figure 6. The maximum number of biomarkers within this window were taken as a single sample for each ID. This was referred to as a substantive sample. Other work was carried out to include categorical variables as well as combining some biomarkers to reduce the missingness as discussed below. Finally, where available a delta between biomarker values when a patient first presented with AAV symptoms to the beginning of maintenance therapy was obtained. These delta biomarkers were appended to the substantive sample to form the delta analysis dataset. Both the substantive sample and delta analysis datasets then formed the input to the supervised data analytics models in Chapter 4. Note a third dataset was created from the substantive dataset using Principal Component Analysis. This is also described in Chapter 4.

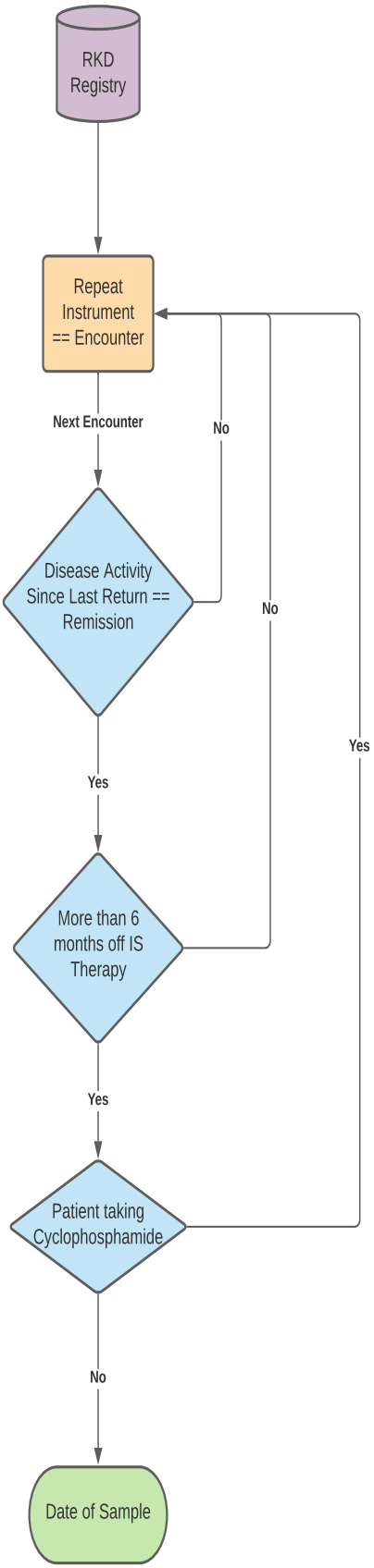

Figure 5 - Date of Sample Workflow Process

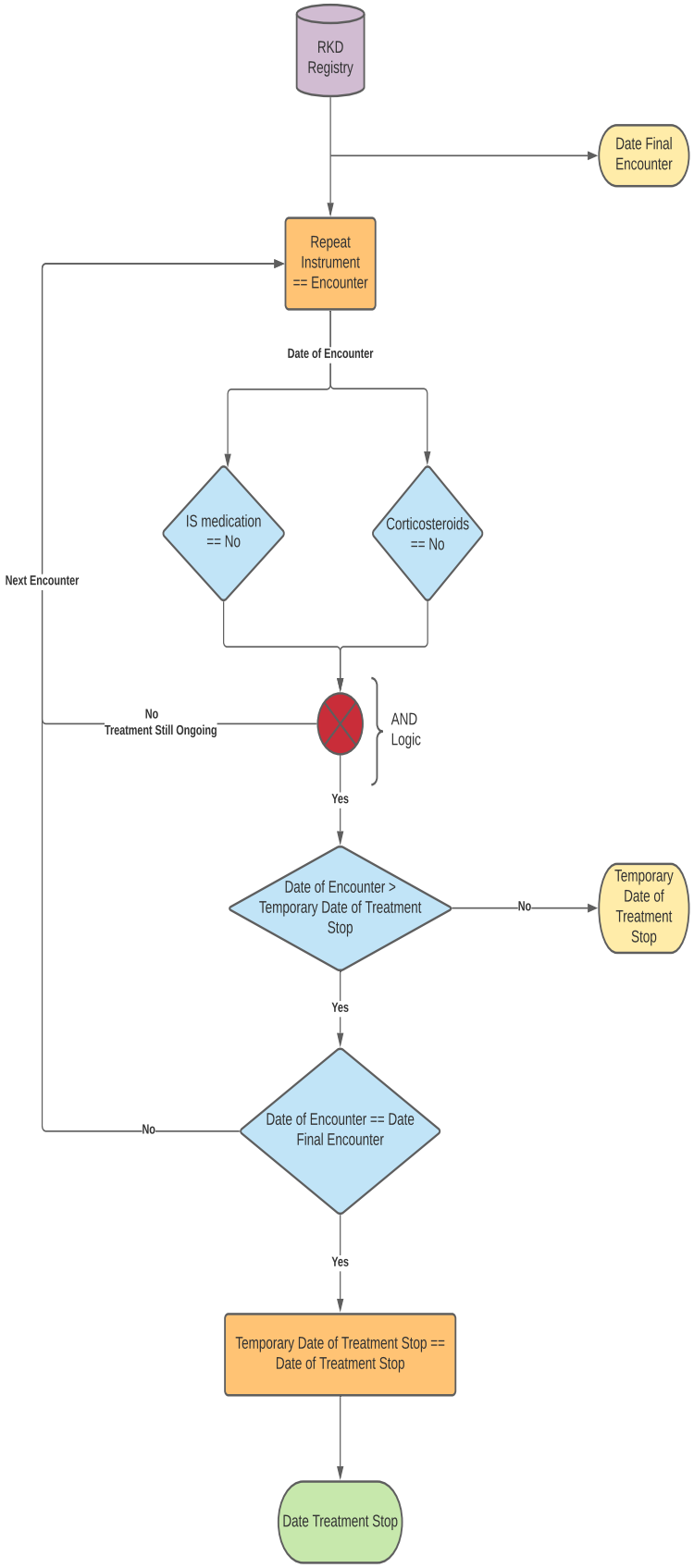

Figure 6 – Date of Treatment Stop Workflow

Date of Sample

The date of sample is that date at which maintenance therapy has begun. The criteria for the beginning of maintenance therapy are that the patient is in remission, the date of sample is greater than 6 months after the date of diagnosis, the patient is no longer taking cyclophosphamide medication and if the patient was taking intravenous RTX, the last dose was greater than 6 months ago. Samples were not taken within the first 6 months of therapy as the medication taken during the induction phase of therapy would result in much larger variations of biomarker values. The goal of the research is to optimise the maintenance treatment phase of therapy by sub-clinically quantifying patients at risk of relapse once in remission, thus the induction phase is not of interest.

Date of Treatment Stop

The date of treatment stop is the date at which maintenance therapy has been stopped and samples can no longer be used after this date. The workflow for finding this date is shown above in Figure 6. A point of note is that this workflow does not distinguish between LTROT and relapse patients. While a LTROT patient may have a simple disease course whereby they will never stop treatment for a short period of time before resuming therapy the converse is always true for relapse patients. A patient who relapses will have a remitting-relapse disease course and may discontinue and resume therapy many times. The above workflow in Figure 6 is designed to find the very last time any patient took any medication and this date becomes the date of treatment stop.

Obtaining a Substantive Sample

The next step in the data cleaning and biomarker extraction is to extract a sample within the maintenance window timeframe for each patient that has the most biomarker data available. This is referred to as a substantive sample. Each patient with AAV will have a single substantive sample which will be used as an input to the data analysis. This sample contains the maximum number of non-missing biomarkers. Ideally this sample corresponds to the date of sample as it is preferable to be able to classify patients as early as possible in the maintenance therapy stage. A common trend in the dataset was biomarker missingness. A technique employed to overcome this was to average samples over a 6-month period. In this way more biomarker samples were included as different biomarkers are taken during consecutive encounters. For example, a patient might have Creatinine measured during their first encounter but their C-Reactive Protein (CRP) and not Creatinine taken during their second. If these two observations occur within 6-months of each other their values can be aggregated into a single input sample as it was clinically confirmed that these values should not change significantly with maintenance therapy within a 6-month period. This method of data extraction and aggregation is referred to as the sliding window method.

Figure 8 shows the workflow for selecting the substantive sample. A key differentiation in this workflow is the timeframe from which the sample can be taken for LTROT and relapsing patients. A LTROT patient can have their substantive sample taken anytime from the date of sample to the date of treatment stop. A relapsing patient cannot have a sample taken after their first relapse, hence the timeframe is from the date of sample until the last encounter before the patient relapsed.

To help decrease the data missingness, particularly for patients who had few encounters between the date of sample and the date of treatment stop, this window was expanded to include any sample taken up to three years post diagnosis. The time frame of three years was chosen and clinically validated as the typical length of disease course is 6 months of induction therapy plus 2-2.5 years of maintenance therapy.

An example of the sliding window methodology is shown in Figure 7. All encounters within the respective window are found. Encounters within a 6-month period are grouped together and the biomarker values are merged. If a biomarker is filled in three times in this period, the average of the three values is taken. Similarly, if a biomarker is filled in twice, then the average of the two values is taken and if only filled in once in 6 months this value is taken. If the value is missing across all instances, then it is left blank. This process is repeated for all groups and the group with the largest number of unique biomarkers is selected. As can be seen in Figure 7, the first three encounters had 20 unique biomarker values taken when the samples were merged versus a maximum of 15 if only a single encounter had been taken. If multiple 6-month groups have the same number of unique biomarkers available, then the group closest to the date of sample is taken as this is the sample closest to the beginning of maintenance therapy.

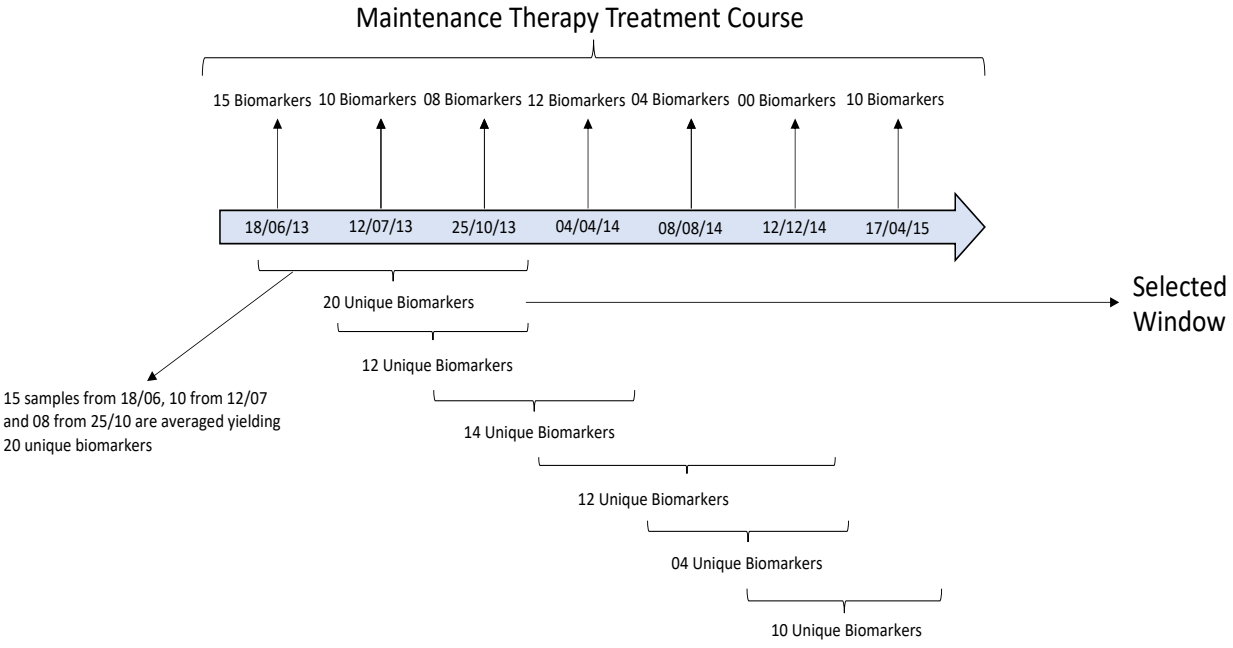

Figure 7 - Example Sliding Window

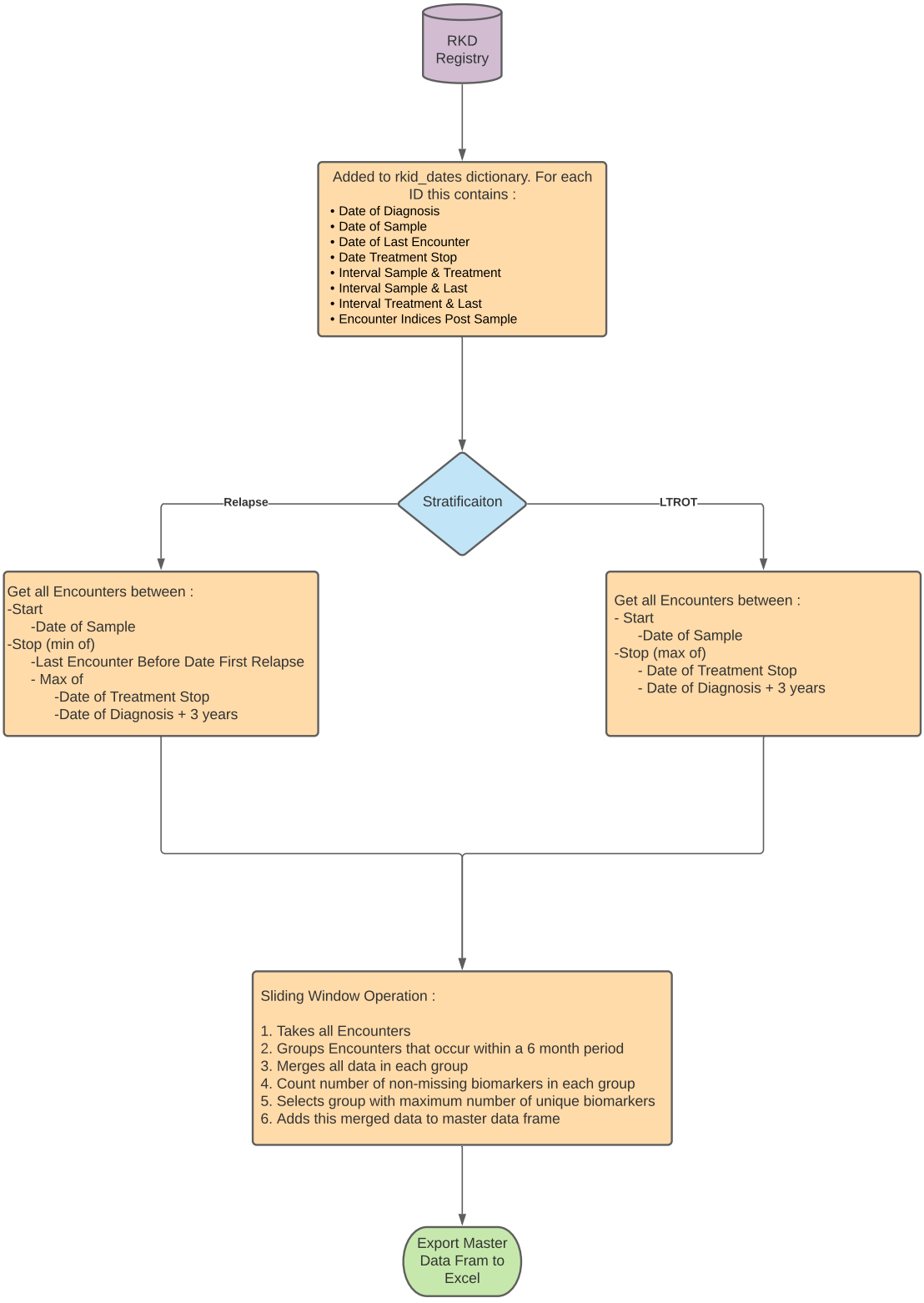

Figure 8 - Sliding Window Workflow

Categorical and Combining Biomarkers

Not all biomarkers are continuous measures of body function. In fact some categorical biomarkers such as ANCA antibodies being directed against PR3 were found to be more predictive in the literature. For this reason the categorical variables Gender, Smoking, Disease Subtype, Treatment Groups and ANCA specificity were added to the substantive sample. The age at diagnosis was also added to the substantive sample. The Disease Subtype or Clinical ANCA subtype logged whether the patient had MPA, GPA or EGPA. ANCA specificity was specified as one of four groups namely, MPO, PR3, MPO & PR3 and ELISA negative.

Patients were divided into groups depending on the Immunosuppressive Therapy they received. Those who received Cyclophosphamide whether it was pulse or intravenous were assigned to the group 'CYC'. Those who received a combination of Cyclophosphamide (pulsed or intravenous) and RTX were assigned to the group 'CYC & RTX'. Those who received RTX only were assigned to the group 'RTX'. Those who received any other type of treatment excluding those mentioned previously were assigned to the group 'Other' and finally those who received no induction treatment were assigned to the group 'No Induction'.

The biomarkers Anti-PR3 and Anti-MPO levels were merged into one biomarker with the maximum value of either being taken. The majority of patients are either PR3 or MPO positive with very few being both PR3 and MPO positive. In patients who are both PR3 and MPO the level would be similar in both and either could be used. In PR3 patients the MPO level is likely negligible and not of use and vice versa for MPO. These markers were also merged with the 'ANCA IF' entry in the RKD Registry. This entry logs whether there was a negative, increase or decrease in ANCA levels. If the 'ANCA IF' test returns a negative then a value of 0 can be logged in the combined Anti-PR3/MPO biomarker. This reduced the level of missingness in this column significantly as a value of 0 would not have been logged otherwise.

Similarly the biomarkers eGFR and 'eGFR (calculated)' were merged. These values should be identical however in the case that one was filled and the other not, the filled sample was taken. If both were filled then the value in 'eGFR (calculated)' was taken as this was deemed clinically to be more accurate.

The biomarkers 'Urine sCD163 (ng/ml), Duoset' and 'Urine sCD163 (ng/ml), Euroimmun' were also combined as they are simply different methods of measuring the quantity of CD163 present in the patients urine.

The variables Urinalysis Protein, Urinalysis Blood and ANCA titre were mixed variables with some entries being categorical and other entries being numerical. The reason for this is that these tests can return a 'Negative' result. In this case a negative result corresponds to there being no

presence of the target in this test. For example a negative ANCA titre means there was no ANCA found in this sample. For all these tests, a 'Negative' result was replaced with a value of 0.

Obtaining a Delta Sample

Delta samples were obtained for all ID's where possible. A delta sample consists of the difference in a biomarker value from when the patient first presents with symptoms of AAV to when the date of sample is obtained. The delta can be taken anytime either one month before the date of diagnosis to one month after. It was corroborated medically that induction treatment would not have yet resulted in a significant change in biomarker value in this time frame. The hypothesis of using delta samples is that a change or lack of change in biomarker value is more indicative of relapse than the nominal biomarker value. The reasoning is that a delta sample gives information on if and how a patient is responding to treatment. For example, a patient can present with an abnormal creatinine level when first diagnosed. This creatinine level may still be abnormal at the beginning of maintenance therapy however a large difference between these values would show that this patient has responded well to induction therapy and therefore could be less likely to relapse.

As per Figure 9 below, the prevalence of ID's with a delta sample available is small. Hence the utilization of deltas in this research is limited to 69 samples.

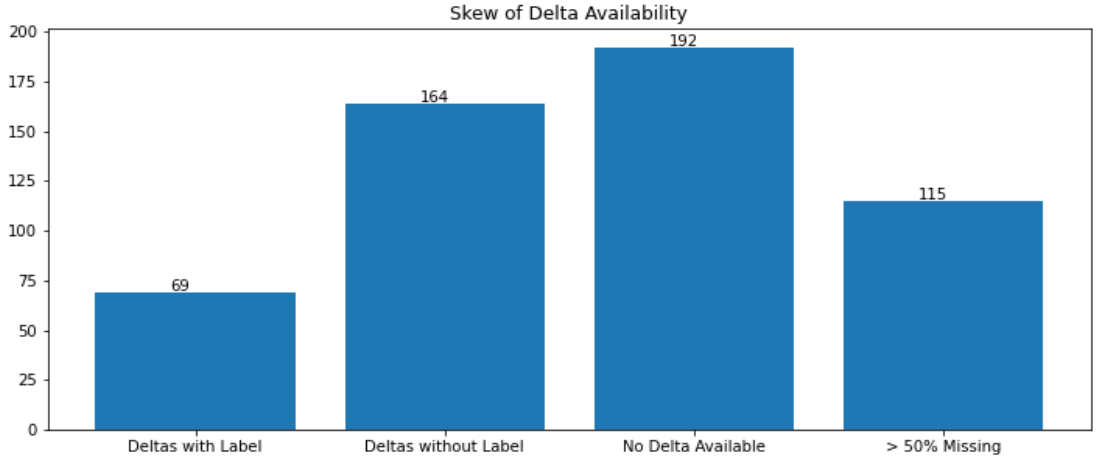

Figure 9 - Delta Availability

Chapter 4

Supervised Learning Approach

The chosen data analytical technique employed was to adopt a supervised learning approach. This approach utilizes labelled data only. As per Figure 4, this results in only approximately 22% of patients being included in this research. Both classification and regression problems were analysed using the labelled data. The goal of the classification was to stratify each patient into one of two categories, LTROT or relapse, as early as possible during their maintenance therapy phase. Three inputs to the model were studied. The substantive sample of biomarker values taken as early as possible during the maintenance phase of their therapy, the PCA groupings as outlined in the section Principal Component Analysis and the delta analysis. The dependent variable was the label of either LTROT or Relapse.

The regression problem utilised only the substantive sample input. The target was the Relapse Rate i.e. predict how many relapses a patient is likely to experience throughout their entire disease course. This analysis was performed on both LTROT and relapsing patients and on relapsing patients alone. The analysis on the relapsing patients alone was limited in size as only 59 patients were labelled as having a relapse.

Pre-processing

Pre-processing is a necessary step in machine learning to ensure the data is correctly formatted for the algorithm employed. The techniques used in this analysis include standardisation, imputation and dimensionality reduction. Standardisation is a common technique used to modify the range of input features to a common scale. Imputation was a required step to account for the data missingness. Dimensionality reduction refers to reducing the number of input variables in the training data.

Dataset and Features

RKD ID	Weight (KG)	CRP	Creatinine	eGFR	Hb	Total white cell count x10^9/L	Neutrophil count x10^9/L	Lymphocyte count x10^9/L	Neutrophil / Lymphocyte ratio	Eosinophil Count x10^9/L	Gender	Stratification
1	78.266667	3.000000	123.666667	56.000000	13.700000	6.733333	5.600000	0.633333	9.031746	0.033333	Male	Relapse
101	66.000000	6.000000	191.000000	30.000000	12.900000	7.000000	3.600000	2.000000	1.800000	0.400000	Male	LTROT
103	54.500000	265.500000	110.000000	42.500000	10.900000	8.550000	7.150000	0.450000	22.714286	0.150000	Female	Relapse
105	80.900000	2.000000	58.000000	90.000000	14.300000	13.300000	11.000000	1.600000	6.875000	0.000000	Female	Relapse
107	65.650000	1.333333	76.333333	65.333333	13.366667	3.933333	2.966667	0.433333	6.916667	0.100000	Female	Relapse

Figure 10 - RKD Registry Clean Dataset

Figure 10 above shows a subset of the cleaned dataset obtained from the RKD database. In total there are 109 observations of the 21 variables presented in Table 1. The continuous variables were standardised to have a mean about 0 and the categorical variables were treated as factors. Stratification, shown in Figure 10, and Relapse Rate were the two dependent variables.

Imputation

As shown in Figure 3 and Figure 11, there was a large number of biomarker values missing, even among those included in the analysis. While there are several options when dealing with data missingness such as a complete case analysis or use of a model which can handle missing data, imputation was conducted to overcome this missingness. Imputation is simply replacement. Imputation techniques are structured on the idea that a data sample of a variable can be replaced by a new randomly chosen sample from an estimate of the distribution of this variable (Donders, et al., 2006). Imputation was performed for several reasons. A complete case analysis would reduce the number of data samples and biomarkers which could be included in the analysis. Simple techniques such as mean imputation and the missing-indicator method have been shown to provide biased results, even when the data is missing completely at random (MCAR) (Donders, et al., 2006). Finally, the data missingness observed in the RKD Registry is missing at random (MAR). Unlike MCAR, where the probability an observation is missing is not related to any other patient characteristic, the missing biomarker values can be considered to be random conditional on other patient characteristics. Thus, multiple imputation was performed whereby several imputed datasets are created and different imputations are based on a random draw from different estimated underlying distributions (Donders, et al., 2006). Averaging over multiple imputations generally lowers the variance as well as allowing one to average the standard errors resulting in an unbiased replacement of the missing data.

Figure 11 below shows the missingness among the biomarkers included in the analysis. Some biomarkers such as Total White Cell Count (White Cell in plot) and Hemoglobin (Hb) have complete cases whereas the Urinalysis biomarkers of both protein and blood have a large degree of missingness. This is likely due to some tests being easier to perform on patients than others. Overall, given the below biomarkers, over 83% of the data is not missing. Approx. 45% are missing the Urinalysis biomarkers and there is no observable pattern to the data missing in the right hand plot of Figure 11. Thus, it can be concluded that the data is indeed MAR.

Figure 12 below, shows the estimated variable distributions for each of the imputations run. There were 10 imputations run and overall the estimate distributions mirror the observed distributions as per Figure 12. The estimated distributions of the Lymphocyte and Neutrophil biomarkers appear to deviate the most from the observed distribution, however for the reasons outlined previously, the final complete data set was assumed to be unbiased and treated as such.

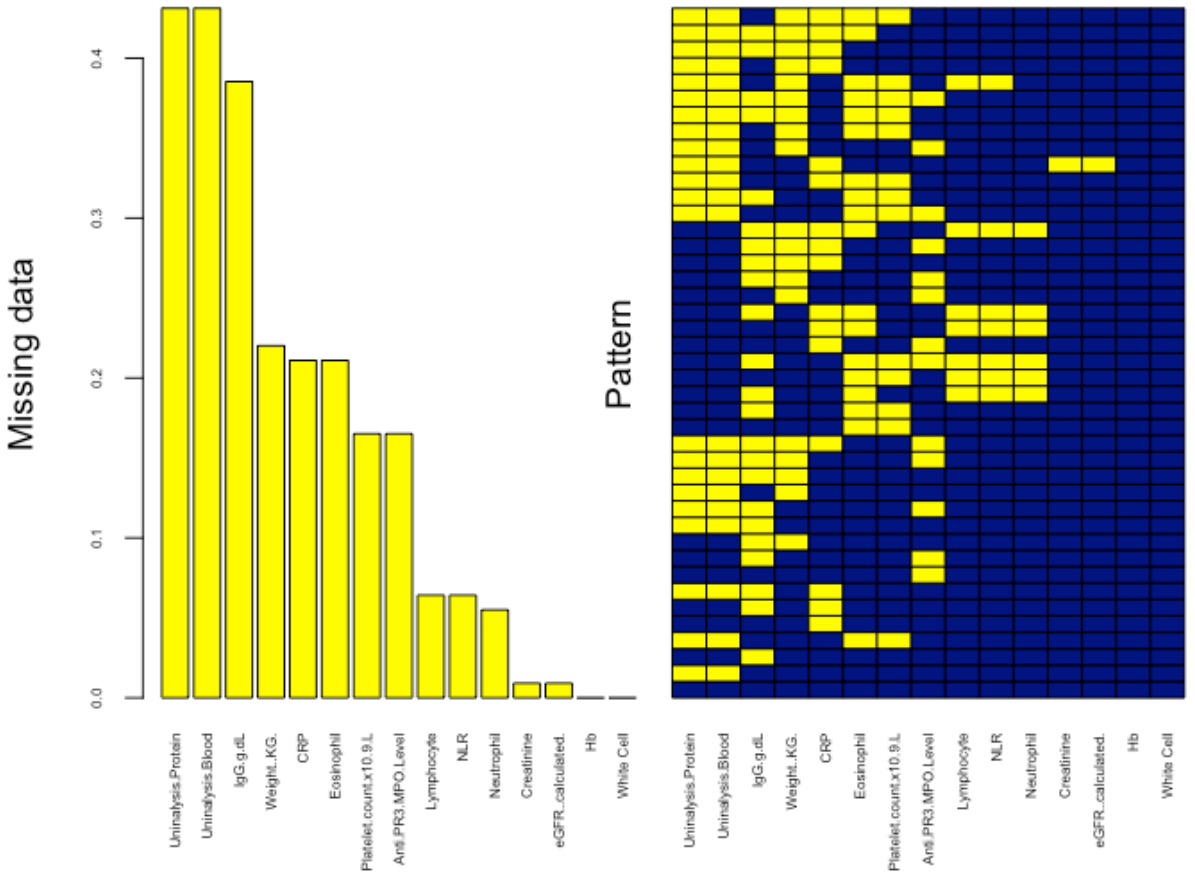

Figure 11 - Biomarker Missing Plot

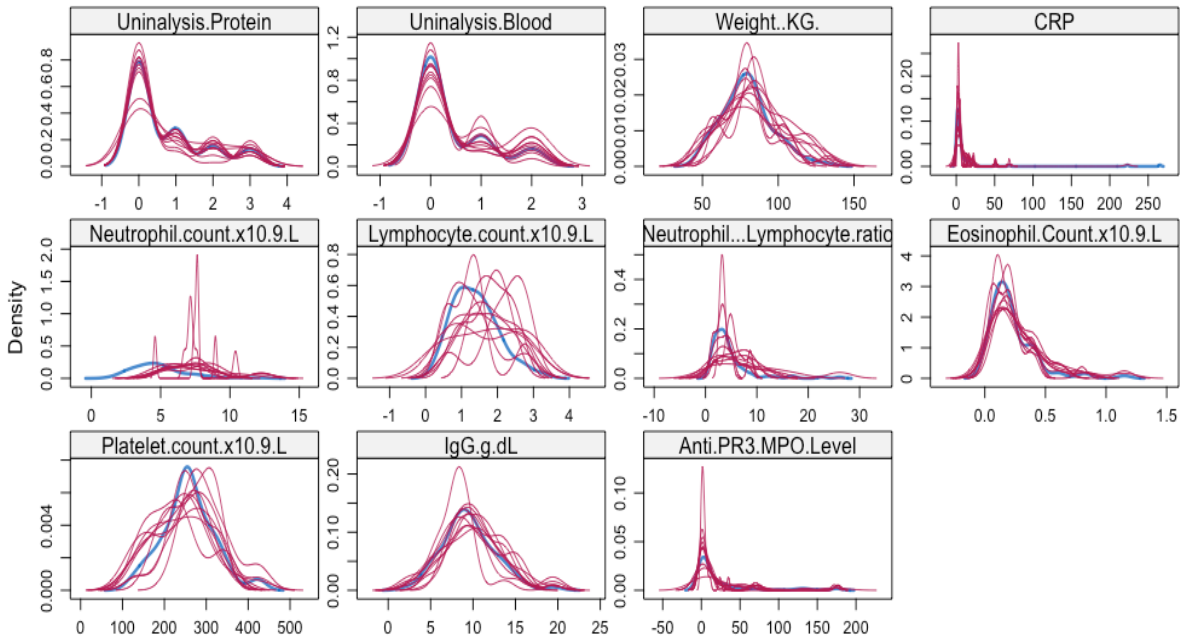

Figure 12 – Imputed Variable Distribution Estimate Plots

Multicollinearity

The issue of multicollinearity is one that required careful consideration during the pre-processing of the data. Multicollinearity arises when two or more highly correlated variables are assessed

simultaneously by a model (Vatcheva, et al., 2016). As can be seen below in Figure 13, there exists some large collinearities between variables A Pearson coefficient of 0.5 or greater was deemed to be strongly correlated, between 0.2 and 0.5 moderately correlated and below 0.2 to be uncorrelated or negligibly correlated. Note, in Figure 13 below, only the largest correlations are shown as all Pearson coefficients are scaled by value. Hence, any values not shown can be deemed to have negligibly small correlation

Multicollinearity causes unstable and biased standard errors leading to unstable p-values for assessing statistical significance of predictors in a regression analysis according to Vatcheva, et al. (Vatcheva, et al., 2016). This also holds true for the feature importance of a classification model. Multicollinearity among the predictor variables can obscure the computation and identification of key independent effects of collinear predictor variables on the outcome variable because of the overlapping information they share (Vatcheva, et al., 2016). Any interpretations or conclusions drawn from the feature importance of a model containing mulitcollinearity are not reliable however the overall result is. In the context of the problem at hand, a significant aspect of this research is to identify which biomarker or combination of biomarkers are significant predictors of relapse. Therefore, ignoring the multicollinearity and focusing only on the accuracy of the model is not an option in this case.

There are several methods for dealing with multicollinearity in a model. The simplest one is to drop the most highly correlated variables as the information they convey is captured by the other variables. Mason, 1987, states however that it is not correct to drop variables from the model unless they are redundant to the program logic and underlying theory (Mason, 1987). Moreover, given the limitation of the data and biomarker availability it is prefereable to utilise all biomarker information unless they are perfectly correlated.

Multicollinearity was handled in two primary methods during the course of this dissertation. The first was the use of a Lasso Logistic Regression model (see Chapter 4 : Lasso Logistic Regression for more information). A Lasso Logistic Regression model can be used to identify feature importance and remove non-predictive and collinear variables from the model. The strength of this technique is its ease of implementation however it can yield unreliable interpretations of features when multicollinearity is involved. The reason for this is that different features would be identified as important for different models due to the high correlations between variables. The second method was to perform a Principal Component Analysis (PCA). PCA is a technique which splits variables into their principal components which are independent of each other. This technique was optimal in maintaining maximal biomarker information while subsequently addressing the issue of multicollinearity as described in the following section.

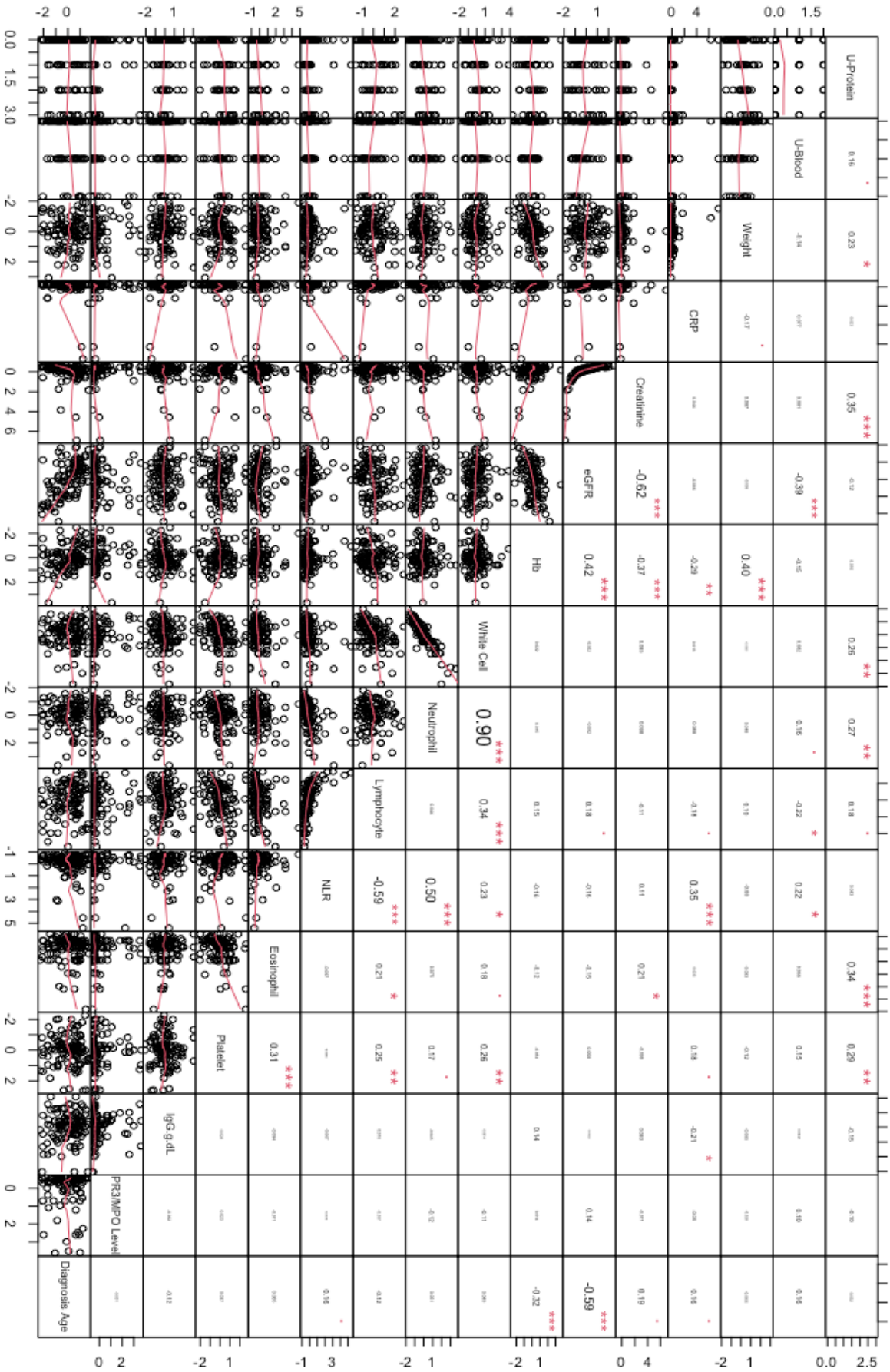

Figure 13 - Multicollinearity Plot

Principal Component Analysis

As discussed in the previous section PCA was used to reduce the multicollinearity between input variables into the classification model. PCA is a technique, not only for handling multicollinearity, but also for reducing the dimensionality of datasets (Jolliffe & Cadima, 2016). Given the extremely small size of the dataset, it was preferable to decrease the number of input features to the predictive model. Minimizing the number of input features reduces the reliance on large datasets and returns more reliable results. The use of PCA not only maintains interpretability of the feature importance but also minimizes information loss. It does so by creating uncorrelated variables, referred to as principal components (PC), while preserving as much variability or statistical information as possible (Jolliffe & Cadima, 2016). In essence PC's are variables that are linear functions of those in the original dataset, that maximise variance but are uncorrelated with each other. Finding such new variables reduces to solving an eigenvalue/eigenvector problem (Vatcheva, et al., 2016).

PCA can only be performed on continuous variables hence the categorical variables and the ordinal variables of Urinalysis Blood and Urinalysis Protein were ignored in this analysis. Usually PCA is performed on the entire dataset however, in order to maintain a higher level of interpretability it was decided to group highly correlated and biologically similar variables together and perform multiple PCA's.

As can be seen in Figure 13, the biomarkers Total White Cell Count, Neutrophil Count, Lymphocyte Count, Neutrophil Lymphocyte Ratio (NLR), Eosinophil Count and Platelet Count were all reasonably highly correlated with at least one other biomarker in that group. Neutrophil Count and Lymphocyte Count were reasonably correlated with each other and highly correlated with NLR. This is because the latter is simply the ratio of Neutrophil and Lymphocyte Count. Total White Cell Count is highly or reasonably correlated with all other biomarkers in this group. The reason for this is due to the fact that Neutrophils, Lymphocytes and Eosinophils are all different sub groups of white blood cells. Platelet is another name for a red blood cell and the Platelet Count was reasonably correlated with the other blood biomarkers hence its inclusion in this group.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.5017	1.3462	0.9927	0.8251	0.46880	0.21617
Proportion of Variance	0.3758	0.3020	0.1643	0.1135	0.03663	0.00779
Cumulative Proportion	0.3758	0.6779	0.8421	0.9556	0.99221	1.00000

Table 2 - Blood Biomarker PCA Group

Table 2 above shows the proportion of variance among the PC's. It can be seen that 95.5% of the statistical information is maintained with only 4 PC's, hence reducing the number of input features by two. In addition the collinearity between the variables has been removed as per the definition of PC's.

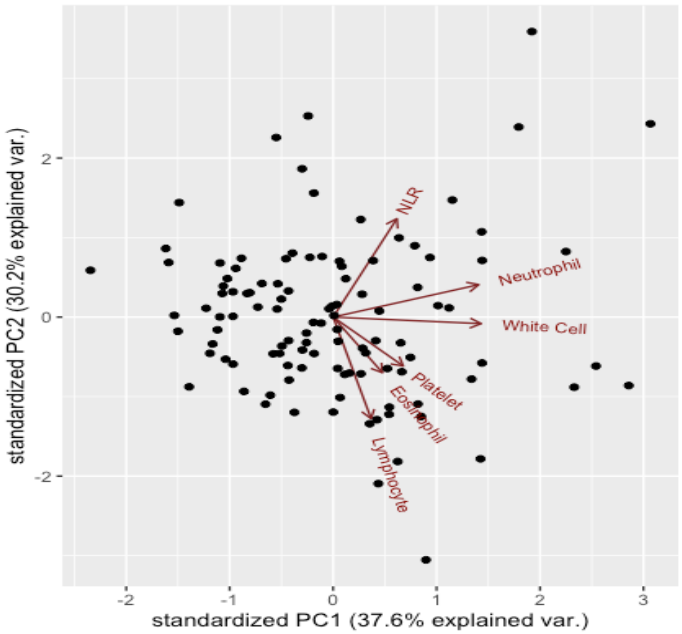

Figure 14 - PCA Blood Biomarker Variation Explanation

Figure 14 highlights how a PCA analysis can maintain interpretability of the dataset. It is clear that Neutrophil Count and White Cell Count are the most important biomarkers for explaining the variance of PC1. Similarly it is NLR and the Lymphocyte Count that explain the largest proportion of variance for PC2. Conversely, White Cell Count has little explained variance for PC2. A similar analysis can be performed for all PC's in this group thus maintaining interpretability of results when the 4 PC's with the largest cumulative proportion are inputted into the model.

A second PCA grouping, consisting mostly of personal characteristics, was also formed containing the variables CRP, eGFR, Creatinine, Weight, Hemoglobin and Age at Diagnosis. These were grouped together as they were highly correlated with at least one other biomarker in the group as per Figure 13 and there is biological reasoning for the collinearity. eGFR and Creatinine are measures of kidney function and as a result were strongly correlated. eGFR was also strongly correlated with Age at Diagnosis. eGFR is a summary statistic of kidney health which takes personal characteristics such as Age, Gender and Creatinine levels into account hence it was unsurprising that these were strongly correlated. Age at Diagnosis, Weight and Hemoglobin were moderately correlated also. People of a higher weight tend to have higher levels of Hemoglobin and weight increases with age hence this relationship was also unsurprising. Finally CRP and Hemoglobin were moderately correlated. CRP is found in blood plasma and whose circulating rise is a marker for inflammation hence the biological connection between the two.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.5382	1.1654	0.9236	0.8940	0.62789	0.47898
Proportion of Variance	0.3943	0.2263	0.1422	0.1332	0.06571	0.03824
Cumulative Proportion	0.3943	0.6207	0.7629	0.8961	0.96176	1.00000

Table 3 - Personal Characteristic PCA Group

Table 3 above shows the distribution of variance among the PC's. As per Table 3, there is 96% information retention in the first 5 PC's, hence the number of input features can be reduced by one and collinearity between variables is removed.

A PC explanation plot, similar to Figure 14, for the above grouping is included in the Appendix. IgG levels and Anti PR3/MPO levels were not included in any PCA grouping as they were not moderately or highly correlated with any other variables in the PCA groupings. Both the PCA groupings and a direct analysis on the biomarkers was performed and compared to conclusively identify the predictive biomarkers.

This following section introduces the predictive models used during the classification and regression problems and reviews their strengths and weaknesses. The below models were selected due to their ease in implementation as well as interpretation of feature importance.

Bayesian Logistic Regression

Introduction to Algorithm

Bayesian Inference for Logistic Regression consists of three core steps common to all Bayesian analyses. The first step is to specify the log-likelihood function for the model. In this case, the log likelihood function is a logistic regression function. The second is to form prior distributions for all unknown parameters and the final step is to use Bayes Theorem, shown below, to find the posterior distributions over all parameters.

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

The posterior distributions were found using a common class of algorithms called Markov Chain Monte Carlo (MCMC). MCMC starts with the predefined prior distributions from which samples are taken. Markov chains then guide these samples towards the posterior distribution using the logistic regression log-likelihood function and Bayes Theorem (Kana, 2020).

The prior distributions of the unknown parameters were left as uninformative binomial distributions. Binomial distributions were chosen as the mean does not have to be centered at 0 and the range

of the standard deviations can be large. Thus, this distribution is very uninformative and unlikely to bias the posterior distribution. While, providing informative prior distributions of parameters can lead to higher accuracies it can also lead to biases. Utilising uninformative priors allows for the Bayesian model to provide unbiased posterior distributions which result in conclusive feature importance's.

The advantages of Bayesian Logistic Regression are that they can handle incomplete datasets, they can prevent overfitting of data and there is no need to remove contradictions from the data (Ray, 2019). Another significant advantage of Bayesian Logistic Regression is that it provides confidence of its estimate predictions (Kana, 2020). Conversely, the selection of a prior can be difficult and can lead to biases of the posterior distribution and it can be computationally expensive (Ray, 2019).

Final Dataset and Features

Substantive Sample Dataset

Of the 21 biomarkers available for analysis, shown in Table 1, only 7 biomarkers, which are shown in Table 4 below, were kept in the final model. Biomarkers were removed iteratively if deemed not important. Pairs of biomarkers with a high collinearity were removed and subsequently re-added to ensure only the optimal biomarkers were kept.

U-Protein	Hb	White Cell	Eosinophil	PR3/MPO Level	Disease Sub	Treatment	Stratification
3	0.258	-0.219	-1.027	-0.567	MPA	CYC	Relapse
0	-0.207	-0.108	0.992	0.266	MPA	Other	LTROT
0	-1.370	0.539	-0.385	-0.329	MPA	CYC	Relapse
0	0.607	2.522	-1.210	-0.507	EGPA	Other	Relapse
0	0.064	-1.388	-0.660	0.551	MPA	CYC	Relapse
1	0.491	0.435	-0.109	-0.555	MPA	CYC	Relapse

Table 4 - Final Input Dataset

Figure 15 below, shows the MCMC trace plot of the 7 biomarkers kept in the model. The trace plot shows that all chains did converge with the average of the 4 chains being similar for all variables. This is ideal behaviour of the Bayesian Logistic Regression model. Note the trace plots of the

subsequent PCA and delta groupings also showed good behaviour and are included in the Appendix.

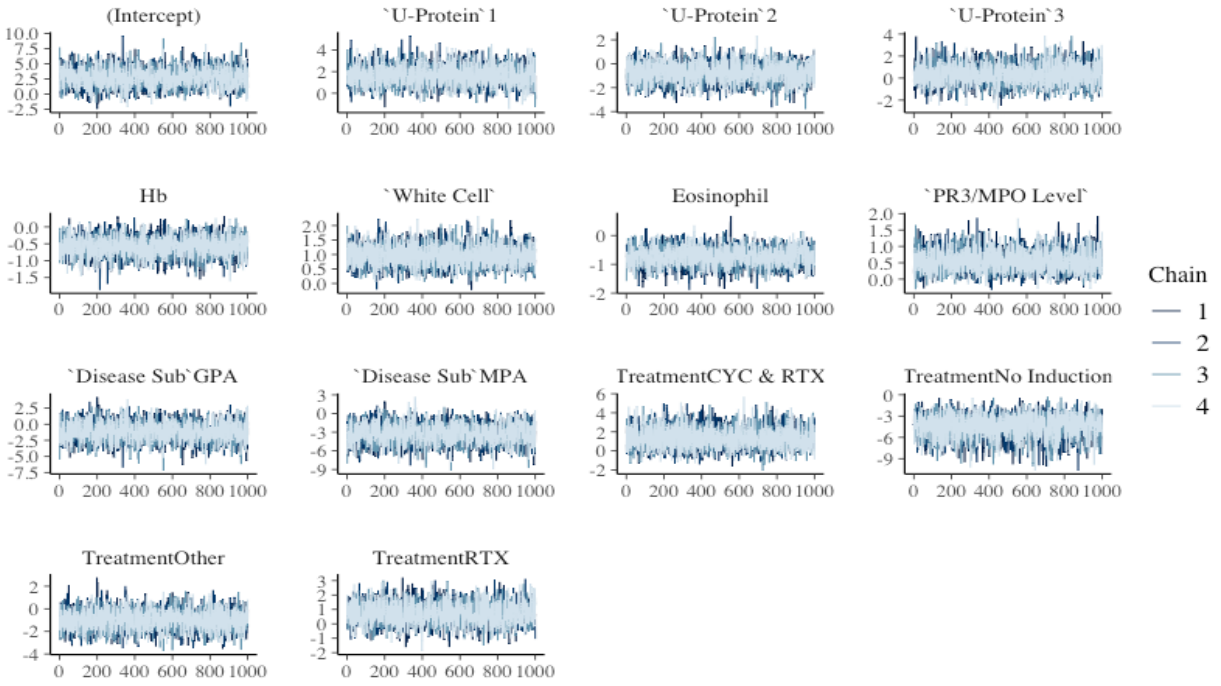

Figure 15 - Bayesian Logistic Regression MCMC Trace Plot Important Features

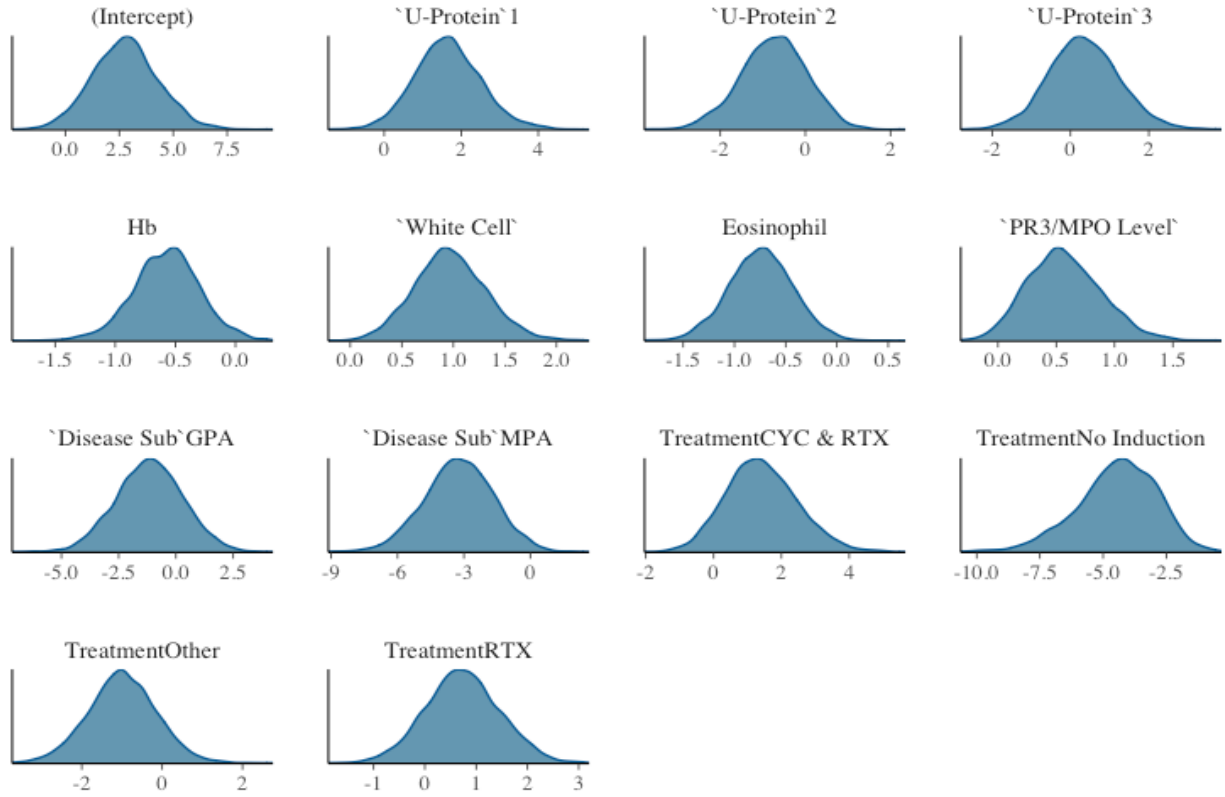

Figure 16 - Bayesian Logistic Regression Posterior Distribution Density Plots Important Features

Figure 16 above shows the posterior distributions of the features deemed important by the Bayesian Logistic Regression algorithm. A mean greater than zero signifies this feature increases the risk of relapse and vice versa for a mean lower than zero. The mean of the intercept is greater than zero which reflects the imbalance in the dataset with more Relapsing patients than LTROT present in the final dataset.

From Figure 16 it can be concluded that patients with a lower amount of protein (`U-Protein1`), a higher total white cell count (`White Cell`), a higher ANCA titre (`PR3/MPO Level`), those who received a CYC and RTX (`TreatmentCYC&RTX`) and those who received only RTX (`TreatmentRTX`) during induction treatment are more likely to relapse. Conversely, those with a higher Hemoglobin (`Hb`), a higher Eosinophil Count, those with a clinical subtype of MPA AAV (`Disease SubMPA`) and those who received No Induction treatment (`TreatmentNoInduction`) are less likely to relapse. Both `U-Protein1` and `TreatmentCYC&RTX` are the biggest indicators of relapse while `TreatmentNoInduction` is the most predictive of LTROT.

PCA Groupings Dataset

Table 5 and Table 6 below show the final inputs to the Bayesian Logistic Regression models for the PCA groupings. All blood biomarker PC's were retained by the model however only 3 of the 5 PC's from the personal characteristic PC's were deemed important. There was significant overlap in information with the Substantive Sample group with only the IgG biomarker being included as an additional biomarker in this analysis. This method resulted in 12 input features to the model with only Urinalysis Blood deemed uninformative by the model and subsequently being dropped. The trace plot shown in Figure 35, located in the Appendix, showed convergence as well as consistent averaging of the chains.

Blood_PC1	Blood_PC2	Blood_PC3	Blood_PC4	Char_PC2	Char_PC3	Char_PC5
-0.393	-2.011	0.152	0.074	-0.106	-0.214	-0.242
-0.560	1.187	0.322	-1.265	0.158	-0.409	-0.781
2.663	-3.007	-2.621	2.069	4.777	4.437	0.157
3.102	-1.202	2.115	0.506	0.023	-0.142	-0.044
-2.269	-1.791	-0.197	-0.761	0.489	-0.545	-0.479
-0.009	-0.895	0.957	-1.134	-0.905	0.107	-0.850

Table 5 - PCA Inputs to Bayesian Logistic Regression part 1 of 2

U-Protein	IgG.g.dL	PR3/MPO Level	Disease Sub	Treatment	Stratification
0	-1.368	-0.552	MPA	CYC	Relapse
0	-0.058	0.276	MPA	Other	LTROT
0	-1.461	-0.316	MPA	CYC	Relapse
0	-0.959	-0.552	EGPA	Other	Relapse
0	-0.366	0.561	MPA	CYC	Relapse
1	-0.879	-0.552	MPA	CYC	Relapse

Table 6 - PCA Inputs to Bayesian Logistic Regression part 2 of 2

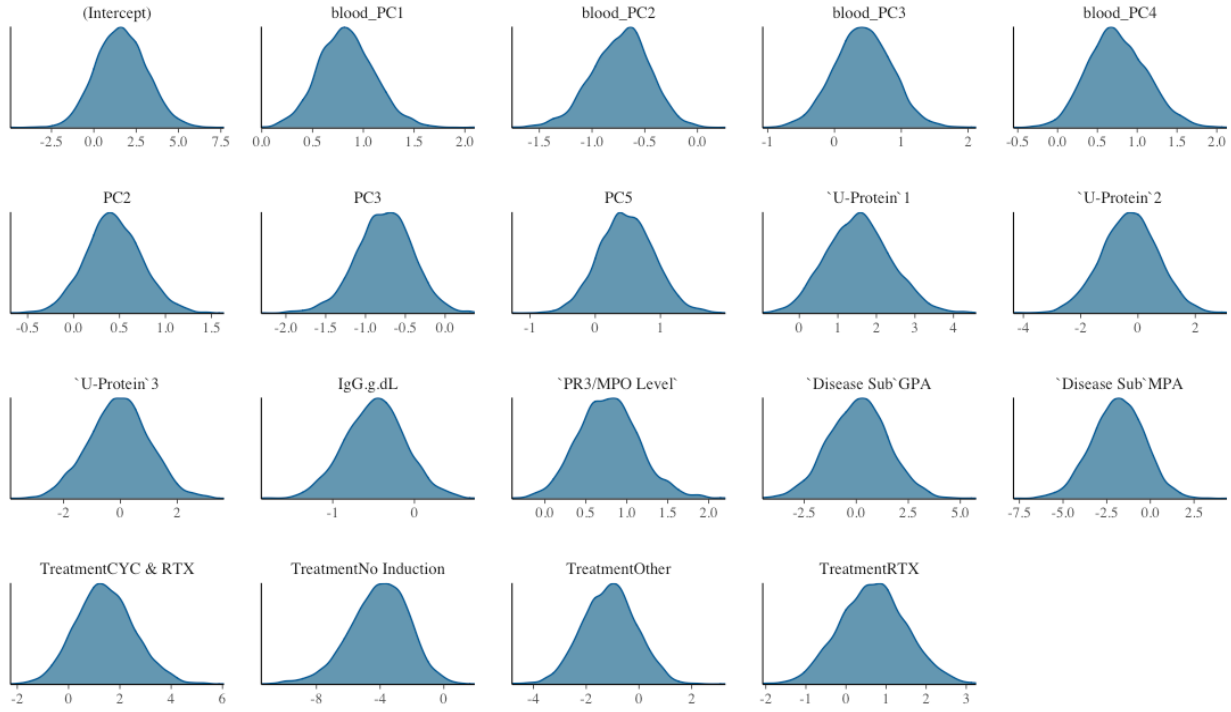

Figure 17 - Bayesian Logistic Regression Posterior Distribution Density Plots PCA Group

Analysing Figure 17 above, once again the mean of the posterior distribution of the intercept is greater than 0, reflecting the imbalance in the dataset with more Relapse patients present. `blood_PC1`, `blood_PC3`, `blood_PC4`, `PC2`, `PC5`, `U-Protein 1`, `PR3/MPO Level`, `TreatmentCYC & RTX` and `TreatmentRTX` are all positive indicators of Relapse. Conversely, all other biomarkers are indicators of LTROT. There is consistency between the important features of both the PCA and Substantive Sample datasets which shows that the issue of multicollinearity does not appear to have a large effect on the data. This is discussed further in the section Discussion.

Delta Analysis Dataset

Table 7 and Table 8 below were combined with 21 biomarkers available in the Substantive Sample to form the inputs for the delta analysis. These two tables show all the biomarkers for which there was sufficient data available to obtain a delta. As per Figure 9, this analysis was limited in sample size to just 69 samples due to the availability of deltas. This resulted in 69 observations of 36 potential features. Of the 36 potential features, 18 were kept in the final model shown in Figure 18 below. Identical to the analysis on the Substantive Sample dataset, biomarkers were removed iteratively if deemed not important by the Bayesian Logistic Regression model. Biomarkers with a high collinearity were removed and subsequently re-added to ensure only the optimal biomarkers were kept.

CRP_delta	Creatinine_delta	eGFR_delta	White Cell_delta	Neutrophil_delta
0.441	0.542	-1.032	-0.310	-0.256
1.464	-0.862	0.743	-0.026	-0.217
-4.012	-0.397	0.010	-0.612	-0.973
0.804	-0.452	-0.183	1.461	1.696
1.044	0.674	-0.145	-0.095	0.133
0.441	-0.008	-0.550	-0.624	-0.839

Table 7 - Delta Inputs to Bayesian Logistic Regression part 1 of 2

Lymphocyte_delta	NLR_delta	Eosinophil_delta	Platelet_delta	ANCA_delta	Hb_delta
-0.116	-0.082	-0.026	0.219	0.571	-0.460
0.547	-0.328	-0.839	0.917	0.442	-0.210
1.154	-1.999	-0.396	-1.211	0.485	1.390
-0.375	3.238	-0.237	-0.003	-0.951	-0.493
-0.590	0.228	-0.554	-0.504	0.244	-1.360
-0.073	-0.383	1.030	-0.054	-0.040	1.040

Table 8 - Delta Inputs to Bayesian Logistic Regression part 2 of 2

Although the dataset is once again imbalanced with a higher number of Relapse than LTROT patients, the mean of the posterior distribution of the intercept is less than 0. Usually, this would indicate a prevalence of LTROT patients in the dataset, however in this case, there are some extremely strong positive features which are indicative of relapse which are likely skewing the mean of the intercept to be below zero.

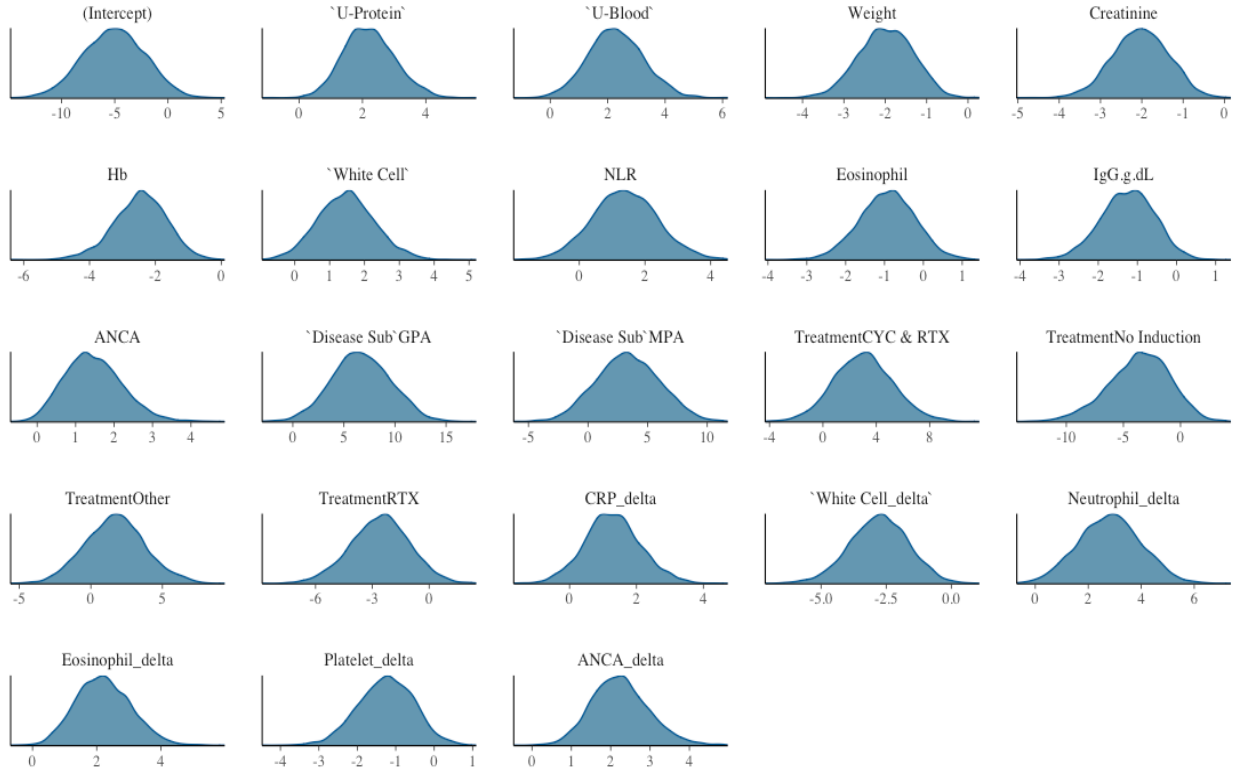

Figure 18 - Bayesian Logistic Regression Posterior Distribution Density Plots Delta Group

As per Figure 18 `U-Protein`, `U-Blood`, `White Cell`, `NLR`, `ANCA`, `Disease Sub GPA`, `Disease Sub MPA`, `Treatment CYC & RTX`, `CRP_delta`, `Neutrophil_delta`, `Eosinophil_delta` and `ANCA_delta` are all positive indicators of relapse. All other biomarkers are negative indicators of relapse i.e. they are indicators of LTROT. As can be seen in Figure 18 some posterior distributions such as `ANCA_delta` do not contain 0 in the range of their distribution. This is very significant as the model has identified this feature of having a negligible chance of being non-predictive. In other words one can be almost 100% confident that this feature is predictive of being of LTROT or Relapse.

Comparison of Results

This section compares the results of all three datasets using the Bayesian Logistic Regression model against a dummy model which predicts the most popular class. Metrics are shown below in Table 9 and comparative plots are subsequently shown also.

Substantive Sample					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.762
Prediction	LTROT	29	10	Specificity	0.844
	Relapse	16	54	Sensitivity	0.644
PCA Groupings					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.762
Prediction	LTROT	30	11	Specificity	0.823
	Relapse	15	53	Sensitivity	0.667
Delta Analysis					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.957
Prediction	LTROT	27	1	Specificity	0.975
	Relapse	2	39	Sensitivity	0.931
Dummy Comparison					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.587
Prediction	LTROT	0	0	Specificity	1.00
	Relapse	45	64	Sensitivity	0.00

Table 9 - Bayesian Logistic Regression Comparison of Results

Table 9 above shows a comparison of the Substantive Sample, PCA group and Delta Analysis versus a dummy model. It is clear from all three datasets that the Bayesian Logistic Regression model significantly outperforms the dummy the comparison with a minimum accuracy of 76.2% for the PCA group compared to 58.7% for the dummy model. There is no difference in accuracies between the PCA group and Substantive Sample group despite the number of features of the PCA group being larger than the Substantive Sample. It can be concluded from this that the statistical information or variance captured by the additional PC's are either captured by the features of the Substantive Sample or do not help in the classification of LTROT from relapse. The accuracy of the Delta Analysis is extremely high at 95.7%. This dataset is clearly significantly easier to classify than either the PCA group or Substantive Sample datasets and highlights the importance of deltas in the classification of LTROT and Relapse. There is a possibility of overfitting occurring on the Delta Analysis dataset due to the large number of features and small sample size and this is discussed further in the section Discussion. The metrics of Specificity and Sensitivity show that the Relapse class is significantly easier to classify than the LTROT class. This could be due to the imbalance in the dataset but it also reflects the complexity and difficult nature in predicting LTROT among patients who have ANCA-associated vasculitis. Overall, the results in Table 9 are conclusive that the information contained in the RKD Registry can be used to differentiate LTROT from Relapse with the inclusion of deltas being of significant importance.

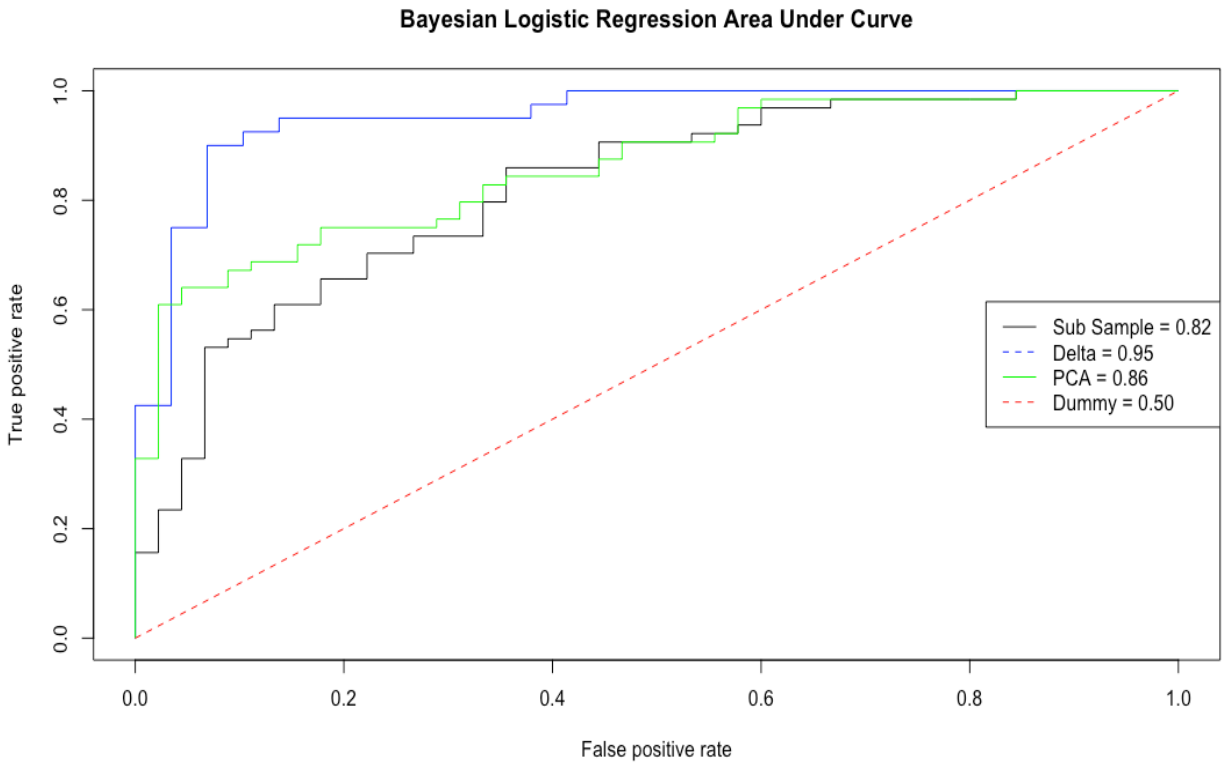

Figure 19 - Bayesian Logistic Regression AUC Comparison

Figure 19 above reflects the findings of Table 9. The PCA grouping slightly outperforms the substantive sample when predicting the LTROT label. This increase in performance is marginal however it could suggest that some personal characteristics are important when classifying this label as this is the main differentiation between the two datasets. Once again the impressive predictive performance of the Delta Analysis is clear to see in Figure 19. An Area Under the Curve (AUC) of 0.95 is close to perfect performance. When compared to the dummy model it is clear that the Bayesian Logistic Regression model is performing well and finding patterns within the dataset which can stratify AAV patients into the categories of LTROT and Relapse.

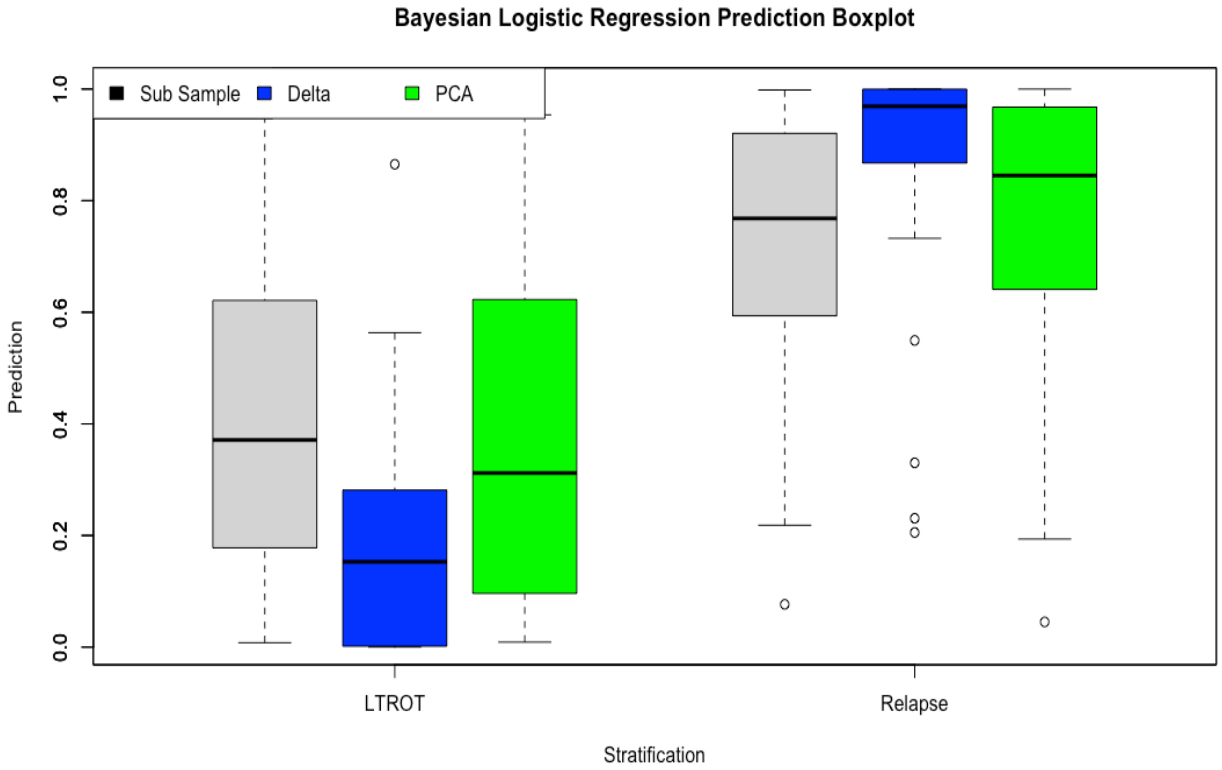

Figure 20 - Bayesian Logistic Regression Classification Certainty Boxplot Comparison

Figure 20 above shows a comparative boxplot of each models predictions. Predictions greater than 0.5 are classified as Relapse and the opposite is true for predictions less than 0.5. The closer the prediction is located to 0.5 the less certain that model was about its predictions of that class. The delta analysis in blue shows a clear divide between its predictions of LTROT and Relapse with several outliers. This divide reflects the great performance of this model. The range of the LTROT prediction are from approx. 0.0 to 0.3 which highlights this models confidence in its predictions. Similarly the range of Relapse predictions is also very narrow. Conversely, it is clear that the Bayesian Logistic Regression is significantly less confident in its predictions of both the Substantive Sample group in grey and the PCA group in green. These boxplots have a much wider range in their predictions which shows the models uncertainties in some of its predictions.

The above analyses was performed on the entire dataset due to the restriction in the size of the dataset. Figure 21 below shows a comparison in how the models generalise when the dataset is split into a training and a test set. It is clear from Figure 21 that the models do not generalise well. Both models suffer a significant drop in accuracy when the datasets are split into a training and a test set. The accuracy of the Substantive Sample model drops from 76% to a maximum of 66% on the 90:10 split while the accuracy of the Delta Analysis model drops from 96% to a maximum of 76% on the 80:20 split. It can be concluded however that the Delta Analysis is significantly more accurate than the Substantive Sample. As can be seen in the 80:20 split of Figure 21, there is no overlap in the boxplots which suggests that even taking standard deviations into account, the Delta Analysis is the more accurate model. From Figure 21, it is clear that the standard deviation of all models is quite large meaning there is a large amount of uncertainty in the true accuracy of the test set. There are two possible reasons for this. The first is simply due to the small size of the datasets. The Substantive Sample with a 70:30, 80:20 and 90:10 split has test set sizes of 33, 22 and 11. The Delta Analysis with a 70:30, 80:20 and 90:10 split has test set sizes of 21, 14 and 7 respectively. Given such small test set and training set sizes the models did not have sufficient data to train properly and also even a small number of misclassified labels could result in large errors. The second possible reason would be due to overfitting on the entire dataset. Given a high number of features and small dataset it is easy to overfit models. When a model is overfit it performs very poorly on unseen datasets as it does not capture the overall behaviour of the model. These issues are discussed further in the Discussion.

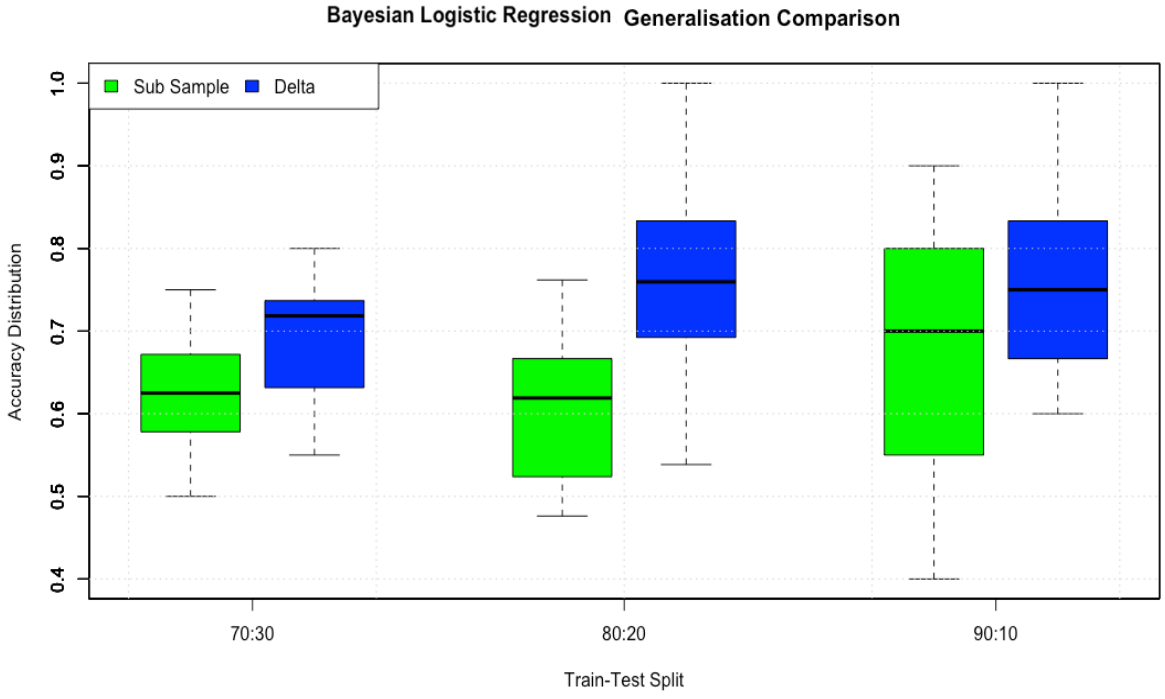

Figure 21 - Bayesian Logistic Regression Generalisation Comparison

Figure 22 below shows the models insensitivity to the imputation performed. The accuracies observed in each of the three datasets showed good resistance to imputation and highlighted the effectiveness of the multiple imputation performed. From Figure 22 it can be concluded that imputation did not impact the accuracies or the results observed for the Bayesian Logistic Regression model.

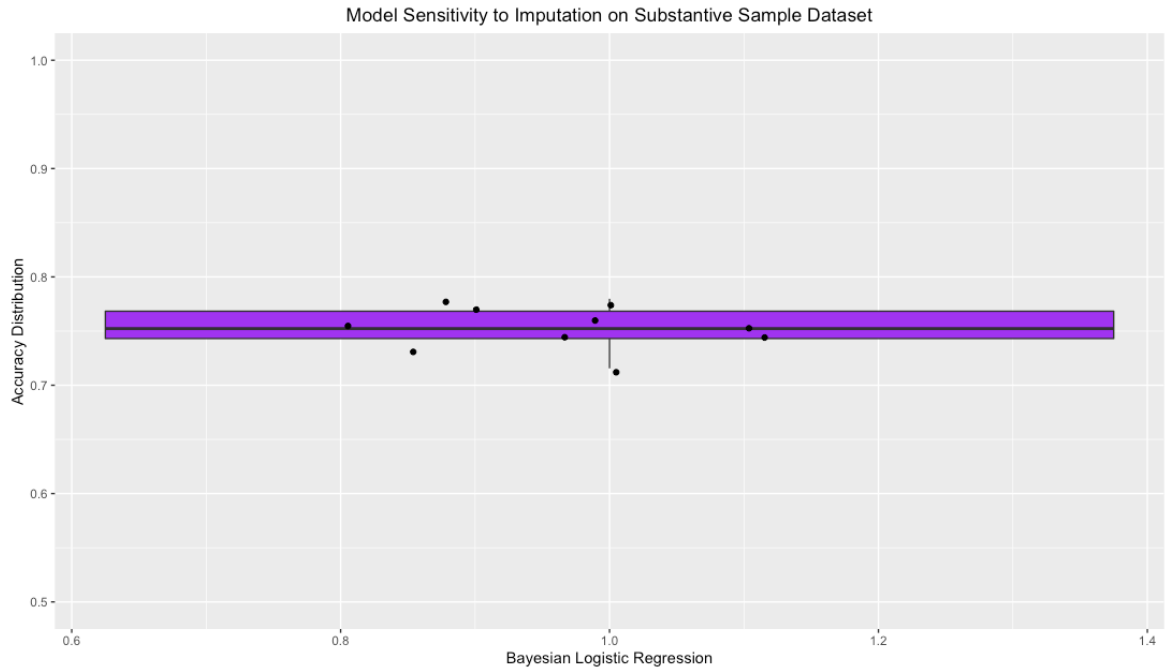

Figure 22 – Bayesian Logistic Regression Model Sensitivity to Imputation on Substantive Sample

Lasso Logistic Regression

Introduction to Algorithm

Least Absolute Shrinkage and Selection Operator (Lasso) Logistic Regression is a technique employed which modifies the log-likelihood by adding a continuous penalty function of the parameters (Makalic & Schmidt, 2010). This penalty function, commonly referred to as a L1 penalty, simultaneously performs parameter shrinkage and pertinent variable selection resulting in a sparse model (Makalic & Schmidt, 2010). The L1 penalty, shown below, generates a sparse model which is easier to interpret.

$$L1 = \sum_{i=1}^{P} |\beta_i|$$

The advantage of Lasso Logistic Regression over basic Logistic Regression is the tendency for Logistic Regression to overestimate parameters resulting in poor predictions. The disadvantage of Lasso is the unpredictability of feature selection among highly correlated predictors (Makalic & Schmidt, 2010). When a group of features are highly correlated, the Lasso algorithm tends to

randomly include one predictor from the group and ignore the other predictors. As there was significant collinearity found among the predictors as per Figure 13, this was an important issue however it was addressed through the use of PCA. Another solution to address the issue of multicollinearity would be to utilise a Ridge Logistic Regression model. A Ridge Logistic Regression model uses a L2 penalty which results in features which are small in value. This approach was not favourable however as there is little interpretability of feature importance and there is no dimensionality reduction both of which are highly relevant to this research.

Final Dataset and Features

Unlike Bayesian Logistic Regression, Lasso models do not differentiate between different factors of categorical variables. For example, the Bayesian Logistic Regression compared MPA and GPA to EGPA, thus this categorical was split into two features as per Figure 18. Lasso Logistic Regression considers all levels of categorical features as a single feature, thus categorical features were inputted in R as a numeric feature.

Substantive Sample Dataset

Of the 21 biomarkers, 9 biomarkers, shown below in Table 10 and Table 11, were retained in the final model. Similarly to the Bayesian Logistic Regression, biomarkers were removed iteratively if deemed not important. Pairs of biomarkers with a high collinearity were removed and subsequently re-added to ensure only the optimal biomarkers were kept.

Table 10 and Table 11 also show the weight of that biomarkers coefficient in the Lasso Logistic Regression model. Coefficient values greater than 0 indicate an increased risk of relapse, while values less than 0 decrease the risk of relapse. Higher coefficient values also have an increased weight in the model which can be used to determine feature importance. From Table 10 and Table 11 it can be concluded that the presence of protein in urine, a higher White Cell count, a higher Platelet count and increases in ANCA titre all increases the risk of relapse. The converse is true of Haemoglobin count, Eosinophil count, IgG and the disease subgroup of EGPA.

Biomarker	Intercept	U-Protein	Hb	White Cell	Eosinophil
Coefficient	2.72	0.15	-0.35	0.44	-0.412

Table 10 - Lasso Substantive Sample Model Coefficients Part 1 of 2

Biomarker	Platelet	IgG	ANCA titre	Disease Sub	Treatment
Coefficient	0.27	-0.377	0.21	-0.92	-0.05

Table 11 - Lasso Substantive Sample Model Coefficients Part 2 of 2

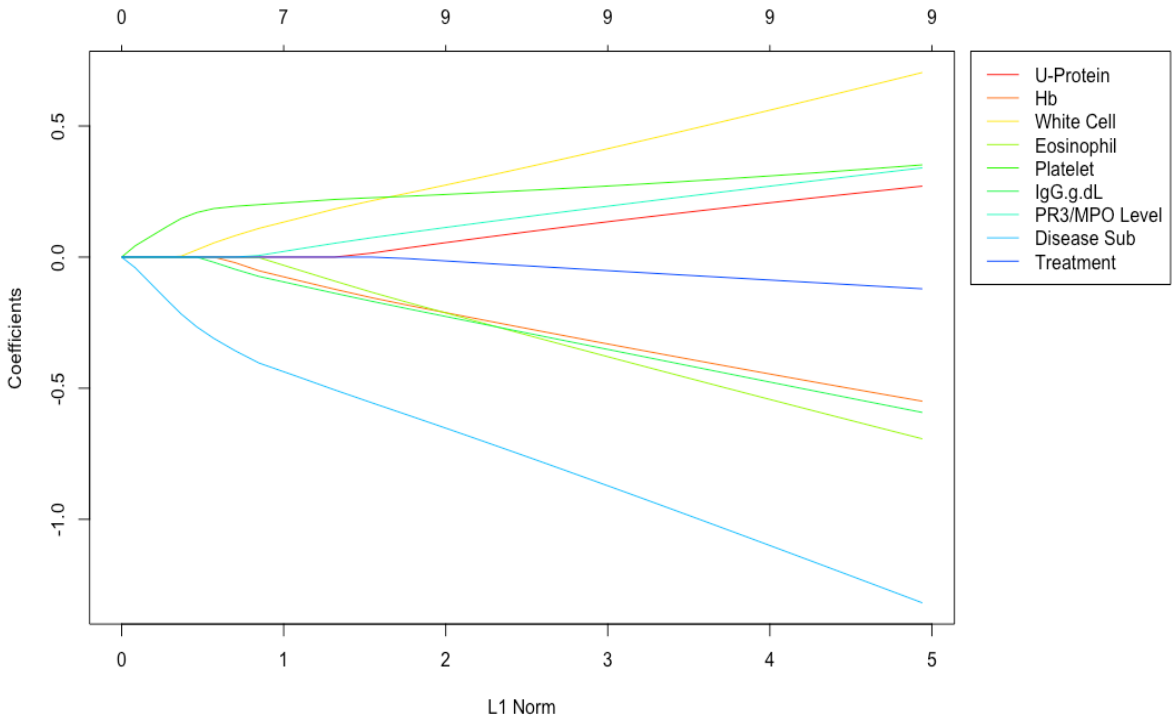

Figure 23 - Lasso Substantive Sample Feature Importance

Figure 23 above shows a plot of the feature importance for the Substantive Sample dataset. Consistent with the tables above it is clear that a high White Cell count is the most important risk factor for relapse. Conversely, patients who are diagnosed with subtype EGPA are less likely to relapse. This can be concluded as Disease Sub is negatively associated with relapse and the EGPA corresponds to the zeroth level of the Disease Sub biomarker.

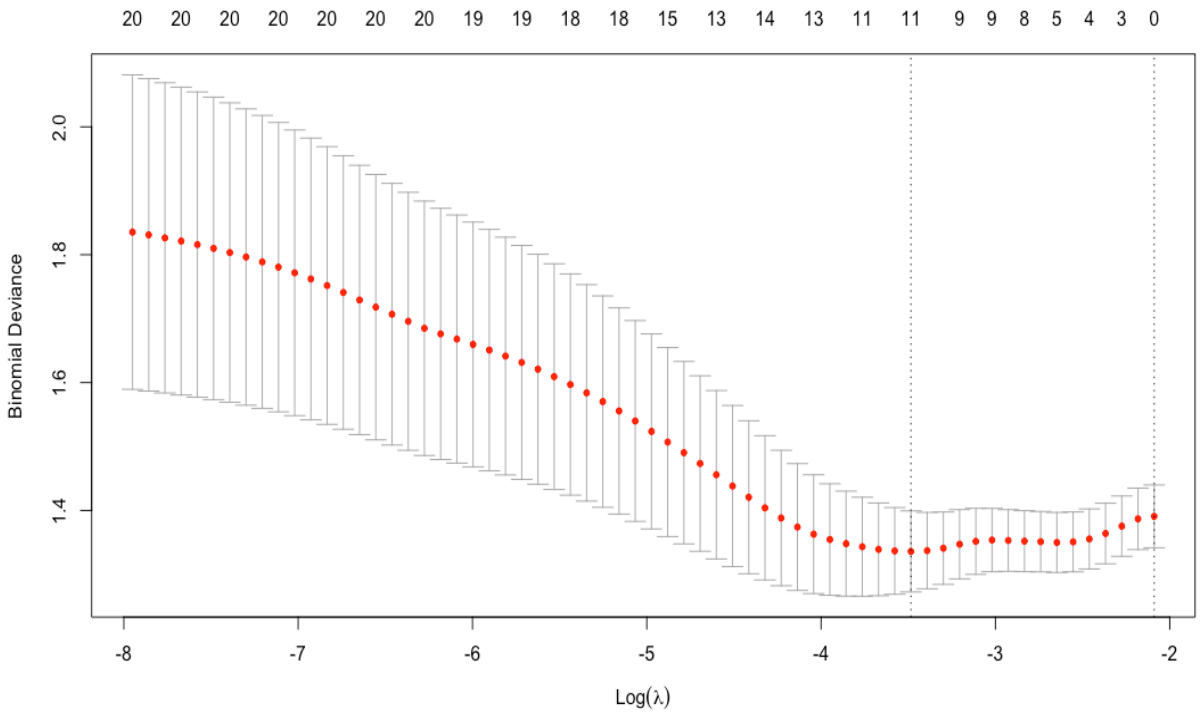

Figure 24 - Lasso Substantive Sample Feature Selection Cross Validation

Figure 24 above shows a cross validation plot generated for the Lasso Logistic Regression model. The cross validation plot has identified 11 biomarkers, as per the top axis, as the optimal number of features to minimize the binomial deviance. Despite Figure 24 identifying 11 biomarkers as optimal there was significant collinearity between Neutrophil and White Cell count and also Lymphocyte and White Cell count. As a result, keeping all three biomarkers in the model was suboptimal and did not result in an increase in accuracy. Neutrophil Count and Lymphocyte Count were dropped as a result however either could be included as a replacement for White Cell while maintaining similar accuracy. The bottom axis shows the log of the coefficient shrinkage λ. The value of λ which minimizes the error rate was 0.03.

PCA Groupings Dataset

Biomarker	Intercept	U-Protein	IgG	ANCA Titre	Disease Sub
Coefficient	2.563	0.073	-0.343	0.333	-0.867

Table 12- Lasso PCA Model Coefficients Part 1 of 2

Biomarker	Blood_PC1	Blood_PC2	Blood_PC4	PC1	PC2	PC3	PC5
Coefficient	0.507	0.360	-0.355	-0.031	0.210	-0.164	0.512

Table 13 - Lasso PCA Model Coefficients Part 2 of 2

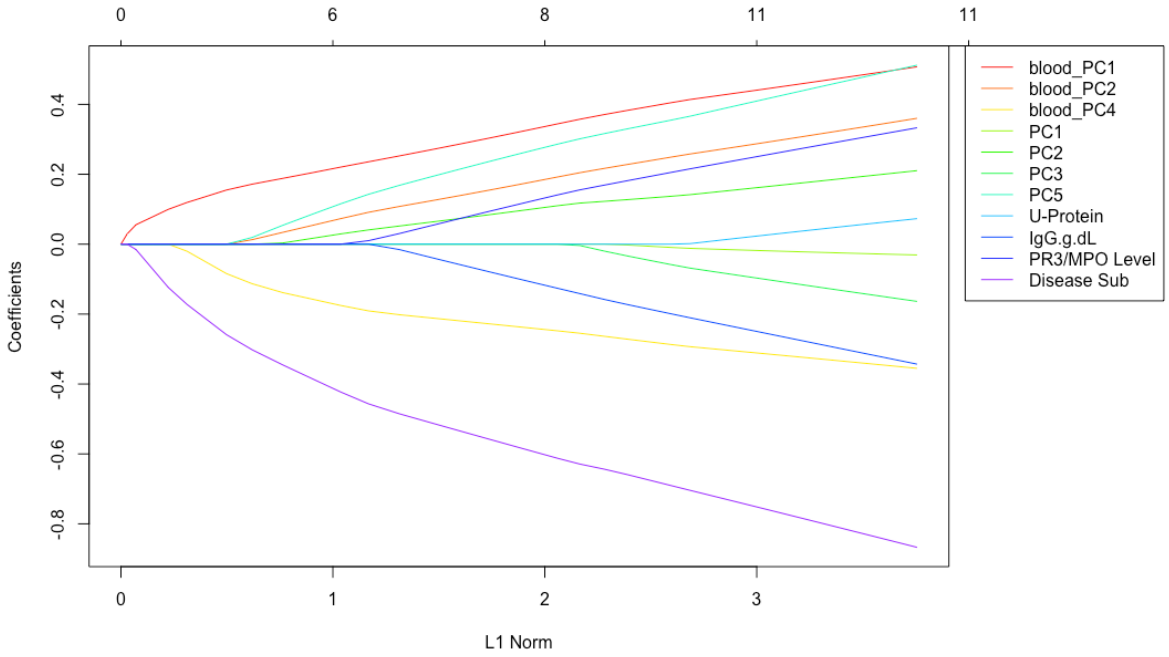

Figure 25 - Lasso PCA Feature Importance

Figure 25, Table 12 and Table 13 show the feature importance for the PCA Grouping Dataset. This model utilizes 11 biomarkers in total including three PC's from the blood grouping and 4 PC's

from the personal characteristic group. Of the 5 blood PC's 2 were deemed not important while only one personal characteristic PC was dropped from the model. There is consistency in the feature importance between the PCA Grouping and the Substantive Sample with the addition of the IgG biomarker being included in the PCA Grouping model being the main difference.

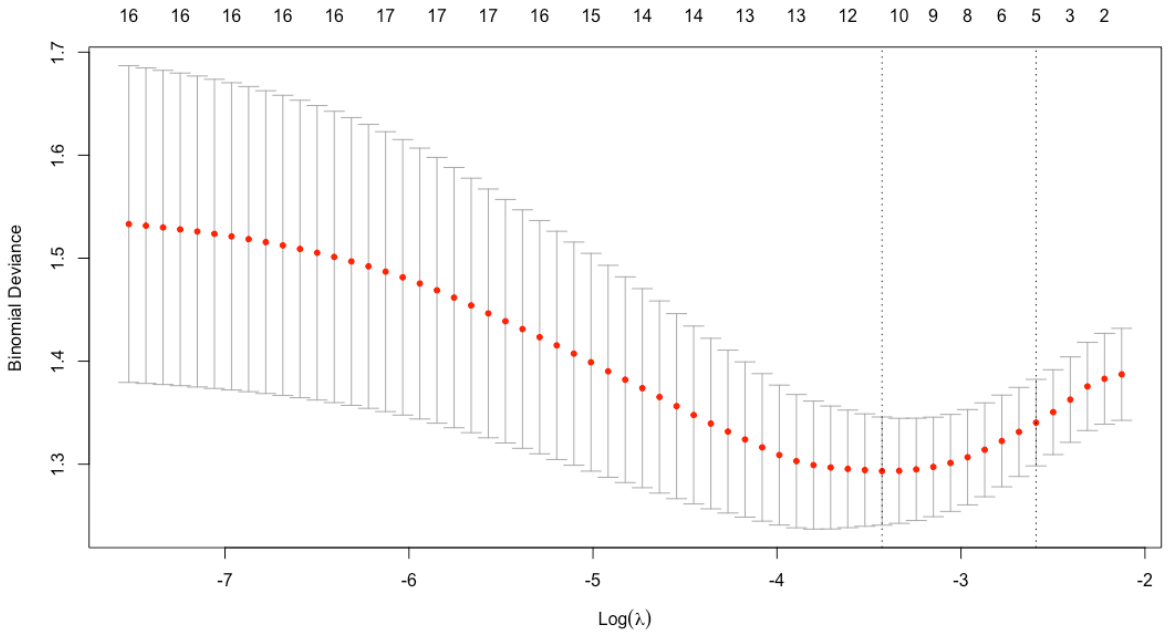

Figure 26 - Lasso PCA Feature Selection Cross Validation

Figure 26 is also similar and consistent with the findings in the Substantive Sample dataset with 11 biomarkers minimizing the binomial deviance at a value of λ equal to 0.03. Unlike the Substantive Sample dataset however all 11 important biomarkers are kept in the model. This highlights the advantage of utilising the PCA approach as by definition the PC's are independent of each other meaning there is no collinearity between them.

Delta Analysis

Biomarker	Intercept	U-Protein	Hb	White Cell	Eosinophil	IgG
Coefficient	1.741	0.443	-0.356	0.981	-0.234	-0.535

Table 14 - Lasso Delta Model Coefficients Part 1 of 2

Biomarker	Disease Sub	Creatinine Delta	eGFR Delta	Neutrophil Delta	Eosinophil Delta	Platelet Delta
Coefficient	-0.609	-1.246	-0.579	1.012	0.897	-0.853

Table 15 - Lasso Delta Model Coefficients Part 2 of 2

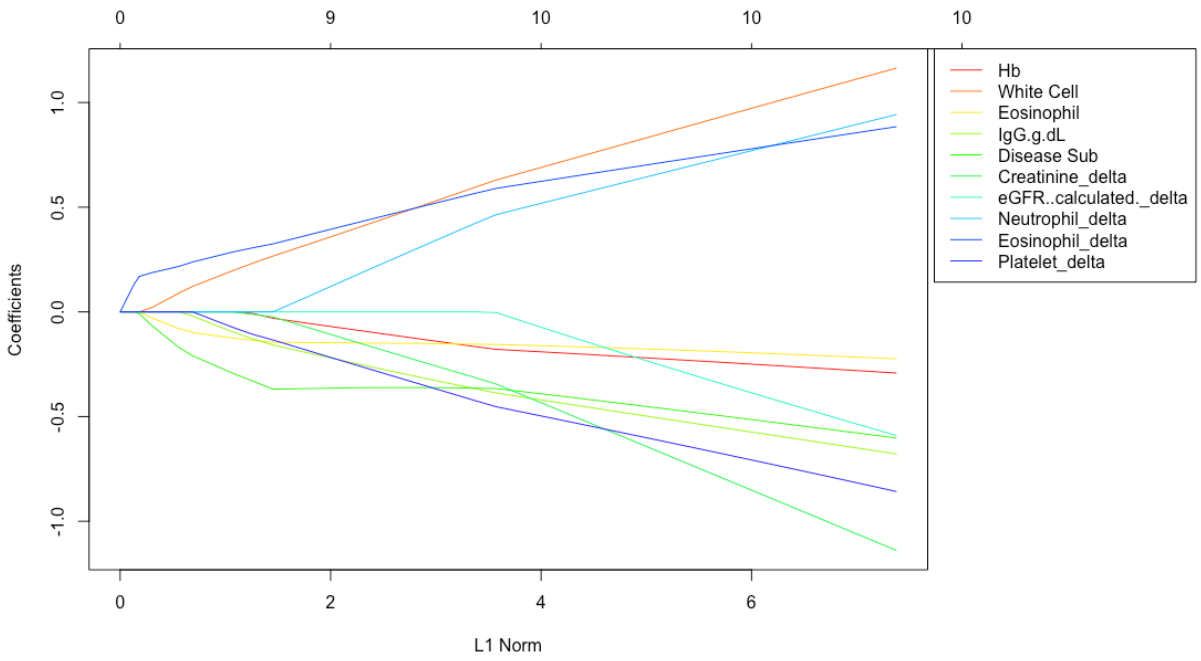

Figure 27 - Lasso Delta Feature Importance

Figure 27, Table 14 and Table 15 show the feature importance for the Delta Analysis model. It is clear that the inclusion of delta features was deemed important by the model as of the 11 biomarkers included in the model, 5 were deltas. Moreover, the nominal value of the delta coefficients are higher than the other biomarkers with only White Cell count having a comparable weighting to the deltas. The Creatinine delta is the single biggest indicator of LTROT while the Neutrophil delta is the largest indicator of relapse.

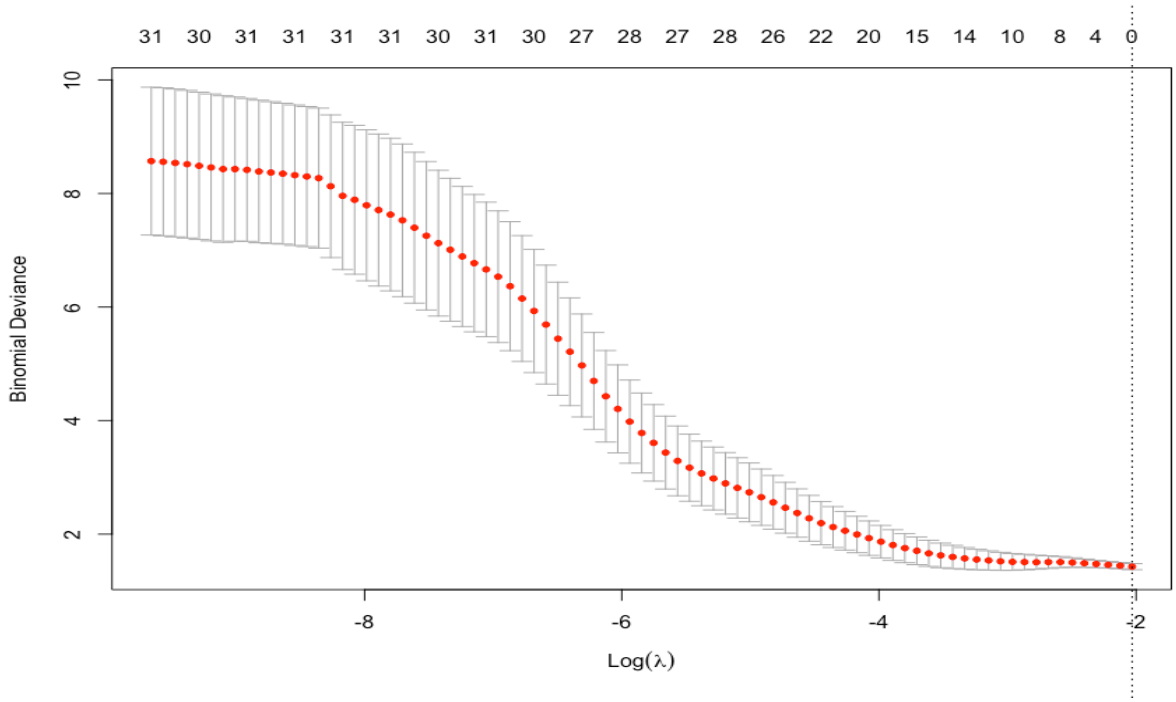

Figure 28 - Lasso Delta Feature Selection Cross Validation

Figure 28 shows poor behaviour when performing cross validation on the Delta Analysis dataset. The Lasso Logistic Regression is unable to find an optimal number of biomarkers to minimize the binomial deviance. Figure 28 states that the optimal model is one which predicts all patients as relapse patients and that the outcome of LTROT/Relapse is independent of the biomarkers inputted into the model. The simple explanation for this result is due to the large number of features inputted into the model compared to the relatively small dataset. Inputting 36 biomarkers with only 69 samples led to the cross validation being unable to find an optimal solution. As a result, 11 biomarkers with the highest coefficient weightings were selected as this does seem to give some trade-off between the bias and the variance, as per Figure 28, however it can be concluded that the Lasso Logistic Regression is not optimal when analysing the Delta Analysis dataset.

Comparison of Results

The results of the three datasets were compared against a dummy model which predicted the most frequent class, in this case Relapse. Table 16 below is a comparison of the classification metrics for the datasets of Substantive Sample, PCA Grouping and Delta Analysis versus the dummy model.

As can be observed in Table 16 the Lasso Logistic Regression model on all three datasets significantly outperforms the dummy model in predicting the classes of LTROT and Relapse when compared to the dummy model. The accuracy of the Lasso Logistic Regression is higher as well as more balanced scores of sensitivity and specificity. It can be therefore concluded that the machine learning techniques applied were successful in their goal of predicting the two classes and that the statistical information contained in the biomarkers can be utilised to predict the classes of LTROT and Relapse.

The PCA grouping dataset was the best performing model in terms of accuracy, specificity and sensitivity although the accuracies of all three groups are comparable. This is an interesting result as it suggests that there is greater statistical information included in the PCA Grouping when compared to the Substantive Sample despite both datasets having very similar features.

Similar to the Bayesian Logistic Regression it is clear that Relapsing patients are easier to classify as the specificity scores were consistently higher among all three groups. The Delta Analysis was not significantly greater in terms of accuracy as was seen in the Bayesian Logistic Regression model, however this is expected given the poor behaviour of the model and the limitations of the dataset when performing cross validation as discussed in the previous section.

Substantive Sample					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.725
Prediction	LTROT	29	14	Specificity	0.781
	Relapse	16	50	Sensitivity	0.644
PCA Groupings					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.752
Prediction	LTROT	29	11	Specificity	0.8281
	Relapse	16	53	Sensitivity	0.644
Delta Analysis					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.725
Prediction	LTROT	18	8	Specificity	0.800
	Relapse	11	32	Sensitivity	0.621
Dummy Comparison					
Confusion Matrix		Reference		Metrics	
		LTROT	Relapse	Accuracy	0.587
Prediction	LTROT	0	0	Specificity	1.00
	Relapse	45	64	Sensitivity	0.00

Table 16 - Lasso Regression Comparison of Results

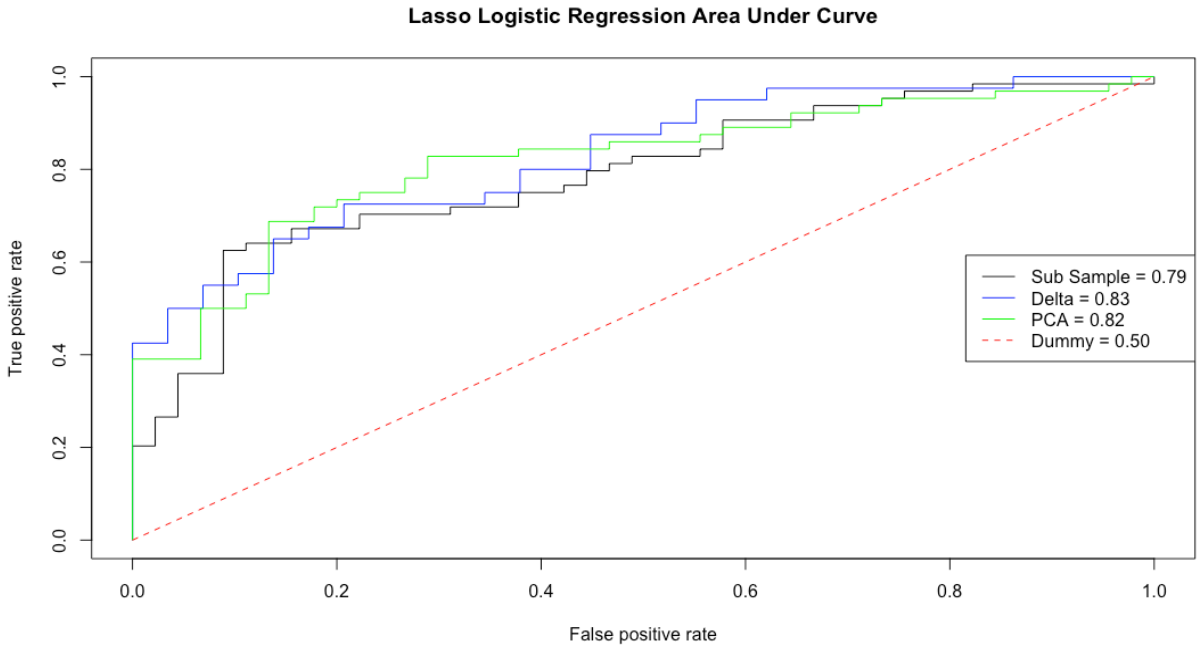

Figure 29 - Lasso Logistic Regression AUC Comparison

Figure 29 above shows a plot of the AUC of all three models compared to the dummy model. This is another metric for analysing the accuracy of the model by comparing the True Positive Rate to the False Positive Rate. The Delta Analysis dataset has the highest AUC however the AUC of all three models are comparable. This is somewhat contradictory to the findings of Table 16 which suggested that the PCA Grouping was the optimal dataset however the AUC takes into account the size of the dataset, therefore overall it can be concluded that similar to the Bayesian Logistic Regression the Delta Analysis is the optimal dataset however the differences in this case are minimal.

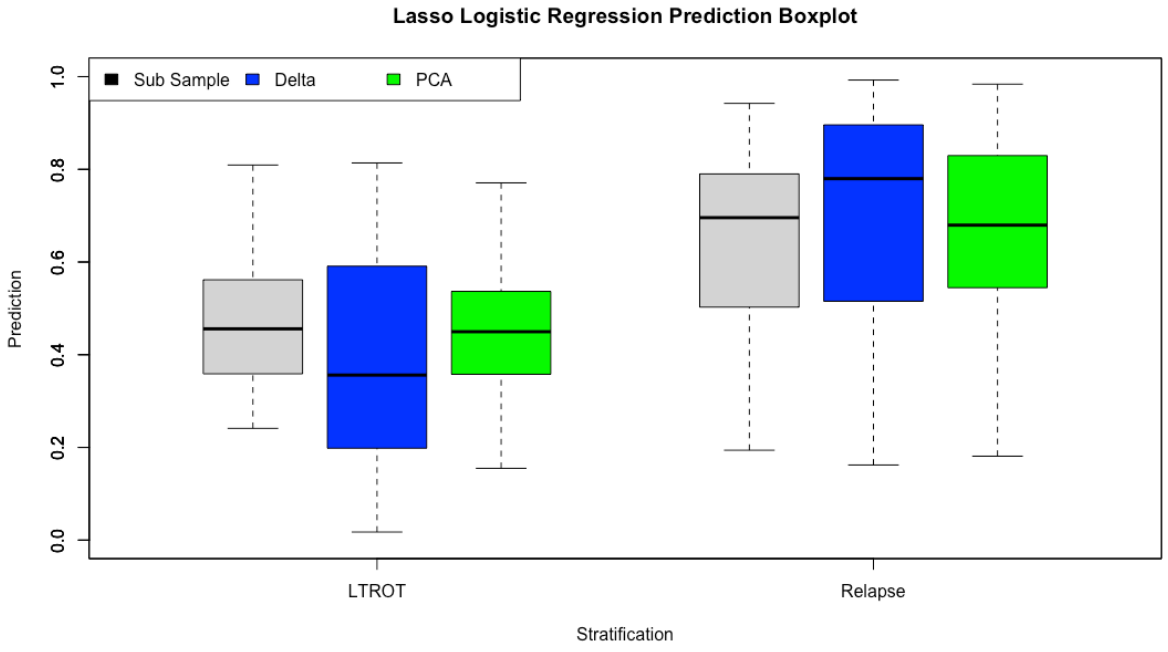

Figure 30 - Lasso Logistic Regression Classification Certainty Boxplot Comparison

Figure 30 above shows a boxplot comparison of the model certainties. Predictions greater than 0.5 are classified as Relapse and the opposite is true for predictions less than 0.5. The closer the prediction is located to 0.5 the less certain that model was about its predictions of that class. The blue Delta Analysis boxplot had a wide range in its certainty of its predictions of the LTROT stratification. This is good performance as it was correctly able to label a significant number of both LTROT and Relapses with a high degree of confidence and other incorrect predictions were less confident. In comparison the PCA Groupings model was significantly more uncertain in its predictions. The small size of the boxplot in Figure 30 whose lower quartile is located at approximately 0.4 suggests that its most confident predictions are only 0.1 unit away from being predicted as Relapse compared to 0.3 units of confidence for the Delta Analysis. The median of the Delta Analysis is also located further from 0.5 which indicates that a greater number of its predictions were more confident when compared to the Substantive Sample and the PCA Groupings.

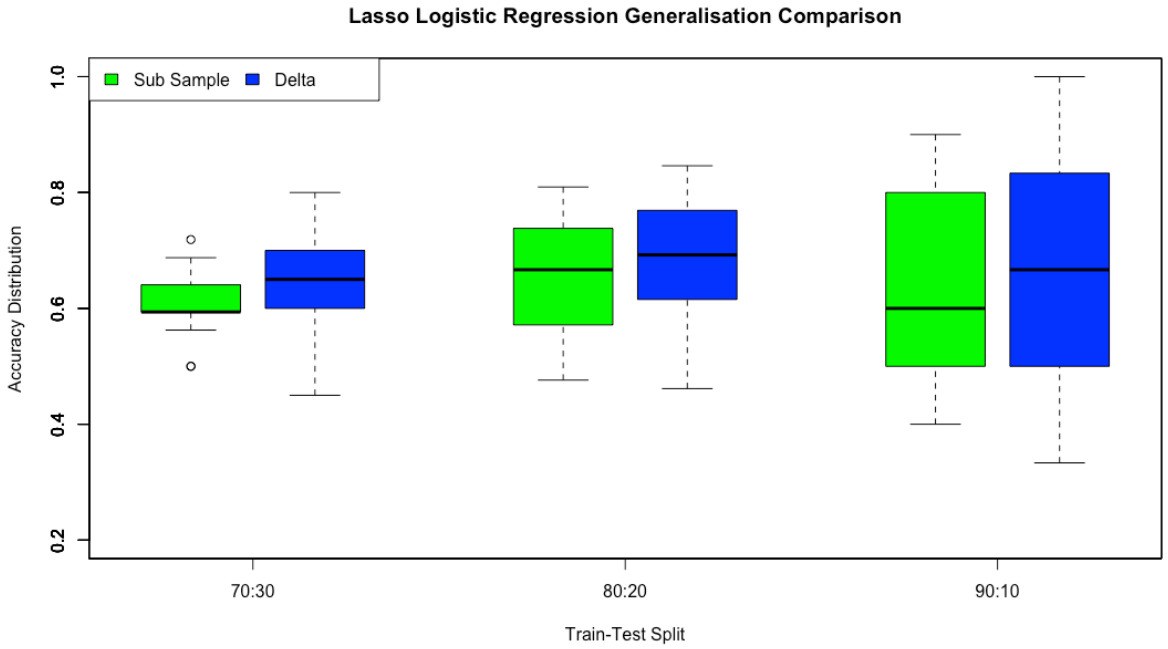

Figure 31 - Lasso Logistic Regression Generalisation Comparison

The Delta Analysis model is also superior when split into a training and test set as per Figure 31. Although the performance of both models is significantly degraded when generalised, the delta analysis on average achieves higher means and similar standard deviations. Given the limitation to the size of its dataset this once again highlights the predictive power of the delta features. The likely reasons for the degradation in performance when generalising the model to an unseen dataset is once again overfitting of the model and small dataset sizes.

Figure 37, located in the Appendix, shows that the models used are not sensitive to the imputation. There is very little variance in the accuracies of the different imputed datasets as can be observed thus confirming that imputation did not affect the accuracy of the Lasso Logistic Regression model.

Linear Regression

Introduction to Algorithm

Linear Regression refers to the Ordinary Least Squares (OLS) estimate shown below. This method minimizes the error between the unknown parameters of a model and the observed data.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

The OLS is the simplest method for producing a model which minimizes the variance between the fitted model and the data points (Powell, 2021). The advantage of the Linear Regression model is its simplicity in its interpretability as it not only provides feature importance but confidence intervals and p-values reaffirming the significance of those predictors. The disadvantage of the Linear Regression mode is the bias-variance trade-off. This trade-off is of particular relevance to this research due to the comparatively large number of features and small number of data points. Without due care this could result in overfitting and overestimation of accuracy. This was addressed through the use of iterative step forward and step back models which removed non-predictive features.

In order to reduce the number of variables in the model, Backward Stepwise Linear Regression with Akaike Information Criterion (AIC) was performed. AIC is an estimator of prediction error which attempts to find balance between underfitting and overfitting in the bias-variance trade-off (Chowdhury & Turin, 2020). Backward Stepwise Linear Regression starts with the full model and iteratively deletes variables which reduce the AIC the least. This is repeated until any variable deleted significantly decreases the AIC (Chowdhury & Turin, 2020).

Final Dataset and Features

The dataset utilised for the Linear Regression algorithm consisted of unlabelled ID's and all biomarkers with less than 50% data missingness. A subset of this dataset is shown below in Table 17. In total there were 19 biomarkers, 6 of which are shown in Table 17, with sufficient data that could be used in the analysis. The target of this analysis is different to the previous research. The dataset and features inputted to the Linear Regression algorithm were used to predict the Relapse Rate. The Relapse Rate, shown below in Table 17, was defined as the number of relapses a patient suffered throughout their disease course. The disease course being the length of time from diagnosis to their last encounter as discussed in Methodology and Data Preparation. There were two analyses run. The first looked at all patients for whom a Relapse Rate could be calculated and

sufficient biomarker data was available. The second analysed only those patients who suffered a relapse, thus having a Relapse Rate greater than 0.

In both analysis, patients with disease subtype of EGPA were filtered from the analysis. There was a significant class imbalance whereby only 5% of the patients belonged to this group. In addition to this, disease subtype was identified as important in both models, hence patients with disease subtype EGPA were filtered for a more meaningful exploration of the importance of this variable.

U-Blood	Weight	Creatinine	eGFR	Hb	White Cell	Relapse Rate
-0.637	-0.052	-0.314	0.189	0.403	-0.314	0.114
-0.637	-0.959	-0.377	-0.321	-0.549	-0.130	0.000
-0.637	1.065	-0.509	0.517	0.668	-0.308	0.000
-0.637	-0.736	0.144	-0.798	-0.020	-0.209	0.000
0.644	-1.378	-0.407	-0.371	-1.077	0.405	0.167
-0.637	0.095	-0.761	1.339	0.720	2.285	0.053

Table 17 - Relapse Rate Subset of Dataset

All Usable Data Points

Table 18 shows the features retained by the Stepwise Linear Regression model. The estimate column shows the effect each parameter has on the overall model, i.e. if the coefficient estimate is greater than 0 this parameter increases the Relapse Rate and vice versa for estimates less than 0.

The model intercept is greater than 0, meaning the longer the disease course of a patient, the increased incidence of relapse. In other words, all other parameters being equal, given an intercept value of 0.583, for a patient who's disease course is 10 years they can expect to suffer 5.83 ± 1.63 relapses. The standard error of the intercept is of the same order of magnitude as the estimate meaning that there is quite a large variance in the expected number of relapses. On the other hand, the P value of the intercept is significantly below 0.05 hence, this parameter is very relevant to the model.

The White Cell Count is the only other variable with an estimate greater than zero. This suggests that patients with an increased White Cell Count are more likely to suffer an increased number of relapses. The standard error of this estimate is an order of magnitude lower than the estimate

hence one can be confident the true value of this estimate lies close to that presented in Table 18. The P-value also shows a high level of significance.

Neither the NLR nor ANCA Specificity Other show significance as their P-values lie above the 0.05 threshold, hence they are not significant features of the model. ANCA Specificity PR3 and Disease Subtype MPA both have similar values in Table 18 below with estimates less than 0. It is important to note that these parameters are categorical parameters, hence their estimate values are in comparison to another feature in that category. For ANCA Specificity PR3 it is being compared to MPO and for Disease Subtype MPA it is compared to GPA. Interpreting Table 18, it can be concluded that patients with phenotype PR3 suffer 0.434 ± 0.162 less relapses than those with MPO. It is important to distinguish that this does not necessarily contradict the findings in the literature that PR3 positive patients are more likely to relapse. Similarly patients with the MPA subtype can expect to suffer 0.441 ± 0.152 less relapses compared to GPA throughout their disease course. Once again the standard errors of both variables are of the same order of magnitude as the estimates hence there could be significant deviation of the estimate from its true value. That being said, once again the importance of both features cannot be underestimated given their P-values are significantly below the 0.05 significance threshold.

Finally, while ANCA Specificity and Disease Subtype are categorical variables meaning their estimates will be multiplied by either a 1 or 0, the White Cell Count is a continuous variable hence its estimate could have a larger weighting over the overall model. The true range of White Cell Count is in fact (-1.98 , 4.54) which when multiplied by its estimate yields (-0.413 , 0.944). While White Cell count has been standardised to have a mean of 0, it cannot be denied that abnormally high or low values of White Cell Count have the largest weighting over the Relapse Rate of patients in this dataset.

	Estimate	**Std. Error**	**t-value**	$\mathbf{Pr}(>\lvert t\rvert)$
Intercept	0.583	0.163	3.571	0.000
White Cell Count	0.208	0.059	3.507	0.001
NLR	-0.089	0.059	-1.496	0.136
ANCA Specificity Other	-0.31	0.172	-1.801	0.072
ANCA Specificity PR3	-0.438	0.162	-2.711	0.007
Disease Subtype MPA	-0.441	0.152	-2.901	0.004

Table 18 - Stepwise Linear Regression All Data Point Model Parameters

Relapsing Patients Only

	Estimate	Std. Error	t-value	$\mathbf{Pr}(>\|t\|)$
Intercept	4.032	1.057	3.815	0.000
White Cell	1.129	0.372	3.035	0.003
NLR	-0.889	0.403	-2.229	0.03
ANCA Specificity Other	-1.651	1.515	-1.090	0.280
ANCA Specificity PR3	-3.769	1.091	-3.456	0.001
Disease Subtype MPA	-3.284	1.067	-3.079	0.003

Table 19 - Stepwise Linear Regression Relapsing Patients Model Parameters

Table 19 above shows the parameter estimates and their significance in the Stepwise Linear Regression model. It is clear that when the non-relapsing patients have been removed there is a much stronger relationship between the parameters and the Relapse Rate. Unexpectedly, the parameter estimates are much larger as they are not being dampened by the non-relapsing patients however the weighting of the estimates (positive/negative), the relative magnitude of their standard errors and their significance are all comparable to those found in Table 18. The only parameter which deviates significantly from Table 18 is the Neutrophil Lymphocyte Ratio (NLR). As per Table 19, this parameter is significant as its P-value has fallen below the 0.05 threshold and it is negatively associated with Relapse Rate. Despite this the estimate and standard error are of the same magnitude, hence the true value of the estimate could deviate from the value presented in Table 19.

Similar to White Cell Count, NLR is also a continuous variable with a range of (-1 , 6.4) after standardisation. When multiplied by the estimate this results in a output range of (0.9 , -5.65). Comparably the White Cell Count output range of this model is (-2.25 , 5.13). Thus, it can be concluded that both of these features have the greatest weighting over the model.

Comparison of Results

Stepwise Linear Regression		
	All Usable Data Points	Relapsing Patients Only
R^2	0.053	0.299

Table 20 - Stepwise Linear Regression Comparison of Results

Table 20 shows a comparison of the R^2 metric for both datasets. The R^2 or coefficient of determination is a measure for the proportion of variance in the response variable Relapse Rate which can be explained by the regression variables in Table 18 and Table 19 (Date, 2020). R^2 is calculated using the below formula whereby the Residual Sum of Squares (RSS) captures deviance of the data from the model and the Total Sum of Squares (TSS) captures the deviance of the model from the mean.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Clearly, when Linear Regression is performed on Relapsing Patients Only, there is a much larger proportion of the variance explained by the variables. The reason for this can be attributed to Figure 32 below. As can be seen in Figure 32, any Relapse Rate greater than 0 is deemed to be an outlier with the mean and median of the dataset being located at 0. There are only 59 patients who have suffered a relapse, therefore all other 325 data points have a relapse rate of 0. This means the mean of the dataset will be approximately 0 and the TSS will be very small. This causes the Linear Regression predictions to be pulled towards 0 and as a result the parameters are unable to explain the variance in Relapse Rate as there appears to be so little.

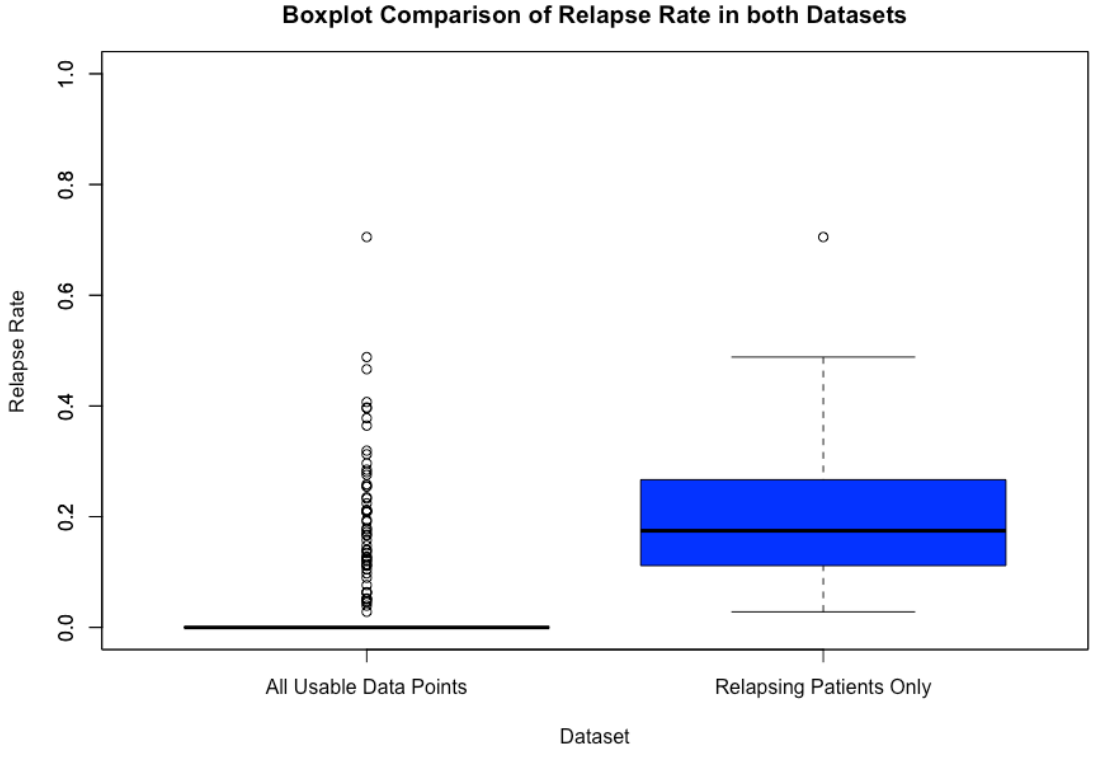

Figure 32 - Boxplot Comparison of Relapse Rate in both datasets

Conversely, an R^2 of 0.299 is quite a positive result when performed on the Relapsing Patients dataset. Clearly there is some proportion of the statistical information leading to an increased Relapse Rate captured by the parameters. That said, given the small dataset and the small value of R^2, there are limitations to the conclusions which can be drawn.

Discussion

ANCA-associated vasculitis is a rare auto-immune disease which if left untreated is fatal. That said, the treatment of AAV is also extremely challenging. The use of immunosuppressive therapies have reduced the mortality rate to about 2.3%, however it remains an issue that approximately 50% of people suffering from ANCA-associated vasculitis will relapse (Karangizi & Harper, 2018). While the disease of AAV leads to chronic symptoms, the immunosuppressive treatments, in particular their cumulative toxicity, have many equally harmful side effects. The primary challenge of ANCA-associated vasculitis is therefore to balance this risk of relapse with the cumulative toxicities of the treatment at the maintenance therapy stage. The goal of this research was to tackle this problem through the application of machine learning techniques. Success of this research is defined by accuracy of a model significantly greater than a dummy baseline model and conclusive identification of biomarkers which are predictive of relapse.

Two classification algorithms, namely Bayesian Logistic Regression and Lasso Logistic Regression, were applied to three datasets to identify whether physiological biomarkers could be used to distinguish between two classes; LTROT and Relapse. The results of the research are conclusive in that the statistical information contained in the biomarkers can be used to distinguish between these two classes with an accuracy significantly greater than a dummy baseline model. The minimum accuracy achieved was 72.5% for the Substantive Sample and Delta Analysis datasets with the Lasso Logistic Regression algorithm. This is significantly greater than the 58.7% baseline accuracy of the dummy model which only predicted the most frequent class. While this result is extremely positive and conclusively confirms the above hypothesis, there are still many considerations to take into account.

The accuracy of the Bayesian Logistic Regression algorithm was superior to the Lasso Logistic Regression algorithm across all three datasets. There are likely multiple reasons for this such as the models identifying different features as important however, the biggest difference between the inputs to these models was their treatment of categorical variables. While the Bayesian Logistic Regression algorithm treated categorical variables as factors and compared different levels of the factors against one another, the Lasso Logistic Regression treated these variables as numerical. The benefit of treating the variables as numerical is that there is only a single feature for each categorical variable. Conversely, for a categorical variable such as Disease Subtype with three levels, MPA, GPA and EGPA, the Bayesian Logistic Regression algorithm will treat this as two comparative variables, EGPA vs MPA and EGPA vs GPA. This can be observed in Figure 16, Figure 17 and Figure 18. It is possible that the increased number of features resulted in overfitting of the Bayesian Logistic Regression model, however it is far more likely that it is more meaningful to compare different levels of categorical variables. Both the results and the findings in the

literature corroborate this. The results show that the accuracy of the Bayesian Logistic Regression model is consistently greater than the Lasso Logistic Regression and the feature importance of the Bayesian Logistic Regression includes more categorical features. This suggests that there is statistical information being missed by the exclusion and method in which the Lasso algorithm handles categorical features. In literature, it was far more common to compare categorical groups for risk of relapse. For example, the risk of relapse among patients who are PR3 positive compared to those who were not was consistently higher (Berti & Specks, 2019). Similarly, studies which compared different induction treatment groups also found that the risk of relapse can depend on the type of induction treatment received. The Lasso Logistic Regression model does include categorical features however it does not compare different levels of categorical features which likely explains the discrepancy in accuracies between the models.

While the Bayesian Logistic Regression model was clearly superior, the issue of poor generalisation was common to both. Generalisation refers to the splitting of the data into a training and test set and is a good measure to see if a model has been overfitted to the training data. The idea behind generalisation is to test the accuracy of a model on unseen data. This is equivalent to testing the model on a new patient presenting with ANCA-associated vasculitis. Figure 21 and Figure 31 show that the accuracy of both Bayesian and Lasso Logistic Regression models is significantly degraded. The two reasons for this were presented as being small test set size and overfitting. The likelihood is that both factors play a significant role in the poor generalisations of the model. The size of the test set means that even mis-labelling only three or four patients can result in poor accuracies of approx. 50% when the data is split 90:10. Not only this but the size of the training and test sets can significantly impact the overfitting of categorical variables. ANCA specificity, which is the proteinase against which ANCA is targeted (commonly PR3 or MPO), was commonly overfitted. Patients with PR3 and MPO were identified as being more likely to relapse when compared to PR3. This contradicted the literature which states that PR3 patients are more inclined to relapse. On further analysis, there were only a very small number of patients belonging to this group, with more having suffered a relapse. Hence, this result was clearly due to the model overfitting to a small subgroup. Care was taken to address this issue by ensuring that different levels of categorical variables were well populated. For small subgroups of categorical variables which had few entries, subgroups were combined under a common heading such as ‘Other’. Where possible biological significance of features under this other group was maintained however this was not always possible. The issue of biologically important features is addressed later in the Discussion. Despite this, due to the small number of datapoints any given test or training set with a random sample could easily have been overfitted to one or two subgroups of a variable resulting in large overfitting and poor accuracy of the model.

Given the relatively large number of features compared to the size of the dataset it also must be concluded that overfitting is occurring with the results on the full dataset. In truth, the true accuracy likely lies somewhere between the accuracy on the full dataset and the accuracy of the generalised test set. There is not sufficient data available to be able to perform a proper train-test split. By running many iterations with different train-test splits it was hoped that on average the accuracy might even out however the standard deviations remained quite large. It is clear that the Delta Analysis dataset suffered the most from overfitting. An accuracy of 95.7% is too good to be true however, the generalised accuracy of this model was still extremely high with a mean accuracy of 76% which is large enough to confirm the original hypothesis. Overall, even with the degradation in performance due to overfitting and small sample sizes, the generalized models outperformed any baseline comparison. While this research is in its infancy, the variance in the biomarker data can distinguish with high accuracy between patients who have relapsed from those who are LTROT.

This result is meaningful under the assumption that every patient will fall under the label of LTROT or Relapse, however with a complicated disease such as ANCA-associated vasculitis this is rarely the case. The labels of LTROT and Relapse were chosen as they represent opposite ends of the spectrum with regards to patient health. These labels assume that patients will either relapse with a new onset of the disease or they will be cured and never suffer from AAV again. Unfortunately, AAV is defined by a remitting-relapsing nature with the vast majority of patients falling somewhere between these two labels. These patients will often have low levels of active disease which are treated with immunosuppressive therapy but are not considered as having a full blown relapse resulting in a restart of induction therapy. Clearly, only a small proportion of the population present in the RKD registry have been included in this research as a result. This challenge is common to interdisciplinary research, where the results while meaningful to the statistician only captures one small aspect of a much larger problem. The results of this research therefore represent a first step in the right direction. It can be conclusively concluded that biomarkers can be utilised to distinguish between these sets of patients with the label of LTROT or Relapse as defined in this research, however implementing this in a medical setting will require further research.

In addition to the above classification problem it was analysed to see if the Relapse Rate could be regressed to. The aim of this was to see if the biomarker data could be used to distinguish which patients suffer multiple relapses. The power of this analysis was limited in terms of both dataset size and performance degradation by a class imbalance. When the dataset was limited to only Relapses an R^2 of 0.299 was obtained. This is promising however a dataset of 59 is too small to have confidence in the result as outliers could have a significant weighting on the result. Conversely when performed on all patients, for whom Relapse Rate could be calculated, the R^2 showed little to no correlation and there was no simple linear model which could predict the target.

The reason for this was the imbalance in the number of Relapse Rates which were zero compared to those which were not. Overall, the results of this analysis, particularly when compared to the classification results, are disappointing. An R^2 of 0.299 suggests that 30% of the variance is explained by the variables which means that a much higher proportion is not.

Combining this result with the previous analysis in the context of AAV, means that yes, one can use biomarker data to distinguish LTROT from Relapse however, one cannot determine with confidence how many times a patient will Relapse from the biomarker data obtained at the beginning of the maintenance therapy stage. The biomarker samples used in this analysis represent one snapshot in time. For a patient who has relapsed, their biomarkers will have largely different values after their relapse. Thus, it is unsurprising that there is only a small correlation between Relapse Rate and the biomarker samples taken at this moment in time. It is possible that the distinguishing factor between a single relapse and multiple relapses could depend on biomarker values after rather than before the first relapse. It could also be worthwhile to include patients who have relapsed multiple times in the classification analysis more than once. One sample taken before their first relapse and another taken after their first relapse when their maintenance therapy has been reintroduced. This would increase the number of data points available for analysis and could give a clearer indication of the biomarker levels which are indicative of relapse.

While the accuracy of the classification model was important, the most relevant aspect of this research was the identification of biomarkers which were predictive of relapse. The ideal candidate biomarker was one which was biologically significant, independent of other factors such as gender and easily obtained. It can be assumed that all biomarkers included in the models are easily obtainable. Given the large amount of data missingness, logically the most common and easily sampled biomarkers will be included. Independence from other factors was addressed largely through the methods employed to handle multicollinearity and a biologically significant biomarker is one that not only plays a role in the pathogenesis of ANCA-associated vasculitis but is one which can provide consistent measurements. A summary of the ten most relevant biomarkers are included in Table 22 in the following chapter. In short, the delta biomarkers, white cell markers (White Cell Count, Eosinophil, Neutrophil and Lymphocyte), ANCA titre, disease subtype and induction treatment received were all very important biomarkers. There is significant literature and research in existence discussing the predictive power of ANCA titre and disease subtypes, however the use of deltas and the findings regarding white cell markers are novel.

The introduction of the delta biomarkers were extremely significant in increasing the accuracy of the classification. The power of the delta markers are that they provide a sub-clinical measurement of how a patient has responded to induction treatment. In essence, while a patient can present as non-symptomatic, with a BVAS of 0 and apparently be in remission, these delta markers capture

how the patient has responded to treatment and if there is some underlying low activity of disease. This is particularly relevant for markers such as ANCA titre. As discussed in the Literature Review, this marker while not sufficiently predictive of relapse to require a change in therapy, it is sufficiently predictive to warrant closer monitoring. A patient suffering from ANCA-associated vasculitis could have ANCA present at both diagnosis and maintenance. Some physicians may see the presence at maintenance indicative that the patient will relapse, however if the ANCA titre has decreased significantly since diagnosis this patient is actually much more inclined to be a LTROT as per Figure 18. Similarly, comparing the White Cell count, to the White Cell count delta in Figure 18, shows why the inclusion of the delta is so significant. If one were to exclude the delta and look at White Cell count on its own, the assumption is simply that an increased number of white cells is indicative of relapse. This was consistent across both models and all datasets. However, if the White Cell count has decreased from diagnosis, then the patient is likely not to relapse. Thus, both aspects of the treatment response and biological significance are captured by the delta analysis.

The inclusion of White Cell count is an interesting marker. There is biological significance in the role that the white cells, namely neutrophils, lymphocytes and eosinophil, have in the pathogenesis of AAV however the measurements of them may not be consistent. An elevated White Cell count is symptomatic of active disease however, white cells are very short lived in the human body. A white blood cell will only live for 1 to 3 days. One week a patient may have a very high white cell count and the next it could be normal. That being said, the consistency at which the model found white cells to be significant is too large to be ignored. This once again represents the challenge of interdisciplinary research whereby while all mathematical signs point towards significance, biologically it does not make sense. The importance of this feature is therefore unlikely to be an elevated white cell count on any given day, instead it is likely that persistent elevation of White Cells are predictive of relapse. This research did aim to take one sample from as early in the maintenance stage as possible, however these samples actually represent a 6 month window over which biomarker values were averaged. While this may not be the case for every patient, the explanation that makes both mathematical and biological sense is that a persistent elevation of White Cells over a 6 month period is predictive of relapse. An interesting note in this analysis is that while Eosinophils, Lymphocytes and Neutrophils are all White Cells, only Lymphocyte and Neutrophils were highly correlated with White Cells. Eosinophils on the other hand were maintained in the model and should be kept alongside one of White Cell count, Lymphocyte count or Neutrophil count in any future models. The most important aspect of inclusion of any of these biomarkers however should be the persistent elevation of their count.

White Cell count raises another interesting aspect of interdisciplinary research. Mathematically, one is inclined to take all biomarkers at face value and trust the outputs of the models. In this

research however it is important to consider the significance of why a certain biomarker has been identified as important. Take induction treatment as an example. In all of Figure 16, Figure 17 and Figure 18 those patients who received no induction treatment were far less likely to relapse than those who had received induction treatment. In other words patients who received no therapy were less likely to relapse. While one might assume that it is therefore better for no patients to receive induction therapy as they are less likely to relapse, the truth is that patients who received no induction treatment actually had a much milder form of the disease. Similar findings could be said for the biomarker of protein in urine in Figure 16 and Figure 17. When only a small amount of protein was found in the urine ('U-Protein1'), these patients were more likely to relapse. The opposite was true for patient with a medium amount of protein in urine ('U-Protein2'). Finally, for those with a large amount of protein in urine ('U-Protein3'), this was not indicative of LTROT or Relapse. Protein in urine is indicative of active disease, therefore it is interesting that the more protein found in urine the less likely it was for the patient to relapse. Logically one would assume the converse. It is possible that the model is simply overfitting to these groups, however it is also possible that the model is capturing how different treatment strategies can effect relapse. A small amount of protein in urine may not be sufficient to require a change in treatment while a medium or large amount could be. When only a medium amount of protein was found the treatment could be shown to be effective however when there is a large amount of protein the treatment may or may not be effective as it is possible that it was given too late. Thus, a small amount of protein in urine could be a symptom of low underlying disease activity leading to relapse which if treated could prevent relapse. This corroborates statements by Salama, 2020, who stated that patients in remission can often have underlying low activity of AAV. While this is merely a hypothesis, the importance of exploration of these features in an interdisciplinary environment is very important so that the not only the importance but the relevance of the features is understood.

Finally, a positive of this research is the consistency of the feature importance across both models and all three datasets. The same key biomarkers such as ANCA titre, disease subtype and white cell were included in almost all models. It is clear that there is biological information contained in these biomarkers can be transformed into statistical information for distinguishing between LTROT and Relapse patients with AAV. There were two main difficulties in determining biomarker importance. The first was the presence of multicollinearity. This made it very difficult to conclusively identify singular biomarkers which were predictive of relapse. The PCA groupings were easily the most consistent biomarker identification given the independence of the PC's however while easily interpretable to a statistician, adoption of PCA in a medical setting is unlikely. It is much easier for a local physician to understand that persistent elevation of white cells is indicative of relapse than explaining the covariance of a PCA matrix. The second difficulty was the performance of the imputation. Clearly from Figure 22 and Figure 37, the use of imputation

did not affect the accuracy of the results. That is to be expected as multiple imputation maintains the statistical variance of the dataset. However, when the imputation was combined with high collinearity it made for difficult and unreliable interpretations of feature importance. Generalisations can easily be made among feature importance. For example, the presence of white cell counts in the model was consistently predictive of relapse, however often one of White Cell count, Neutrophil count and Lymphocyte count would be far more predictive than the other two depending on the imputed dataset. It was clear that biomarkers with less missingness performed more consistently than those with higher missingness. A higher haemoglobin was consistently predictive of LTROT whereas IgG was sometimes included and other times not. The use of imputation in a medical setting is debated and in general imputation should not be performed when missingness exceeds 20%-30%. Given the limitations of the dataset, this was stretched to 50% hence inconsistencies in the interpretation of feature importance is not unsurprising.

Overall the findings of this research are extremely positive in the utilization of the RKD registry database to optimise the treatment of AAV. The accuracies of the classification models even when generalised to a small test set exceed any dummy baseline comparison. It can be conclusively stated that the statistical information in the biomarker dataset can be utilised to distinguish between the classes LTROT and Relapse. There are limitations to the secondary goal of important biomarker identification. Multicollinearity and imputation make for difficulty in interpretability of importance beyond findings in current literature however general observations such as importance of elevated white cell counts can be conclusively made. The inclusion of deltas is novel to this research and their role in the classification of LTROT from relapse cannot be understated. The importance of the interdisciplinary aspect of this research is also integral to the understanding of the biomarker importance and to the future implementation and deployment of any model in a medical setting. These findings are very encouraging for future work. This research represents a very positive first step to the employment of machine learning in optimising the treatment plans of patients with ANCA-associated vasculitis however it is clear that more data, maturation of the RKD registry and further research will be required before its utilisation in a medical setting.

Chapter 5

Future Work

This research has created a foundation for the integration of machine learning techniques with the RKD Registry database to improve patient treatment plans for those suffering from ANCA-associated vasculitis. This project has created a framework for the cleaning and extraction of relevant biomarker data and its application to a Bayesian Logistic Regression model with the target of LTROT or Relapse as the outcome. The success of this work is the classification accuracy, particularly when delta variables are included, however the limitation of this work is the conclusive identification of important biomarkers. The reason for this limitation is largely due to data availability. Both in terms of biomarker missingness and the size of the dataset. That being said, the RKD registry is a well curated dataset and there was much data excluded from this analysis that could be utilised in future works.

There are two main avenues to research in the future that would allow for the inclusion of more data with the RKD registry in its current state. They are to adjust the labels of LTROT and Relapse or to remove the labels entirely. The work undertaken in this project had strict definitions of LTROT and Relapse for reasons outlined in the Literature Review. As mentioned in the discussion however these strict labels only reflect the two extremes of the spectrum of patients who suffer from ANCA-associated vasculitis. Simply one can re-define the labels more loosely to include more data however this approach will likely remove the time element from the analysis. The time element being that for a patient to be considered LTROT, they must be at least one year off all treatments. This analysis then results in one simply classifying active disease for which there already exists the BVAS scoring system. Thus, any re-definition of the labels will have to be done with careful consideration. Another approach could be to utilise a semi-supervised approach. Semi-supervised classification is a data analytical technique which trains a supervised classifier from both labelled and unlabelled data (Bouveyron, et al., 2019). Semi-supervised learning can not only help to better define the decision boundary on the spectrum of LTROT vs. Relapse but it can also classify points as being on either side of this decision boundary (Bouveyron, et al., 2019). This approach is similar to that of the current research however it does not exclude those patients which do not fall under the label of LTROT or Relapse, thus capturing the entire spectrum of patients without removing the time element.

The other avenue to explore is to remove the labels LTROT and Relapse entirely and to undertake an unsupervised learning approach. An unsupervised approach looks at clustering similar groups of patients together who have similar features. In this approach, an ideal implementation could be the use of a unsupervised algorithm, such as k-means, where there are two main groups, one

LTROT and one Relapse, and the distance to these groups provide a classification of whether a new patient will be an LTROT or a Relapse. This implementation is presumptuous as there is no guarantee that there will be a clear division of LTROT and Relapse when the unlabelled data is included. On the other hand the benefit of this analysis could be the discovery of unknown groups or patterns in the biomarkers.

The final future work that could be carried out would be to simply wait for patients to either fall into one of the categories of either LTROT or Relapse. ANCA-associated vasculitis is a complex disease and as stated the labels utilised in this research do not capture the true nature of the disease with majority of patients lying somewhere between LTROT and Relapse. That being said, as the disease course of some these patients progresses it is likely that the number of both LTROT and Relapse patients will increase. The RKD Registry, while a well curated database, is still largely in its infancy given the length of the disease course. In general it takes a minimum 2 years from diagnosis for a patient to finish therapy and a further year for them to become an LTROT. Given the RKD Registry began in 2012, only patients from 2012-2018 at the very latest are included in this analysis. Some patients diagnosis dates predate 2012 however the further before 2012 these patients are diagnosed the less biomarker availability there tends to be. It is likely that in 3 or 4 years the number of patients included in this analysis could be significantly increased.

Future work should also be carried out to reduce biomarker missingness. While it is not conclusive that certain biomarkers are predictive of relapse, it is conclusive that deltas, white cells, induction treatment received and disease subtype are all important features. In particular for new patients the importance of a delta sample cannot be understated. To reduce biomarker missingness biological samples and tests will be required. A disappointing aspect of this research was the exclusion of the CD163 biomarker. This biomarker, obtainable in a urine sample, has shown great promise in the diagnosis of a renal relapse and could be very important in the prediction of relapse in the future (O'Reilly, et al., 2016). That said, reducing the biomarker missingness is the key to be able to conclusively state biomarker importance and should be the focus of any work before any future application of data analytical techniques.

The final step in any future work must look at deploying the predictive model in a medical setting. The aim of this research is to one day be able to personalise the treatment plans of patients. These treatment plans will always require medical experts hence future work must be done to create explainable models which can be easily be deployed by physicians. This will require interdisciplinary training and cooperation to be successful, hence future and current work should be undertaken to include key medical experts in all aspects of the interdisciplinary research.

Conclusion

In the most fundamental aspect, this research was an exploratory analysis of the RKD Registry database. The RKD Registry is a dedicated database with the aim of monitoring patient biomarkers to aid in the treatment and health of those suffering with ANCA-associated vasculitis. The largest challenge facing patients with AAV is balancing the risk of relapse with the cumulative toxicities of treatments. Through application of data analytical techniques, it was hoped that the optimal trade-off between these two could be found by analysing patient data during their maintenance treatment stage. In conclusion, this research was successful in achieving this goal.

Summary of Model Classification Accuracy				
	Substantive Sample	PCA Grouping	Delta Analysis	Dummy
Bayesian Logistic Regression	76.2%	76.25%	95.7%	58.7%
Lasso Logistic Regression	72.5%	75.2%	72.5%	58.7%

Table 21 - Summary of Model Classification Accuracy

Table 21 above shows the model accuracies obtained on three different datasets. A minimum accuracy of 72.5% was achieved while a maximum accuracy of 95.7% was achieved. Comparative to the Dummy baseline it is clear that the statistical information contained in the biomarkers of the RKD Registry can be used to distinguish between patients who have relapsed and patients who are off all treatment for a minimum time period of one year and who have never relapsed. Ignoring the limitations of the research, such as overfitting, poor generalisation and multicollinearity, these results conclusively confirm the primary hypothesis.

It was also hypothesised that the Relapse Rate, i.e. the number of relapses a patient experienced throughout their disease course, could be regressed to. This research achieved an optimal R^2 value of 0.299. This result was not conclusive given the relatively low coefficient of determination and small dataset size. Give only 59 data points and the aforementioned R^2 value it cannot be determined with confidence that the statistical information analysed in this project can be used to predict the Relapse Rate.

Secondary to the primary goal was to identify feature importance which contribute to an increased risk of relapse. Table 22 below shows the ranking of each of the consistently important biomarkers in their relative model and dataset. For example, Disease Subtype was the most important feature of the Bayesian Logistic Regression model in both the Substantive Sample and Delta Analysis

datasets. This table was created by simply sorting the mean of the posterior distribution of the Bayesian Logistic Regression model (as shown in Figure 18 for example) and by sorting the weights of the Lasso Logistic Regression model coefficients (as shown in Table 15 for example). For categorical variables in the Bayesian Logistic Regression model it was the average of the means of the posterior distribution that was taken.

Summary of Biomarker Importance						
	Ranking					
Model	Bayesian Logistic Regression			Lasso Logistic Regression		
Dataset	Substantive Sample	PCA Grouping	Delta Analysis	Substantive Sample	PCA Grouping	Delta Analysis
U – Protein	3	7	5	6	6	5
Hb	6	6	4	4	5	6
White Cell	3	4	8	2	2	2
Eosinophil	5	4	9	3	2	7
ANCA	6	3	9	5	4	N/A
Disease Subtype	1	2	1	1	1	4
Treatment	2	1	2	7	N/A	N/A
White Cell delta	N/A	N/A	3	N/A	N/A	1
Eosinophil delta	N/A	N/A	5	N/A	N/A	3
ANCA delta	N/A	N/A	5	N/A	N/A	N/A

Table 22 - Summary of Biomarker Importance

Given a large proportion of biomarker missingness and a subsequent reliance on imputation combined with high collinearity between variables it was difficult to identify individual feature importance conclusively. Clearly from Table 22 the categorical variables of induction treatment and Disease Subtype were important markers. White cell markers such as White Cell count, Neutrophil count, Lymphocyte count and Eosinophil count all were consistently important however due to the multicollinearity between them and reliance on imputation it is difficult to extract which one exactly is the most predictive. ANCA titre was consistently shown as important however it is the addition of the delta biomarkers that resulted in the greatest accuracy. Once again the delta markers suffer from multicollinearity and a large reliance on imputation. Overall, if this analysis were to be repeated, the biomarkers shown in Table 22 would be a good place to start.

In conclusion, despite the research presented in this paper being too raw to be deployed in a medical setting, the classification accuracies obtained conclusively show that machine learning techniques can sub-clinically classify patients who have relapsed from those who are LTROT before the termination of maintenance treatment.

Reflection

Personally, I found this project to be extremely rewarding yet very challenging. I think that undertaking a project such as this, where the outcome is required to be beneficial to an interdisciplinary cohort, requires flexibility and resilience. I found myself lucky that I got the opportunity to begin this project as early as possible during the course year. In order for a interdisciplinary project to be successful one must immerse themselves wholly in both sides of the project as early as possible. I found that my understanding of not only the project but the disease ANCA-associated vasculitis changed several times throughout this project but that I was far more knowledgeable because of it. For me, my greatest learning from this project was to not only consider what you are doing but to consider why.

Often as a data scientist it is easy to jump to the next best latest algorithm however working with the CDIG and listening to their meetings allowed me to better understand their goals. Similar to the discussion on the spectrum of labels of LTROT vs. Relapse, the goals of the medical team and of the computer science team are often at opposite extremes. For example, the importance of white cells was found to be important by the model however, the medical side of this project were largely disinterested with this finding due to their biological irrelevance. It was only on further discussion that the persistent elevation of white cells was discussed. Similarly, adoption of an unsupervised approach, while statistically very interesting, is largely unusable in a medical setting. I found myself not only having to consider the goal of this research but also who would be implementing this research were it to be successful as there is no GP in the world with the capability to run a complex unsupervised algorithm in their local practice. Lastly, it is easy to trust

a model and in particular a model outputs. This project forced me to question why certain features are being identified as important and to wonder if this finding is merely a proxy for something else. No better example of this was the ‘No Induction’ treatment group. This group were significantly less likely to relapse than any other despite not receiving any treatment. Not only did this contradict the literature, logically it did not make sense. However, this in fact was simply a proxy for a subgroup of patients who did not require induction treatment as they had a much milder form of disease. Being inquisitive in my findings was therefore another valuable skill learnt.

By aligning the goals of this research with that of the CDIG and being resilient, flexible and inquisitive I did thoroughly enjoy and find satisfaction in the outcome of this project. I enjoyed learning and listening at the CDIG meetings and consider myself lucky to be able to contribute to research which can have a real world impact on those with ANCA-associated vasculitis. My experience was extremely positive, rewarding and the skills I learnt will definitely benefit my future in the area of biomedicine and data analytics.

Bibliography

Berti, A. & Specks, U., 2019. Remission maintenance in ANCA-associated vasculitis: does one size fit all?. *Expert Review of Clinical Immunology,* 15(12), pp. 1273-1286 .

Bouveyron, C., Celeux, G., Murphy, B. & Raftery, A., 2019. *Model-based Clustering and Classification for Data Science, with Applications in R.* 1st Edition ed. Cambridge: Cambridge University Press.

Chowdhury, M. Z. I. & Turin, T. C., 2020. Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health,* pp. 1-7.

Date, S., 2020. *towards data science.* [Online]
Available at: https://towardsdatascience.com/the-complete-guide-to-r-squared-adjusted-r-squared-and-pseudo-r-squared-4136650fc06c
[Accessed 4 August 2021].

Donders, A. R. T., van der Heijden, G. J., Stijnen, T. & Moons, K. G., 2006. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology,* Volume 59, pp. 1087-1091.

Edner, N. M. et al., 2020. Follicular helper T cell profiles predict response to costimulation blockade in type 1 diabetes. *Nature Immunology,* Volume 21, pp. 1244-1255.

Geetha, D. & Jefferson, J. A., 2019. ANCA-Associated Vasculitis: Core Curriculum 2020. *American Journal of Kidney Diseases,* 75(1), pp. 124-137.

Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W., 2020. How Machine Learning Will Transform Biomedicine. *Cell,* 181(1), pp. 92-101.

Hara, A. et al., 2018. Risk Factors for Relapse of Antineutrophil Cytoplasmic Antibody-associated Vasculitis in Japan: A Nationwide, Prospective Cohort Study. *The Journal of Rheumatology,* 45(4), pp. 1-8.

He, P. et al., 2020. Prevalence and risk factors of relapse in patients with ANCA-assocaited vasculitis receiving cyclophosphamide induction: a systematic review and meta-analysis of large observational studies. *Rheumatology,* 60(3), pp. 1067-1079.

Hogan, S. L. et al., 2019. Understanding Long-term Remission Off Therapy in Antineutrophil Cytoplasmic Antibody-Associated Vasculitis. *Kidney International Reports,* Volume 4, pp. 551-550.

Jolliffe, I. T. & Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of The Royal Society A,* 374(2065), pp. 1-16.

Kana, M., 2020. *towards data science.* [Online]
Available at: https://towardsdatascience.com/introduction-to-bayesian-logistic-regression-7e39a0bae691
[Accessed 27 July 2021].

Karangizi, A. H. & Harper, L., 2018. Small vessel vasculitides. *Medicine,* 46(2), pp. 98-106.

Kemna, M. J. et al., 2015. ANCA as a Predictor of Relapse: Useful in Patients with Renal Involvement But Not in Patients with Nonrenal Disease. *American Society of Nephrology,* Volume 26, pp. 537-542.

Kemna, M. J. et al., 2017. Seasonal Influence on the Risk of Relapse at a Rise of Antineutrophil Cytoplasmic Antibodies in Vasculitis Patients with Renal Involvement. *The Journal of Rheumatology,* 44(4), pp. 473-481.

Kitching, A. R. et al., 2020. ANCA-associated Vasculitis. *Nature Reviews : Disease Primers,* 6(71), pp. 1-27.

Little, M., 2019. *TCD Data Protection Impact Assesment Template 2.0,* Dublin: Avert Health Group Trinity College Dublin.

Makalic, E. & Schmidt, D. F., 2010. Review of Modern Logistic Regression Methods with Application to Small and Medium Sample Size Problems. *Lecture Notes in Computer Science ,* 7-10 December, Volume 6464, pp. 213-222.

Mason, G., 1987. Coping with Collinearity. *The Canadian Journal of Program Evaluation,* 2(1), pp. 87-93.

McClure, M. E. et al., 2020. Long-term maintenace rituximab for ANCA-associated vasculitis: relapse and infection prediction models. *Rheutomology ,* Volume 00, pp. 1-11.

O'Reilly, V. P. et al., 2016. Urinary Soluble CD163 in Active Renal Vasculitis. *Journal of the American Society of Nephrology,* 27(9), pp. 2906-2916.

Powell, A., 2021. *towards data science.* [Online]
Available at: https://towardsdatascience.com/ordinary-least-squares-regression-da96dde239d5
[Accessed 27 July 2021].

Ray, S., 2019. *A Quick Review of Machine Learning Algorithms.* Faridabad, India, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

REDCap, 2021. *REDCap.* [Online]
Available at: https://www.project-redcap.org/
[Accessed 02 03 2021].

Salama, A. D., 2020. Relapse in Anti-Neutrophil Cytopasm Antibody (ANCA)-Associated Vasculitis. *Kidney Infection Reports,* Volume 5, pp. 7-12.

Sanders, J. et al., 2004. Risk factors for relapse in anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis: Tools for treatment decisions?. *Clincal and Experimental Rheumatology,* 22(36), pp. 94-101.

Stafford, I. S. et al., 2020. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *Digital Medicine,* 3(30).

Trejo, M. A. C. W. et al., 2019. Renal Relapse in antineutrophil cytoplasmic autoantibody-associated vasculitis: unpreictable but predictive of renal outcome. *The Journal of Rheumatology,* Volume 58, pp. 103-109.

Trinity Health Kidney Centre, 2021. *Rare Kidney Disease Registry and Biobank Data Management Plan,* Dublin: Trinity Health Kidney Centre.

Vatcheva, K. P., Lee, M., McCormick, J. B. & Rahbar, M. H., 2016. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Journal of Epidemiology,* 6(2).

Walsh, M. et al., 2012. Risk Factors for Relapse of Antineutrophil Cytoplasmic Antibody-Associated Vasculitis. *Journal of Arthritis & Rheumatism,* 64(2), pp. 542-548.

Appendix

Ethics Declaration

'RKD data is saved, if this is necessary, on a password encrypted device. RKD data is not emailed to yourself or anyone or stored on cloud services without being encrypted. RKD data is not shared with anyone else or discussed with anyone else. Demonstrations, reports and publications about the project will not display actual individual level patient data'

Data Cleaning Workflow Extra Process's

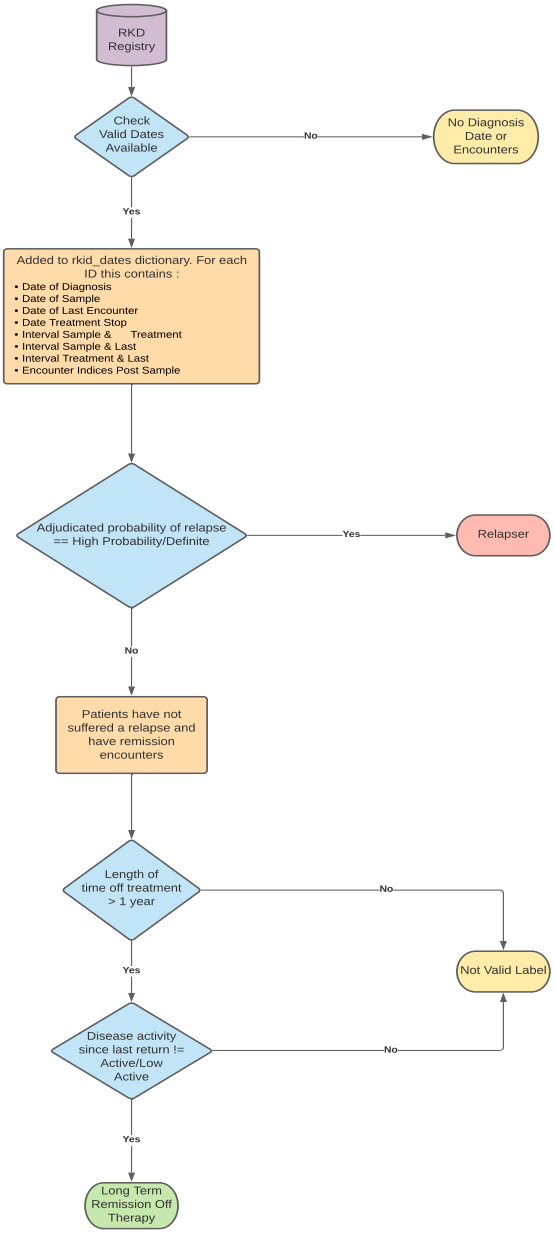

Figure 33 - Patient Stratification Workflow

Full Biomarker List

Biomarker Name	Description	% Missingness	Included	Reason
Serum sCD25 (pg/mL)	Biological marker of macrophage activation	100.00	No	Data Missingness > 50%
Anti-GBM Level	Measure of Kidney Damage	95.38	No	Data Missingness > 50%
Serum YKL-40 (ng/mL)	Secretory Protein Which Potentially Play a Role in Tissue Remodelling	94.62	No	Data Missingness > 50%
Serum sCD206 (ng/mL)	Biological marker of macrophage activation	94.62	No	Data Missingness > 50%
Urinary Calprotectin (ng/mL)	Protein Biomarker Present During Intestinal Inflammation	93.85	No	Data Missingness > 50%
Rheumatoid factor	Measure of Proteins Produced When the Immune System Attacks Health Tissue	93.08	No	Data Missingness > 50%
Urine Glutaric acid Results	Measure of Acidity of Body Tissues	86.15	No	Data Missingness > 50%
Urine Betaine Results	Biological Measure	86.15	No	Data Missingness > 50%

Urine Dimethylglycine Results	Biological Measure	86.15	No	Data Missingness > 50%
Urine Citric acid Results	Measure of Vitamin D	86.15	No	Data Missingness > 50%
Urine TMAO Results	Measure of Cardio Function	86.15	No	Data Missingness > 50%
Urine Succinate Results	Biological Measure	86.15	No	Data Missingness > 50%
Urine Oxoglutaric acid Results	Diagnostic Test for Overgrowth of Harmful Gut Flora	86.15	No	Data Missingness > 50%
Urine Maltose Results	Measure of Glucose in blood	86.15	No	Data Missingness > 50%
Urine Glycolic acid Results	Biological Measure	86.15	No	Data Missingness > 50%
Urine N-Phenylacetylglycine Results	Biological Measure	86.15	No	Data Missingness > 50%
Urine MyoInositol Results	Measure of Blood Sugar	86.15	No	Data Missingness > 50%

Absolute CD19 count cells/uL	Biological Marker for B cell disorders	85.38	No	Data Missingness > 50%
Complement (C4)	Measure of Autoimmune Disease Activity	85.38	No	Data Missingness > 50%
Complement (C3)	Measure of Autoimmune Disease Activity	85.38	No	Data Missingness > 50%
ESR	Measure of Inflammation	83.85	No	Data Missingness > 50%
AST	Marker of Liver Disease	83.85	No	Data Missingness > 50%
Visual Analogue Health Scale	Measure of Quality of Life	78.46	No	Data Missingness > 50%
5L Anxiety Depression Level	Patient Reported Outcome	77.69	No	Data Missingness > 50%
5L Mobility Level	Patient Reported Outcome	77.69	No	Data Missingness > 50%
5L Pain - Discomfort Level	Patient Reported Outcome	77.69	No	Data Missingness > 50%

5L Selfcare Level	Patient Reported Outcome	77.69	No	Data Missingness > 50%
5L Usual Activities Level	Patient Reported Outcome	77.69	No	Data Missingness > 50%
sCD163 (serum, ng/mL)	Measure of CD163 in blood	77.69	No	Data Missingness > 50%
Monocyte count x10^9/L	Biological Measure	76.92	No	Data Missingness > 50%
Total cholesterol mM	Biological Measure	66.92	No	Data Missingness > 50%
BMI	Body Mass Index	60.00	No	Data Missingness > 50%
IgA g/dL	Immunoglobulin A - Antibody present in Blood	58.46	No	Data Missingness > 50%
IgM g/dL	Immunoglobulin M - Antibody present in Blood	58.46	No	Data Missingness > 50%
sCD163 (urine, ng/mL)	Measure of CD163 in Urine	57.69	No	Data Missingness > 50%

Urine PCR / ACR mg/mmol	Protein to Creatinine Ratio	56.92	No	Data Missingness > 50%

Table 23 - Full Biomarker List

Principal Component Variation Explanation Plot

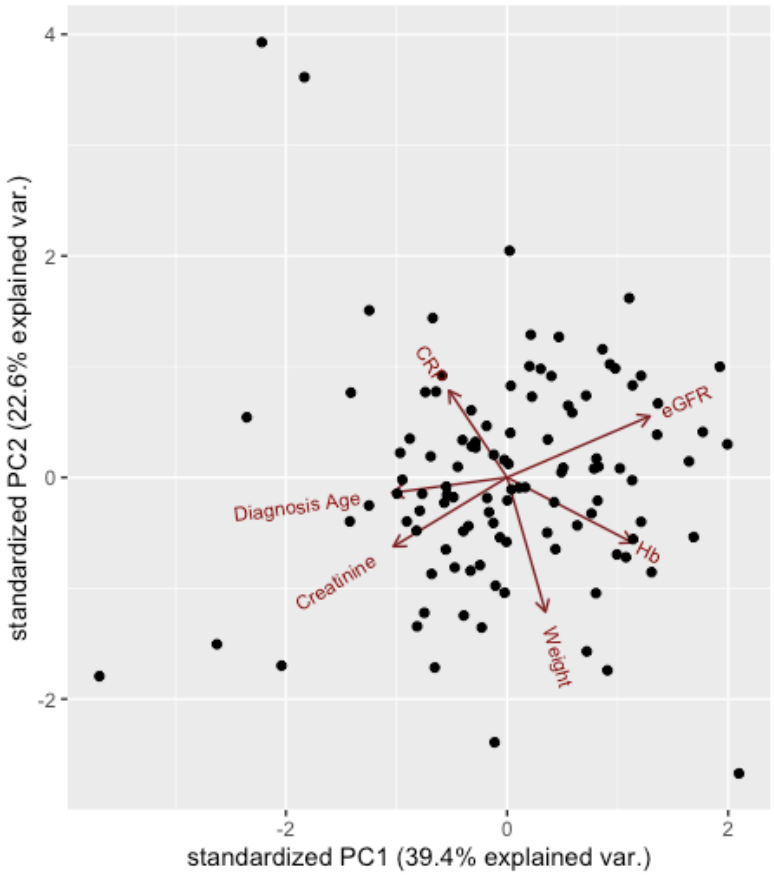

Figure 34 - Personal Characteristics PCA Variation Plot

Bayesian Logistic Regression Trace Plots

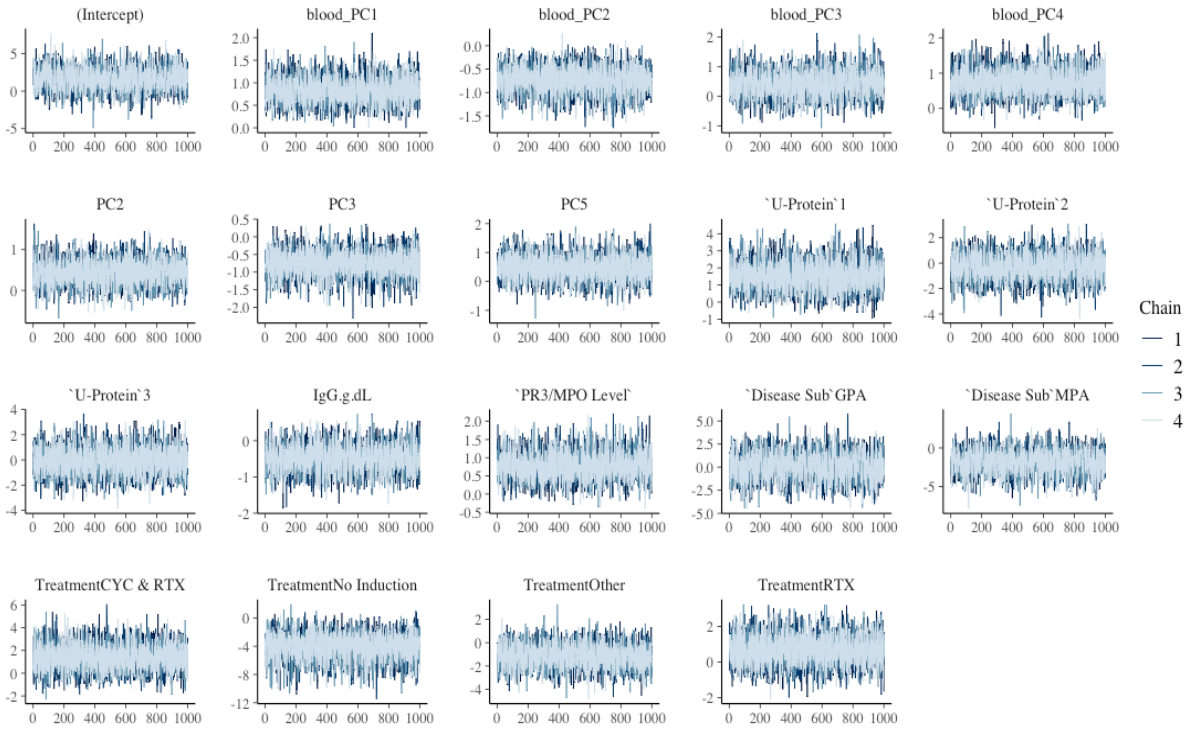

Figure 35 - Bayesian Logistic Regression MCMC Trace Plot PCA Group

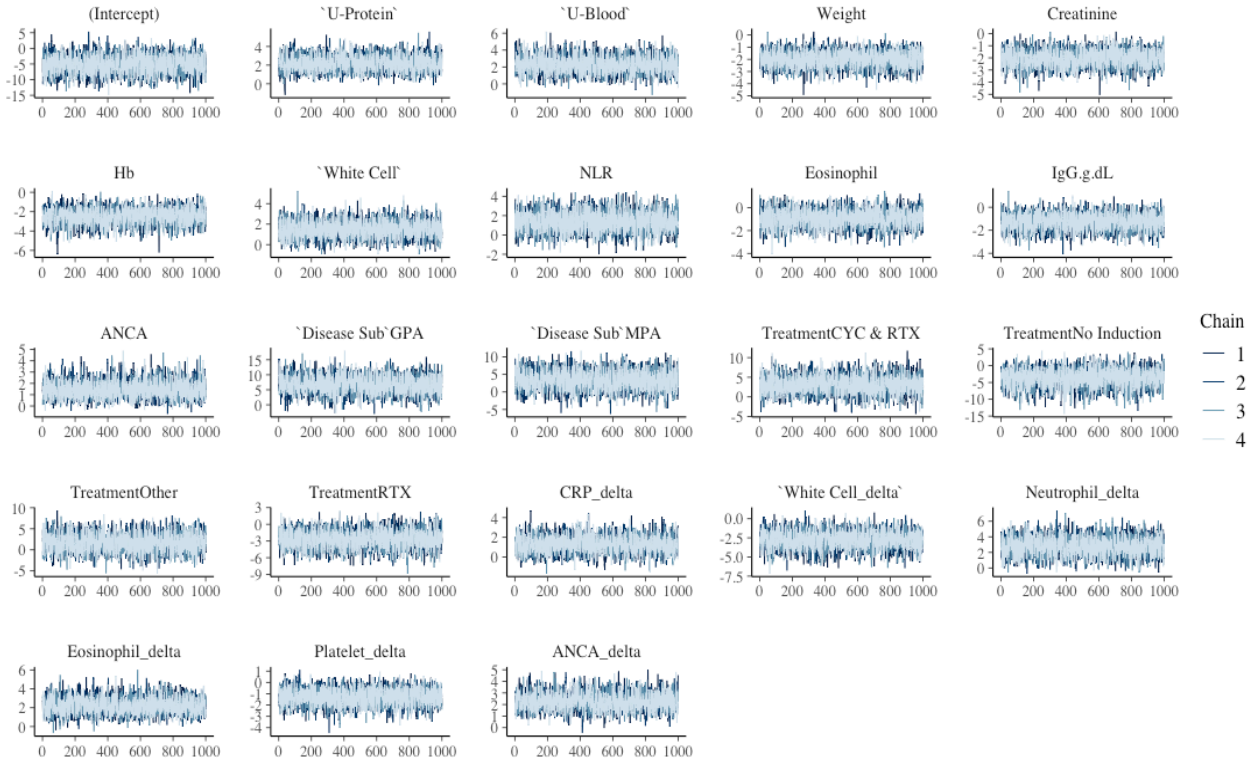

Figure 36 - Bayesian Logistic Regression MCMC Trace Plot Delta Group

Lasso Logistic Regression Imputation Plot

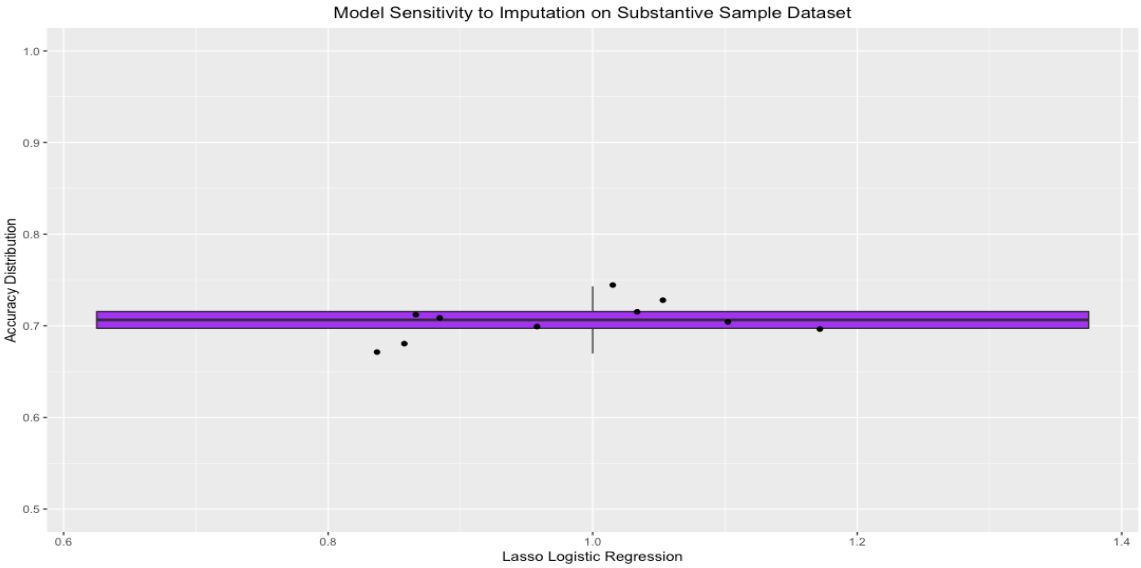

Figure 37 - Lasso Logistic Regression Model Sensitivity to Imputation on Substantive Sample