

Development of machine learning based computational tools for stratified healthcare and personalised medicine

Bodhayan Prasad

Faculty of Life & Health Sciences



A thesis submitted for the degree of

Doctor of Philosophy (PhD)

June 2023

I confirm that the word count of this thesis is less than 100,000 words

*To all my teachers,
especially my parents.*

Contents

Acknowledgements	vii
Abbreviations	ix
Declaration	xv
Abstract	xvii
Chapter 1 General Introduction.....	1
1.1 Introduction	3
1.2 The need for precision medicine	4
1.2.1 Stratified healthcare	4
1.2.2 Personalised medicine	5
1.3 Different datasets used in this thesis	5
1.3.1 PMC database (in-house dataset)	5
1.3.2 UK Biobank (public dataset).....	6
1.4 Machine Learning (ML)	6
1.4.1 Unsupervised Learning	7
1.4.2 Supervised Learning.....	7
1.5 Gaps in the field and the problem statement of this thesis	7
1.6 Aims of the thesis	8
1.7 Thesis outline	9
1.8 Posters, publications, patents and funding	9
1.8.1 Journal publications	9
1.8.2 Patents filed from the work done in this thesis	10
1.8.3 Commercialisation funding secured from the work done in this thesis	10
1.8.4 Conference posters	10
1.8.5 Other relevant peer-reviewed publications and patents that are not included in this thesis	10
1.9 References	11

Chapter 2	Literature Review	15
2.1	Introduction	17
2.2	Premise and its scope	18
2.2.1	Clustering approaches for stratified healthcare.....	18
2.2.2	Predictor developments in personalised medicine	19
2.3	Popular ML models applied in the fields of stratified healthcare and personalised medicine	19
2.3.1	Linear models.....	19
2.3.2	Decision Tree (DT)	20
2.3.3	Random Forest (RF).....	20
2.3.4	Gradient Boosting algorithm.....	20
2.3.5	Support Vector Machine (SVM).....	21
2.3.6	Naïve Bayes (NB)	21
2.3.7	Neural Networks (NNs)	21
2.4	Stratified healthcare and personalised medicine in practice	22
2.4.1	Diabetes Meletus (DM).....	22
2.4.2	Cardio-vascular disease (CVD)	22
2.4.3	Rheumatoid arthritis (RA)	23
2.4.4	Inflammatory bowel disease (IBD).....	23
2.4.5	Chronic obstructive pulmonary disease (COPD).....	23
2.4.6	Chronic kidney disease (CKD)	24
2.4.7	Cancer	24
2.4.8	Mental health.....	24
2.4.9	Role of multimorbidity.....	24
2.5	Limitations in research	26
2.6	Research gaps, open challenges and opportunities	27

2.7	Conclusion.....	28
2.8	References	28
Chapter 3	Data-driven patient stratification of UK Biobank cohort suggests five endotypes of multimorbidity	35
3.1	Introduction	37
3.2	Material and methods	39
3.2.1	Datasets	41
3.2.2	Statistical, computational and bioinformatic analyses	41
3.2.3	Multiple correspondence analysis	41
3.2.4	Cluster validation	41
3.2.5	Network analysis	42
3.3	Results	42
3.3.1	Cluster analysis and its validation	42
3.3.2	Exploratory data analysis on patients with multimorbidity	43
3.3.3	Disease–disease interaction network.....	46
3.4	Discussion	47
3.5	References	51
Chapter 4	ATRPred: A machine learning based tool for clinical decision making of anti-TNF treatment in rheumatoid arthritis patients.....	57
4.1	Introduction	59
4.2	Design and implementation.....	61
4.2.1	Ethics statement	61
4.2.2	Patient recruitment and selection criteria.....	61
4.2.3	Sample collection and collation of clinical information	61
4.2.4	Plasma protein profile	62
4.2.5	Statistical, computational and bioinformatics analyses	62
4.2.6	Feature selection with machine learning.....	63

4.2.7	Machine learning based model development.....	64
4.2.8	ATRPred tool development	64
4.3	Results	65
4.3.1	Exploratory data analysis on plasma proteins.....	66
4.3.2	Anti-TNF response feature selection and classifier	68
4.3.3	Plasma protein model for clinical decision making	71
4.3.4	Enrichment analysis with Gene Ontology (GO) terms and KEGG pathways	74
4.3.5	Network analysis.....	74
4.4	Discussion	75
4.5	References	79
Chapter 5	muSignAI: An algorithm to search for multiple omic signatures with similar predictive performance	87
5.1	Introduction	89
5.2	Algorithm development.....	90
5.3	Algorithm evaluation.....	91
5.4	References	94
Chapter 6	General discussions	97
6.1	The need for computational tools.....	99
6.2	Summary and key findings.....	100
6.2.1	Stratified healthcare in multimorbidity	100
6.2.2	Personalised medicine in RA	101
6.2.3	Computational tool for multiple signature detection	101
6.3	Reasoning and critical review of ML models applied in the thesis.....	102
6.3.1	Clusters of multimorbidity (MulMorPip)	102
6.3.2	Endotypes of RA.....	103
6.3.3	Anti-TNF treatment response predictor (ATRPred)	104

6.3.4 Multiple signature prediction algorithm (muSignAI)	105
6.4 Challenges and limitations	105
6.5 Future perspectives and directions	106
6.6 Conclusions	107
6.7 References	107
Appendix I Supplementary data for Chapter 3.....	109
Appendix II Supplementary data for Chapter 4	113

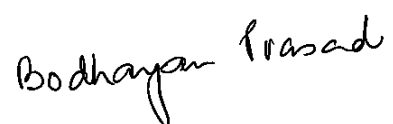
Acknowledgements

It is a great pleasure to thank my supervisors Dr. Priyank Shukla and Prof. Anthony J. Bjourson, without whom, I would not be able to see this day. Further, I would also like to thank my colleagues at C-TRIC, especially Dr. David Gibson for co-supervising one of my research articles; Dr. Paula McClean and Dr. Andrew English for including me for validating DiaStrat cohort research; Dr. Victoria McGilligan for making me understand CVD; Dr. Elaine Murray for making me understand Mental Health; and Dr. Taranjit Rai for being a mentor and friend. I am also thankful to the Doctoral College, especially Prof. Alison Gallagher, whose constant support has kept me sane during my emotional roller-coaster days.

I am also thankful to ResLife Magee, especially Mary Murphy, Isobel Smith and Geraldine Black, who constantly helped me during my stay at Coppin House. I am also thankful to Noel Christy and Caroline McNutt from the international student experience department, who made my PhD days enjoyable and chatty. I would also like to mention the help from UUSU, especially from Collette Cassidy and Emer Smith for motivating and making me student counsellor and Globetrotter Magee Society Chairperson.

I am grateful for the funding support of Vice-Chancellor's Research Scholarship (VCRS), which enabled me to study at Ulster University. Moreover, funding and mentoring support from InnovateUK's NxNW ICURe programme and InvestNI's PoC Stage 1 programme, has helped me in gaining experience in market research and commercialisation aspect. The research was carried out with the support from the European Union Regional Development Fund (ERDF), EU Sustainable Competitiveness Programme for Northern Ireland & the Northern Ireland Public Health Agency.

I would like to thank my internal examiner Prof. Huiru (Jane) Zheng and external examiner Professor Pier Luigi Martelli for a rigorous defense and Prof. Ibrahim Banat for agreeing to chair my viva at a very short notice. Finally, I've been thankful for all my teachers in life, especially my parents and hence I would like to dedicate my thesis to all of them.



Bodhayana Prasad

Abbreviations

<i>Acronym</i>	<i>Definition</i>
AMS	The Academy of Medical Sciences
ACS	Acute Coronary Syndrome
AKI	Acute Kidney Injury
ATRPred	Anti-TNF Treatment Response Predictor
AUC	Area Under the ROC Curve
bDMARD	biologic DMARD
BHSCT	Belfast Health & Social Care Trust
BLDAS	Baseline DAS
BP	Biological Process
BSR	British Society for Rheumatology
CAD	Coronary Artery Disease
CC	Cellular Component
CC BY 4.0	Creative Commons Attribution 4.0 International Public License
CCI	Charlson Comorbidity Index
CCC	Charlson's Comorbidity Classification
CD	Crohn's Disease
CDC	Centers for Disease Control and Prevention
cDMARD	conventional DMARD
CHD	Coronary Heart Disease

CHF	Congestive Heart Failure
CKD	Chronic Kidney Disease
COPD	Chronic Obstructive Pulmonary Disease
CRUK	Cancer Research UK
CV	Cross-Validation
CVD	Cardio-vascular Disease
DAS	Disease Activity Score
DAS-28	DAS across 28 joints
DM	Diabetes Meletus
DMcx	DM with chronic complications
DMARD	Disease Modifying Anti-Rheumatic Drug
DTC	Decision Tree Classifier
ECG	Electro-cardiogram
ECR	Electronic Care Record / Early Career Researcher
EMT	Epithelial-Mesenchymal Transition
ES	Effect Size
ESR	Erythrocyte Sedimentation Rate
EU	European Union
EULAR	European League Against Rheumatism
GDP	Gross Domestic Product
GLM	Generalised Linear Model

GO	Gene Ontology
GP	General Practitioner
GWAS	Genome-Wide Association Studies
FDR	False Discovery Rate
FI	Feature Importance
IBD	Inflammatory Bowel Disease
ICD-10	International Classification of Diseases 10 th revision
ICURe	Innovation to Commercialization of University Research
IHD	Ischemic Heart Disease
IMD	Index of Multiple Deprivation
k-NN	k-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LiverMild	Mild Liver disease
LiverSevere	Severe Liver disease
LoD	Limit of Detection
LOOCV	Leave-One-Out Cross-Validation
MACE	Major Adverse Cardiac Event
MCA	Multiple Correspondence Analysis
MCC	Multiple Chronic Conditions / Matthews Correlation Coefficient
MF	Molecular Function
MI	Myocardial Infarction

ML	Machine Learning
MRC	Medical Research Council
MulMorPip	Multimorbidity analysis Pipeline
muSignAl	multiple Signature Algorithm
NB	Naïve Bayes
NHS	National Health Service
NICE	National Institute of Health and Care Excellence
NN	Neural Network
NPX	Normalised Protein Expression
NxNW	North/North-West
OA	Open Access
ORECNI	Office of Research Ethics Committees Northern Ireland
PCA	Principal Component Analysis
PEA	Proximity Extension Array
PMC	Personalised Medicine Centre
PoC	Proof of Concept
PPI	Protein-Protein Interaction
PPIN	PPI Network
PRESS	Predicted Sum of Squares
PRS	Polygenic Risk Score
PUD	Peptic Ulcer Disease

PVD	Peripheral Vascular Disease
RA	Rheumatoid Arthritis
ROC	Receiver Operator Characteristics
SDG	Sustainable Development Goals
T2DM	Type 2 DM
TNF	Tumor Necrosis Factor
ML	Machine Learning
MDS	Multidimensional Scaling
NIH	National Institutes of Health
NKF	National Kidney Foundation
QC	Quality Control / Quality Check
UA	Unstable Angina
UC	Ulcerative Colitis
UK	United Kingdom
UN	United Nations
UREC	Ulster University Research Ethics Committees
WBC	White Blood Cells
WHO	World Health Organization
WHSC	Western Health & Social Care Trust

Declaration

Note on access to contents

I hereby declare that, with effect from the date on which the thesis is deposited in the Doctoral College of Ulster University, I permit:

1. The Librarian of the University to allow the thesis to be copied in whole or in part without reference to me on the understanding that such authority applies to the provision of single copies made for study purposes or for inclusion within the stock of another library.
2. The thesis to be made available through the Ulster Institutional Repository and/or EThOS under the terms of the Ulster eTheses Deposit Agreement which I have signed.

IT IS A CONDITION OF USE OF THIS THESIS THAT ANYONE WHO CONSULTS IT MUST RECOGNISE THAT THE COPYRIGHT RESTS WITH THE AUTHOR AND THAT NO QUOTATION FROM THE THESIS AND NO INFORMATION DERIVED FROM IT MAY BE PUBLISHED UNLESS THE SOURCE IS PROPERLY ACKNOWLEDGED.

Abstract

The medical field is getting flooded with high dimensional datasets, with the advent of high throughput technologies in the omics era. This calls for more sophisticated data analyses technology. Machine learning provides a generic solution to address all kinds of datasets, including high dimensional datasets.

Stratified healthcare involves stratifying patients into different risk groups or ‘endotypes’, which helps in optimising their disease management, whereas personalised medicine involves tailoring the diagnosis and treatment for the patients as per the individual’s biological makeup or use of ‘biomarkers’ in terms of genomic, epigenetic, transcriptomic, proteomic, or other omics profiles. An unsupervised machine learning technique helps in identification of endotypes, whereas a supervised machine learning technique helps in identification of biomarkers.

An unsupervised learning pipeline denoted as MulMorPip, was developed and applied to stratify patients with multimorbidity, in order to identify clusters based on disease diagnosis and interactions (Chapter 3). We have found evidence for five endotypes in patients with multimorbidity using this unsupervised approach. Further, two endotypes of RA were discovered using an unsupervised learning technique and a predictor denoted as ATRPred was developed for the prognosis of anti-TNF treatment response of rheumatoid arthritis patients using a supervised learning technique (Chapter 4). Furthermore, an algorithm muSignAl, was developed that can report multiple signatures with similar predictive power in case of high dimensionality data (Chapter 5).

The thesis has attempted to build computational methods/tools using various machine learning techniques for stratified healthcare and personalised medicine approaches in multimorbidity. These tools can be extended to similar applications in other disease conditions.

Chapter 1 General Introduction

Contribution

Bodhayan Prasad carried out all the research unless otherwise stated.

Dr. Priyank Shukla and Prof. Anthony J. Bjourson supervised the research and proofread.

1.1 Introduction

Over the past 25 years the application of evidence-based medicine has been based mainly on averages (as generally done by a statistician) and it has frequently failed to optimally improve the quality of life for the patients [1]. The need for a stratified approach in this regard can be achieved via identifying distinct patient groups, who are homogenous in terms of their genotype, treatment response, phenotypic data and various other features. Stratified healthcare involves stratifying patients into smaller groups for better disease management, whereas personalised medicine is a branch of science that deals with personalising the diagnosis and treatment for specific individual patients. Stratified healthcare involves effects of medicine at a population level whereas, personalised medicine involves individualised unique diagnosis, treatment, drug targeting and development [2]. Since every individual is unique and several treatment pathways are usually available for clinicians to prescribe, personalised medicine offers the promise of bridging the gap between the two. Stratified Medicine is one of the priority research areas for Innovate UK, the Medical Research Council, Academy of Medical Science UK [3].

Inflammation is a common etiology in many different diseases. This includes diabetes mellitus, rheumatoid arthritis, CVD, cancer, etc. Inflammation is a defense mechanism involving white blood cells (WBC) and substances produced by them; to protect from infection with micro-organisms, such as fungi, bacteria and viruses or abnormal somatic cells that may arise throughout the life course. However, in some diseases like arthritis, the immune system triggers an inappropriate inflammatory response, even when there are no such micro-organisms. These types of diseases are called auto-immune diseases. Most of the current therapies do not address the nexus between inflammation and the existence of multiple chronic conditions in a given patient, as most of the clinical trials exclude patients with multimorbidity and concentrate on single index disease conditions. Inflammation is a major risk factor in the development of chronic conditions such as neurodegeneration, cancer diabetes, cardiovascular disease (CVD), arthritis, chronic obstructive pulmonary disease (COPD), obesity and inflammatory bowel disease (IBD). Chronic inflammation involves multiple signaling pathways with potential paracrine and autocrine networks. Some of the pathways associated with the initiation and progression of inflammation include p38 MAPK, IL-6/JAK/STAT3, PI3K and Hippo pathway [4].

Chapter 1: General Introduction

The United Nations (UN) Sustainable Development Goal (SDG) No. 3 focuses on good health and well-being [5]. Better disease management ensures good health and well-being. Stratifying patients into smaller groups can pave the way to assess different risks better and thereby help in managing disease optimally. Further, medical science is developing at a faster rate, with many new treatments and medicines being developed. However, we also need a way to personalize these treatments for the patients. This will not only reduce the cost spent on unwanted treatments that cause harm or have no clinical benefit, but also improve the quality of life for patients as delayed response can worsen the patient's conditions. Thus, stratified healthcare and personalised medicine aspires to address the UN SDG No. 3 in the most effective manner.

Machine learning (ML) is being widely exploited across the domains of stratified healthcare and personalised medicine [6]. ML based algorithms provide solutions for prediction or classification problems using supervised or unsupervised approach, respectively. With the continued deployment of advanced high-throughput omics technologies (especially Next Generation Sequencing or NGS) in clinical practice, ML offers the promise of significant opportunities to assist with the analysis of the terabytes of clinical and omics data being generated from patients. The computational methods developed would have a potential translational impact by assisting clinical researchers in moving the field of personalised medicine further towards clinical utility.

1.2 The need for precision medicine

With the increasing use of technology in medical sciences, a lot of optimisations are now becoming possible, leading to precision medicine. Precision medicine not only helps in optimising performance of our healthcare (stratified healthcare) [7-9], but also tailoring treatment as per the patients' biological makeup (personalised medicine) [10-12]. Some of the major advantages in stratified healthcare and personalised medicine is expanded below:

1.2.1 Stratified healthcare

Stratified healthcare involves finding the key differences at the population level, so that they can be grouped into molecular sub-classes (or endotypes), so that better optimisation of their health can be achieved. For example, GPs at primary care can identify high-risk patients and

speed-up their treatment in secondary care. In public healthcare, such as the National Health Service (NHS) in the UK, patients are experiencing higher waiting time, due to lack of resources. Stratified healthcare has the potential of optimising healthcare, so that most can be achieved, based on limited resources. A recent study [13], has shown the disparity in healthcare utilization based on burden of different chronic conditions. Improving on these is useful in achieving universal health coverage.

1.2.2 Personalised medicine

Personalised medicine involves tailoring treatment as per the patient's biological make-up. A lot of medicines are being developed and approved for different diseases. However, only a few prognostic tests are available to personalise these treatments. Success of drugs like warfarin, PQ and imatinib, which only works on certain genetic profile, without the side effect has instilled the interest, amongst the research community, to identify such factors [14]. Personalising these medicines, will not only help in saving cost of the drugs, but also saving patients from refractory condition of their diseases and slowing down their disease progression.

1.3 Different datasets used in this thesis

In this thesis, computational analysis was done on the following in-house and public datasets:

1.3.1 PMC database (in-house dataset)

Patients were recruited in different disease areas at the Personalised Medicine Centre (PMC) of Ulster University. These diseases include those that have a mostly inflammatory etiology – rheumatoid arthritis (RA), diabetes, cardio-vascular disease (CVD), cancer and mental health. Ethical approvals were obtained from Office for Research Ethics Committees Northern Ireland (ORECNI), Ulster University Research Ethics Committee (UREC), Belfast Health and Social Care Trust (BHSCT) and Western Health and Social Care Trust (WHSC). Formal written informed consent was obtained from all participants enrolled in the research studies to permit publication of anonymised clinical data.

The analysis reported in this thesis was done on patients recruited for RA. It is a cohort of 144 patients (cases). Four Olink Proteomics panels (Immunology, CVD-I, CVD-II and

Chapter 1: General Introduction

Inflammatory panels) composed of 92 proteins each (i.e., 4 X 92 features) were examined at baseline (T0-prior to the specific treatment) as well as at T0 + 6 months post treatment, to classify RA patients for response to treatment based on EULAR (European League Against Rheumatism) guidelines.

1.3.2 UK Biobank (public dataset)

UK Biobank (<https://www.ukbiobank.ac.uk>) is an online biomedical resource of about 500,000 participants recruited across Great Britain (England, Scotland and Wales), UK during 2006-10. These participants were aged 40-69 years at the time of recruitment. The database contains genomic, imaging and biological assay data along with meta-data such as demographic details. Further, consent was taken from these participants to access their electronic care record (ECR) and the database is updated periodically. Access to UK Biobank resources was obtained through application number 48433 [15].

1.4 Machine Learning (ML)

Recent advances in hardware technology have made the application of ML possible in practice, where large data is involved. ML could be useful for identifying complex patterns from gene expression, variant-calling and methylation profiles of patients for stratification [16-19]. ML can broadly be classified into supervised and unsupervised learning, as shown in Figure 1.1 [20].

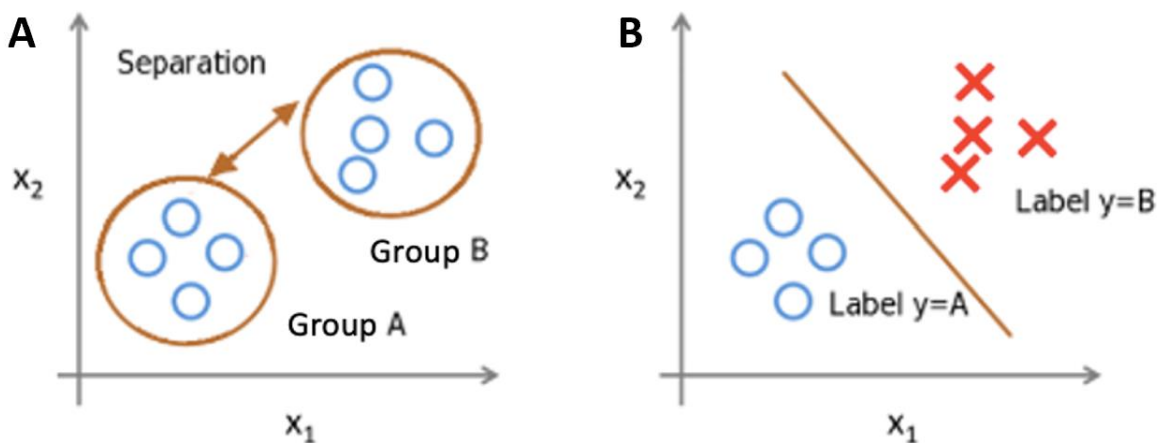


Figure 1.1 Machine Learning techniques. x_1 and x_2 are features of dataset. The circle or cross represents data points in the scatter plot. A and B are data groups or labels for

unsupervised or supervised learning respectively [20]. **(A)** Unsupervised Learning labels the data into group A and B based on their Euclidean separation. **(B)** Supervised Learning finds a discriminating pattern (i.e. line) dividing two labels A and B.

1.4.1 Unsupervised Learning

We often see data associated with a broad class, for example a disease. Not all patients respond to their treatment the same way in the broad class, which makes it difficult for the clinicians to tailor the treatment as per their need. In such a case, one can try using unsupervised learning techniques to identify if there exists any molecular subclass, formally known as endotypes, which can be used to stratify patients and tailor a unique treatment plan for each endotype, eventually leading to better disease management for all the patients. Figure 1.1A shows an example case of unsupervised learning. We can clearly view a separation between two clusters of data points (patients). Thus, we can stratify the data into two sub-class i.e., Group A and B. In a real-world scenario of multiple features, the ML algorithm will try to identify the separation in multi-dimensional space.

1.4.2 Supervised Learning

Some of the data (e.g., biological assays, omics, etc.) in medical settings are often labelled, for example, diseased or healthy, responders or non-responders. Supervised learning uses these labels to identify patterns in the data that can be used as biomarkers for prognostic or diagnostic applications. This is helpful for the clinicians to make better informed decisions about prescription and treatment pathways. Figure 1.1B shows the dataset with feature X_1 and X_2 (e.g. Expression of protein 1 and 2) and labels A and B (responder and non-responder to a drug). The line separating the two labels is a discriminatory pattern identified by the ML algorithm and used as a model to predict the labels. In a real-world scenario, there will be multi-dimensional data or feature sets and the ML algorithm tries to identify the best hyperplanes, separating the groups.

1.5 Gaps in the field and the problem statement of this thesis

Despite lots of advancements in the medical informatics field, a lot of questions still remain unanswered. With the advent of modern omics technology and its costs going down year by

Chapter 1: General Introduction

year, a lot of data is being generated in the field [21]. This dataset opens up a lot of opportunities for identifying patterns that may be helpful in further stratifying the patients and designing tailor-made therapies for each of the strata. This calls for increasing demand for computational and bioinformatic techniques, especially in the omics domain.

This thesis focuses on development of various ML based computational tools that can address two major open challenges:

- Stratified healthcare challenge: Identification of molecular subclasses or endotypes of patients, that can help towards stratified healthcare and thereby lead to better patient management.
- Personalised medicine challenge: Identification of biomarkers for personalised medicine, that can help in deciding the most effective treatment pathway.

1.6 Aims of the thesis

In this thesis I have tried to identify clusters (viz. endotypes) and features (viz. biomarkers), using unsupervised and supervised ML based approach, respectively. This study develops different computational tools that can help in stratified healthcare and personalised medicine. The study is extended to topological or network analysis techniques to find out complex patterns in the dataset. Further, Gene-Set Enrichment Analysis was undertaken to help understand the underlying biology. These pipelines developed can also be extended and applied to different disease conditions. In summary, I have focused on the following three aims:

1. To identify endotypes of patients, that can lead to better patient management (stratified healthcare challenge).
2. To identify biomarkers of treatment response that can help in making better informed decisions (personalised medicine challenge).
3. To apply network and/or gene-set enrichment analysis to understand the underlying biology of endotypes and biomarkers identified in 1 and 2.

1.7 Thesis outline

The remainder of this thesis is as follows:

- Chapter 2 sets the premise and its scope with a detailed literature review on different ML models and methods that are being deployed in medical sciences, followed by their implementation for stratified healthcare and personalised medicine in practice. It ends with discussions on limitations and research gaps in current medical research.
- In Chapter 3, we present our published findings wherein we have stratified UK Biobank multimorbid patients into five endotypes, using unsupervised ML approach. To further understand the differences between those endotypes, we have also presented disease-disease interaction networks underlying the patients with multimorbidity.
- In Chapter 4, we present our published work of identifying RA endotypes, using an unsupervised ML approach, which were not related to anti-TNF therapy response and development of a prognostic test to identify responders to anti-TNF treatment in RA, using a supervised ML approach.
- In Chapter 5, we present our published work of developing an algorithm that can identify multiple signatures in high dimensionality data.
- Finally, we have summarised the aforesaid research, talked about our limitations and suggested some future developments that can be carried out, in Chapter 6.

1.8 Posters, publications, patents and funding

The research works carried out on the thesis are published/submitted as following:

1.8.1 Journal publications

- **Prasad B**, Bjourson AJ, Shukla P. Data-driven patient stratification of UK Biobank cohort suggests five endotypes of multimorbidity. *Brief Bioinform.* 2022 Nov 19;23(6):bbac410. doi: 10.1093/bib/bbac410. PMID: 36209412; PMCID: PMC9677496.
- **Prasad B**, McGeough C, Eakin A, Ahmed T, Small D, Gardiner P, Pendleton A, Wright G, Bjourson AJ, Gibson DS, Shukla P. ATRPred: A machine learning based

Chapter 1: General Introduction

tool for clinical decision making of anti-TNF treatment in rheumatoid arthritis patients. PLoS Comput Biol. 2022 Jul 5;18(7):e1010204. doi: 10.1371/journal.pcbi.1010204. PMID: 35788746; PMCID: PMC9321399.

- **Prasad B**, Bjourson AJ, Shukla P. muSignAl: An algorithm to search for multiple omic signatures with similar predictive performance. Proteomics. 2023 Jan;23(2):e2200252. doi: 10.1002/pmic.202200252. Epub 2022 Oct 3. PMID: 36076312.

1.8.2 Patents filed from the work done in this thesis

- A UK wide patent filed for proteomic signature in ATRPred against application number 2108522.0 dated 15.06.2021.

1.8.3 Commercialisation funding secured from the work done in this thesis

- ATRPred, a proteomic based signature to predict response to anti-TNF therapy in rheumatoid arthritis patients, was selected in the top 12 projects cohort G for funding market research Cohort G of NxNW ICUR programme (<https://www.nxnwpartners.org/our-programmes/icure>), funded by Innovate UK. Bodhayan Prasad as an early career researcher (ECR) lead the project for 16 weeks (24 May 2021 - 9 Sep 2021).

1.8.4 Conference posters

- **Prasad B**, Gibson D, McGeough C, Eakin A, Ahmed T, Small D, Gardiner P, Pendleton A, Wright G, Bjourson AJ, Shukla P 2019, 'Stratifying patients with inflammatory diseases based on treatment response', Mathematical and statistical explorations in disease modelling and public health, Bengaluru, India, 1/07/19 - 6/07/19.

1.8.5 Other relevant peer-reviewed publications and patents that are not included in this thesis

- English AR, **Prasad B**, McGuigan DH, Horigan G, O'Kane M, Bjourson AJ, Shukla P, Kelly C, McClean PL. Simvastatin is associated with superior lipid and glycaemic

control to atorvastatin and reduced levels of incident Type 2 diabetes, in men and women, in the UK Biobank. *Endocrinol Diabetes Metab.* 2022 May;5(3):e00326. doi: 10.1002/edm2.326. Epub 2022 Mar 4. PMID: 35243827; PMCID: PMC9094470.

- Shukla P, Pandey P, **Prasad B**, Robinson T, Purohit R, D'Cruz LG, Tambuwala MM, Mutreja A, Harkin J, Rai TS, Murray EK, Gibson DS, Bjourson AJ. Immunoinformatics analysis predicts B and T cell consensus epitopes for designing peptide vaccine against SARS-CoV-2 with 99.82% global population coverage. *Brief Bioinform.* 2022 Jan 17;23(1):bbab496. doi: 10.1093/bib/bbab496. PMID: 34962259; PMCID: PMC8769887.
- A UK wide patent filed for peptide-based epitope in PVPred against application number 2102598.6 dated 24.02.2021.
- A peptide-based vaccine against SARS-CoV-2 using PVPred, was selected for funding market research under the Proof of Concept (PoC) Stage 1 (<https://www.investni.com/support-for-business/proof-of-concept>), funded by Invest NI. Bodhayan Prasad as an early career researcher (ECR) lead the project for 3 months (24 Oct 2021 - 24 Jan 2022).

1.9 References

1. Fortin, M., Lapointe, L., Hudon, C., Vanasse, A., Ntetu, A.L. and Maltais, D., 2004. Multimorbidity and quality of life in primary care: a systematic review. *Health and Quality of life Outcomes*, 2(1), p.51.
2. Erikainen S, Chan S. Contested futures: envisioning "Personalized," "Stratified," and "Precision" medicine. *New Genet Soc.* 2019 Jul 12;38(3):308-330.
3. AMS (n.d.), Stratified Medicine. Available at: <https://acmedsci.ac.uk/policy/policy-projects/Stratified-medicine> (Accessed: 25 March 2022).
4. Yeung, Y.T., Aziz, F., Guerrero-Castilla, A. and Arguelles, S., 2018. Signaling pathways in inflammation and anti-inflammatory therapies. *Current pharmaceutical design*, 24(14), pp.1449-1484.
5. UN (n.d.), The 17 goal. Available at: <https://sdgs.un.org/goals> (Accessed: 25 March 2022).
6. Cirillo D and Valencia A. Big data analytics for personalized medicine. *Curr Opin Biotechnol.* 2019 Aug;58:161-167.

Chapter 1: General Introduction

7. Dalal V, Carmicheal J, Dhaliwal A, Jain M, Kaur S, Batra SK. Radiomics in stratification of pancreatic cystic lesions: Machine learning in action. *Cancer Lett.* 2020 Jan 28;469:228-237. doi: 10.1016/j.canlet.2019.10.023. Epub 2019 Oct 17.
8. Ellahham S. Artificial Intelligence: The Future for Diabetes Care. *Am J Med.* 2020 Aug;133(8):895-900. doi: 10.1016/j.amjmed.2020.03.033. Epub 2020 Apr 20.
9. Koski E, Murphy J. AI in Healthcare. *Stud Health Technol Inform.* 2021 Dec 15;284:295-299.
10. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. *J Am Coll Cardiol.* 2017 May 30;69(21):2657-2664.
11. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, Fuchs BD, Meadows L, Lynch M, Donnelly PJ, Pavan K, Fishman NO, Hanson CW 3rd, Umscheid CA. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice. *Crit Care Med.* 2019 Nov;47(11):1485-1492.
12. Boutet A, Madhavan R, Elias GJB, Joel SE, Gramer R, Ranjan M, Paramanandam V, Xu D, Germann J, Loh A, Kalia SK, Hodaie M, Li B, Prasad S, Coblenz A, Munhoz RP, Ashe J, Kucharczyk W, Fasano A, Lozano AM. Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. *Nat Commun.* 2021 May 24;12(1):3043.
13. Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. *Fertil Steril.* 2018 Jun;109(6):952-963.
14. Mahumud RA, Gow J, Mosharaf MP, Kundu S, Rahman MA, Dukhi N, Shahajalal M, Mistry SK, Alam K. The burden of chronic diseases, disease-stratified exploration and gender-differentiated healthcare utilisation among patients in Bangladesh. *PLoS One.* 2023 May 2;18(5):e0284117.
15. Bjourson AJ et al. (2019), Computational analyses of genotypic and phenotypic data with treatment response from patients with multimorbidity and the role of inflammation as a driver of multimorbidity. Available at: <https://www.ukbiobank.ac.uk/enable-your-research/approved-research/computational-analyses-of-genotypic-and-phenotypic-data-with-treatment-response>

[response-from-patients-with-multimorbidity-and-the-role-of-inflammation-as-a-driver-of-multimorbidity](#) (Accessed: 29.04.2022).

16. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. *Nat Genet.* 2019 Jan;51(1):12-18.
17. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* 2019 Nov 19;11(1):70. doi: 10.1186/s13073-019-0689-8.
18. van den Akker J, Mishne G, Zimmer AD, Zhou AY. A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. *BMC Genomics.* 2018 Apr 17;19(1):263.
19. Lian H, Han YP, Zhang YC, Zhao Y, Yan S, Li QF, Wang BC, Wang JJ, Meng W, Yang J, Wang QH, Mao WW, Ma J. Integrative analysis of gene expression and DNA methylation through one-class logistic regression machine learning identifies stemness features in medulloblastoma. *Mol Oncol.* 2019 Oct;13(10):2227-2245.
20. Ng A, Machine learning course. Available at: <https://www.coursera.org/learn/machine-learning> (Accessed: 07.03.2021).
21. Perez-Riverol Y, Zorin A, Dass G, Vu MT, Xu P, Glont M, Vizcaíno JA, Jarnuczak AF, Petryszak R, Ping P, Hermjakob H. Quantifying the impact of public omics data. *Nat Commun.* 2019 Aug 5;10(1):3512.

Chapter 2 Literature Review

Contribution

Bodhayan Prasad carried out all the research unless otherwise stated.

Dr. Priyank Shukla and Prof. Anthony J. Bjourson supervised the research and proofread.

2.1 Introduction

Traditionally, medicines have been developed for broad disease classes. Many of these medicines don't work on the patients and they are shifted to another treatment. Further, a lot of medicine are getting approved on regular basis. With the advent of precision medicine, clinicians can now choose a medicine, based on their patients' genes, biochemicals and environments, in order to better optimise the treatments. Precision medicine tries to move from traditional one-size fits-all medicine, by initially stratifying the patients in smaller sub-groups, and then finally personalising their treatment, as shown in Figure 2.1 [1].

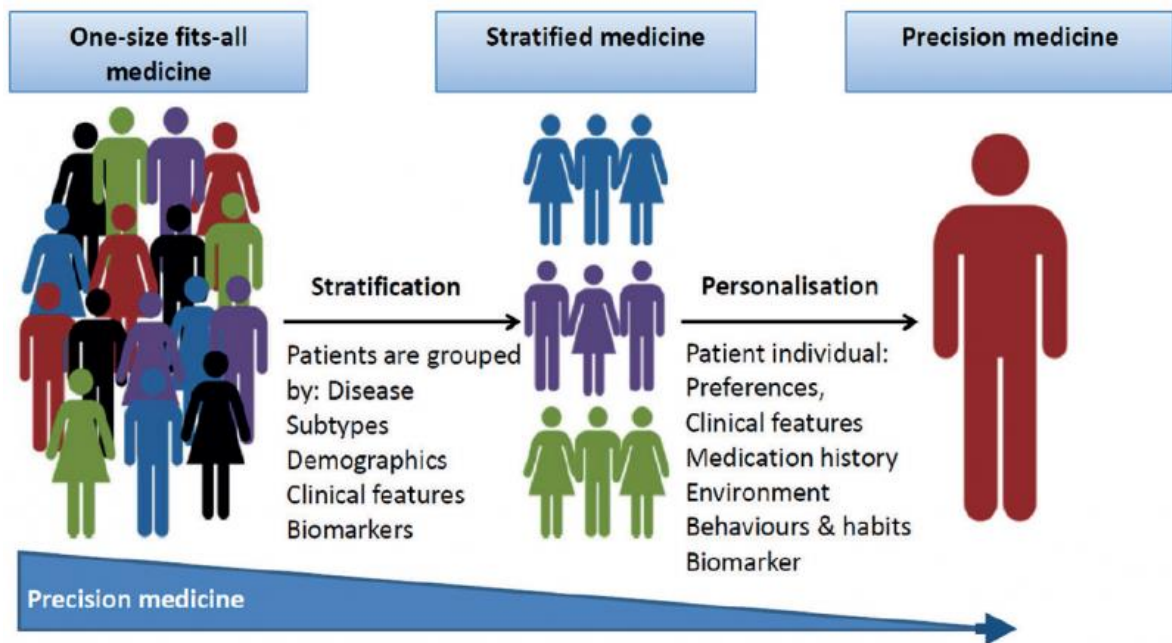


Figure 2.1 The shift of traditional medicine into precision medicine (from stratification to personalisation) [1].

Stratified healthcare involves identifying key differences at population level to stratify (or sub-group) patients. Whereas personalised medicine involves personalising (or tailoring) the treatment, as per different clinical and environmental biomarkers. Stratified healthcare groups patients with similar demographics and medical history and tries to predict risks of any disease incidence, early diagnosis and disease progression, so that clinicians can better manage and prioritise treatment for the high-risk group. Further, within these groups, personalised medicine tries to identify treatment with best outcome.

2.2 Premise and its scope

In this literature review, we will start by discussing different machine learning (ML) models, that is popular amongst the medical research community, followed by their implementation and usage in practice. Finally, we will try to identify their limitations and any research gaps that we see, which we would like to address in upcoming chapters.

We will be assessing each of these different ML methods in two broad contexts viz., unsupervised and supervised approach, that is used in stratified healthcare and personalised medicine respectively. Stratified healthcare tries to sub-group patients into broad groups using mostly demographic variables like age, sex, ethnicity, diet, etc. [2]. Therefore, it uses unsupervised ML or clustering methods to identify patient heterogeneity. In contrast, personalised medicine uses molecular profiling to tailor the right therapeutic strategy. This is manifested as a ‘companion test’ or clinical biomarker for assessing therapeutic response [3]. Therefore, personalised medicine deploys supervised ML or predictors to prognose. These ML manifestations in two broad contexts are briefly introduced in the following sub-sections.

2.2.1 Clustering approaches for stratified healthcare

Stratified healthcare identifies different clusters of patients. This involves an unsupervised approach to ML. Unsupervised ML techniques try to discover patterns in the data, so that it can be sub-divided into smaller groups. Some of the most popular techniques include hierarchical clustering, k-means clustering and Gaussian mixture models.

In the study by Mossotto et al [4], in order to check if clinical features in inflammatory bowel disease (IBD) segregate into two clusters of Chron’s disease (CD) and ulcerative colitis (UC), principal component analysis (PCA) for linear feature association as well as multidimensional scaling (MDS) for non-linear feature association were applied. A study conducted by Pen et al [5] used a consensus clustering approach on gene expression dataset of rheumatoid arthritis in order to have statistically robust partitions of data. Orange et al [6] used k-means clustering for patient stratification in juvenile-onset systemic lupus erythematosus (SLE). Martin-Gutierrez et al [7] stratified patients with Sjögren's syndrome (SS) along with SLE and using k-means clustering and clinical trajectory analysis.

2.2.2 Predictor developments in personalised medicine

One of the aims in personalised medicine is to predict response to different treatments. This involves a supervised approach to ML. Supervised ML techniques try to investigate the relationship between biomarkers and the outcome variables, in order to create predictive models. Broadly supervised ML approaches can be divided into classification and regression. Classification models try to predict categorical outcome variable e.g. treatment response, whereas regression models are used when outcome variable is continuous e.g. disease activity. Some of the most popular techniques include linear, logistic, tree-based, ensemble.

In the study conducted by Mossotto et al [4], hierarchical clustering with Hamming distance and average linkage was applied for visualizing the relationship between paediatric IBD patients and traits. In study by Pen et al [5], support vector machines (SVM) using leave-one-out cross-validation in R were used to predict sub-types of gene expression in RA. Orange et al [6] assessed classification and parameter selection using a balanced random forest (BRF) and sparse partial least squares-discriminant analysis (sPLS-DA) in R and validated using 10-fold cross-validation. They used logistic regression for testing the associations between and immune phenotype. Martin-Gutierrez et al [7] also used BRF plots and sPLS-DA in R for classification and parameter identification. Further, logistic regression was applied for association analysis.

2.3 Popular ML models applied in the fields of stratified healthcare and personalised medicine

Several ML techniques have been applied for different chronic inflammatory diseases [8]. We have tried to summarise different ML models used for clustering and/or predictors development in the following sections. A recent review [9] has highlighted a few popular ML methods in the medical arena, which are discussed in the following sub-sections.

2.3.1 Linear models

Linear models have been quite popular amongst the research community as it is the simplest ML model, unlike the so-called ‘blackbox’ of other more sophisticated ML models. In linear regression, for example, the final model is a linear combination of its features and therefore,

Chapter 2: Literature review

we can really understand what it is the model doing, in order to predict [10]. This model has been shown to work quite accurately in many medical applications, such as drug-target interactions [11]. Therefore, many researchers switch to more sophisticated model, only when linear regression accuracies are not on par.

Unlike linear regression, logistic regression (another linear model) provides the framework for classification problems, wherein label/target is a categorical variable. It applies the linear regression, followed by an activation, that changes continuous output from the linear regression into categories [12]. In medical field, we often come across case-control studies, which comprise binary classification, wherein logistic regression comes to the rescue, and can provide promising results [13]. However, these linear models cannot handle clearly discrete or skewed continuous responses [14]. A generalised linear models (GLMs) extend linear modelling to include wide variety of response type.

2.3.2 Decision Tree (DT)

Decision tree (DT) is getting very popular amongst the research community, as the model is simple and similar to linear regression it is easy to visualise. Especially when the data has many categorical variables or skewed continuous variables (often found in medical datasets), DT presents a better way to manage it [15]. DT has shown to be the most flexible, intuitive and powerful data analytics tool for exploring complex datasets in stratified healthcare [16].

2.3.3 Random Forest (RF)

Random forest (RF) falls under the umbrella of bagging methods in ML, where several DTs become part of the overall model and output is generated based on majority voting coming from these models. It is one of the powerful tree-based models and amongst the first choice as a general-purpose classification and regression method [17], when simple ML models, like linear regression/DT, do not provide adequate performance. RF has been seen quite useful in a number of medical applications [18].

2.3.4 Gradient Boosting algorithm

In contrast to bagging ML models, gradient boosting falls under the umbrella of boosting algorithms, wherein, a weak ML model is improved (boosted) by sequentially adding models

and updating its weight [19]. These weak ML models can be any of the ML models. For example, with DTs as the weak learners, we get gradient boosting trees, which usually outperforms RF (bagging DTs) [20].

2.3.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) learns by examples for labelling objects [21]. SVMs are gaining popularity amongst the medical community due to their simplicity and flexibility in addressing a range of classification problems [22]. Especially, it is known to be quite effective in image classification [23] as well as medical data mining [24].

2.3.6 Naïve Bayes (NB)

Naïve Bayes (NB) harnesses then power of Bayer’s rule in order to classify. NB uses a strong assumption that the dataset contains independent feature-set, which is often violated in practice, nevertheless, it provides competitive performance [25, 26].

2.3.7 Neural Networks (NNs)

Artificial neural networks (ANNs) and deep learning (DL) are both NN-based ML techniques that are used to analyze large amounts of data and identify patterns that can inform decision-making. However, the main difference between ANNs and DL is the complexity of the models. ANNs are a type of model that is designed to mimic the structure and function of the human brain. They consist of layers of artificial neurons that are connected to one another, and each neuron is designed to perform a specific task. In contrast, DL models are more complex, and they are designed to learn from large amounts of data. They typically consist of multiple layers of artificial neurons, known as deep neural networks, and each layer is designed to perform a specific task [27]. Further, DL methods have the ability to learn from big data sets and have a better performance on complex tasks, such as image and speech recognition [28].

Deep learning (DL) algorithms provide the most powerful ML tools, especially seen in the field of computer vision. DL methods, such as recurrent neural networks (RNN) and convolutional neural networks (CNN) have been applied to medical imaging recognition for both stratified healthcare [29, 30] and personalised medicine [31, 32].

2.4 Stratified healthcare and personalised medicine in practice

Some of the major diseases, along with the current challenges in stratified healthcare and personalised medicine are described below:

2.4.1 Diabetes Meletus (DM)

DM involves an increase in blood sugar levels [33]. Insulin is a hormone made by our pancreas, which helps in maintaining blood sugar levels. When a patient develops insulin resistance, it leads to DM. This can be further divided into insulin dependent diabetes (Type 1 DM) and insulin independent diabetes (Type 2 DM). However, some people with T2DM may still need insulin. DM is diagnosed using HbA1c or C-peptide level in blood.

Stratified healthcare research [34] has identified 7 molecular subclass or endotype of T2DM patients using unsupervised machine learning approaches. Further, some diabetes patients are prescribed multiple different medications. This includes Metformin medication; however, it still has about 40% non-responders. Different GWAS studies have been undertaken to personalise this diabetes medicine, thus, a lot still needs to be done. Additionally, diabetes patients are also put on different statin treatment like simvastatin or atorvastatin, which can potentially be personalised to treat comorbidities. Thus, advances in stratified healthcare and personalised medicine of DM and comorbidities can lead to the better management and tailoring treatment for the patients by providing a holistic approach towards managing the blood sugar levels for them, however it still has a lot of improvements are possible.

2.4.2 Cardio-vascular disease (CVD)

CVD involves conditions of heart and blood vessels. It is a broad disease class and includes all types of heart and circulatory disorders like coronary heart disease (CHD) / coronary artery disease (CAD) / ischemic heart disease (IHD), acute coronary syndrome (ACS), congenital heart disease, hypertension, stroke and vascular dementia. Further, an ACS can be due to unstable angina (UA) or heart attack (STEMI or NSTEMI, based on electrocardiogram (ECG) graph). Stratified healthcare research [35] has led to risk stratify CVD patients into low, medium, high and very high risk, based on EuroSCORE II.

Major Averse Cardiac Event (MACE) involves ACS, Arrhythmias, Heart failure (and LV impairment), stent re-stenosis, recurrent chest pain, peripheral artery disease, stroke and ministroke (TIA), all cause death and death from cardiac origin. It is important to predict a recurrent MACE event within a year or even between 1-4 years, in order to prioritise and personalise the treatment for CVD patients.

2.4.3 Rheumatoid arthritis (RA)

RA is an autoimmune disease. One third of these patients do not respond to conventional disease-modifying anti-rheumatic drug (cDMARD) and are prescribed biologic DMARD (bDMARDs), with disease response is accessed after 3-6 months post treatment [36]. This treatment is often costly and about one-third again do not respond [36]. There are many treatments including TNF inhibitors, JAK inhibitors, IL6 inhibitors, B-Cell inhibitors, etc. But there is currently no reliable way to personalise treatment in RA yet. A prognostic biomarker to identify the right treatment for a patient would be helpful in restraining disease progression. A review [37] advocates the need for meaningful clinical endotypes for personalising RA treatment.

2.4.4 Inflammatory bowel disease (IBD)

IBD involves damage along the lining and tissues of the digestive tract. Crohn's disease (CD) and ulcerative colitis (UC) are two kinds of IBD. Anti-TNF therapy is one of the popular methods for treating IBD. Research [38] has tried to identify current challenges in effectiveness of anti-TNF treatment in UC. An effective prognostic biomarker to personalise this treatment would control the disease progression in patients. Recent research [39] has reviewed different predictors of primary response to biologic treatment, such as anti-TNF.

2.4.5 Chronic obstructive pulmonary disease (COPD)

COPD is a broad class of lung conditions, which causes difficulty in breathing. There is a need to stratify these patients into smaller sub-groups that can each then receive their own tailored care, personalised to their healthcare requirements or needs. Stratified healthcare research [40] has identified molecular subclasses or endotypes of COPD, using unsupervised machine learning on clinical data. Since these endotypes are clinically different, it implies

that they respond differently to their treatment. This classification can now be used in the development of personalised treatment, tailored to each endotypes.

2.4.6 Chronic kidney disease (CKD)

Acute Kidney Injury (AKI) often leads to CKD [41]. CKD involves progressively decreasing working capability of kidney compared to normal. Kidney function is measured using Estimated Glomerular Filtration Rate (EGFR) calculated using creatinine levels present and urea in blood. CKD is known to be dependent on socioeconomic factors, lifestyle, ethnicity etc. [42] and hence one needs to personalise the treatment based on these modifiable risk factors as well.

2.4.7 Cancer

Cancer involves uncontrolled growth of cells in specific parts of the human body, forming tumors. A benign tumor gradually becomes malignant during the metastasis stage, where it starts spreading to other parts of the body, including by a process known as epithelial-mesenchymal transition (EMT). The four most common types of cancer are breast, lung, prostate and bowel cancers [43]. However, there are more than 200 types of cancer and personalised medicine research has been carried out to diagnose and treat each of these cancers in a particular way [44].

2.4.8 Mental health

Major mental health conditions involve depression, anxiety, OCD, PTSD, etc. Several GWAS have been carried out to identify SNPs that are indicative of different mental health disorders [45]. However, a lot of improvement in the performance of these biomarkers is still required. In addition, inflammation has been implicated in mental health disorders [46].

2.4.9 Role of multimorbidity

An acute event involves rapid onset like ACE, which stays for hours/days, whereas chronic condition is developed slowly over months/years. A lot of acute events may also progress to chronic conditions, like AKI progresses to CKD. Co-occurrence of multiple chronic conditions (MCCs) in a patient is often referred to as multimorbidity [47]. With an increase

in human longevity and changing lifestyle, it is one of the most common conditions in ageing patients. However, most of the clinical trials are still biased towards single index disease conditions. Thus, there is a need for a tailored approach in stratified healthcare for patients with multimorbidity.

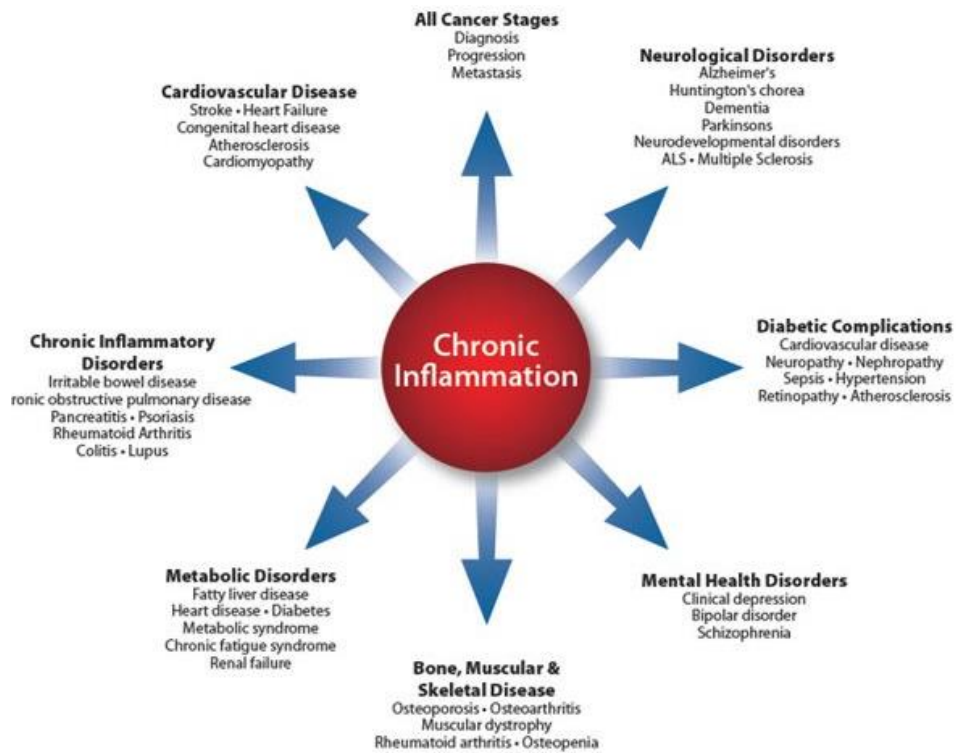


Figure 2.2 Chronic inflammation as a common risk factor for patients with multimorbidity [48].

Figure 2.2 [48] shows that, in most cases, inflammatory response is found to be associated with patients with multimorbidity. Since all chronic conditions involve inflammation, these chronic conditions often interact with each other. Further, inflammation can occur due to multiple pathways, such as p38 MAPK, IL-6/JAK/STAT3, PI3K and Hippo pathway [49], making these patients biologically more complex. Unfortunately, most of these diseases are treated based on research and clinical trials involving patients having single index disease conditions. However, with the advancement of medical science, longevity has increased and age being a major risk factor for multimorbidity [50], it has become a norm in today's aging population. Thus, we need to stratify these multimorbid patients, which can be helpful towards their management as well as improving their quality of life.

Chapter 2: Literature review

Assessing multimorbidity becomes very important when approaching stratified healthcare. With multimorbidity, many treatment goals exist for a clinician to consider. Different diseases lead to chronic inflammation in the patient. The level of this inflammation dictates the disease progression in patients. Assessing chronic inflammation, interaction of proteins involved, and interaction between diseases can help in the development of new treatment regimens for multimorbidity, as the drugs that are currently being approved are tested on patients with single disease conditions (excluding the patients with multimorbidity). Given that multimorbidity has become a norm, targeting the markers of chronic inflammation in different disease conditions could be a way forward in precision medicine for multimorbid patients.

2.5 Limitations in research

Although a lot of interest has been seen amongst the medical research community in exploiting the power of ML in addressing the research questions, they are still underutilised. Many research papers stick to statistical tests (such as the Student's t-test) to show significance and do not carry on to how well these significant biomarkers are able to perform. Further, many weak biomarkers are often seen in medical dataset. ML could help in combining these weak biomarkers into a single strong biomarker, which can have a significant performance improvement.

Furthermore, many ML based research papers use generalised computational tools. These tools may not be optimised for the dataset at hand. Most medical datasets are of similar type and a generally used computational tool might end up taking huge computing resources for the calculation and hence a computational tool developed specifically for the application can be more appropriate.

Many ML models applied in medical research are linear (such as polygenic risk scores), as they are easy to comprehend. However, biological interactions are often complex and hence a more sophisticated ML model might capture the real interaction responsible for the phenotype, more appropriately.

2.6 Research gaps, open challenges and opportunities

With a flood of datasets and availability of good computing resources, the medical field is becoming more and more data-driven investigation. Despite the exponential growth in the usage of ML applications in medical sciences, there still exists a lot to be addressed in stratified healthcare as well as personalised medicine.

A lot of research is being carried out in stratified healthcare. However, most of this research involves either a single disease, for example cancer [51], or a disease with its comorbidity, for example diabetes and its complications [52]. However, with the increase in longevity, most of the elder populations now live with multiple chronic conditions (or multimorbidity). This shows a gap in multimorbidity research. A pivot in research directions for patients with multimorbidity has the potential to improve the quality of life for these patients [53]. Furthermore, a lot of multimorbidity research is being done, using simple descriptive analytics. This calls for a more sophisticated approach, such as ML to address multimorbidity, which is complex as such.

Even though a lot of research is carried out in personalised medicine, it is still at a very early stage. For example, a review on personalised medicine in RA (rheumatoid arthritis) [54] elucidates that the interpatient heterogeneity is significant in terms of their efficacy towards biological therapies and hence proposes the need for robust biomarkers. Furthermore, RA shows heterogeneity at molecular and cellular levels [55], which can have clinical implications. This calls for a stratified approach towards RA patients, as well.

A lot of computational tools are being developed in the field of bioinformatics, in order to tackle the flood of data in the medical field. However, traditional computational tools do not give optimal performance, as the given computational tools show different performances with different kinds of datasets. For example, in medical sciences, we often find datasets that are multi-dimensional. This is due to the capture of multiple feature-sets but only for a few patients, as recruitment is a long and tedious process. Bioinformatics tools need to be tailored to address the complexity of datasets as such. This poses a need to develop several robust pipelines that are tailored to the medical field.

2.7 Conclusion

Stratified healthcare and personalised medicine involve low and high precision respectively and a holistic improvement in healthcare requires both. ML serves as a powerful tool for the analysis and has been extensively applied at both levels of precision to improve the quality of life for the patients. Stratified healthcare involves applying unsupervised approaches in ML to identify sub-classes (or clusters), which can then help in better management and prioritising the treatment, based on risk. Further, personalised medicine involves supervised approaches in ML to predict an outcome, such as treatment response, so that optimal results can be achieved.

Both stratified healthcare and precision medicine research has been extensively applied in all the disease classes, including the role of multimorbidity. There exist different kinds of ML models to answer the question involving stratified healthcare and precision medicine. Usually, the researchers in the medical field prefer the simplest models viz. GLM and DT, as they are easy to comprehend. However, if the adequate performance is not met, more sophisticated models viz. RF and gradient boosting tree are applied. Furthermore, a wide range of data types like speech and image can be seen in the medical arena, where ML models like DL become extremely useful.

Despite the exponential growth in the usage of ML for addressing medical research questions, it is still underutilised. ML presents a more powerful tool as compared to the traditional statistical tests, generally used by the medical research community. Further, there is a need for computational tools that are optimised for medical datasets.

2.8 References

1. Trends (2018), Building the foundation for personalized medicine. Available at: <https://trends-magazine.com/building-the-foundation-for-personalized-medicine-2/> (Accessed: 11.01.2023)
2. Tang A, Oskotsky T, Sirota M. Personalizing routine lab tests with machine learning. *Nat Med.* 2021 Sep;27(9):1514-1515.
3. Elborn JS. Personalised medicine for cystic fibrosis: treating the basic defect. *Eur Respir Rev.* 2013 Mar 1;22(127):3-5.

4. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci Rep*. 2017 May 25;7(1):2427.
5. Orange DE, Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J, McNamara M, Cummings R, Andersen KM, Mirza S, Figgie M, Ivashkiv LB, Pernis AB, Jiang CS, Frank MO, Darnell RB, Lingampali N, Robinson WH, Gravallese E; Accelerating Medicines Partnership in Rheumatoid Arthritis and Lupus Network; Bykerk VP, Goodman SM, Donlin LT. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis Rheumatol*. 2018 May;70(5):690-701.
6. Robinson GA, Peng J, Dönnies P, Coelewijn L, Naja M, Radziszewska A, Wincup C, Peckham H, Isenberg DA, Ioannou Y, Pineda-Torra I, Ciurtin C, Jury EC. Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: patient stratification using a machine-learning approach. *Lancet Rheumatol*. 2020 Jul 29;2(8):e485-e496.
7. Martin-Gutierrez L, Peng J, Thompson NL, Robinson GA, Naja M, Peckham H, Wu W, J'bari H, Ahwireng N, Waddington KE, Bradford CM, Varnier G, Gandhi A, Radmore R, Gupta V, Isenberg DA, Jury EC, Ciurtin C. Stratification of Patients With Sjögren's Syndrome and Patients With Systemic Lupus Erythematosus According to Two Shared Immune Cell Signatures, With Potential Therapeutic Implications. *Arthritis Rheumatol*. 2021 Sep;73(9):1626-1637.
8. Peng J, Jury EC, Dönnies P, Ciurtin C. Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges. *Front Pharmacol*. 2021 Sep 30;12:720694.
9. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, Maojo V, Pazos A, Fernandez-Lozano C. A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J*. 2021 Aug 12;19:4538-4558.
10. Schober P, Vetter TR. Linear Regression in Medical Research. *Anesth Analg*. 2021 Jan;132(1):108-109.
11. Buza K, Peška L, Koller J. Modified linear regression predicts drug-target interactions accurately. *PLoS One*. 2020 Apr 6;15(4):e0230726.

Chapter 2: Literature review

12. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med*. 2011 Oct;18(10):1099-104.
13. Schober P, Vetter TR. Logistic Regression in Medical Research. *Anesth Analg*. 2021 Feb 1;132(2):365-366.
14. Faraway JJ. Generalized linear models. In *International Encyclopedia of Education* 2010 (pp. 178-183). Elsevier.
15. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015 Apr 25;27(2):130-5.
16. Banerjee M, Reynolds E, Andersson HB, Nallamothu BK. Tree-Based Analysis. *Circ Cardiovasc Qual Outcomes*. 2019 May;12(5):e004879. doi: 10.1161/CIRCOUTCOMES.118.004879. Erratum in: *Circ Cardiovasc Qual Outcomes*. 2019 Jun;12(6):e000056.
17. Biau G, Scornet E. A random forest guided tour. *Test*. 2016 Jun;25:197-227.
18. Qi Y. Random forest for bioinformatics. In *Ensemble machine learning: Methods and applications* 2012 Jan 19 (pp. 307-323). Boston, MA: Springer US.
19. Griesbach C, Säfken B, Waldmann E. Gradient boosting for linear mixed models. *Int J Biostat*. 2021 Jan 13;17(2):317-329.
20. Hastie T, Tibshirani R, Friedman J, Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction*. 2009:337-87.
21. Noble WS. What is a support vector machine?. *Nature biotechnology*. 2006 Dec 1;24(12):1565-7.
22. Pisner DA, Schnyer DM. Support vector machine. In *Machine learning* 2020 Jan 1 (pp. 101-121). Academic Press.
23. Othman MF, Abdullah NB, Kamal NF. MRI brain classification using support vector machine. In *2011 Fourth international conference on modeling, simulation and applied optimization* 2011 Apr 19 (pp. 1-4). IEEE.
24. Janardhanan P, Sabika F. Effectiveness of support vector machines in medical data mining. *Journal of communications software and systems*. 2015 Mar 20;11(1):25-30.
25. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. *Encyclopedia of machine learning*. 2010;15:713-4.

26. Rish I. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence 2001 Aug 4 (Vol. 3, No. 22, pp. 41-46).
27. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature*. 2015 May 28;521(7553):436-44.
28. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016 Nov 10.
29. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol*. 2017 Sep;10(3):257-273.
30. Jaber MI, Song B, Taylor C, Vaske CJ, Benz SC, Rabizadeh S, Soon-Shiong P, Szeto CW. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Res*. 2020 Jan 28;22(1):12.
31. Le AH, Liu B, Huang HK. Integration of computer-aided diagnosis/detection (CAD) results in a PACS environment using CAD-PACS toolkit and DICOM SR. *Int J Comput Assist Radiol Surg*. 2009 Jun;4(4):317-29.
32. Klang E, Barash Y, Margalit RY, Soffer S, Shimon O, Albshesh A, Ben-Horin S, Amitai MM, Eliakim R, Kopylov U. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. *Gastrointest Endosc*. 2020 Mar;91(3):606-613.e2.
33. NIH (n.d.), What is Diabetes? Available at: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes> (Accessed: 19.04.2022).
34. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, Vikman P, Prasad RB, Aly DM, Almgren P, Wessman Y, Shaat N, Spégel P, Mulder H, Lindholm E, Melander O, Hansson O, Malmqvist U, Lernmark Å, Lahti K, Forsén T, Tuomi T, Rosengren AH, Groop L. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*. 2018 May;6(5):361-369.
35. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U. EuroSCORE II. *Eur J Cardiothorac Surg*. 2012 Apr;41(4):734-44; discussion 744-5.
36. Curtis JR and Singh JA. Use of biologics in rheumatoid arthritis: current and emerging paradigms of care. *Clin Ther*. 2011 Jun;33(6):679-707.
37. McInnes IB, Buckley CD, Isaacs JD. Cytokines in rheumatoid arthritis - shaping the immunological landscape. *Nat Rev Rheumatol*. 2016 Jan;12(1):63-8.

Chapter 2: Literature review

38. Pugliese D, Felice C, Papa A, Gasbarrini A, Rapaccini GL, Guidi L, Armuzzi A. Anti TNF- α therapy for ulcerative colitis: current status and prospects for the future. *Expert Rev Clin Immunol*. 2017 Mar;13(3):223-233.
39. Gisbert JP, Chaparro M. Predictors of Primary Response to Biologic Treatment [Anti-TNF, Vedolizumab, and Ustekinumab] in Patients With Inflammatory Bowel Disease: From Basic Science to Clinical Practice. *J Crohns Colitis*. 2020 Jun 19;14(5):694-709.
40. Burgel PR, Paillasseur JL, Roche N. Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities. *Biomed Res Int*. 2014;2014:420134.
41. Sato Y, Takahashi M, Yanagita M. Pathophysiology of AKI to CKD progression. *Semin Nephrol*. 2020 Mar;40(2):206-215.
42. NKF (n.d.), Social determinants of kidney disease. Available at: <https://www.kidney.org/atoz/content/kidneydiscauses> (Accessed: 29.04.2022).
43. WHO (2022), Cancer. Available at: <https://www.who.int/news-room/fact-sheets/detail/cancer> (Accessed: 19.04.2022).
44. CRUK (n.d.), What is cancer? Available at: <https://www.cancerresearchuk.org/about-cancer/what-is-cancer> (Accessed: 19.04.2022).
45. Horwitz T, Lam K, Chen Y, Xia Y, Liu C. A decade in psychiatric GWAS research. *Mol Psychiatry*. 2019 Mar;24(3):378-389.
46. Gialluisi A, Bonaccio M, Di Castelnuovo A, Costanzo S, De Curtis A, Sarchiapone M, Cerletti C, Donati MB, de Gaetano G, Iacoviello L, Moli-Sani Study Investigators. Lifestyle and biological factors influence the relationship between mental health and low-grade inflammation. *Brain, behavior, and immunity*. 2020 Mar 1;85:4-13.
47. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med*. 2009 Jul-Aug;7(4):357-63.
48. Natura Health, Inflammation... Are You Inflamed? Available at: <http://www.nurturahealth.com.au/blog/inflammation-are-you-inflamed> (Accessed: 07.03.2021).

49. Yeung, Y.T., Aziz, F., Guerrero-Castilla, A. and Arguelles, S., 2018. Signaling pathways in inflammation and anti-inflammatory therapies. *Current pharmaceutical design*, 24(14), pp.1449-1484.
50. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, Meinow B, Fratiglioni L. Aging with multimorbidity: a systematic review of the literature. *Ageing Res Rev*. 2011 Sep;10(4):430-9.
51. Jackson SE, Chester JD. Personalised cancer medicine. *Int J Cancer*. 2015 Jul 15;137(2):262-6.
52. Bailes BK. Diabetes mellitus and its chronic complications. *AORN J*. 2002 Aug;76(2):266-76, 278-82; quiz 283-6.
53. Makovski TT, Schmitz S, Zeegers MP, Stranges S, van den Akker M. Multimorbidity and quality of life: Systematic literature review and meta-analysis. *Ageing Res Rev*. 2019 Aug;53:100903.
54. Isaacs JD, Ferraccioli G. The need for personalised medicine for rheumatoid arthritis. *Ann Rheum Dis*. 2011 Jan;70(1):4-7.
55. Zhao J, Guo S, Schrodi SJ, He D. Molecular and Cellular Heterogeneity in Rheumatoid Arthritis: Mechanisms and Clinical Implications. *Front Immunol*. 2021 Nov 25;12:790122.

Chapter 3 Data-driven patient stratification of UK Biobank cohort suggests five endotypes of multimorbidity

Contribution

Bodhayana Prasad performed data analysis, data visualization, data interpretation, built the analysis pipeline MulMorPip and wrote the first draft of the manuscript.

Dr. Priyank Shukla and Prof. Anthony J. Bjourson conceived and supervised the project, and helped in data interpretation, reviewing and editing of the manuscript.

This chapter is a slightly modified version of below-mentioned open access research article published in Oxford's Briefings in Bioinformatics journal, under Creative Commons Attribution 4.0 (CC BY 4.0) license (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted reuse, distribution and reproduction in any medium, provided the original work is properly cited.

Prasad B, Bjourson AJ, Shukla P. *Data-driven patient stratification of UK Biobank cohort suggests five endotypes of multimorbidity*. Brief Bioinform. 2022 Nov 19;23(6):bbac410. doi: 10.1093/bib/bbac410. PMID: 36209412; PMCID: PMC9677496.

3.1 Introduction

Multimorbidity generally refers to the occurrence of more than one chronic disease [1]. Chronic diseases are those that are persistent and long-lasting and include arthritis, diabetes and high blood pressure amongst many others. These chronic conditions can be physical non-communicable diseases of long duration such as cardiovascular disease or cancer, a mental health condition of long duration such as a mood disorder or dementia or an infectious disease of long duration such as HIV or hepatitis C [2]. For example, the study by Pieringer and Pitchler in 2011 [3] and the Center for Disease Control and Prevention report [4] on patients with arthritis has reported that 24% suffered from cardiovascular diseases, 19% respiratory conditions, 16% diabetes and 24% depression.

Older population, i.e. people over the age of 40, are more likely to develop multiple chronic conditions (multimorbidity). Hospitals in the UK see around 40–50% older patients [5]. Total long-term care expenditure in 2017 was £48.2 billion (~2% UK GDP), of which approximately two-thirds (66%) was financed by the government and 31% by people who directly paid for the services and medication [6], thus making multimorbidity an economic challenge as well.

With the progress of medical science, patients' longevity has increased. Global life expectancy now sits around age 72 – more than double that of 100 years ago [7]. But the increased longevity has led to the rise of multimorbidity in patients [8]. In 1900, top three causes of death were infectious diseases like pneumonia and flu, tuberculosis, and gastrointestinal infection [9]. By 2010, these were replaced by cancer, heart disease and cerebrovascular disease [10]. Further, the mortality from all causes has declined by 54% from 1900 to 2010 [11]. With the advent of modern medicine, life expectancy has been gradually increasing. The data for 2018–20 show that the life expectancy at birth for UK has now reached 79 years for males and 82.9 years for females [12]. The difference between the genders is also gradually decreasing as the male life expectancy is increasing at a faster rate than females [13]. Additionally, healthy life expectancy data show that it is 62.9 years for males and 63.3 years for females for 2017–19 [14]. The difference between the life expectancy and healthy life expectancy is years a patient spends in poor health. This difference is about 19.1 years (64 years in good health) for 2012–14 and was 18.1 years (62.5 years in good health) in 2000–02 [15]. For both the sexes, years in poor health from age 65

Chapter 3: MulMorPip

has increased by 1.4 years for females and 1.5 years for males in 2012–14 as compared to 2000–02 [15]. Hence, multimorbidity is not just a special case, but a norm in today's world.

There exist many metrics to measure multimorbidity and/or comorbidity beyond a simple disease count [16]. One of the first study in this field by Charlson et al. in 1987 [17] suggested Charlson Comorbidity Index (CCI) as a weighted metric for multimorbidity, giving weights to 17 broad disease classes based on severity that can decrease longevity. Thus, a higher CCI means the patient is more multimorbid and hence prone to die early. The CCI was later standardized using the International Classification of Diseases-10 (ICD-10) by Quan et al. in 2005 [18]. The CCI is the most widely used multimorbidity measurement metric [16]. Therefore, for the current scope of our study we have used it for describing multimorbidity in UK Biobank cohort and have focused on the patients having at least two disease classes defined by the Charlson's comorbidity classification.

Comorbidity is often used interchangeably with multimorbidity, but there is a subtle difference between the two. Traditionally, a patient with multimorbidity visits specialists for each disease in secondary care settings. These patients are labelled with one disease as the major disease alias index disease or condition and the rest of the conditions are labelled as comorbidities. This approach makes the specialist treat and mostly consider the major condition. In contrast, multimorbidity is where multiple chronic conditions are studied together with their interactions with each other and analysed under a single umbrella, like how generalist practice in a primary care setting. Thus, index-comorbidity regime deals majorly with one index disease, whereas multimorbidity looks at multiple chronic conditions together along with their interactions with each other [19]. There is a need to renew the relationship between specialists and generalists, who have different but complimentary skills to personalise the treatment of patients with multimorbidity.

At public healthcare systems, such as the National Health Service (NHS) in the UK, patients often enter a long waitlist [20]. There is a current need to explore more sophisticated and stratified or personalised treatment approach that can prioritise treatments for the high-risk patients, especially those with multimorbidity. This has led to stratified or personalised medicine becoming one of the priority research areas for Innovate UK, the Medical Research Council (MRC) UK and the Academy of Medical Sciences UK [21]. The need for a personalised medicine approach in this regard can be achieved by clustering patients using

unsupervised machine learning (ML) techniques and then characterizing them using various demographic and clinical data. Here we propose an analytical approach (MulMorPip) based on multiple correspondence analysis (MCA), which is a multivariate technique within unsupervised ML field. MulMorPip is a step towards equipping the clinicians with patient stratification based on multimorbidity and understanding disease–disease interactions within multimorbid groups, which can eventually help them in better decision-making, prioritizing and personalising the treatment plans for multimorbid patients.

3.2 Material and methods

A summary flowchart of the bioinformatics analysis pipeline, namely multimorbidity analysis pipeline (MulMorPip), is presented in Figure 3.1, and all the code of the pipeline has been made available in the following public repository <https://github.com/ShuklaLab/MulMorPip>.

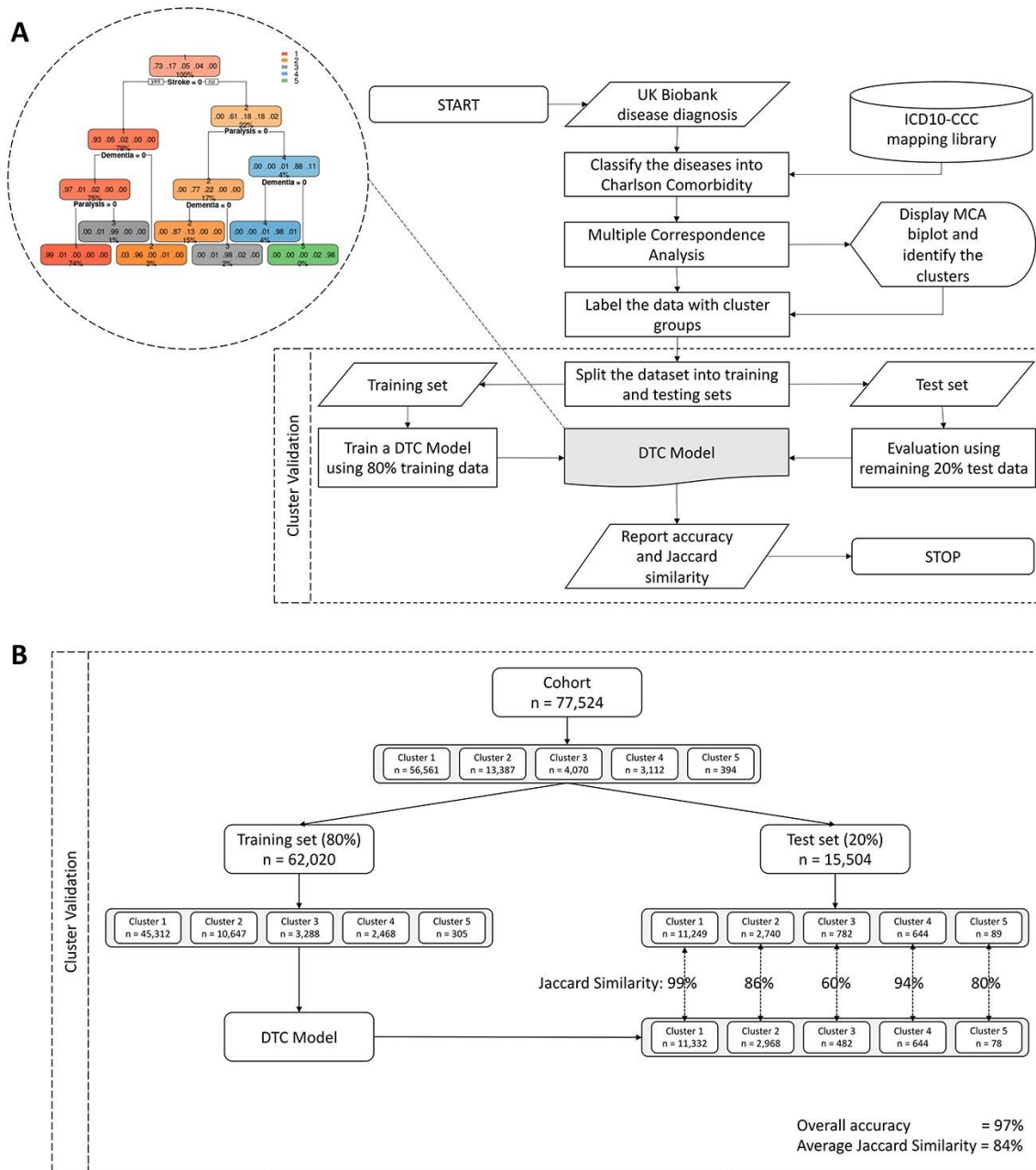


Figure 3.1 (A) Flowchart presentation of multimorbidity analysis pipeline (MulMorPip). Oval shapes represent start/stop, parallelogram boxes represent input/output, rectangular boxes represent computation process and cylinder represents library/database. DTC model has been zoomed-out and presented in the dotted circle. **(B)** Detailed presentation of cluster validation. ICD10 = International Classification of Diseases 10th Revision, CCC = Charlson's comorbidity classification, DTC = decision tree classifier, MCA = multiple correspondence analysis.

3.2.1 Datasets

UK Biobank (<https://www.ukbiobank.ac.uk>) has recruited about 500 000 patients from Great Britain (England, Scotland and Wales). These participants gave consent for access to their electronic care record. We obtained the UK Biobank data via Application No. 48433. We collected the ICD-10 disease diagnosis of the patients from the UK Biobank field id 41270 (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41270>). Dataset belonging to participants who later revoked their consent was excluded.

3.2.2 Statistical, computational and bioinformatic analyses

All statistical and computational analyses were carried out in R v3.6.1 [22]. The t-test and chi-square test to check for demographic variables were performed in the ‘base’ package. The UK Biobank data was loaded with the help of ‘ukbtools’ package [23] and transformed into subsequent Charlson’s broad disease classes using the ‘icd’ package [24]. MCA was carried out using the ‘FactoMineR’ package [25] and plotted using the ‘factoextra’ package [26]. Data splitting was done using the ‘caret’ package [27]. ML model for decision tree was made using ‘rpart’ package [28] and plotted using ‘rpart.plot’ package [29]. Network analysis was performed using Cytoscape software [30].

3.2.3 Multiple correspondence analysis

The ICD-10 summary diagnoses of the UK Biobank patients were grouped into Charlson’s broad disease classification. The patients with multimorbidity were obtained using disease count of greater than 1 for the Charlson’s broad disease classification. This subset of patients was then used for performing MCA. The MCA plot was then rotated using matrix multiplication $M = \begin{bmatrix} 1 & 1.8 \\ 1.8 & -1 \end{bmatrix}$ to make clusters vertical, which were then partitioned using the x-axis (cut-offs: 0, 0.7, 1.4 and 2.1) and labelled with cluster numbers.

3.2.4 Cluster validation

We divided the data into training set (80%) and test set (20%), trained a decision tree classifier (DTC) with the training set and used the DTC model to predict the test set. We compared the predictions with the original cluster values using overall accuracy and Jaccard similarity scores. Finally, we plotted the predictions as well as original cluster values in a

separate MCA plot of 20% test set. Random number generation seed was set to 200, prior to carrying out validation.

3.2.5 Network analysis

The prevalence of each disease was calculated for each of the identified clusters and size of the network disease node was then made proportional to it. Further, the co-occurrence of two diseases was obtained and used in defining the thickness of the network edges. Disease interaction networks were then plotted in circular topology.

3.3 Results

3.3.1 Cluster analysis and its validation

We selected the UK Biobank participants belonging to two or more broad disease classes as per Charlson's classification of multimorbidity. The count of broad disease classes in this multimorbid cohort ($n=77\,524$) varied from 2 to 13 (Supplementary Table 3.1). This classification was used to carry out MCA, which showed five distinct clusters (Figure 3.2A). Plot of variable categories (i.e. 17 broad disease classes defined by Charlson) against principal dimensions showed dependence of clusters on paralysis, followed by stroke and dementia (Figure 3.2B). In order to validate these clusters, we set aside 20% test data and confirmed that the basic demographic features are representative of the 80% training set (Supplementary Table 3.2). We then trained a DTC model on the remaining 80% data. The flowchart for the validation scheme is presented in Figure 3.1A (lower panel), and the corresponding sample sizes and results are shown in Figure 3.1B. We chose DTC because the disease data is categorical, and DTC is extensively used with categorical data and gives a simple and meaningful decision tree for decision-making. The DTC model obtained using the 80% training set is presented in Figure 3.1A (zoomed dotted circle). The DTC was seen to use the same three disease classes (i.e. paralysis, stroke and dementia) to define its branching. The performance of the DTC model in terms of confusion matrix is presented in Supplementary Table 3.3. The prediction of 20% test set using the DTC model gave an overall accuracy of 97% (Figure 3.1B). However, since the number of patients in each cluster was significantly different, we computed Jaccard similarity for each cluster and obtained an average Jaccard similarity score of 84% (Figure 3.1B). To visually compare the original

clustering results with the validation results, we further did MCA on 20% test set and coloured each data point (patients) using their original cluster membership as well as predicted cluster membership, which showed a huge overlap between the original clustering results with the validation results (Supplementary Figure 3.1). The overall high prediction accuracy of 97%, average Jaccard similarity score of 84% and MCA plot of 20% test set validate the existence of five multimorbid clusters in the UK biobank cohort.

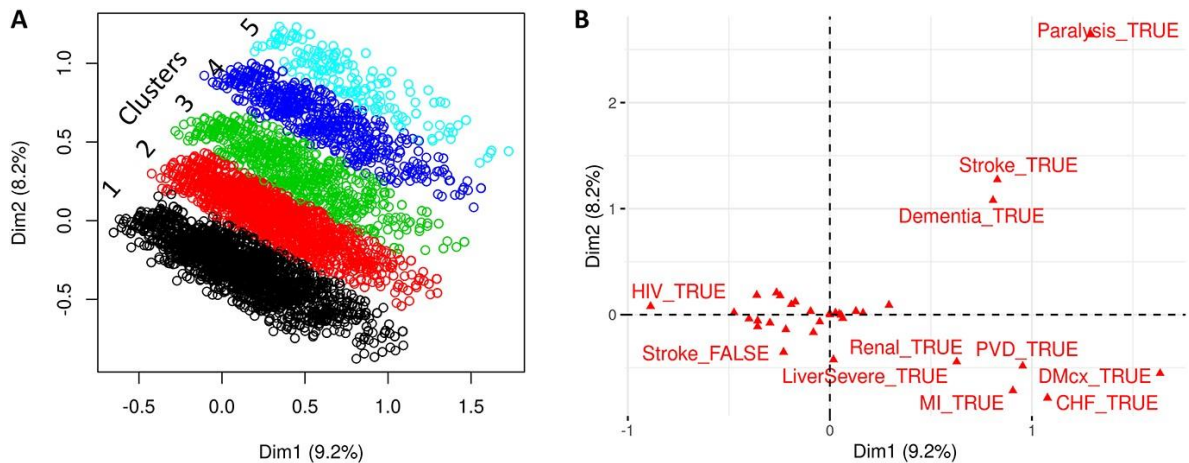


Figure 3.2 Multimorbid clusters: multiple component analysis (MCA) for Charlson's comorbidity classification of 77 524 patients with multimorbidity in UK Biobank. **(A)** MCA plot showing five different clusters. **(B)** Coordinates of variable categories in the two principal dimensions of MCA plot. Variance explained by MCA dimensions are mentioned under parenthesis. Variables far from $x,y = 0,0$ have been labelled. MI = myocardial infarction, CHF = congestive heart failure, PVD = peripheral vascular disease, DMcx = DM with chronic complications, TRUE = disease present, FALSE = disease absent.

3.3.2 Exploratory data analysis on patients with multimorbidity

A total of 77 524 multimorbid patients were seen in five stacked oblong clusters, when first two principal dimensions were plotted (Figure 3.2A). This contained 72.96% patients in cluster 1, 17.27% patients in cluster 2, 5.25% patients in cluster 3, 4.01% patients in cluster 4 and 0.51% patients in cluster 5 (Figure 3.1B). Principal dimensions dictating the clustering of patients were seen to be dependent on existence of paralysis, stroke and dementia in patients (Figure 3.2B). Figure 3.3 shows basic demographic features underlying each of the disease clusters. A decreasing trend of the proportion of females was noted as we move from cluster 1 to 5 (Figure 3.3A). The life expectancy shows an increasing trend from cluster 1 to

5 (Figure 3.3B). Patients in cluster 2 to 5 are functionally not much active due to high number of stroke and paralysis. Cluster 5 has the highest life expectancy with patients having both paralysis and stroke and therefore might be bed-ridden leading to poor quality of life. Age is a major risk factor for dementia [31], and highest number of dementia cases was noted in cluster 5 (Figure 3.4A) which had the highest life expectancy (Figure 3.3B). An increasing trend towards multimorbidity signified by increase in CCI was noted as we move from cluster 1 to cluster 5 (Figure 3.3C). Figure 3.3D shows the Index of Multiple Deprivation (IMD) as formalized by England for the patients in each cluster. The IMD scores for the first three clusters are similar, followed by an increasing trend, with the highest IMD for cluster 5.

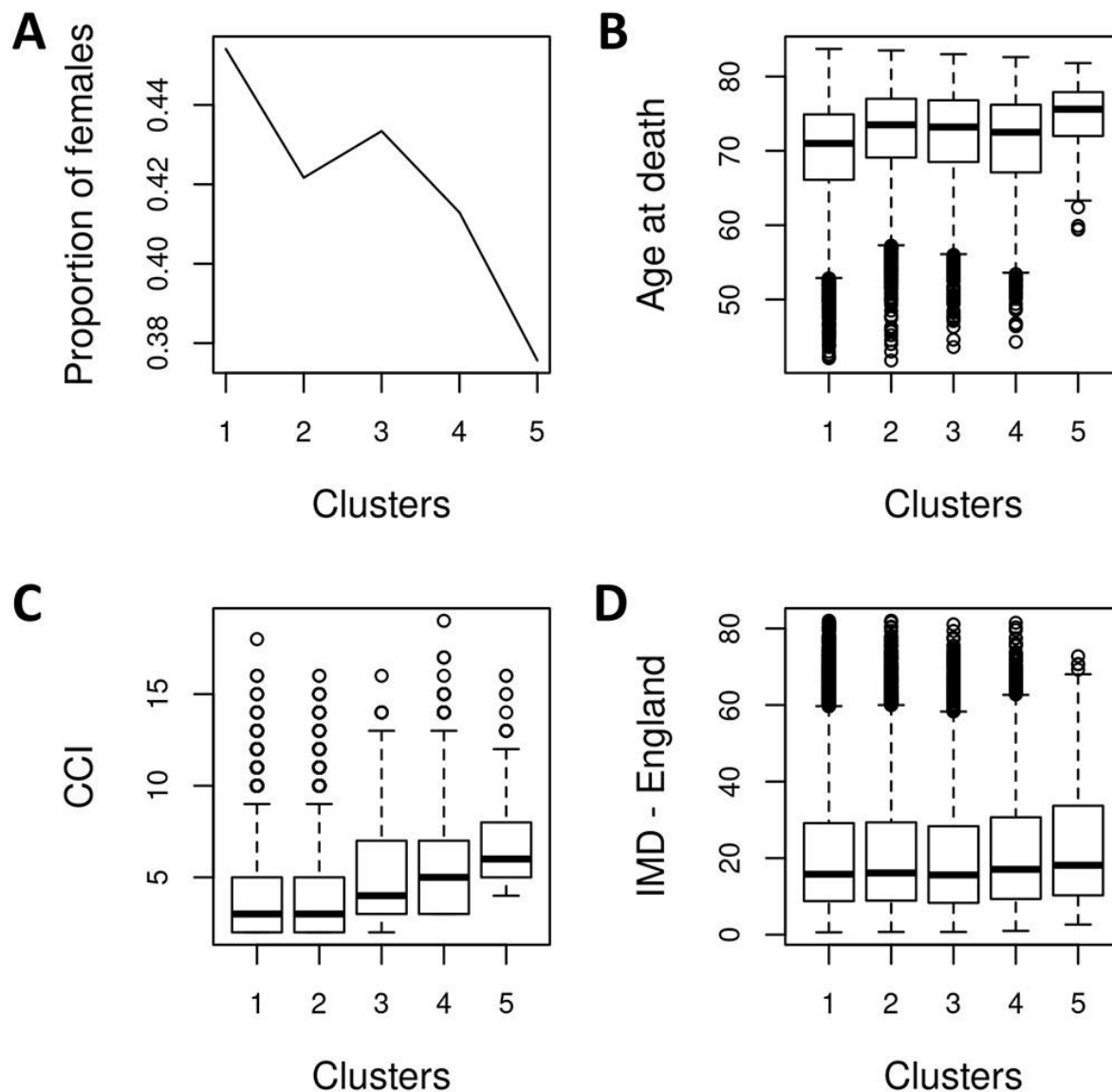


Figure 3.3 Characterization of clusters based on demographic data. **(A)** Plot of gender distribution shows a decreasing proportion of females from cluster 1 to 5. Box plots of **(B)** age at death, **(C)** Charlson Comorbidity Index (CCI) and **(D)** Indices of Multiple Deprivation (IMD) show significant differences between clusters. IMD classification of England was used.

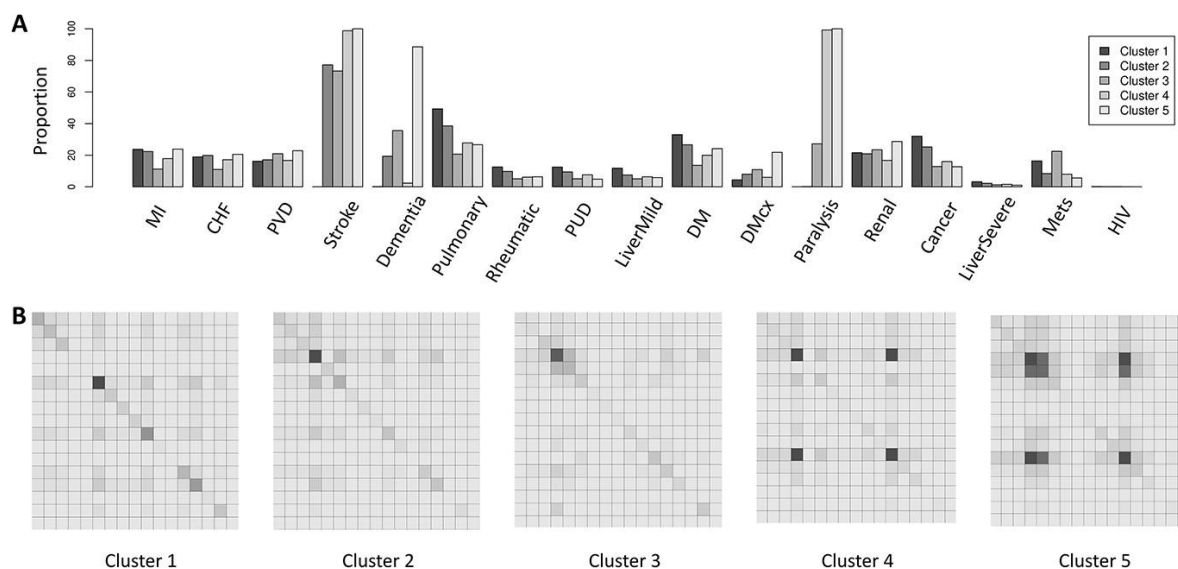


Figure 3.4 **(A)** Prevalence of 17 broad disease classes as per Charlson's comorbidity classification. **(B)** Heatmap showing the co-occurrence of 17 broad disease classes as per Charlson's comorbidity classification. Order of disease from top to bottom and left to right are myocardial infarction (MI), congestive heart failure (CHF), peripheral vascular disease (PVD), stroke, dementia, pulmonary disease, rheumatic disease, peptic ulcer disease (PUD), mild liver disease (LiverMild), diabetes mellitus (DM), DM with chronic complications (DMcx), paralysis, renal disease, cancer, severe liver disease (LiverSevere), metastasis and HIV. Darker shade represents higher co-occurrence.

Figure 3.4A shows a relatively smaller prevalence of diseases in cluster 1, which gradually increases in number of conditions as we move towards cluster 5. Out of all the Charlson's broad disease classes, pulmonary disease has the maximum prevalence in cluster 1, as 49.32% of patients have pulmonary disease, followed by diabetes mellitus (33.01%) and cancer (32.04%). Cluster 2 is majorly dominated by stroke (77.18%) followed by pulmonary disease (38.59%), diabetes mellitus (26.62%) and cancer (25.20%). Cluster 3 shows the dominance of 73.27% stroke, followed by dementia (35.55%) and paralysis (27.20%). Cluster 4 is dominated by paralysis (99.26%), followed by stroke (98.78%) and pulmonary

disease (27.76%). However, dementia in this cluster is negligible (only 2.31%). Finally, all the cluster 5 patients have both paralysis and stroke. Dementia is also one of the major diseases with 85.8% prevalence in cluster 5. Figure 3.4B shows a co-occurrence matrix for disease classes in each of the identified disease clusters. Cluster 4 has higher cases of patients having both stroke and paralysis, whereas cluster 5 showed co-occurrence of stroke, paralysis and dementia (Figure 3.4B). Since clusters 1 to 5 were majorly defined by paralysis, stroke and dementia, we went on first visually inspecting their presence in each cluster (Supplementary Figure 3.2). This was followed by investigation of the prevalence and co-occurrence of their subclasses (Supplementary Figure 3.3). Hemiplegia (G81) in paralysis, cerebral infarction (I63), other cerebrovascular disease (I67), sequelae of cerebrovascular disease (I69) in stroke, vascular dementia (F01), unspecified dementia (F03) and delirium (F05) in dementia were found to be more prevalent (Supplementary Figure 3.3).

3.3.3 Disease–disease interaction network

Figure 3.5 shows a disease–disease interaction network that was obtained for each of the five identified disease endotypes of patients with multimorbidity. Cluster 1 clearly shows dominance of pulmonary disease (largest node) and its strong interaction (thick edges) with diabetes mellitus, cancer, renal disease, peripheral vascular disease, congestive heart failure and myocardial infarction. Unlike other clusters, in cluster 1 there is smaller prevalence of multiple diseases and general interaction (co-occurrence) between them. Cluster 2 is majorly containing the patients with stroke and showing strong interaction with pulmonary disease, diabetes mellitus and cancer. Like cluster 2, cluster 3 is also dominated by stroke, but unlike cluster 2 a strong interaction between stroke and dementia can be easily seen in cluster 3. In general, the interaction pattern of cluster 3 is very different from cluster 2 although both are driven by stroke. Cluster 4 and 5 show strong prevalence and interaction between paralysis and stroke. Finally, cluster 5 has a prominent triad of paralysis, stroke and dementia, showing their strong prevalence and interaction.

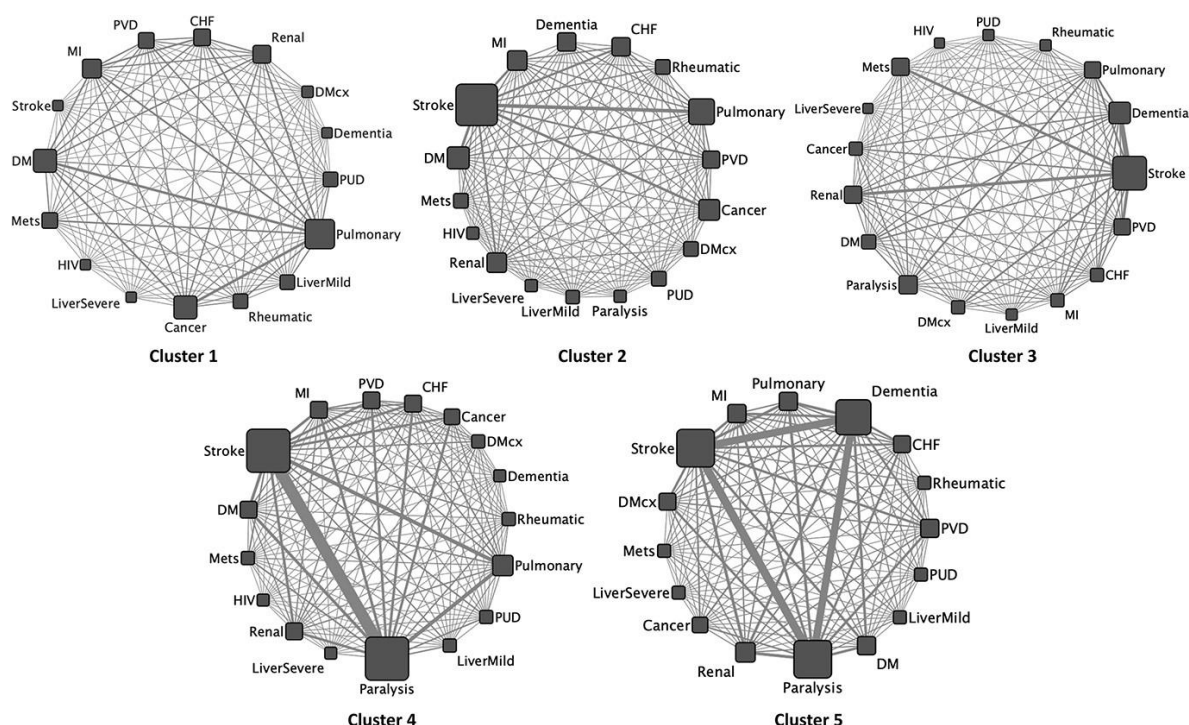


Figure 3.5 Disease–disease interaction network of 17 broad disease classes as per Charlson’s comorbidity classification. Node size is proportional to the disease prevalence and edge thickness is proportional to the disease co-occurrence. CHF=congestive heart failure, DM=diabetes mellitus, DMcx=DM with chronic complications, Mets=metastasis, MI=myocardial infarction, PUD=peptic ulcer disease, PVD=peripheral vascular disease.

3.4 Discussion

Research in the healthcare sector is mostly focused on individual long-term conditions in a structured and standardized way. The traditional approach of treating each disease individually puts the patients with multimorbidity under multiple treatments and often this brings its own issues as multiple medications (polypharmacy) can conflict and cause unwanted side effects. Furthermore, patients with multimorbidity are often excluded from clinical trials. As a result, medicines are developed and tested with a single disease focus. There has been recent consensus amongst clinicians and researchers that this trend may not be appropriate for a patient with multimorbidity. The elderly are the highest consumers of prescribed medications and over 50% of whom suffer from multimorbidity. This results in higher medication and undesirable sequelae. Further, patients have different demographics and diverse genetic makeup. Prescribing the same treatment to everyone adds to the burden

Chapter 3: MulMorPip

of higher medication and poor drug compliance. This is leading to an increased dissonance between the existing healthcare regimes and the need for the patients they serve. Healthcare should be more holistic and person-centred and hence there is a need to understand multimorbidity better and explore sophisticated, personalised diagnosis and treatments for the same. Otherwise, in future multimorbidity will become more challenging for clinicians, patients and the system.

Multimorbidity needs to be managed more efficiently by general practitioners (GPs) or geriatricians in the primary care setting, as the specialists in secondary care tend to focus on only one index disease condition. Managing multimorbidity is tricky as there are so many diseases which require treatment together. The effectiveness of treating patients with multimorbidity should be assessed not just by disease specific indices but by indices such as quality of life, which includes not only symptoms and physical function but also mental health and longevity. National Institute for Health and Care Excellence (NICE) guidelines [32] have also provided suggestions for clinical assessments and management, wherein they are suggesting tailoring the approach to care. Further, they have also provided guidelines to assess the frailty of patients with multimorbidity. Although NICE guideline of tailoring care for multimorbid patient exists [32], there is little guidance available for managing patients with multimorbidity. This calls for the need to develop efficient and effective strategies for screening and stratifying patients with multimorbidity. Implementation of the analytical approach (MulMorPip) developed in our study suggests five endotypes of multimorbidity which can aid GPs in prioritizing treatment and better management of patients with multimorbidity.

A robust individual becomes frail with age, leading to multimorbidity, which can further lead to disability. Although this simplification is generally true, they (frailty, multimorbidity and disability) may exist independently as well as have some intersections [33]. Multimorbidity may modify the health outcomes and lead to an increase disability or a decreased quality of life or frailty [34]. Cluster 1 patients do not have paralysis and very few have stroke, whereas all cluster 5 patients have both paralysis and stroke (Supplementary Figure 3.2), suggesting functional impairment to be a major cluster driving feature. An increasing trend in the cases of dementia (Figure 3.4A) was noted along with the increase in the life expectancy (Figure 3.3B). A high degree of multimorbidity in dementia patients was

noted in Cluster 5 (Figures 3.4B and 3.5), which was the oldest cohort. Recent research [35] shows the need for identifying modifiable risk factors and pathways common to multimorbidity that can aid in delaying the age-dependent deterioration in patients. The proportion of female participants in cluster 1 is highest (45%) compared to other four clusters where generally a decreasing trend was noted (Figure 3.3A). This is in line with the literature [36], which suggested that females have relatively lower risk for multimorbidity as the CCI increases from cluster 1 to cluster 5 (Figure 3.3C). Further the previous studies [36, 37] also suggest that people with low socioeconomic status are more likely to develop multimorbidity, which we confirmed with IMD - England (Figure 3.3D). This is because socioeconomic status is often related to eating habits and lifestyle [38].

An interesting pattern can be seen in cluster 3. Patients in this cluster majorly either have paralysis or both stroke and dementia (Supplementary Figure 3.2). Contrary to cluster 4 and 5, cluster 3 patients never had stroke and paralysis together. In fact, stroke patients in cluster 3 had dementia as the major comorbidity (Figure 3.5). While validating the clusters, the minimum Jaccard similarity of 60% was seen for cluster 3 (Figure 3.1B). Upon further investigation, we found that most of them were getting misclassified as cluster 2, probably due to a similar comorbidity pattern seen for cluster 2 and 3 (Figure 3.4B). However, stroke's comorbidity with dementia can be seen as the major differentiating factor between the two. We confirmed the same by extracting all the cluster 3's misclassified 305 patients as cluster 2 (Supplementary Table 3.3). They all were found to have no dementia. Stroke can lead to dementia, specifically vascular dementia [39]. Since dementia was present in all five clusters, we investigated the prevalence and co-occurrence of subtypes of dementia (Supplementary Figure 3.3). Cluster 1 predominantly contained delirium (F05) and did not show any preferential comorbidity pattern with any other diseases. Cluster 2 and 3 which were predominated by stroke subtypes—cerebral infarction (I63) and other cerebral vascular disorders (I67)—showed different preferences in terms of their comorbidity pattern with dementia subtypes. While cluster 2 dementia subtypes did not show any preferential comorbidity pattern with any other diseases, cluster 3 dementia subtypes were mostly partnering with I63 and I67. Cluster 4 and 5 were predominated by stroke subtypes—I63, I67 and sequelae cerebrovascular disease (I69). However, while cluster 4 dementia subtypes did not show any preferential comorbidity pattern with any other diseases, cluster 5 dementia subtypes were mostly partnering with I63, I67 and I69.

Chapter 3: MulMorPip

Stroke most often leads to paralysis [40]. Further, the location of injury dictates the type of paralysis. A spinal cord stroke can lead to tetraplegia (quadriplegia) or paraplegia (ICD-10: G82), whereas a brain injury can lead to hemiplegia (ICD10: G81), i.e. left-side or right-side paralysis for right or left hemisphere injury [40]. Hemiplegia (G81) was the most prevalent type of paralysis in both clusters 4 and 5, and it was noted to be comorbid with both stroke and dementia in cluster 5, whereas in cluster 4 it was noted to be comorbid with only stroke (Supplementary Figure 3.3B).

Multimorbidity involves a complex interaction between genetics, biobehavioural and socioenvironmental factors. Further, the absence of disease is linked to the balance of proinflammatory and anti-inflammatory activities that can vary across the time course [41]. For patients with multimorbidity, multiple chronic conditions often interact with each other. Thus, finding such interactions and/or associations can contribute to the integrative healthcare approach for patients with multimorbidity. We have characterized a disease–disease interaction network (Figure 3.5) for each of the five subgroups or endotypes of patients with multimorbidity. These networks were dominated by interaction between stroke, paralysis and dementia.

A recent study [42] worked on investigating the heterogeneity of diabetes and showed seven distinct clusters of the disease using only six variables. Such stratification of patients can help clinicians to better understand the disease subtypes, their progression and interaction with other diseases, and eventually inform a more personalised treatment pathway for each subtype. Our analytical approach (MulMorPip) is a step towards stratifying multimorbid patients into five endotypes using a very unbiased dataset of UK Biobank. We were further able to validate these endotypes of disease clusters using ML techniques. These endotypes may be considered by a specialist in secondary care, to stratify patients more efficiently for various treatments. For example, a paralysis specialist may want to classify their multimorbid patients into clusters 2 to 5 (Figure 3.4). These endotypes of patients might be at different stages of their disease progression and/or respond differently to different drugs as they are fundamentally different in terms of disease–disease interaction (Figure 3.5). Further, we also speculate that these endotypes might be related to risk of early onset and prognosis of certain diseases and thus can be helpful in stratification and prioritizing treatment for the high-risk

patients. In terms of future directions, further research is needed to investigate the onset of diseases, their progression and treatment response in these endotypes.

Our study on multimorbidity is limited to the analysis of selected variables, namely disease diagnosis, gender, age and IMD. Future studies can extend on genomic, imaging, biochemical and other datasets present in the UK Biobank. Also, another limitation is that the UK Biobank cohort is heavily dominated by white ethnicity. Therefore, our results may not be generalisable to other ethnicities such as Asian, African or mixed, and thus would require independent validation studies in these cohorts. While our analytical approach (MulMorPip) shows a strong promise of a specific clinical application of patient stratification problem in the field of personalised medicine, it can be adapted and improvized for much wider applications in the field of bioinformatics.

3.5 References

1. Salive ME. Multimorbidity in older adults. *Epidemiol Rev* 2013;35:75–83.
2. The Academy of Medical Sciences. Multimorbidity: a priority for global health research. <https://acmedsci.ac.uk/file-download/82222577> (Accessed: 25 March 2022).
3. Pieringer H, Pichler M. Cardiovascular morbidity and mortality in patients with rheumatoid arthritis: vascular alterations and possible clinical implications. *QJM* 2011;104(1):13–26.
4. Centre for Disease Control and Prevention. Comorbidities. https://www.cdc.gov/arthritis/data_statistics/comorbidities.htm (Accessed: 25 March 2022).
5. National Health Services. Hospital Admitted Patient Care Activity 2020–21. <https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2020-21> (Accessed: 25 March 2022).
6. Office for National Statistics. Healthcare expenditure, UK Health Accounts: 2017. <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2017> (Accessed: 25 March 2022).
7. Roger M, Ortiz-Ospina E, Ritchie H. Life Expectancy. <https://ourworldindata.org/life-expectancy> (Accessed: 25 March 2022).

Chapter 3: MulMorPip

8. Marengoni A, Vetrano DL. Multimorbidity: disease of society? *Lancet Healthy Longevity* 2021;2(8):e451–2.
9. Centers for Disease Control and Prevention. Leading Causes of Death, 1900–1998. https://www.cdc.gov/nchs/data/dvs/lead1900_98.pdf (Accessed: March 2022).
10. Centres for Disease Control and Prevention. Morbidity and Mortality Weekly Report (MMWR). <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6208a8.htm> (Accessed: 25 March 2022).
11. Tippet R. Mortality and Cause of Death, 1900 v. 2010. <https://www.ncdemography.org/2014/06/16/mortality-and-cause-of-death-1900-v-2010> (Accessed: 25 March 2022).
12. Office of National Statistics. National Life Tables – Life Expectancy in the UK: 2018 to 2020. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2018to2020> (Accessed: 25 March 2022).
13. McIntyre N. When Will Men Live as Long as Women? By 2032, Say Experts. <https://www.theguardian.com/inequality/2018/feb/13/when-will-men-live-as-long-as-women-by-2032-say-experts> (Accessed: 25 March 2021).
14. Office for National Statistics. Health State Life Expectancies, UK: 2017 to 2019. <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/healthstatelifeexpectanciesuk/2017to2019> (Accessed: 25 March 2022).
15. Office for National Statistics. Health and Life Expectancies. <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies> (Accessed: 25 March 2022).
16. Stirland LE, González-Saavedra L, Mullin DS, et al. Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice. *BMJ* 2020;368:m160.
17. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40(5):373–83.

18. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43(11):1130–9.
19. Gibson DS, Drain S, Kelly C, et al. Coincidence versus consequence: opportunities in multi-morbidity research and inflammation as a pervasive feature. *Expert Rev Precis Med Drug Dev* 2017;2(3):147–56.
20. National Health Services. Guide to NHS Waiting Times in England. <https://www.nhs.uk/nhs-services/hospitals/guide-to-nhs-waiting-times-in-england> (Accessed: 25 March 2022).
21. The Academy of Medical Sciences. Stratified Medicine. <https://acmedsci.ac.uk/policy/policy-projects/Stratified-medicine> (Accessed: 25 March 2022).
22. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org>.
23. Hanscombe KB, Coleman JRI, Traylor M, et al. ukbtools: An R package to manage and query UK Biobank data. *PLoS One* 2019;14(5):e0214311.
24. Wasey JO, Lang M. R Core team. ICD: comorbidity calculations and tools for ICD-9 and ICD-10 codes. R package version 409. <https://CRAN.R-project.org/package=icd>.
25. Le S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw* 2008;25(1):1–18.
26. Kassambara A, Mundt F. factoextra: extract and visualize the results of multivariate data analyses. R package version 107. <https://CRAN.R-project.org/package=factoextra>.
27. Kuhn M. caret: classification and regression training. R package version 60–88. <https://CRAN.R-project.org/package=caret>.
28. Therneau T, Atkinson B. rpart: recursive partitioning and regression trees. R package version 41–15. <https://CRAN.R-project.org/package=rpart>.
29. Milborrow S. rpart.plot: plot 'rpart' models: an enhanced version of 'plot.rpart'. R package version 310. <https://CRAN.R-project.org/package=rpart.plot>.

30. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13(11):2498–504.
31. Alzheimer's Society. Risk Factors for Dementia. https://www.alzheimers.org.uk/sites/default/files/pdf/factsheet_risk_factors_for_dementia.pdf (Accessed: 25 March 2022).
32. National Institute for Health and Care Excellence. Multimorbidity: Clinical Assessment and Management. <https://www.nice.org.uk/guidance/ng56> (Accessed: 25 March 2022).
33. Espinoza SE, Quiben M, Hazuda HP. Distinguishing comorbidity, disability, and frailty. *Curr Geriatr Rep* 2018;7(4):201–9.
34. Le Reste JY, Nabbe P, Manceau B, et al. The European General Practice Research Network presents a comprehensive definition of multimorbidity in family medicine and long term care, following a systematic review of relevant literature. *J Am Med Dir Assoc* 2013;14(5):319–25.
35. Calderón-Larrañaga A, Vetrano DL, Ferrucci L, et al. Multimorbidity and functional impairment-bidirectional interplay, synergistic effects and common pathways. *J Intern Med* 2019;285(3):255–71.
36. Schäfer I, Hansen H, Schön G, et al. The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. First results from the multicare cohort study. *BMC Health Serv Res* 2012;12:89.
37. Kuo RN, Lai MS. The influence of socio-economic status and multimorbidity patterns on healthcare costs: a six-year follow-up under a universal healthcare system. *Int J Equity Health* 2013;12(1):1–11.
38. Sodjinou R, Agueh V, Fayomi B, et al. Obesity and cardio-metabolic risk factors in urban adults of Benin: relationship with socio-economic status, urbanisation, and lifestyle patterns. *BMC Public Health* 2008;8(1):1–13.
39. Vijayan M, Reddy PH. Stroke, vascular dementia, and Alzheimer's disease: molecular links. *J Alzheimers Dis* 2016;54(2):427–43.
40. Christopher & Dana Reeve Foundation. Stroke (Cerebral Vascular Accident (CVA) and Spinal Stroke). <https://www.christopherreeve.org/living-with-paralysis/health/causes-of-paralysis/stroke> (Accessed: 16 August 2021).

41. Sturmberg JP, Bennett JM, Martin CM, et al. 'Multimorbidity' as the manifestation of network disturbances. *J Eval Clin Pract* 2017;23(1):199–208.
42. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* 2018;6(5):361–9.

Chapter 4 ATRPred: A machine learning based tool for clinical decision making of anti-TNF treatment in rheumatoid arthritis patients

Contribution

Bodhayan Prasad performed data analysis, data visualization, data interpretation, built the software ATRPred and wrote the first draft of the manuscript.

Dr. Priyank Shukla and Prof. Anthony J. Bjourson conceived and supervised the project, and helped in data interpretation, reviewing, and editing of the manuscript.

Dr. David Gibson conceived the project, and helped in data interpretation, reviewing, and editing of the manuscript.

Cathy McGeough, Amanda Eakin and Tan Ahmed helped in data curation, contributed to discussions and proofread the manuscript.

Dawn Small, Dr. Philip Gardiner, Dr. Adrian Pendleton and Dr. Gary Wright helped in patient recruitment, contributed to discussions and proofread the manuscript.

This chapter is a slightly modified version of below-mentioned open access research article published in PLoS Computational Biology journal, under Creative Common Attribution 4.0 (CC BY 4.0) license (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted reuse, distribution and reproduction in any medium, provided the original work is properly cited.

Prasad B, McGeough C, Eakin A, Ahmed T, Small D, Gardiner P, Pendleton A, Wright G, Bjourson AJ, Gibson DS, Shukla P. *ATRPred: A machine learning based tool for clinical decision making of anti-TNF treatment in rheumatoid arthritis patients*. PLoS Comput Biol. 2022 Jul 5;18(7):e1010204. doi: 10.1371/journal.pcbi.1010204. PMID: 35788746; PMCID: PMC9321399.

4.1 Introduction

Rheumatoid Arthritis (RA) is a chronic autoimmune condition characterised by relapsing joint pain, inflammation, and damage along with systemic effects and elevated morbidity. Without effective treatment, RA patients suffer greater risk of disability [1]. Initially, RA patients are treated with non-steroidal anti-inflammatory drugs and conventional disease modifying anti-rheumatic drugs (DMARDs). Patients, refractory to conventional DMARDs, are subsequently prescribed biologic DMARDs [2], among which anti-tumour necrosis factor (anti-TNF) therapies are common, which includes adalimumab, etanercept, infliximab, certolizumab or golimumab—a monoclonal anti-TNF antibody. However, not all patients respond well to anti-TNF therapy. Approximately 10–30% do not respond initially and 23–46% lose the responsiveness over time [3]. A recent article suggests that at least 6% of RA patients on biologics suffer from a refractory condition of the disease [4]. This suggests the existence of molecular sub-classes within the broad disease class. These molecular sub-classes are known as endotypes. Unlike phenotype which involves only observable characteristics, an endotype has direct relation with disease process as it involves inflammatory parameters and specific biological mechanisms. A recent paper from McInnes et al [5] advocates the need for clinically meaningful RA endotypes to stratify patients for therapeutics.

Clinicians generally decide to prescribe anti-TNF therapy based on their disease severity, progression, and other comorbidities. Recent research suggests that the clinicians often switch between different treatments empirically because of a lack of suitable predictive tests [6]. A major downside of this approach is that for patients who remain unresponsive to attempted biologic treatments, inadequate suppression of ongoing disease activity elevates the risk of permanent joint damage and disability [7]. This argues for the need to develop a better prognostic model, that can predict a patient's responsiveness towards the anti-TNF therapy.

Furthermore, RA is known to affect at least 1% of the European population [8]. A recent epidemiological study has reviewed prevalence of RA in different countries of every continent and reports that the prevalence is still close to 1% in many European countries [9]. Additionally, biologic treatments remain relatively costly and continue to rank among the highest grossing drugs. Humira (adalimumab) for example, alone generated 20 billion US

dollars of revenue worldwide in 2018 [10]. A very recent study [11] has pointed out various hidden access barriers to biologic treatment in the European Union (EU).

Thus, there is a strong clinical as well as health-economic need for more personalised prognostic models which can determine likelihood of response to anti-TNF therapy [12]. Several studies using different omics profiles have attempted to predict response to anti-TNF therapy [13]. Literature review shows that the researchers have identified serum proteomic biomarkers for response to anti-TNF therapy [14] including one based on autoantibody and cytokine profiles [15]. Biomarkers have also been found specific to infliximab drug response [16] and etanercept drug response [17]. Further differentiated responses have been noted for adalimumab and infliximab [18]. Also, clinical efficacy can be intensified with infliximab using therapeutic drug monitoring approaches [19]. Several multi-omics approaches have also been used to predict anti-TNF efficacy [20]. For example, an integrated multi-omics approach of previously known DNA, RNA, and protein biomarkers [21], and a more recent approach which combines transcriptomic and genomic analysis [22]. However, none of these studies have presented a robust scoring scheme/model for drug responsiveness that can help in decision making under a clinical setting; rather they relied on only p-values.

We have strictly followed European League Against Rheumatism (EULAR) criteria for patient recruitment, as it is known to have good construct, criterion, and discriminatory validity [23]. Further, to stratify a patient's potential response to treatment, a proteomic profile (which is highly variable) may better reflect current disease state than transcriptomic (variable) or genomic (constant) profiles. With the advent of new high-throughput proteomics technology such as multiplexed proximity extension assay (PEA), it is now possible to profile a patients' plasma proteins with high accuracy and sensitivity [24]. This study was designed to identify a robust protein signature which can predict a patient's response to anti-TNF therapy using a highly sensitive protein detection platform. This study investigates whether plausible endotypes with clinical relevance can be detected in the plasma proteome and if further stratification can predict future response to anti-TNF treatment. Machine Learning (ML) based algorithms, which have been widely exploited for prediction and/or classification problems in bioinformatics, were deployed to mine targeted proteome data. This could help clinicians to optimise treatment selection, reduce spend on

biologics in unresponsive patients and overall improve quality of life for non-responsive RA patients.

4.2 Design and implementation

4.2.1 Ethics statement

Office for Research Ethics Committees Northern Ireland (ORECNI) (11/NI/0188), Ulster University Research Ethics Committee (UREC) (REC/11/0366), Belfast Health and Social Care Trust (11098AB-SS) and Western Health and Social Care Trust (WT/11/35) approvals were obtained for the study. All methods were performed in accordance with the relevant guidelines and regulations. Formal written informed consent was obtained for all participants in the study, allowing for publication of anonymised clinical data.

4.2.2 Patient recruitment and selection criteria

A total of one hundred and forty-four ($N = 144$) Rheumatoid arthritis (RA) patients who were unresponsive to conventional DMARDs and naïve to biologic DMARDs were recruited from rheumatology biologic clinics at Altnagelvin Hospital, Londonderry and Musgrave Park Hospital, Belfast, Northern Ireland. The study inclusion criteria were: i) RA patients fulfilling EULAR classification criteria [25,26], ii) about to receive anti-TNF treatment as part of routine clinical practice, iii) fulfil the BSR 2001 criteria for anti-TNF therapy [27], iv) had a DAS28 score of >5.1 when assessed for treatment (before baseline), and v) reached 6 months of follow-up. Patients who stopped anti-TNF temporarily during the first six months or discontinued therapy prior to the 6 months' follow-up for reasons other than inefficacy were excluded.

4.2.3 Sample collection and collation of clinical information

The study was supported by a patient advisory group who met regularly throughout the study to advise on study design, recruitment literature and results dissemination. Eligible patients were invited by mailed patient information sheets, a minimum of 48 hours before a routine care appointment. Written informed consent was obtained and blood samples were collected prior to anti-TNF treatment. Blood samples were then processed to plasma by centrifugation, aliquoted and stored at -80°C until shipped to Olink Proteomics, Uppsala, Sweden for

Chapter 4: ATRPred

proximity extension assay (PEA) analysis. Clinical and demographic information were collated from medical records and clinic databases. The disease activity score across 28 joints (DAS-28) based on erythrocyte sedimentation rate (ESR) was recorded at baseline and after six months of anti-TNF therapy. Patients were classified as responders and non-responders at six months as per British Society for Rheumatology (BSR) response criteria [28]. Further, a patient whose drug was changed from anti-TNF to a different class by clinicians were also classified as non-responders. Out of $N = 144$ patients recruited 55 were either lost to follow-up or were given other biologic DMARDs (such as Tocilizumab, Rituximab, etc). The recruits lost were unable to make 6 months follow-up appointments, or complete composite data required to calculate DAS score were not available.

4.2.4 Plasma protein profile

Patients' plasma samples were analysed by multiplexed PEA [29] provided by Olink Proteomics (<https://www.olink.com>). Following four Proseek Multiplex panels comprising 92 proteins each were used for analyses: cardiovascular panels II and III, immune response panel and the inflammatory panel. Each panel was quantified by real-time PCR using the Fluidigm BioMark HD platform. In each panel run, 92 samples, 1 negative control and 3 positive controls were analysed. Controls were used for determining the assay limit of detection (LoD) values as well as allowing normalization of measurements into ddCq ($\Delta\Delta Cq$: double delta quantification cycle in qPCR) values. The ddCq values are then log2-transformed to promote normal distribution for subsequent analysis. Olink proteomics returned protein expression data in exponential scale called normalised protein expression (NPX), such that the real expression values are proportional to 2^{NPX} . Each protein's NPX values are relative quantification and hence they cannot be compared across different proteins [30]. Therefore, to obtain comparable results for all proteins [31] and as a pre-processing step for machine learning inputs, each of them is separately scaled into a standard normal distribution $\sim N(0, 1)$. A total of 352 proteins passed the initial quality control (QC) and were subsequently used for the statistical and machine learning based analysis.

4.2.5 Statistical, computational and bioinformatics analyses

All statistical and computational analyses were carried out in R v3.6.1 [32]. The t-test or chi-square test (as appropriate) to check for statistical significance of demographic and clinical

features, and the principal component analysis (PCA) of Olink proteomics data, were performed in the base R package. Quality control (QC) of protein NPX datasets involved discarding protein values which were flagged with a QC warning (sample did not pass quality control for a given protein panel). Also, NPX values were removed if below the limit of detection (LoD) level for a given protein PEA, resulting in $< 2\%$ of missing values. Since missingness was very small, it was imputed using k-Nearest Neighbour (k-NN) method using the RANN package [33]. PCA result was validated with leave-one-out cross-validation (LOOCV) using sinkr package [34]. General ML pre- and post- processing methods were derived from caret [35] and e1071 package [36]. Further, we deployed generalised linear models (GLMs), using the glmnet package [37], to create an intuitive mathematical formulation with a linear combination of protein expression values. Receiver operator characteristic (ROC) curves were obtained via pROC package [38]. Finally, Youden Index [39] was used to choose the best point in ROC curve to calculate thresholds for model score to obtain sensitivity and specificity values. Box plot and beeswarm plot were drawn using beanplot package [40] and beeswarm package [41] respectively, and gplots package [42] and ggrepel package [43] were used for presenting the results. The final model selection was done based on Area Under the ROC Curve (AUC) metric, which is the most preferred metric for the classification problems. Enrichment analysis and Protein-Protein Interaction (PPI) network analysis was performed using STRING database [44]. The Gene Ontology (GO) terms were summarised using REVIGO [45] with its default parameters and the PPI networks were visualised using Cytoscape [46], an open-source software commonly used for network-based analysis. The Pearson's correlation coefficient between the protein features was computed using stats namespace under base R package. This was followed by hierarchical clustering and plotting using the heatmaply package [47].

4.2.6 Feature selection with machine learning

A total of 500 simulations were run by randomly splitting the dataset into 80%:20% and a GLM was learned on 80% training data and tested on 20% test data. If the GLM model had better than random performance (i.e., $AUC > 0.5$), the feature selected in the model was then appended to a feature list. Thus, the importance of a feature reflects its frequency in the feature list. For example, a frequency of 0.8 for a feature represents that the feature showed up in 80% of the 500 simulated models. It is worth mentioning here, that multiple proteomics

signature, having different feature set, are possible [48]. However, getting all the signatures and its performance can be computationally expensive due to large number of combinations possible. Therefore, we went with a deterministic approach of stepwise feature selection, by calculating feature importance (FI) as described above, using a fixed seed value of 200 for 500 simulations.

4.2.7 Machine learning based model development

Our dataset involved 89 samples; hence we chose 5-fold double alias nested cross-validation (CV) for the development of the predictive model [49]. This CV scheme for testing ensures no bias in the selection of completely independent model-blind test-set [50]. Model evaluation was done first by having only gender and baseline DAS and then including protein features one-by-one as per the frequency obtained during feature selection in decreasing order. Mean AUC of training and test sets were measured after fitting a GLM, which was optimised for lambda hyperparameter by 10-fold CV within the training set. The GLM was an Elastic Net with alpha of 0.9, which implements regression with 90% LASSO and 10% Ridge regularization. The aim was to select non-correlated protein, which is achieved by LASSO regularization; a popular method used for feature selection. However, 10% of Ridge regularization was kept to overcome LASSO's limitation to saturate with fewer features. The protein feature set having the highest test set AUC, without the decrease in training set AUC, was selected and the model performance was noted. Finally, with these protein features along with gender and baseline DAS, the model was trained on the whole data and the beta or regression coefficients were computed.

4.2.8 ATRPred tool development

An R-based package was developed for implementing the above-mentioned ML model with the help of devtools package [51]. An input file template along with sample input files of a responder as well as a non-responder are also included in the examples folder present within the package. The R function `antiTNFresponse()` reads the input and normalises the same with the internal 89 patient data to get comparable numbers for feature sets and finally scores the patient for response to the anti-TNF therapy. It then calculates the patient's probability to respond anti-TNF treatment and predicts if the patient will be a responder or non-responder.

This tool is provided as an open-source GitHub repository at <https://github.com/ShuklaLab/ATRPred>.

4.3 Results

The main demographic and clinical features of the patients are shown in Table 4.1. Gender and DAS values at both baseline and 6 months, were found to be statistically significant ($p < 0.05$) between responders and non-responders. The anti-TNF response rate of 67% in our study is almost identical to the 68% reported in a larger study [52]. However, neither this study [52] nor any other study has reported any gender difference as per the author's knowledge. This deference might be due to gender selective confounders like smoking history for which unfortunately the data was not available.

Table 4.1. Demographic and clinical features of rheumatoid arthritis patients. Gender and DAS values (both at baseline and 6 months) were found to be statistically significant between responders and non-responders. RF = Rheumatoid Factor, ACPA = Anti-citrullinated protein/peptide antibody, Anti-CCP = Anti-cyclic citrullinated peptides, DMARD = Disease-modifying antirheumatic drugs and DAS28-ESR = Disease activity score with 28-joint counts and erythrocyte sedimentation rate.

Cohort Characteristics	Responders (N = 60)	Non- Responders (N = 29)	Combine d (N = 89)	P-value
Gender, female, n (%)	51 (85.0)	17 (58.6)	68 (76.4)	*0.006
Age at baseline, mean (s.d.), years	60.6 (11.8)	61.1 (10.3)	60.8 (11.3)	0.848
Disease duration, mean (s.d.), years	8.7 (7.9)	11.1 (10.8)	9.5 (9.0)	0.299
RF Seropositivity, n (N) [#]	38 (48)	18 (25)	56 (73)	0.65
ACPA/anti-CCP Seropositivity, n (N) [#]	34 (42)	16 (23)	50 (65)	0.46

Chapter 4: ATRPred

Concurrent conventional DMARD at baseline, n (%)	55 (91.6)	26 (89.7)	81 (91.0)	-
Concurrent conventional DMARD at 6 months, n (%)	38 (63.3)	14 (48.3)	52 (58.4)	-
Adalimumab, n (%)	40 (66.7)	12 (41.4)	52 (58.4)	-
Etanercept, n (%)	17 (28.3)	12 (41.4)	29 (32.6)	-
Infliximab, n (%)	0 (0.0)	1 (3.4)	1 (1.1)	-
Certolizumab, n (%)	2 (3.3)	2 (6.9)	4 (4.5)	-
Golimumab, n (%)	1 (1.7)	2 (6.9)	3 (3.4)	-
DAS28-ESR at baseline, mean (s.d.)	5.7 (1.2)	4.8 (1.4)	5.4 (1.3)	*0.006
Δ DAS28-ESR at 6 months, mean (s.d.)	-3.0 (1.1)	-0.2 (1.1)	-2.1 (1.7)	*4.8e-14

*significant ($p < 0.05$)

#where data was available

4.3.1 Exploratory data analysis on plasma proteins

Principal Component Analysis (PCA) for all $n = 89$ patients was performed to visualise potential endotypes based on plasma proteome profile. The elbow plot of first 30 PCs showed the drop of explained variance to less than 1% at PC 20 (Supplementary Figure 3.1A). Therefore, we carried out LOOCV of the first 20 PCs, which gave top 20, 6, and 4 PCs with minimum predicted sum of squares (PRESS) for naïve, approximate, and pseudoinverse approaches, respectively (Supplementary Figure 3.1B). Although the naïve approach has limitations [53], all three LOOCV approaches suggested that at least the first 4 PCs are important. The first two principal components (PC1 and PC2) did not show any segregation; however, the third principal component (PC3) was able to subdivide patients into two distinct clusters i.e., endotypes (Figure 4.1). The demographic and clinical features for each cluster are shown in Table 4.2. A statistically significant difference ($p < 0.05$) in baseline DAS and gender was noted between the two clusters. Age, disease duration and anti-TNF biologic

treatment response were not significantly different between the two clusters. The association between baseline DAS and gender within the clusters is illustrated in Figure 4.1. The plot indicates a relatively higher baseline DAS and a higher proportion of females in the cluster positioned in the upper/positive PC3 quadrant. It appears that the two endotypes clearly distinguish patients based on disease activity and are gender dependent.

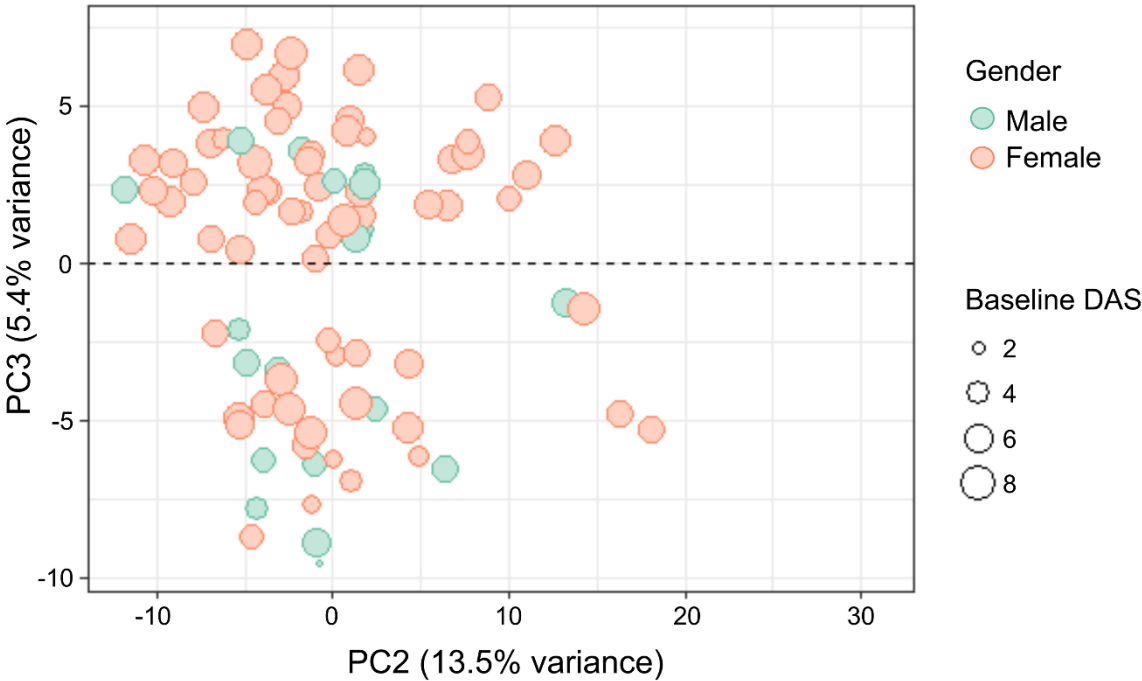


Figure 4.1 Principal component analysis (PCA) plot of rheumatoid arthritis patients (n = 89) using 352 plasma protein Normalised Protein Expression (NPX) values reveals two molecular sub-classes or endotypes with respect to positive and negative third principal component (PC3) values. Endotype 1 is with PC3 values > 0 and endotype 2 is with PC3 values < 0. Each data point represents a patient, where the size of the dot is proportional to the disease activity score (DAS) of the patient at baseline.

Table 4.2. Demographic and clinical features of two molecular sub-class or endotypes presented in Figure 4.1. Gender and baseline DAS values were found to be statistically significant between the two endotypes. DAS28-ESR = Disease activity score with 28-joint counts and erythrocyte sedimentation rate.

Cohort Characteristics	Endotype 1 (N = 55)	Endotype 2 (N = 34)	P-value
Gender, female, n (%)	46 (83.6)	22 (64.7)	*0.041
Age at baseline, mean (s.d.), years	61.2 (11.1)	60.1 (11.8)	0.648
Disease duration, mean (s.d.), years	10.1 (8.5)	8.5 (9.8)	0.480
DAS28-ESR at baseline, mean (s.d.)	5.7 (1.1)	5.0 (1.4)	*0.022
Δ DAS28-ESR at 6 months, mean (s.d.)	-2.3 (1.6)	-1.9 (1.8)	0.248
Responders, n (%)	38 (64.7)	22 (69.1)	0.668

*significant ($p < 0.05$)

4.3.2 Anti-TNF response feature selection and classifier

A quick summary of the computational pipeline built for the discovery of plasma protein signature is presented in Figure 4.2A and the detailed ML analysis schema for model development is presented in Figure 4.2B; both are discussed in more detail in methods section. The feature set available for building the ML classifier includes demographic and clinical data (viz. gender, age, disease duration, baseline DAS (BLDAS) and Δ DAS at 6 months) as well as 352 QC passed proteins' normalised NPX values. Since gender and BLDAS were found to be statistically significant to response to anti-TNF therapy as per Table 4.1, these two features were also included in the signature formulation.

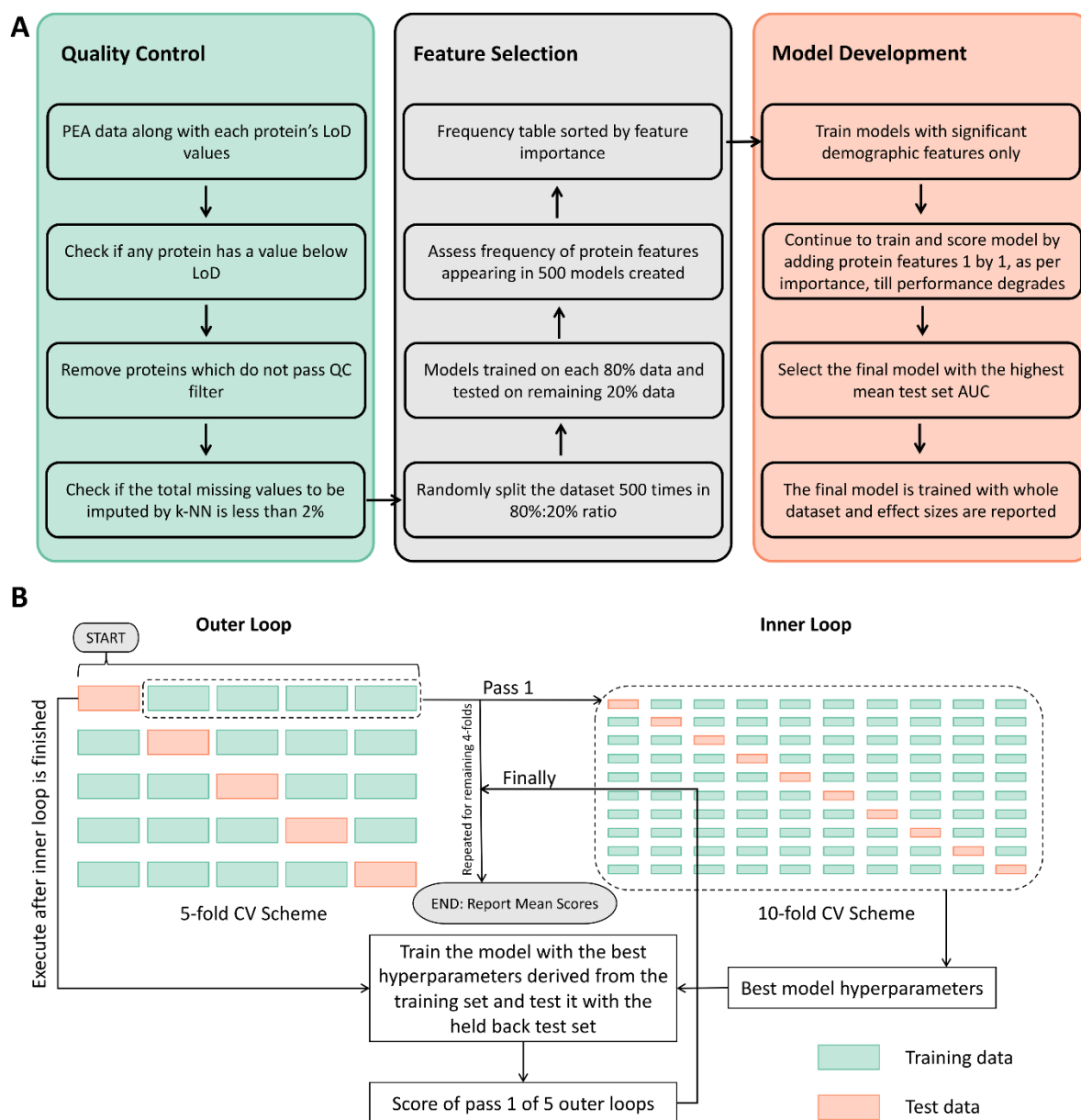


Figure 4.2 (A) Computational pipeline for the development of plasma protein signature. PEA = Protein Expression Analysis, LoD = Limit of Detection, QC = Quality Control, k-NN = k Nearest Neighbour, AUC = Area Under the Curve. **(B)** The Machine Learning (ML) schema. 5-fold nested cross-validation (CV) followed for building the classifier for response to anti-tumour necrosis factor (anti-TNF) treatment in rheumatoid arthritis (RA) patients.

The Feature Importance (FI) of top 30 proteins, along with gender and BLDAS is shown in Figure 4.3A. The graph depicting mean AUC for training as well as test set for each stepwise addition of protein features up to 30 proteins is shown in Figure 4.3B. The threshold of 30 proteins as features was decided after noting the gradual dip in the AUC values for test set

Chapter 4: ATRPred

(Figure 4.3B). A set of 17 proteins gave the maximum mean AUC of 0.86 on test sets, without decreasing the training set AUC. The ROC curves for 5-fold training sets and test sets are shown in Figure 4.3C and 4.3D, respectively. The corresponding best point threshold on ROC curve gave a mean sensitivity of 0.75 and mean specificity of 0.86 on the test sets. The overall mean accuracy was 0.81 on the test set. Further, the mean Matthews correlation coefficient (MCC), popularly used and advocated to assess the quality of binary classification [54], was 0.60, implying a good prediction for each class, viz. responders and non-responders. The summary of mean performance metrics is presented in Supplementary Table 4.1. The final model was trained on the whole dataset and mathematical formulation is presented in the next section.

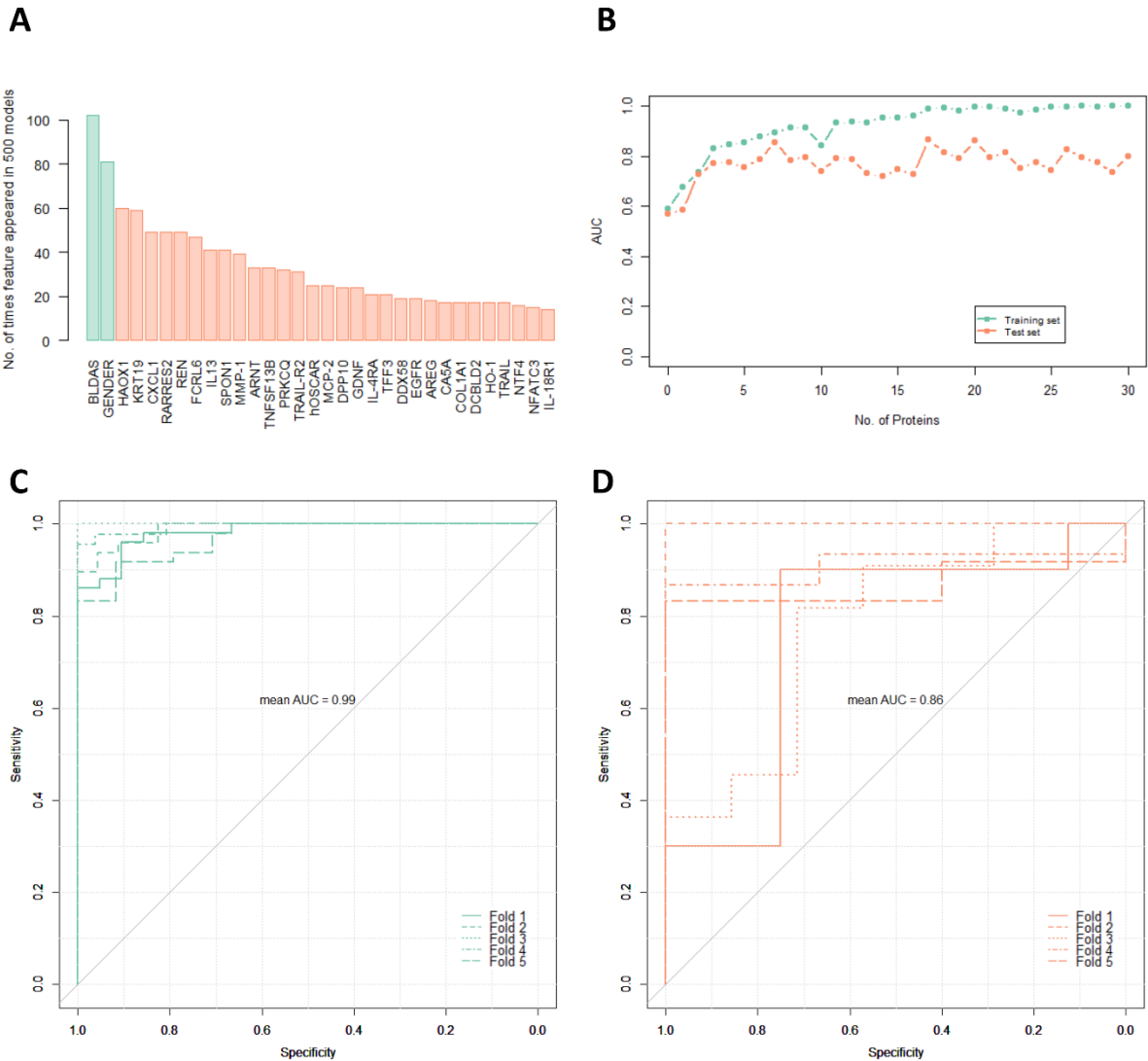


Figure 4.3 (A) Feature importance of top 30 proteins along with significant demographic and clinical features, viz. gender and base line disease activity score (BLDAS). (B) Area Under the Curve (AUC) of training and test set vs. number of protein features. A set of 17 proteins along with gender and BLDAS gave the maximum mean AUC of 0.86 on test set without decreasing the training set's AUC. Receiver Operator Characteristics (ROC) for the 5-fold cross-validation using gender, BLDAS, and 17 protein features of (C) training sets and corresponding (D) test sets.

4.3.3 Plasma protein model for clinical decision making

The final model was trained on the whole dataset and the beta coefficient of each feature obtained from the model was plotted against its feature importance (FI) obtained from the feature selection procedure and presented as Figure 4.4A. Table 5.3 summarises all the model features; gender, BLDAS and seventeen selected proteins along with their Uniprot and Entrez gene IDs, gene names, Feature Importance (FI) and Effect Sizes (ES) or regression/beta coefficients. Further the boxplot of calculated scores along with p-value for the patients is shown in Figure 4.4B. The model score (S) for each patient is given by:

$$S = \sum_{i=1}^n \beta_i x_i + b$$

Where, x_i are model features, β_i are corresponding effect sizes (or regression/beta coefficients) and b is the intercept (or bias). Finally, the patient's response to anti-TNF can be binarised, i.e., 0 for NR and 1 for R, by choosing a threshold (t) and mapping the score to logistic function, which takes the output to a probability of response by patient, $p \in [0,1]$ as per:

$$p = \text{logit}(S - t) = \frac{1}{1 + e^{-(S-t)}}$$

Where, ' t ' is the best point threshold, which was found to be 0.7136 (Figure 4.4B).

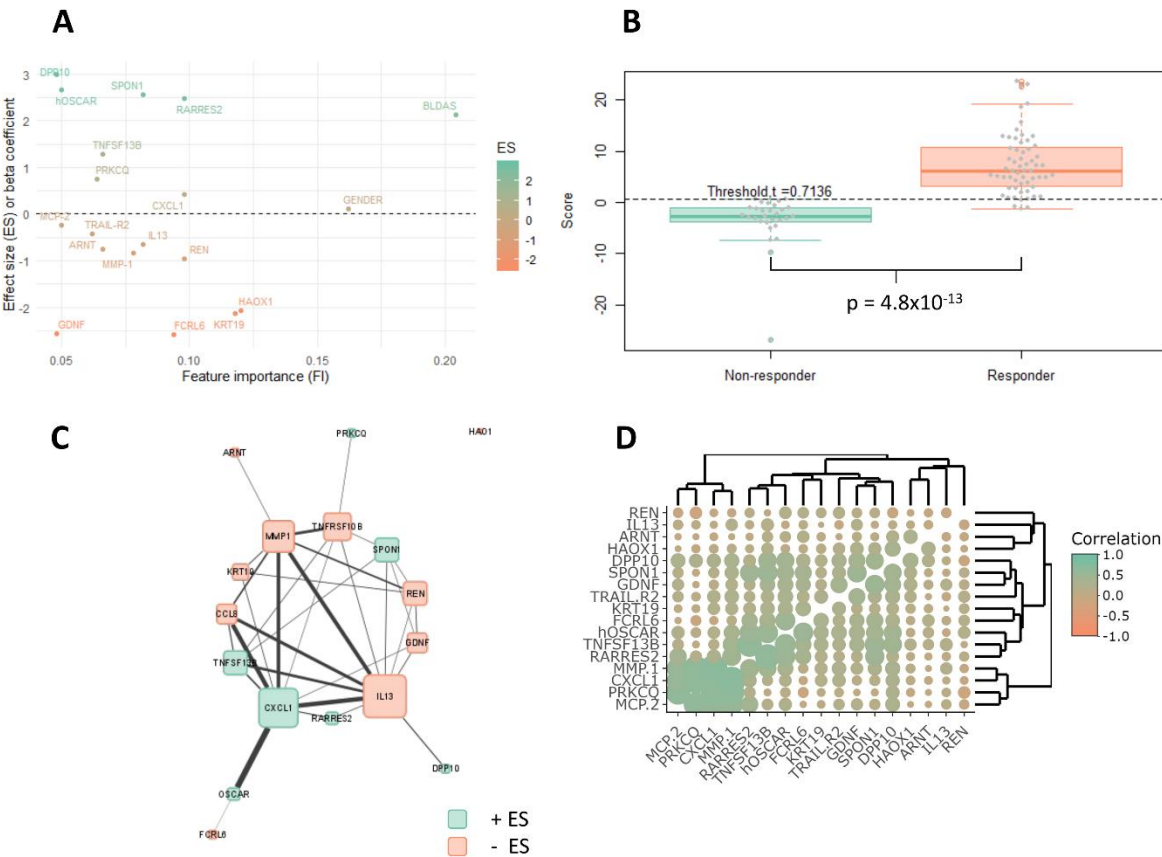


Figure 4.4 (A) Effect sizes (ES) or beta coefficients of regression vs. feature importance, i.e. fraction of 500 models, the feature appeared. (B) Boxplot of model score of each patient. NR = Non-responder, R = Responder. (C) Protein-Protein Interaction (PPI) network obtained from STRING database for 17 featured proteins. The size of the cell depicts the degree of the node i.e. number of connection with the other proteins, whereas the edge thickness represents the STRING database's interaction scores. ES = effect size, as presented in Table 4.3. (D) Pearson's correlation coefficient plot of 17 feature proteins. The size of circle depicts the $-\log_{10}(p\text{-value})$ of the correlation.

Table 4.3 Plasma protein signature, along with gender and baseline DAS (BLDAS) for anti-TNF treatment response prediction. Feature Importance (FI) is defined as the fraction of models a feature appears in. Beta (β) Coefficients are the effect sizes of features obtained from the logistic regression analysis. DAS = Disease activity score.

Features		Uniprot ID	Entrez Gene ID	Gene Name	Olink Panel	FI	β Coeff.
Intercept or bias, b		-	-	-	-	-	3.800
Baseline (BLDAS)	DAS	-	-	-	-	0.21	2.133
Gender (M:1; F:0)		-	-	-	-	0.17	0.116
KRT19		P08727	3880	Keratin 19	IMMUNE	0.13	-2.126
HAOX1/HAO1		Q9UJM8	54363	Hydroxyacid oxidase 1	CVD II	0.13	-2.068
CXCL1		P09341	2919	C-X-C motif chemokine ligand 1	CVD II + INFLAM	0.10	0.421
RARRES2		Q99969	5919	Retinoic acid receptor responder 2	CVD III	0.10	2.488
FCRL6		Q6DN72	343413	Fc receptor like 6	IMMUNE	0.10	-2.595
REN		P00797	5972	Renin	CVD II	0.10	-0.960
IL13		P35225	3596	Interleukin 13	INFLAM	0.09	-0.651
SPON1		Q9HCB6	10418	Spondin 1	CVD III	0.08	2.557
MMP-1/MMP1		P03956	4312	Matrix metalloproteinase 1	INFLAM	0.08	-0.830
ARNT		P27540	405	Aryl hydrocarbon receptor nuclear translocator	IMMUNE	0.07	-0.758
TNFSF13B		Q9Y275	10673	Tumor necrosis factor superfamily member 13b	CVD III	0.07	1.281
PRKCQ		Q04759	5588	Protein kinase C theta	IMMUNE	0.07	0.744
TRAIL-R2/TNFRSF10B		O14763	8795	TNF receptor superfamily member 10b	CVD II	0.07	-0.421
hOSCAR/OSCAR		Q8IYS5	126014	Osteoclast associated, immunoglobulin-like receptor	CVD II	0.05	2.661

MCP-2/CCL8	P80075	6355	C-C motif chemokine ligand 8	INFLAM	0.0 5	-0.243
DPP10	Q8N608	57628	Dipeptidyl peptidase like 10	IMMUNE	0.0 5	2.990
GDNF	P39905	2668	Glial cell derived neurotrophic factor	INFLAM	0.0 5	-2.574

4.3.4 Enrichment analysis with Gene Ontology (GO) terms and KEGG pathways

The 17 protein set, when tested for enrichment with Gene Ontology (GO) terms for Biological Process (BP) using STRING database, gave 72 significant ($FDR < 0.05$) hits as shown in Supplementary Table 4.2. These 72 GO BP terms along with its FDR, when summarised using REVIGO (Supplementary Table 4.3), were mostly involved with inflammatory response or its regulation (Supplementary Figure 4.2). The enrichment for GO terms for Molecular Function (MF) gave 8 significant ($FDR < 0.05$) hits (Supplementary Table 4.4), mostly corresponding to receptor binding. Furthermore, the enrichment for GO terms for Cellular Components (CC) gave 4 significant ($FDR < 0.05$) hits (Supplementary Table 4.5), mostly suggesting extracellular region as the location of proteins. Finally, the enrichment analysis for the KEGG pathway gave 6 significant ($FDR < 0.05$) hits as shown in Supplementary Table 4.6. These hits include, as expected, rheumatoid arthritis pathway. Further, it also included IL-17 signalling pathway as well as NF-kappa B signalling pathway, which are well known for their role in inflammatory response in case of rheumatoid arthritis [55,56], suggesting their pathological role in response to biologic DMARDs as well. It was also interesting to see Measles appearing in these hits. It was recently found through pathway and network analyses of Genome-Wide Association Studies (GWAS) that Measles truly contributes to rheumatoid arthritis [57].

4.3.5 Network analysis

STRING database reports scores for Protein-Protein Interaction (PPI). These scores range from 0 for no evidence of interaction to 1 implying evidence of strong interaction. These scores are computed using different parameters such as co-expression, annotated pathways, neighbourhood, text mining, etc. We obtained the combined PPI scores of all combinations

of our feature proteins. The PPI network thus obtained, was then uploaded in Cytoscape for visualizing the graph in circular layout (Figure 4.4C). The size of the cell corresponds to the degree i.e., number of connections with the other proteins. We note that the cytokine IL13 has the highest degree of connection in the network; connected to 10 other feature proteins (Figure 4.4C). This was closely followed by CXCL1 which was connected to 9 other feature proteins. Further, the edge thickness is proportional to the score from STRING database. Figure 4.4C shows thick edges connecting IL13, CXCL1, CCL8 (alias MCP-2) and MMP1, thus implying high interaction between them. Interestingly, all these proteins are present in the extracellular region (Supplementary Table 4.5) and except CCL8 all other proteins are involved in IL17 signalling pathway (Supplementary Table 4.6). Out of these four highly interactive proteins, only CXCL1 has positive effect size to response to treatment, whereas IL13, CCL8, and MMP1 have negative effect sizes (Table 4.3). Thus, a high expression of CXCL1 and low expression of IL13, CCL8, and MMP1 will lead to a better response to anti-TNF treatment. Further, these four highly interacting proteins have smaller effect sizes compared to other proteins (Figure 4.4A), suggesting they are correlated due to their high PPI scores. We confirmed that indeed MMP1, MCP-2 (alias CCL8) and CXCL1 are significantly and highly correlated (Figure 4.4D). The elastic net regression distributes the weightage among the three proteins due to redundancy, as these variables have similar variations. On the contrary, less correlated features, even if they have low FI, have high effect sizes, since they have independent variation and can contribute more to anti-TNF treatment response prediction.

4.4 Discussion

Rheumatoid arthritis (RA) patients show different pathologies in terms of functional or biological mechanism, treatment response, etc. and hence can be considered as a broad disease class containing different disease entity or sub-class. Therefore, there is a need to further stratify patients based on their distinct functional or pathobiological mechanism, more commonly called as endotypes [58]. A recent review article [59], investigates such pathobiological endotypes in early RA ($n = 85$). They validated 2 proteins, 52 SNPs and 72 gene expression biomarkers, that were predictive of changes in DAS28-CRP, identified from literature review. Out of the 72 transcript biomarkers, they independently replicated 8 biomarkers (SORBS3, AKAP9, CYP4F12, MUSTN, CX3CR1, SLC2A3, C21orf58 and

TBC1D8). Further, the two protein candidates viz. sICAM1 and CXCL13 were also validated as predictors of anti-TNF response. They have also validated 2 SNPs (rs6028945 and rs73055646), that were significantly associated with anti-TNF response. Using 11 biomarkers, this integrative approach showed an anti-TNF response predictability with an AUC of 0.815.

The current study uncovered two distinct endotypes based on the expression profile of 352 plasma proteins, which had significantly different gender proportions and baseline DAS (Figure 4.1 and Table 4.2). Since these endotypes were not significantly different in terms of their anti-TNF treatment response (Table 4.2), there is a possibility of the existence of two distinct RA disease endotypes, which may be important in other aspects of the disease management or other drug response.

Gender is known to be significantly associated with plasma protein profile [60]. Further, DAS28 is also known to be correlated with plasma proteins such as IL37 [61] and CXCL10 [62]. A significantly higher average ESR has been observed in females of age up to 75 years [63]. Considering the above literature, there is another possibility that the two endotypes uncovered in this study may be totally unrelated to RA. Hence, the clinicians may consider keeping a strict vigil on these endotypes, which may be helpful in better informed decision making.

Anti-TNF therapy is also a part of treatment regimens followed in other inflammatory disorders like psoriatic arthritis and inflammatory bowel disease (IBD), which includes Crohn's disease (CD) and ulcerative colitis (UC). Proteomic signature for response to anti-TNF treatment in these disorders have also been studied. About 57 out of 107 targeted proteins were found to be predictive to anti-TNF treatment response with AUC of 0.76 in psoriatic arthritis [64]. In another study [65], 25 potential anti-TNF treatment predictive biomarkers based on significant differential expression between good and poor response were suggested out of 119 investigated proteins in psoriatic arthritis (n = 12). They further went on to investigate 4 out of the 25 proteins as the anti-TNF treatment predictive biomarkers, however, none of these 25 differentially expressed proteins have any intersection with our feature proteins. Another study [66] tried to stratify patients (n = 56) for prognosis or predicting response to anti-TNF therapy in IBD by identifying candidate proteomics biomarkers involved in therapeutic pathways. They suggested overall expression of defensin-

5 α and eosinophil cationic protein was related to responders ($n = 25$) and high expression of cathepsin, IL-12, IL17A and TNF was related to non-response ($n = 31$). Unfortunately, performance of anti-TNF treatment response prediction was not reported. With AUC of 0.86 for a relatively bigger cohort ($n = 89$), our plasma protein signature for the prognosis of anti-TNF therapy responsiveness in RA patients is different and its prediction performance is more accurate than of those described in the studies discussed above.

A robust machine learning based bioinformatics study requires a complete independent test set from the cross-validation set for the evaluation of the predictive model. Conventionally, a single choice of independent test set is implemented, leading to possible biasness towards better performance of the predictive model. To mitigate this issue and being conscious of our limited sample size, we implemented a double or nested cross-validation based ML architecture (Figure 4.2B), which not only ensures an independent test set from the cross-validation sets, but also removes the biasness from choosing the independent test set by averaging the performance for all possible choice of independent test sets.

The feature importance (FI) for the proteins, obtained from the feature selection procedure, suggest the need for the feature to be included in the model. Further, the effect sizes or regression/beta coefficient, obtained from the model training, suggests the contribution of a particular feature protein has on the final score of the patient. However, FI and β -coefficient are not correlated (Figure 4.4A). This is due to the fact that some of the proteins are interacting with each other (Figure 4.4C) and therefore are correlated (Figure 4.4D). All the feature proteins having a lower β -coefficients are mostly correlated with each other and therefore the Elastic-Net regression analysis distributes their weightage due to redundancy. Proteins that can classify patients into responders and non-responders to anti-TNF drugs were filtered down to seventeen (Table 4.3). The model presented is a simple linear combination of gender, BLDAS, and plasma protein expression values that has been implemented to develop an R-based tool ATRPred. Further, the model was 5-fold cross-validated and the mean performance was reported, which although modest, is the highest till date as per the literature review presented and the author's knowledge.

In current clinical practice, RA patients who may not respond to conventional DMARDs are routinely administered anti-TNF therapy, without enough prior knowledge of potential for efficacy. Table 4.3 indicates that gender and BLDAS have the highest discriminatory

Chapter 4: ATRPred

feature importance with respect to future response to anti-TNF therapy. These two features were also significantly different for treatment response to anti-TNF therapy (Table 4.1). It is common knowledge amongst clinicians that the response to biologics is greater when the ESR is higher. This knowledge is also advocated by NICE (National Institute for Health and Care Excellence) guidelines which recommends a cut-off of DAS28-ESR >5.1 . The patients had all fulfilled the criteria (DAS28 >5.1) but at the time they started therapy their disease could have been going through a flare or a dip in disease activity. The former would clearly be expected to respond better, partly from the ‘regression to the mean’ trend. However, significance of female patients in general respond better to biologics than male patients have not been widely reported. Females are less likely to achieve remission with DAS28-ESR partly due to differences in the baseline ESR and the way the DAS28 is calculated [52]. Further, it is known that RA is more commonly found in women than men [67]. In line with this, most of the patients observed by the clinicians in our BioRA cohort were also females (Table 4.1). We have taken these two demographic and clinical features, viz. gender and BLDAS, as confounders and included in our signature summarised in Table 4.3. As per the model performance (Supplementary Table 4.1), we can note that the performance using just the gender and BLDAS has a test set 5-fold mean AUC of 0.57. A random model has an AUC 0.5, hence the clinical decision making using these two demographic and clinical features is only slightly better than random. However, inclusion of the 17 informative plasma proteins increased the test set 5-fold mean AUC to 0.86, resulting in about 51% increase in performance (Supplementary Table 4.1). Thus, our plasma protein signature may prove to be an advancement in the current clinical decision making and treatment regime of anti-TNF therapy for RA patients.

Different genome wide association studies clearly implicate the central role of the immune system in RA. To further investigate the pathways defining the patients’ responsiveness and to understand the biological processes underlying the 17-protein signature, we went on to carry out enrichment analysis and network analysis. Well known rheumatoid arthritis related pathways such as IL-17 and NF-kappa B signalling pathway were found to be significantly enriched in this protein signature. Further, the clustering of significant GO BP terms for the 17 featured protein set suggests that they mostly belong to either inflammatory response or its regulation (Supplementary Figure 4.2). However, our study was limited to the set of proteins obtained from four pre-selected Olink Proteomics’ panels; so, there is a possibility

of selection bias which would influence enrichment analysis. To get an unbiased pathway topology, we extracted a protein-protein interaction network that was built on pre-existing knowledge (Figure 4.4C). We identified four highly interacting proteins IL13, CXCL1, CCL8, and MMP1. IL13, CXCL1 and MMP1 are involved in IL-17 signalling pathway, and their signature in responders suggests a potential role of IL-17 signalling pathway in anti-TNF response. Out of these proteins, only CXCL1 has positive effect size i.e., its higher baseline expression is indicative of future anti-TNF response. Further, CXCL1 is known to contribute to inflammation and present at higher levels during inflammatory flare [68]. Thus, a high pre-treatment CXCL1 expression may act as a sentinel of future good response towards anti-TNF treatment.

We have identified two clusters (Figure 4.1 and Table 4.2) driven by plasma protein profile as plausible endotypes. Unfortunately, they do not correspond to anti-TNF therapy responsiveness, but they are still significantly different in terms of disease activity and gender, and thus possibly play an important role in patient management. For example, since these endotypes are independent of future treatment response, they may indicate pre-biologic treatment pathology sub-groups, which can be investigated in future studies. Further, we have built an ML based classifier ATRPred to predict anti-TNF treatment response of RA patients at earlier timepoint using seventeen proteins feature set along with gender and BLDAS. Our model was rigorously cross-validated and performance on model-blind test sets have been presented. We have provided this tool in the form of a R-based package on an open-source GitHub repository at <https://github.com/ShuklaLab/ATRPred>, which may aid clinicians in deciding about putting an RA patient under anti-TNF therapy. This will help in saving the treatment cost as well as preventing nonresponsive patients to go through refractory condition of the disease leading to poor quality of life.

4.5 References

1. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. *Nat Rev Dis Primers*. 2018;4:18001. pmid:29417936.
2. Mewar D, Wilson AG. Treatment of rheumatoid arthritis with tumour necrosis factor inhibitors. *Br J Pharmacol*. 2011;162(4):785–91. pmid:21039421.

3. Roda G, Jharap B, Neeraj N, Colombel JF. Loss of response to anti-TNFs: definition, epidemiology, and management. *Clin Transl Gastroenterol*. 2016;7(1):e135. pmid:26741065.
4. Caceres V. Common characteristics in RA patients who don't respond to biologics. *The Rheumatologist*. 2019. Available from: <https://www.the-rheumatologist.org/article/common-characteristics-in-ra-patients-who-dont-respond-to-biologics> (Accessed: 12 April 2021).
5. McInnes IB, Buckley CD, Isaacs JD. Cytokines in rheumatoid arthritis—shaping the immunological landscape. *Nat Rev Rheumatol*. 2016;12(1):63–8. pmid:26656659.
6. Freites-Núñez D, Baillet A, Rodriguez-Rodriguez L, Nguyen MVC, Gonzalez I, Pablos JL, et al. Efficacy, safety and cost-effectiveness of a web-based platform delivering the results of a biomarker-based predictive model of biotherapy response for rheumatoid arthritis patients: a protocol for a randomized multicenter single-blind active controlled clinical trial (PREDIRA). *Trials*. 2020;21(1):755. pmid:32867830.
7. Kearsley-Fleet L, Davies R, De Cock D, Watson KD, Lunt M, Buch MH, et al. Biologic refractory disease in rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register for Rheumatoid Arthritis. *Ann Rheum Dis*. 2018;77(10):1405–1412. pmid:29980575.
8. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res*. 2002;4 Suppl 3(Suppl 3):S265–72. pmid:12110146.
9. EULAR. EULAR's position and recommendations. 2011. Available from: https://www.eular.org/myUploadData/files/EU_Horizon_2020_EULAR_position_paper.pdf (Accessed: 12 April 2021).
10. Mikulic M. Global pharmaceutical industry—statistics & facts. Statista: Health and Pharmaceuticals. 2018. Available from: <https://www.statista.com/topics/1764/global-pharmaceutical-industry> (Accessed: 4 September 2020).
11. Inotai A, Tomek D, Niewada M, Lorenzovici L, Kolek M, Weber J, et al. Identifying patient access barriers for tumor necrosis factor alpha inhibitor treatments in rheumatoid arthritis in five central eastern european countries. *Front Pharmacol*. 2020;11:845. pmid:32581804.

12. Hughes LB, Danila MI, Bridges SL. Recent advances in personalizing rheumatoid arthritis therapy and management. *Per Med*. 2009;6(2):159–170. pmid:29788606.
13. Stuhlmüller B, Skriner K, Häupl T. Biomarker zur Prognose des Ansprechens auf eine Anti-TNF-Therapie bei der rheumatoiden Arthritis: Wo stehen wir? [Biomarkers for prognosis of response to anti-TNF therapy of rheumatoid arthritis: Where do we stand?]. *Z Rheumatol*.
14. Thomson TM, Lescarbeau RM, Drubin DA, Laifenfeld D, de Graaf D, Fryburg DA, et al. Blood-based identification of non-responders to anti-TNF therapy in rheumatoid arthritis. *BMC Med Genomics*. 2015;8:26. pmid:26036272.
15. Hueber W, Tomooka BH, Batliwalla F, Li W, Monach PA, Tibshirani RJ, et al. Blood autoantibody and cytokine profiles predict response to anti-tumor necrosis factor therapy in rheumatoid arthritis. *Arthritis Res Ther*. 2009;11(3):R76. pmid:19460157.
16. Ortea I, Roschitzki B, Ovalles JG, Longo JL, de la Torre I, González I, et al. Discovery of serum proteomic biomarkers for prediction of response to infliximab (a monoclonal anti-TNF antibody) treatment in rheumatoid arthritis: an exploratory analysis. *J Proteomics*. 2012;77:372–82. pmid:23000593.
17. Blaschke S, Rinke K, Maring M, Flad T, Patschan S, Jahn O, et al. Haptoglobin- $\alpha 1$, - $\alpha 2$, vitamin D-binding protein and apolipoprotein C-III as predictors of etanercept drug response in rheumatoid arthritis. *Arthritis Res Ther*. 2015;17(1):45. pmid:25884688.
18. Ortea I, Roschitzki B, López-Rodríguez R, Tomero EG, Ovalles JG, López-Longo J, et al. Independent candidate serum protein biomarkers of response to adalimumab and to infliximab in rheumatoid arthritis: an exploratory study. *PLoS One*. 2016;11(4):e0153140. pmid:27050469.
19. Eng GP. Optimizing biological treatment in rheumatoid arthritis with the aid of therapeutic drug monitoring. *Dan Med J*. 2016;63(11):B5311. pmid:27808043.
20. Xie X, Li F, Li S, Tian J, Chen JW, Du JF, et al. Application of omics in predicting anti-TNF efficacy in rheumatoid arthritis. *Clin Rheumatol*. 2018;37(1):13–23. pmid:28600618.
21. Folkersen L, Brynedal B, Diaz-Gallo LM, Ramsköld D, Shchetynsky K, Westerlind H, et al. Integration of known DNA, RNA and protein biomarkers provides prediction of anti-TNF response in rheumatoid arthritis: results from the COMBINE study. *Mol Med*. 2016;22:322–328. pmid:27532898.

22. Aterido A, Cañete JD, Tornero J, Blanco F, Fernández-Gutierrez B, Pérez C, et al. A combined transcriptomic and genomic analysis identifies a gene signature associated with the response to anti-TNF therapy in rheumatoid arthritis. *Front Immunol.* 2019;10:1459. pmid:31312201.
23. van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. *Arthritis Rheum.* 1996;39(1):34–40. pmid:8546736.
24. Lundberg M, Thorsen SB, Assarsson E, Villablanca A, Tran B, Gee N, et al. Multiplexed homogeneous proximity ligation assays for high-throughput protein biomarker research in serological material. *Mol Cell Proteomics.* 2011;10(4):M110.004978. pmid:21242282.
25. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum.* 1995;38(1):44–8. pmid:7818570.
26. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum.* 2010;62(9):2569–81. pmid:20872595.
27. Ledingham J, Deighton C, et al. Update on the British Society for Rheumatology guidelines for prescribing TNFalpha blockers in adults with rheumatoid arthritis (update of previous guidelines of April 2001). *Rheumatology (Oxford).* 2005;44(2):157–63. Erratum in: *Rheumatology (Oxford).* 2006;45(9):1170. pmid:15637039
28. Deighton C, Hyrich K, Ding T, Ledingham J, Lunt M, Luqmani R, et al. BSR and BHPR rheumatoid arthritis guidelines on eligibility criteria for the first biological therapy. *Rheumatology (Oxford).* 2010; 49(6):1197–9. Erratum in: *Rheumatology (Oxford).* 2010; 49(8):1609. pmid:20308121.

29. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One*. 2014;9(4):e95192. pmid:24755770.
30. Olink. Strategies for design of protein biomarker studies. 2018. Available from: <https://www.olink.com/content/uploads/2018/09/Strategies-for-design-of-protein-biomarker-studies-v1.0.pdf> (Accessed: 12 April 2021).
31. Lind L, Elmståhl S, Ingelsson E. Cardiometabolic proteins associated with metabolic syndrome. *Metab Syndr Relat Disord*. 2019;17(5):272–279. pmid:30883260.
32. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. Available from: <https://www.R-project.org>.
33. Arya S, Mount D, Kemp SE, Jefferis G. RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric. R package version 2.6.1. 2019. Available from: <https://CRAN.R-project.org/package=RANN>.
34. Taylor M. Sinkr: Collection of functions with emphasis in multivariate data analysis. R package version 0.6. 2020. Available from: <https://github.com/marchtaylor/sinkr>.
35. Kuhn M. Caret: Classification and Regression Training. R package version 6.0–86. 2020. Available from: <https://CRAN.R-project.org/package=caret>.
36. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7–3. 2019. Available from: <https://CRAN.R-project.org/package=e1071>.
37. Friedman J, Hastie T, Tibshirani T. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22. pmid:20808728.
38. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. pmid:21414208.
39. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*. 2008;50(3):419–430. pmid:18435502.
40. Kampstra P. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software, Code Snippets*. 2008;28(1):1–9.

41. Eklund A. Beeswarm: The Bee Swarm Plot, an alternative to Stripchart. R package version 0.2.3. 2016. Available from: <https://CRAN.R-project.org/package=beeswarm>.
42. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: various R programming tools for plotting data. R package version 3.0.3. 2020. Available from: <https://CRAN.R-project.org/package=gplots>.
43. Slowikowski K. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. R package version 0.9.1. 2021. Available from: <https://CRAN.R-project.org/package=ggrepel>.
44. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. 2016. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 2017;45(D1):D362–D368. pmid:27924014.
45. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one.* 2011;6(7):e21800. pmid:21789182.
46. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498–2504. pmid:14597658.
47. Galili T, O'Callaghan A, Sidi J, Sievert C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics.* 2018;34(9):1600–1602. pmid:29069305.
48. Enroth S, Berggrund M, Lycke M, Broberg J, Lundberg M, Assarsson E, et al. High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Commun Biol.* 2019;2:221. pmid:31240259.
49. Stone M. Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological).* 1974;36(2):111–133.
50. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research.* 2010;11:2079–2107.
51. Wickham H, Hester J, Chang W. devtools: tools to make developing R packages easier. R package version 2.4.0. 2021. Available from: <https://CRAN.R-project.org/package=devtools>.

52. Hyrich KL, Watson KD, Silman AJ, Symmons DP, et al. Predictors of response to anti-TNF-alpha therapy among patients with rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register. *Rheumatology (Oxford)*. 2006;45(12):1558–65. pmid:16705046.
53. Bro R, Kjeldahl K, Smilde AK, Kiers HA. Cross-validation of component models: a critical look at current methods. *Anal Bioanal Chem*. 2008;390(5):1241–51. pmid:18214448.
54. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6. pmid:31898477.
55. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. *Lancet*. 2016;388(10055):2023–2038. Erratum in: *Lancet*. 2016;388(10055):1984. pmid:27156434.
56. Serasanambati M, Chilakapati SR. Function of nuclear factor kappa B (NF-kB) in human diseases-a review. *South Indian Journal of Biological Sciences*. 2016;2(4):368–87.
57. Liu G, Jiang Y, Chen X, Zhang R, Ma G, Feng R, et al. Measles contributes to rheumatoid arthritis: evidence from pathway and network analyses of genome-wide association studies. *PLoS One*. 2013;8(10):e75951. pmid:24204584.
58. Russell CD, Baillie JK. Treatable traits and therapeutic targets: goals for systems biology in infectious disease. *Curr Opin Syst Biol*. 2017;2:140–146. pmid:32363252.
59. Tarn JR, Lendrem DW, Isaacs JD. In search of pathobiological endotypes: a systems approach to early rheumatoid arthritis. *Expert Rev Clin Immunol*. 2020;16(6):621–630. pmid:32456483.
60. Silliman CC, Dzieciatkowska M, Moore EE, Kelher MR, Banerjee A, Liang X, et al. Proteomic analyses of human plasma: Venus versus Mars. *Transfusion*. 2012;52(2):417–24. pmid:21880043.
61. Xia T, Zheng XF, Qian BH, Fang H, Wang JJ, Zhang LL, et al. Plasma interleukin-37 Is elevated in patients with rheumatoid arthritis: its correlation with disease activity and Th1/Th2/Th17-related cytokines. *Dis Markers*. 2015;2015:795043. pmid:26435567.

62. Pandya JM, Lundell AC, Andersson K, Nordström I, Theander E, Rudin A. Blood chemokine profile in untreated early rheumatoid arthritis: CXCL10 as a disease activity marker. *Arthritis Res Ther*. 2017;19(1):20. pmid:28148302.
63. Wetteland P, Røger M, Solberg HE, Iversen OH. Population-based erythrocyte sedimentation rates in 3910 subjectively healthy Norwegian adults. A statistical study based on men and women from the Oslo area. *J Intern Med*. 1996;240(3):125–31. pmid:8862121.
64. Ademowo OS, Hernandez B, Collins E, Rooney C, Fearon U, van Kuijk AW, et al. Discovery and confirmation of a protein biomarker panel with potential to predict response to biological therapy in psoriatic arthritis. *Ann Rheum Dis*. 2016;75(1):234–41. pmid:25187158.
65. Collins ES, Butt AQ, Gibson DS, Dunn MJ, Fearon U, van Kuijk AW, et al. A clinically based protein discovery strategy to identify potential biomarkers of response to anti-TNF- α treatment of psoriatic arthritis. *Proteomics Clin Appl*. 2016;10(6):645–62. pmid:26108918.
66. Wang C, Baer HM, Gaya DR, Nibbs RJB, Milling S. Can molecular stratification improve the treatment of inflammatory bowel disease? *Pharmacol Res*. 2019;148:104442. pmid:31491469.
67. Favalli EG, Biggioggero M, Crotti C, Becciolini A, Raimondo MG, Meroni PL. Sex and management of rheumatoid arthritis. *Clin Rev Allergy Immunol*. 2019;56(3):333–345. pmid:29372537.
68. Silva RL, Lopes AH, Guimarães RM, Cunha TM. CXCL1/CXCR2 signaling in pathological pain: role in peripheral and central sensitization. *Neurobiol Dis*. 2017;105:109–116. pmid:28587921.

Chapter 5 muSignAl: An algorithm to search for multiple omic signatures with similar predictive performance

Contribution

Bodhayan Prasad performed data analysis, data visualization, data interpretation, built the software muSignAl and wrote the first draft of the manuscript.

Dr. Priyank Shukla and Prof. Anthony J. Bjourson conceived and supervised the project, and helped in data interpretation, reviewing and editing of the manuscript.

This chapter is a slightly modified version of below-mentioned open access technical brief published in Wiley's Proteomics journal, under Creative Common Attribution 4.0 (CC BY 4.0) license (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted reuse, distribution and reproduction in any medium, provided the original work is properly cited.

Prasad B, Bjourson AJ, Shukla P. *muSignAl: An algorithm to search for multiple omic signatures with similar predictive performance*. Proteomics. 2023 Jan;23(2):e2200252. doi: 10.1002/pmic.202200252. Epub 2022 Oct 3. PMID: 36076312.

5.1 Introduction

In bioinformatics data analysis, sometimes we encounter a small sample size, for example in case of patient recruitment for rare diseases [1, 2]. However, the samples may frequently have a large number of features, such as those involving high throughput omics experiments [3]. Multiple features from such samples are often highly correlated, for example, due to their involvement in associated biological interactions, and hence multiple different combinations of the correlated features can perform similarly in predicting a given phenotype or outcome. This opens up a possibility of discovering multiple feature sets, that are equally good in predicting the phenotype. These multiple signatures, for example, can help in understanding the relationship between multiple combinations of biological features and phenotypes. They can also help in optimising the cost of biomarker panel development for diagnostic or prognostic applications, by providing more options of signatures with equally good predictive performance. A recent study [4, 5] reported a method to obtain unbiased features in such situations involving low sample size and high feature space. However, the authors did not explore the possibility of recursive search of all possible feature combinations as its complexity is of $O(N^N)$ making it exponentially computer intensive with the increase in features. Taking inspiration from Enroth et al. study [6], we have developed and implemented muSignAI algorithm in R, which recursively explores all feature combinations by systematically deleting the selected features one-by-one to facilitate the discovery of multiple signatures that exhibit similar predictive performance (Figure 5.1).

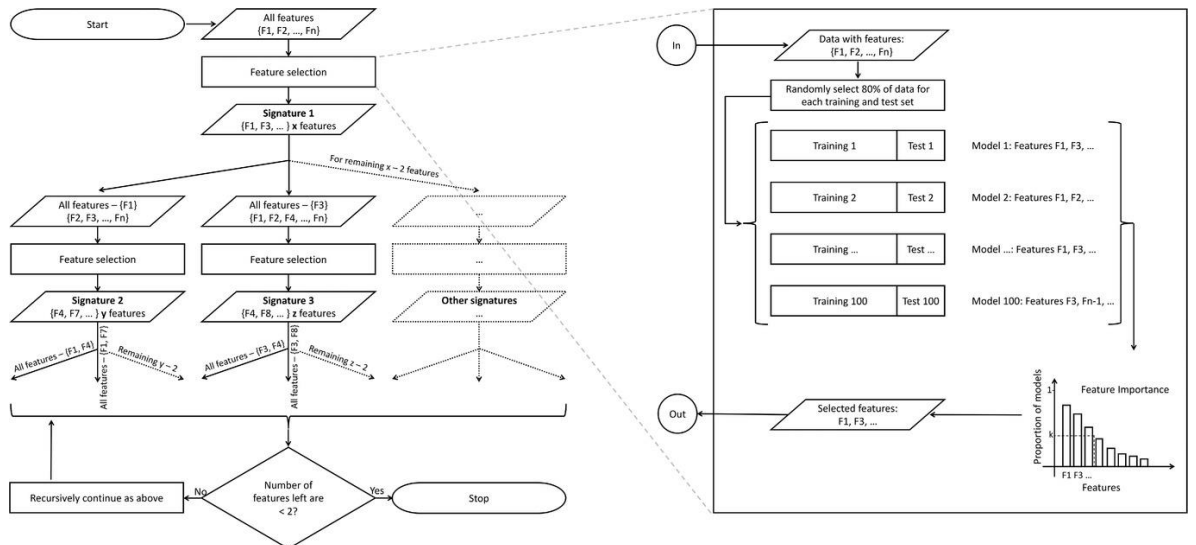


Figure 5.1 Flowchart presentation of the muSignAI algorithm. Oval shapes represent start/stop, circles represent in/out connectors, parallelogram boxes represent input/output, square boxes represent computation process, and rhombus box represent decision process. Signatures retrieved at each pass are presented in bold. The decision of whether the number of features is less than 2 will be taken at every input box. The feature selection box of the main algorithm on the left has been zoomed-out and presented on the right. k is the cut-off for the feature importance (FI) where FI is the proportion of models in which the feature has appeared.

5.2 Algorithm development

Initially, 80% (default value) of the data along with all the features is randomly selected and feature selection is performed using generalised linear models (GLMs) (Figure 5.1, right panel). This process is repeated 100 times (default value) and feature sets showing an area under the receiver operating characteristic curve (AUC) > 0.5 are selected, where > 0.5 ensures better than a random selection. The proportion of models in which the feature has appeared is obtained as a measure of feature importance (FI). A threshold $k = 0.9$ (default value) is applied on the FI to select the first set of features and its predictive performance is calculated. In the next pass, features from this first feature set are removed one-by-one and the above process of feature selection is recursively repeated on the reduced feature set until the algorithm is left with only two features in each leaf at the bottom of the algorithm tree (Figure 5.1, left panel). All the feature sets along with their predictive performance are sent as an output of the function. Default values of all the parameters of the algorithm can be changed by the user. A sample evaluation case run with the final output is available in the example folder of the Github repository <https://github.com/ShuklaLab/muSignAI>, and have been presented and discussed later.

We have used the least absolute shrinkage and selection operator (LASSO) for feature selection [7]. However, LASSO saturates with fewer features [8]. This was overcome by partly including Ridge regularisation, resulting in an Elastic Net model. We deployed GLMs to create an intuitive mathematical formulation with a linear combination of feature values. The GLM was an Elastic Net with alpha of 0.9, which implements regression with 90% LASSO and 10% Ridge regularization. The aim was to select non-correlated features, which

was achieved by LASSO regularization. The muSignAl algorithm is developed in an open-source platform R version 3.6 [9]. The basic data pre-processing was done using the caret package [10]. The model was built using the glmnet package [8]. ROC was built using the pROC package [11]. The algorithm requires a dataframe of features and target variable. The R function muSignAl() reads the input data file along with feature space to search from and target variable. It then outputs multiple signatures along with their performances as a dataframe. The algorithm is provided as a tool on the open-source GitHub repository: <https://github.com/ShuklaLab/muSignAl>.

5.3 Algorithm evaluation

For the evaluation of muSignAl algorithm, we have taken the publicly available dataset from Brunner et al. study [12] which includes 77 patient samples and 91 protein features from Olink Proteomics (<https://www.olink.com>) CVD-II panel. Samples were grouped as a healthy cohort ($n = 18$) and an atopic dermatitis cohort ($n = 59$). We ran muSignAl() with a FI threshold of $k = 0.9$ on this dataset, which generated 1984 signatures as an output; out of which, 158 were unique. Figure 5.2 reports the AUC performance of these 158 signatures; out of which, 47 had greater than 0.95 AUC (Table 5.1). The AUC has been recommended to be used in preference to overall accuracy when evaluating machine learning (ML) algorithms [13, 14], which should equally apply when evaluating different multiple signatures. Other bioinformatics studies also report AUC as one of the major performance metrics when applying ML [15-17]. The above case run took 20.05 h on a PowerEdge R740XD server. However, since the muSignAl algorithm implements a recursive function, the computational run-time may vary depending on the dataset and computational resources. For example, if most of the feature variables present in the dataset are predictive (i.e., significantly associated with a phenotype), the algorithm will identify more signatures and hence will take longer. Similarly, if most of the feature variables present in the dataset are non-predictive, the algorithm will stop quickly.

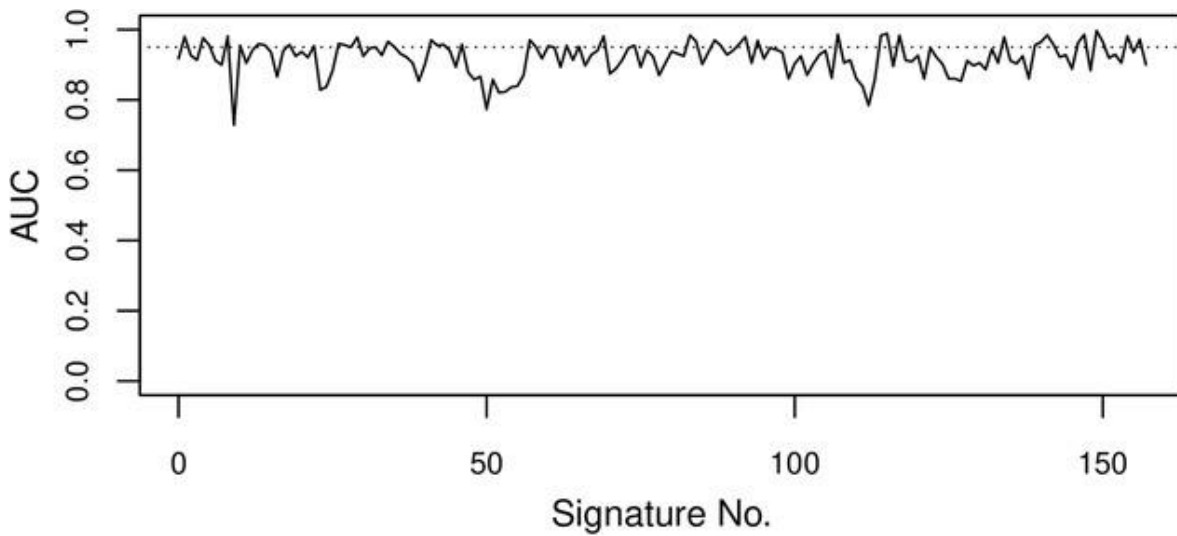


Figure 5.2 AUC values of the first 158 unique signatures. The dotted line shows 0.95 cut-off value for AUC. Forty-seven peaks above the dotted line are the 47 signatures presented in Table 5.1.

Table 5.1 Selected signature proteins (feature sets) based on $AUC > 0.95$.

Signature No.	Signature proteins	AUC
2	IDUA, IL16, DECR1, SORT1, GT	0.98
5	GT, IL16, PARP-1, STK4, SORT1	0.98
6	IL16, STK4, PARP-1, SORT1	0.95
9	STK4, GT, PARP-1, SCF, SORT1	0.98
11	PARP-1, SORT1, STK4, GT	0.95
14	IL16, PARP-1, GT, SORT1, GH	0.96
15	IL16, GT, PARP-1, TIE2	0.96
19	GT, NEMO, PARP-1, TIE2, IL-4RA, SORT1	0.96
23	TIE2, GT, NEMO, PARP-1	0.95
27	GT, NEMO, PARP-1, TIE2, IL-4RA	0.96
28	IL16, PARP-1, GT, TIE2	0.96
30	IL16, PARP-1, GT, SORT1, TIE2	0.98
35	GT, IL16, PARP-1, SORT1, DCN	0.97
36	IL16, PARP-1, SORT1, GT	0.95
42	STK4, MMP-12, PARP-1, GT, SORT1	0.97

43	GT, PARP-1, STK4, SORT1	0.95
44	GT, IL-4RA, NEMO, PARP-1, SORT1, TIE2	0.96
47	MMP-12, PARP-1, GT, STK4, TIE2	0.96
58	PARP-1, STK4, GT, MMP-12, SORT1	0.97
59	GT, MMP-12, NEMO, PARP-1, TIE2	0.95
61	GT, NEMO, PARP-1, TIE2	0.95
64	GT, STK4, PARP-1, SORT1	0.95
66	IL16, GT, PARP-1, SORT1	0.95
70	GT, STK4, PARP-1, SCF, SORT1	0.98
75	GT, PARP-1, SORT1, STK4	0.95
84	DECR1, GT, IL16, PARP-1, SORT1	0.98
85	DECR1, IL16, PARP-1, SORT1, GH	0.97
88	DECR1, IL16, PARP-1, SORT1	0.97
89	IL16, PARP-1, STK4, SORT1	0.95
92	DECR1, MMP-12, PARP-1, SORT1, GT	0.96
93	DECR1, IL16, GT, SORT1	0.98
95	DECR1, IL16, PARP-1, GT	0.97
108	IDUA, IL16, SORT1, GT	0.99
115	IDUA, IL16, ADM, SORT1	0.98
116	IDUA, IL16, SORT1, PARP-1	0.99
118	IDUA, IL16, SORT1	0.98
135	IDUA, IL16, ADM, DECR1, SORT1	0.98
140	CCL17, ADM, DECR1, IL16, PARP-1	0.96
141	CCL17, DECR1, SORT1, PARP-1	0.97
142	DECR1, IL16, PARP-1, SORT1, GT	0.98
143	CCL17, PARP-1, SORT1, IL16	0.96
147	CCL17, DECR1, SORT1	0.96
148	CCL17, IL16, SORT1, PARP-1, ADM	0.99
150	CCL17, DECR1, ADM, PARP-1, SORT1	1.00
151	CCL17, DECR1, PARP-1, SORT1	0.97
155	CCL17, ADM, DECR1, SORT1	0.98
157	CCL17, ADM, SORT1	0.97

Thus, muSignAI algorithm can discover multiple omic signatures with similar predictive performance. It will be useful in analysing multidimensional omic datasets, especially those with low sample sizes often encountered for example in studies of rare diseases. It will be applicable in various bioinformatics driven explorations, such as understanding the relationship between multiple combinations of biological features and phenotypes, and discovery and development of biomarker panels while providing the opportunity of optimising their development cost with the help of equally good multiple signatures.

5.4 References

1. Frésard, L., Smail, C., Ferraro, N. M., Teran, N. A., Li, X., Smith, K. S., Bonner, D., Kernohan, K. D., Marwaha, S., Zappala, Z., Balliu, B., Davis, J. R., Liu, B., Prybol, C. J., Kohler, J. N., Zastrow, D. B., Reuter, C. M., Fisk, D. G., Grove, M. E., ... Montgomery, S. B. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, 25(6), 911–919.
2. Papuc, S. M., Abela, L., Steindl, K., Begemann, A., Simmons, T. L., Schmitt, B., Zweier, M., Oneda, B., Socher, E., Crowther, L. M., Wohlrab, G., Gogoll, L., Poms, M., Seiler, M., Papik, M., Baldinger, R., Baumer, A., Asadollahi, R., Kroell-Seger, J., ... Rauch, A. (2019). The role of recessive inheritance in early-onset epileptic encephalopathies: A combined whole-exome sequencing and copy number study. *European Journal of Human Genetics: EJHG*, 27(3), 408–421.
3. Koh, Y., Park, I., Sun, C. H., Lee, S., Yun, H., Park, C. K., Park, S. H., Park, J. K., & Lee, S. H. (2015). Detection of a distinctive genomic signature in rhabdoid glioblastoma, a rare disease entity identified by whole exome sequencing and whole transcriptome sequencing. *Translational Oncology*, 8(4), 279–287.
4. Shi, L., & Brunius, C. (2018). A brief tutorial on MUVr: Multivariate methods with unbiased variable selection in R. https://gitlab.com/CarlBrunius/MUVr/-/blob/20210719_01/Tutorial/MUVr_Tutorial.pdf.
5. Shi, L., Westerhuis, J. A., Rosén, J., Landberg, R., & Brunius, C. (2019). Variable selection and validation in multivariate modelling. *Bioinformatics (Oxford, England)*, 35(6), 972–980.

6. Enroth, S., Berggrund, M., Lycke, M., Broberg, J., Lundberg, M., Assarsson, E., Olovsson, M., Ståhlberg, K., Sundfeldt, K., & Gyllenstein, U. (2019). High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Communications Biology*, 2, 221.
7. Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, 30, 1– 25.
8. Friedman, J., Hastire, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1– 22.
9. Core Team, R. (2019). R: A language and environment for statistical computing. R foundation for statistical computing. <https://www.R-project.org>.
10. Kuhn, M. (2021). Caret: Classification and regression training. R package version 6.0-90. <https://CRAN.R-project.org/package=caret>.
11. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77.
12. Brunner, P. M., Suárez-Fariñas, M., He, H., Malik, K., Wen, H. C., Gonzalez, J., Chan, T. C. C., Estrada, Y., Zheng, X., Khattri, S., Dattola, A., Krueger, J. G., & Guttman-Yassky, E. (2017). The atopic dermatitis blood signature is characterized by increases in inflammatory and cardiovascular risk proteins. *Scientific Reports*, 7(1), 8707.
13. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145– 1159.
14. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1), 86– 112.
15. Hung, T., Le, N., Le, N., Van Tuan, L., Nguyen, T. P., Thi, C., & Kang, J. H. (2022). An AI-based prediction model for drug-drug interactions in osteoporosis and Paget's diseases from SMILES. *Molecular Informatics*, 41(6), 2100264.
16. Le, N., & Ho, Q. T. (2022). Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods (San Diego, Calif.)*, 204, 199– 206.

17. Mustafić, S., Brkić, S., Prnjavorac, B., Sinanović, A., Porobić Jahić, H., & Salkić, S. (2018). Diagnostic and prognostic value of procalcitonin in patients with sepsis. *Medicinski glasnik: Official publication of the medical association of Zenica-Doboj Canton. Bosnia and Herzegovina*, 15(2), 93– 100.

Chapter 6 General discussions

Contribution

Bodhayan Prasad carried out all the research unless otherwise stated.

Dr. Priyank Shukla and Prof. Anthony J. Bjourson supervised the research and proofread.

6.1 The need for computational tools

With the advent of many computational and bioinformatic tools, stratified healthcare and personalised medicine has been growing very rapidly [1]. However, a lot of research questions are yet to be addressed. In the current era of multi-omics, machine learning (ML) is extensively used for the analyses of data [2]. We have tried to apply various ML techniques, specifically tailored on patient data, to identify various endotypes or biomarkers, that can be helpful in terms of better patient management and/or tailoring their treatment as per their needs. The pipelines that we have developed can be applied to patients recruited for other disease conditions as well.

In Chapter 3, we identified five endotypes of multimorbid patients, using an unsupervised ML approach (MulMorPip). This computational tool has a relevance in the development of stratified healthcare for patients with multimorbidity. Each of these endotypes can be different in terms of their phenotype such as response to treatment, morbidity risk, etc. Therefore, it can potentially help in better optimisation of these phenotypes. For example, a GP in the public healthcare sector could prioritise treatment if he finds these endotypes to be related to morbidity risk. This would additionally necessitate the incorporation of existing and enhanced medication use review tools.

In Chapter 4, we identified two endotypes of rheumatoid arthritis (RA), using an unsupervised ML approach, however these endotypes were not significantly different in terms of their response to anti-TNF treatment and hence, we went on to identify a proteomic based prognostic test (ATRPred) for response to anti-TNF treatment in RA patients, using a supervised ML approach. This computational tool has relevance in the development of personalised treatment for RA patients. More specifically, it can help in saving non-responsive patients from refractory conditions and thereby improve their quality of life. In terms of potential clinical implementation, a consultant rheumatologist could use ATRPred as a prognostic test, before putting their patients into anti-TNF therapy

Chapter 5 provides a computational tool (muSignAI) to identify multiple signatures which have similar predictive power for a phenotype. This tool is most effective and has a relevance in high dimensionality data, where we are likely to find many correlated or similar variables, making them redundant to one another. With the advent of high throughput technologies, we

Chapter 6: General discussions

see a flood of such datasets, especially in the omics field. If multiple signatures are available, for example, then a user can find a signature based on the variables that are being measured or cost effective.

Thus, the computational tools developed within the scope of this thesis are highly relevant to stratified healthcare and personalised medicine and thus may have clinical utility. These pipelines were developed using various ML approaches, that were tailored to the patient dataset. These pipelines are available as GitHub open access repositories for clinical or any further research and development.

6.2 Summary and key findings

We have summarised our research chapters 3-5, along with their key findings below:

6.2.1 Stratified healthcare in multimorbidity

Multimorbidity generally refers to concurrent occurrence of multiple chronic conditions. These patients are inherently at high risk and often lead a poor quality of life due to delayed treatments. With the emergence of personalised medicine and stratified healthcare, there is a need to stratify patients right at the primary care setting. Here we developed multimorbidity analysis pipeline (MulMorPip), which can stratify patients into multimorbid subgroups or endotypes based on their lifetime disease diagnosis and characterize them based on demographic features and underlying disease-disease interaction networks. By implementing MulMorPip on UK Biobank cohort, we report five distinct molecular subclasses or endotypes of multimorbidity. For each patient, we calculated the existence of broad disease classes defined by Charlson's comorbidity classification using the International Classification of Diseases-10 encoding. We then applied multiple correspondence analysis in 77 524 patients from UK Biobank, who had multimorbidity of more than one disease, which resulted in five multimorbid clusters. We further validated these clusters using machine learning and were able to classify 20% model-blind test set patients with an accuracy of 97% and an average Jaccard similarity of 84%. This was followed by demographic characterization and development of interlinking disease network for each cluster to understand disease-disease interactions. Our identified five endotypes of multimorbidity draw attention to dementia, stroke and paralysis as important drivers of multimorbidity stratification. Inclusion of such

patient stratification at the primary care setting can help general practitioners to better observe patients' multiple chronic conditions, their risk stratification and personalization of treatment strategies.

6.2.2 Personalised medicine in RA

Rheumatoid arthritis (RA) is a chronic autoimmune condition, characterised by joint pain, damage and disability, which can be addressed in a high proportion of patients by timely use of targeted biologic treatments. However, the patients, non-responsive to the treatments often suffer from refractoriness of the disease, leading to poor quality of life. Additionally, the biologic treatments are expensive. We obtained plasma samples from $N = 144$ participants with RA, who were about to commence anti-tumour necrosis factor (anti-TNF) therapy. These samples were sent to Olink Proteomics, Uppsala, Sweden, where proximity extension assays of 4 panels, containing 92 proteins each, were performed. A total of $n = 89$ samples of patients passed the quality control of anti-TNF treatment response data. The preliminary analysis of plasma protein expression values suggested that the RA population could be divided into two distinct molecular sub-groups (endotypes). However, these broad groups did not predict response to anti-TNF treatment, but were significantly different in terms of gender and their disease activity. We then labelled these patients as responders ($n = 60$) and non-responders ($n = 29$) based on the change in disease activity score (DAS) after 6 months of anti-TNF treatment and applied machine learning (ML) with a rigorous 5-fold nested cross-validation scheme to filter 17 proteins that were significantly associated with the treatment response. We have developed an ML based classifier ATRPred (anti-TNF treatment response predictor), which can predict anti-TNF treatment response in RA patients with 81% accuracy, 75% sensitivity and 86% specificity. ATRPred may aid clinicians to direct anti-TNF therapy to patients most likely to receive benefit, thus saving cost as well as preventing non-responsive patients from refractory consequences. ATRPred is implemented in R.

6.2.3 Computational tool for multiple signature detection

Multidimensional omic datasets often have correlated features leading to the possibility of discovering multiple biological signatures with similar predictive performance for a phenotype. However, their exploration is limited by low sample size and the exponential

Chapter 6: General discussions

nature of the combinatorial search leading to high computational cost. To address these issues, we have developed an algorithm muSignAl (multiple signature algorithm) which selects multiple signatures with similar predictive performance while systematically bypassing the requirement of exploring all the combinations of features. We demonstrated the workflow of this algorithm with an example of proteomics dataset. muSignAl is applicable in various bioinformatics-driven explorations, such as understanding the relationship between multiple biological feature sets and phenotypes, and discovery and development of biomarker panels while providing the opportunity of optimising their development cost with the help of equally good multiple signatures. Source code of muSignAl is freely available at <https://github.com/ShuklaLab/muSignAl>.

6.3 Reasoning and critical review of ML models applied in the thesis

One of the critical requirements to obtain unbiased results around performance of clusters or predictors developed by ML models is to have a totally independent model-blind set. However, due to the smaller dataset, frequently seen in medical/clinical studies, some of the transformations are often done of whole dataset, leading to the problem of information leakage and resulting in slightly inflated performance for the models. Some of these limitations, along with the rationale of different models developed in the thesis, are described in detail, below:

6.3.1 Clusters of multimorbidity (MulMorPip)

We applied various definition of multimorbidity indices viz. Charlson's Comorbidity Classification (CCC), Cumulative Illness Rating Scale (CIRS), Index of Coexistent Disease (ICED) and Kaplan indices, however, a pattern segregating cohort into five clusters was seen with CCC and hence we went on to further investigate the reasons behind the clustering. Further, the parametric (Student's t-test) and non-parametric test (Mann-Whitney U Test) of significance of demographic dataset were highly significant across clusters, even though the difference didn't look very striking. This might be due to high sample size, which makes it significant even for a very low difference and hence we skipped these tests for significance in Chapter 3.

CCC categorised multimorbidity data of the UK Biobank into Boolean variables of 17 broad disease class, stating either existence or absence of the disease class. We, therefore, chose to apply decision tree classifier (DTC) models during the cluster validation analysis to better understand what decision tree goes to each of the 5 clusters. However, we didn't try to optimise the model such as hyperparameter optimisation, as we thought accuracy was enough to suffice the validation of cluster. However, we did see Jaccard similarity of cluster 3 was as low as 60%, which we discussed due to narrow difference between cluster 2 and 3 and maybe they both belong to the same cluster. Hence, the hyper-parameter optimisation of model could have been resulted in better classification of clusters 2 and 3; and should be considered in future studies and thereby serves as a limitation to MulMorPip algorithm.

Further, as the clusters were defined as per the multiple correspondence analysis (MCA) on the whole dataset, this raises a question of information leakage from training set to test set during the cluster validation analysis. Even if we separately define labels in training and test sets, a proper folded cross-validation would be necessary to ensure real performance of the predictors. Hence, this is the major limitation for the machine learning model created in MulMorPip.

6.3.2 Endotypes of RA

For investigating existence of endotypes of RA, we applied linear method viz. principal component analysis (PCA) on protein expressions of RA cohort. We didn't see any patterns, while visualising first two principal components (PC1 & PC2), however, there was a clear split of data in terms of the third principal component (PC3). Upon investigation and literature review, we did find some biological significance in terms of gender and baseline disease activity for the clusters seen. This suggest that the variation in terms of these variables viz. gender and baseline disease activity, are not the major variation seen in the data but present in amongst the top 3 principal components. We could have further gone to try non-linear methods like t-SNE and/or UMAP, which could have further separated these clusters, but linear methods themselves looked visually significant, that we chose to go with it.

6.3.3 Anti-TNF treatment response predictor (ATRPred)

ATRPred deploys a logistic regression – a linear method for classification problems. The rationale behind using the linear method was to produce a model that is linear combination of predictive biomarkers, so that models could be visually understood and gives a handy experience for the clinicians. Many sophisticated ML models, such as gradient boosting, although are better in terms of their performance, they look like a black box to the clinicians. Further, for the calculation of feature importance, we used alpha at 0.90 for elastic net, as we were aware from our previous analysis that choice of 0.90, 0.95 or 0.99 doesn't make much of a difference, and it is for bypassing LASSO's limitation by including a bit of ridge regression.

Our major aim in ATRPred was to identify an exhaustive list of biomarkers from the protein expression set. Therefore, we started with creating feature importance of each of the proteins. Finally, for building up the model, these proteins were inducted one-by-one and mean performances for training and test set were plotted side by side. This allowed the choice of a model, which doesn't overfit – which a model normally does after having a lot of feature space. However, this procedure again can lead to information leakage problem, leading to inflated performance for the model. Due to the small dataset, we normalised whole protein expression (training and testing set) together, however, this also possesses problem of information leakage. Further, in the outer 5-fold nested cross-validation scheme in ATRPred, sensitivities of beta coefficient of biomarkers proteins could have been analysed, especially when fold-1 and fold-3 are showing low performances on the test set.

One of the major biomarkers found in ATRPred, was CXCL1, which is a small chemokine, known to be involved in inflammation under the immune process. We could see it to be majorly interacting with all the other biomarkers, especially in conjunction with a strong interaction with IL13, which is another small inflammatory chemokine, but with negative effect size. Hence, high expression of CXCL1 and thereby low expression of IL13 can increase responsiveness to anti-TNF in RA patients.

6.3.4 Multiple signature prediction algorithm (muSignAI)

Model-creation part of muSignAI applies methods similar to ATRPred, and hence the problem of information leakage in ATRPred as discussed in the previous section applies to muSignAI as well. We chose an AUC cut-off of 0.95 to list signatures with similar predictive performance, as we were able to achieve this accuracy for about hundred signatures. We checked muSignAI pipeline to produce ATRPred signature as first output and finally choose to apply it on independent dataset. However, we didn't go on to understand the biological signification of results obtained on the independent dataset as our aim was to develop computational tool.

Finally, although we have minimised the computationally intensive problem of searching the entire feature space, by a recursive search, it still can be further optimised. One of such alteration could be maintaining a hash table for all the signatures identified to minimise and inform the direction for signature search.

6.4 Challenges and limitations

There exists a lot of challenges in terms of quality control of the datasets [3]. Further, some of the datasets we used were quite small, especially in terms of sample sizes, however they had many features, making it a high dimensional dataset. This required us to produce the ML models, which are not biased, while taking care of overfitting of the models [4]. Some of the specific limitations of the research presented in the thesis are as follows:

- In Chapter 3, we have used Charlson's comorbidity classification (CCC) [5] to classify patients' disease diagnosis, however, there exist more sophisticated ways to define multimorbidity [6]. During our initial analysis, we tried to cluster UK Biobank participants based on other classification, but the one that showed some pattern was CCC scheme. Further, broad disease classes defined and used by CCC are commonly used by clinicians. Furthermore, Charlson Comorbidity Index (CCI) is a weighted sum of a patient's diseases, with higher value representing severity of multimorbidity.
- Further, in Chapter 3, no hyperparameter optimization for the DTC model in MulMorPip was carried out. This could have resulted in better classification seen in

Chapter 6: General discussions

terms of clusters 2 and 3, and therefore should be considered in future extension of the research.

- In Chapter 4, we did not have a very good sample size ($n=89$). Further, although we used nested CV schema for ensuring a model blind validation set, we do need a larger completely independent cohort to validate our results. This would also require independent validation in larger cohorts composed of different ethnic groups. Variation in the genomes is known to affect drug responses [7]. Further, allele frequencies of these genomic variations differ across the ethnicity in the world [8]. Hence, a drug which works for one ethnicity might not work for another due to genetic variation. For example, ethnic specific SNPs associated with pharmacogenetic cytochrome p450 genes are associated with drug response [9].
- Chapter 5 describes a recursive algorithm to search multiple signatures, which is comparatively less computationally intensive but still can be computationally expensive. For example, it can still take a lot of time if a lot of signature markers are present in the data. Hence, the algorithm is designed to output the signature as soon as found, so that a user can terminate at any point of time, as per their convenience. But this might result in outputting the same signatures by different recursive branches, which can be avoided with the use of hash table to maintain the calculated signatures.

6.5 Future perspectives and directions

Some of the specific improvements in the research presented in the thesis are as follows:

- In Chapter 3, we can further narrow down the patients to one of the major inflammatory conditions like rheumatoid arthritis and try to find any disease-disease interactions with concordant as well as non-concordant disease conditions. This will give further insights towards the role of multimorbidity towards various aspects of index disease conditions, like treatment response. Further, we have only used demographics and disease diagnosis for the analyses, however various other variables present in the UK Biobank such as genomics, imaging, etc., can also be included to identify more patterns, which can be relevant.

- In Chapter 4, we could identify multiple signatures with similar predicting performance, using the algorithm developed in Chapter 5. Further, we can also include genomic data or polygenic risk scores (PRS) to increase the performance of ATRPred predictor.
- We can further use the algorithm developed in Chapter 5 to predict multiple signatures with similar predicting power for different publicly available datasets as well as in-house PMC cohort.

6.6 Conclusions

This thesis presented significant work in the field of stratified healthcare and personalised medicine in different diseases using both supervised and unsupervised approaches in ML. We have tried to adapt the ML techniques to fit the best practices and performances accordingly. Finally, these methods can be extended to similar problem statements in different biological problems like other diseases and their treatment regimes. Some of the specific conclusions are:

- There exists five endotypes of patients in multimorbidity (Chapter 3).
- There exist two endotypes of RA patients that are dependent on gender and baseline DAS, however they are not different in anti-TNF therapy response. (Chapter 4).
- Proteomic biomarkers can be predictive of anti-TNF treatment response in RA patients (Chapter 4).
- Multiple signatures with similar predictive performance are possible in high dimensional dataset (Chapter 5).
- Unsupervised ML algorithms can identify endotypes for stratified healthcare (Chapter 3 and 4).
- Supervised ML algorithms can identify biomarkers for personalised medicine, that are predictive of a phenotype (Chapter 4 and 5).

6.7 References

1. MacEachern SJ, Forkert ND. Machine learning for precision medicine. *Genome*. 2021 Apr;64(4):416-425. doi: 10.1139/gen-2020-0131.

Chapter 6: General discussions

2. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv.* 2021 Jul-Aug;49:107739. doi: 10.1016/j.biotechadv.2021.107739.
3. Kauffmann HM, Kamp H, Fuchs R, Chorley BN, Deferme L, Ebbels T, Hackermüller J, Perdichizzi S, Poole A, Sauer UG, Tollefsen KE, Tralau T, Yauk C, van Ravenzwaay B. Framework for the quality assurance of 'omics technologies considering GLP requirements. *Regul Toxicol Pharmacol.* 2017 Dec;91 Suppl 1(Suppl 1):S27-S35. doi: 10.1016/j.yrtph.2017.10.007.
4. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. *Brief Bioinform.* 2008 Mar;9(2):119-28.
5. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40(5):373-83.
6. Stirland LE, González-Saavedra L, Mullin DS, Ritchie CW, Muniz-Terrera G, Russ TC. Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice. *Bmj.* 2020 Feb 18;368.
7. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. *New England Journal of Medicine.* 2011 Mar 24;364(12):1144-53.
8. Middleton D, Menchaca L, Rood H, Komerofsky R. New allele frequency database: <http://www.allelefrequencies.net>. *Tissue antigens.* 2003 May;61(5):403-7.
9. Ingelman-Sundberg M. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends in pharmacological sciences.* 2004 Apr 1;25(4):193-200.

Appendix I Supplementary data for Chapter 3**Supplementary Table 3.1 Distribution of number of diseases vs. number of patients with multimorbidity in UK Biobank.** Multimorbidity implies occurrence of 2 or more diseases.

No. of diseases	No. of patients
2	44,778
3	18,501
4	8047
5	3647
6	1596
7	675
8	213
9	51
10	15
13	1
Total	77,524

Supplementary Table 3.2 Demographics of multimorbid patient's cohort in UK Biobank. Cohort is further presented based on cluster affiliation (rows) as per Figure 3.2, training-test split (columns) used in cluster validation as per Figure 3.1B, and overall cohort characteristics (last row). P = p-value of either 2-sample proportion test or 2-sample t-test, as appropriate.

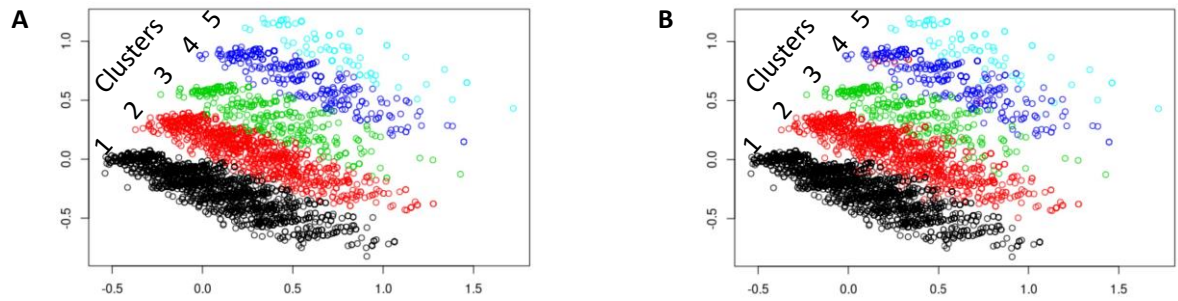
Clusters	Proportion of Females, n (%)				Age at death, mean (s.d.), years				CCI, mean (s.d.)				IMD – England, mean (s.d.), years			
	Train	Test	Total	P	Train	Test	Total	P	Train	Test	Total	P	Train	Test	Total	P
1	20545	5140	25685	0.51	69.9	70.0	69.9	0.59	4.0 (2.2)	4.0 (2.2)	4.0 (2.2)	0.88	20.8	20.9	20.8	0.67

Appendix I

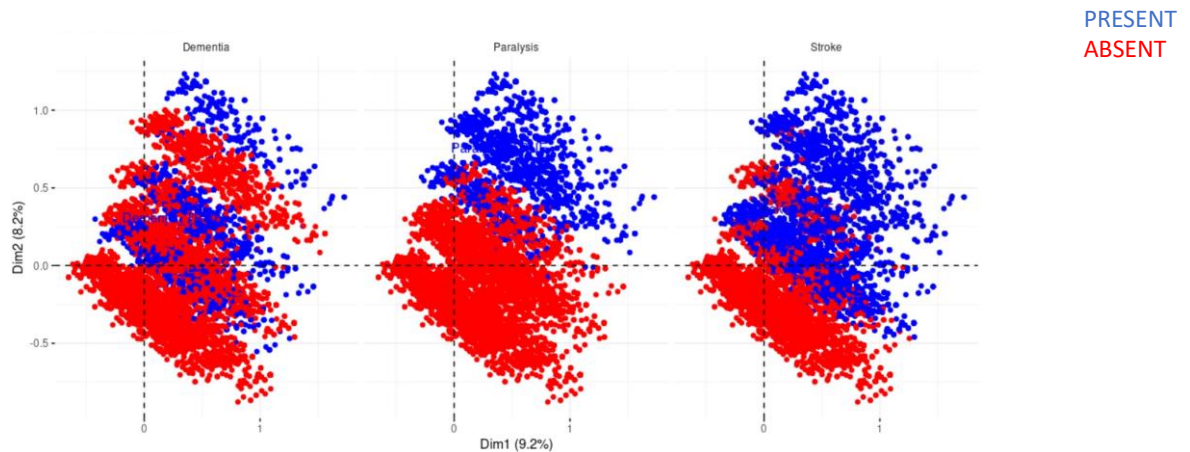
	(45. 3)	(45. 7)	(45. 4)		(6. 9)	(6. 7)	(6. 9)						(15. 8)	(15. 7)	(15. 8)	
2	450 1 (42. 3)	114 4 (41. 8)	564 5 (42. 2)	0. 64	72. 5 (6. 4)	72. 2 (6. 2)	72. 2 (6. 3)	0. 33	4.1 (2. 3)	4.1 (2. 3)	4.1 (2. 3)	0. 78	21. 0 (15. 8)	21. 1 (16. 2)	21. 0 (15. 9)	0. 77
3	144 6 (44. 0)	318 (40. 7)	176 4 (43. 3)	0. 10	71. 9 (6. 9)	71. 6 (6. 8)	71. 9 (6. 9)	0. 43	4.9 (2. 5)	4.8 (2. 5)	4.9 (2. 5)	0. 34	20. 3 (15. 6)	20. 9 (16. 2)	20. 4 (15. 7)	0. 44
4	101 4 (41. 1)	271 (42. 1)	128 5 (41. 3)	0. 68	70. 8 (7. 2)	71. 5 (7. 1)	71. 0 (7. 2)	0. 26	5.5 (2. 6)	5.5 (2. 5)	5.5 (2. 6)	0. 64	22. 0 (16. 3)	21. 5 (16. 3)	21. 9 (16. 3)	0. 55
5	116 (38. 0)	32 (36. 0)	148 (37. 6)	0. 82	74. 9 (4. 6)	74. 3 (4. 8)	74. 7 (4. 6)	0. 52	6.8 (2. 5)	6.9 (2. 3)	6.9 (2. 4)	0. 72	23. 2 (15. 6)	22. 7 (17. 5)	23. 1 (16. 1)	0. 79
Over all	276 22 (44. 5)	690 5 (44. 5)	345 27 (44. 5)	1. 00	70. 7 (6. 9)	70. 7 (6. 7)	70. 7 (6. 9)	0. 95	4.2 (2. 3)	4.2 (2. 2)	4.2 (2. 3)	0. 87	20. 9 (15. 8)	21. 0 (15. 8)	20. 9 (15. 8)	0. 59

Supplementary Table 3.3 Confusion matrix for cluster validation on 20% model-blind test set. Each cell represents number of patients.

Cluster No.	Predicted				
	1	2	3	4	5
Actual	1	11235	14	0	0
	2	97	2639	4	0
	3	0	305	473	4
	4	0	10	5	625
	5	0	0	0	15



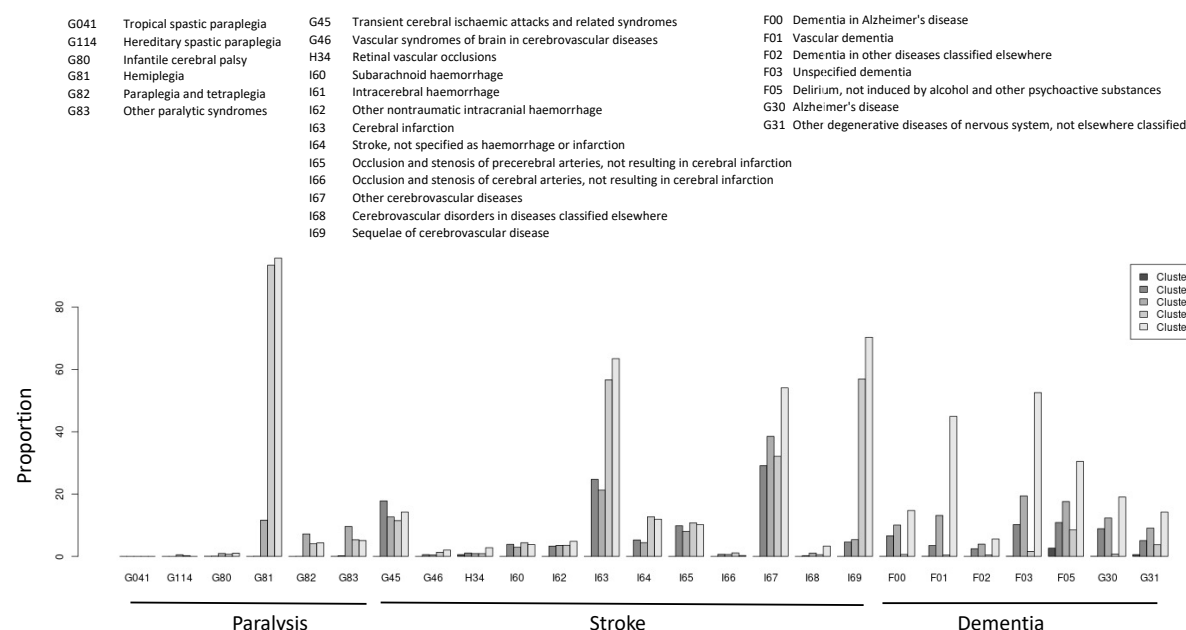
Supplementary Figure 3.1 (A) The MCA plot of 20% test set, coloured by the original cluster numbers. (B) The MCA plot of 20% test set, coloured by the cluster numbers obtained from the DTC model. MCA = multiple correspondence analysis and DTC = decision tree classifier.



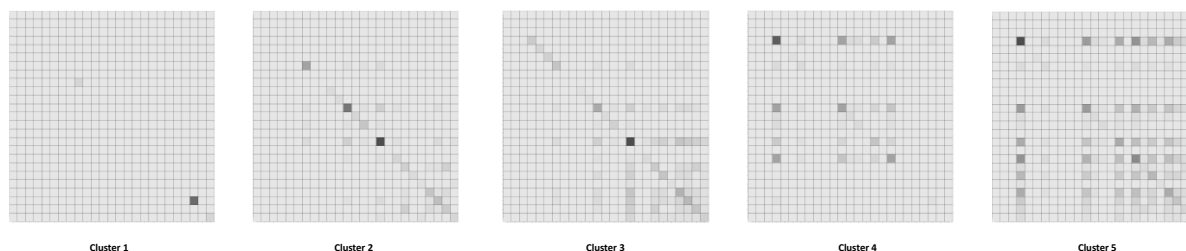
Supplementary Figure 3.2 Multimorbid clusters: MCA for Charlson Comorbidity disease class of 77,524 patients with multimorbidity in UK Biobank. **MCA plot showing 5 different multimorbid clusters labelled for the presence or absence of dementia, stroke, and paralysis.** MCA = multiple correspondence analysis and DTC = decision tree classifier.

Appendix I

A

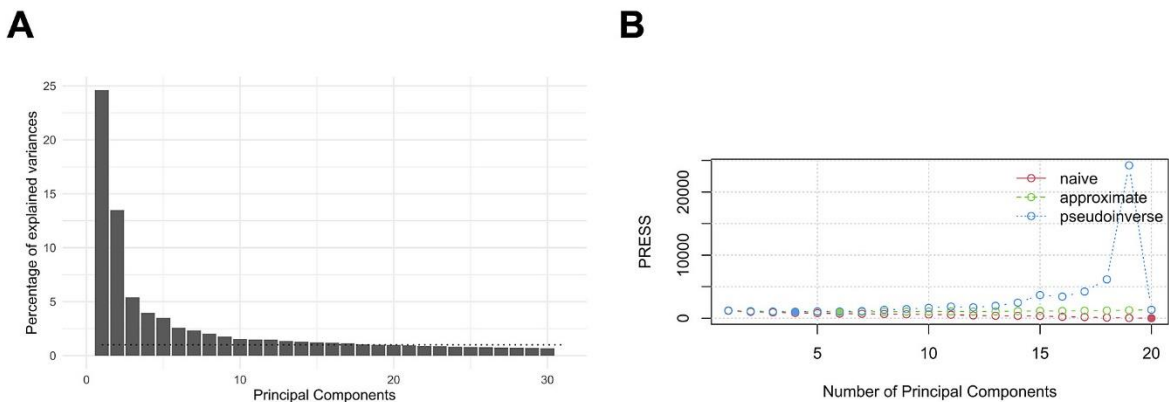


B

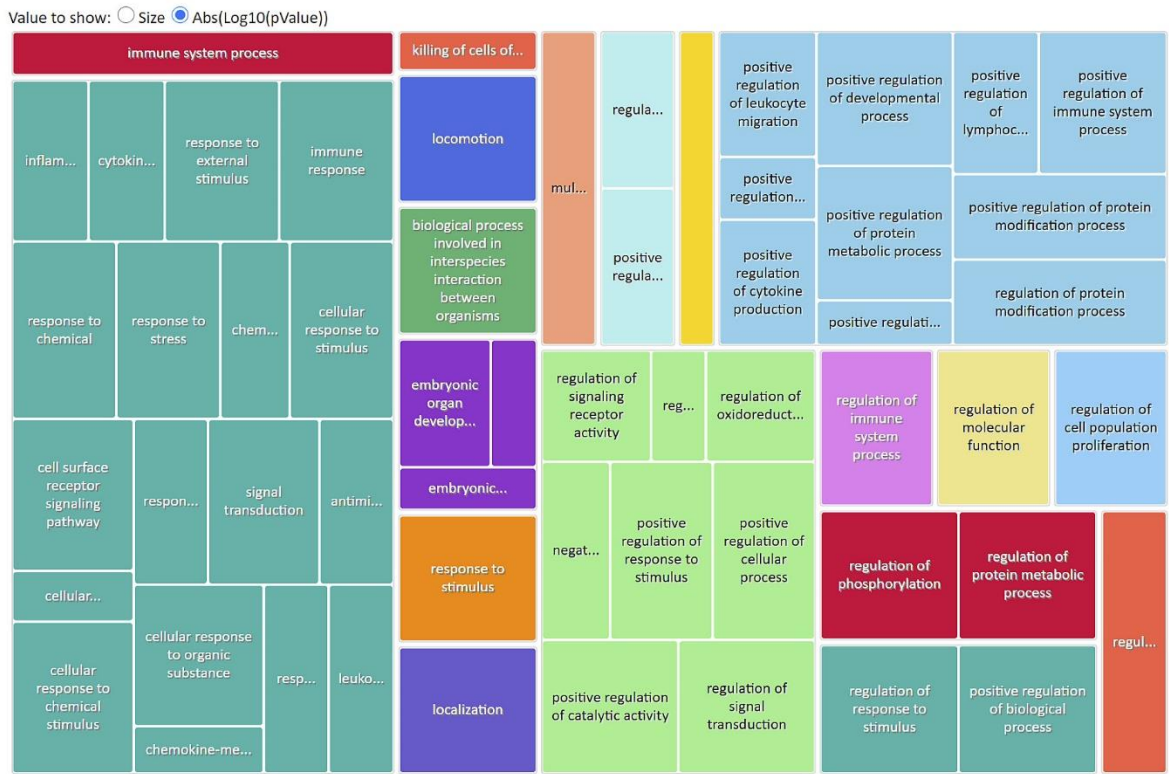


Supplementary Figure 3.3 (A) Prevalence of sub-classes of paralysis, stroke, and dementia. **(B)** Heatmap showing the co-occurrence of sub-classes of paralysis, stroke, and dementia. Order of disease from top to bottom and left to right are: G041, G114, G80, G81, G82, G83, G45, G46, H34, I60, I62, I63, I64, I65, I66, I67, I68, I69, F00, F01, F02, F03, F05, G30, and G31. Darker shade represents higher co-occurrence.

Appendix II Supplementary data for Chapter 4



Supplementary Figure 4.1 (A) Elbow plot for first 30 Principal Components (PCs). Dotted line represents the cut-off of 1% explained variance, crossing between PC 19 and 20. (B) Predicted sum of squares (PRESS) vs. number of PCs for first 20 PCs. Solid dot represents minimum value of PRESS.



Appendix II

Supplementary Figure 4.2 TreeMap summary view of significant Gene Ontology (GO) Biological Process (BP) terms for the 17 featured protein set. Size of each rectangle represents \log_{10} p-value of the GO terms.

Supplementary Table 4.1 The ML classifier performance with 5-fold nested cross-validation and the inclusion of protein features one-by-one with decreasing feature importance along with baseline DAS and gender information. The best model performance with 17 protein features along with baseline DAS and gender information is highlighted in grey.

No. of proteins in the feature set	5-fold AUC		Mean		5-fold Accuracy		Mean		5-fold Sensitivity		Mean		5-fold Specificity		Mean		5-fold MCC		Mean	
	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>
0	0.59	0.57	0.61	0.5	0.51	0.36	0.65	0.59	0.17	-0.05										
1	0.68	0.58	0.67	0.54	0.66	0.48	0.68	0.58	0.32	0.06										
2	0.73	0.73	0.71	0.66	0.68	0.67	0.73	0.71	0.4	0.32										
3	0.83	0.77	0.79	0.68	0.81	0.69	0.78	0.69	0.56	0.35										
4	0.85	0.77	0.81	0.66	0.78	0.7	0.82	0.64	0.6	0.32										
5	0.87	0.77	0.82	0.7	0.81	0.67	0.82	0.72	0.62	0.4										
6	0.88	0.79	0.84	0.68	0.79	0.65	0.86	0.72	0.64	0.39										
7	0.89	0.86	0.88	0.75	0.77	0.59	0.94	0.85	0.73	0.51										
8	0.91	0.78	0.84	0.71	0.92	0.69	0.81	0.7	0.69	0.38										
9	0.91	0.8	0.86	0.71	0.84	0.63	0.87	0.76	0.7	0.39										
10	0.84	0.74	0.79	0.69	0.79	0.61	0.8	0.71	0.57	0.33										
11	0.93	0.79	0.88	0.73	0.86	0.62	0.89	0.8	0.75	0.43										
12	0.94	0.79	0.87	0.72	0.9	0.64	0.86	0.77	0.74	0.4										
13	0.93	0.73	0.88	0.7	0.89	0.57	0.87	0.76	0.74	0.32										
14	0.95	0.72	0.9	0.77	0.93	0.69	0.88	0.81	0.79	0.47										
15	0.95	0.75	0.89	0.67	0.93	0.66	0.86	0.68	0.77	0.31										

16	0.96	0.73	0.9	0.65	0.9	0.51	0.9	0.72	0.79	0.22
17	0.99	0.86	0.95	0.81	0.98	0.75	0.93	0.86	0.89	0.6
18	0.99	0.81	0.97	0.75	0.98	0.66	0.96	0.8	0.93	0.44
19	0.98	0.79	0.95	0.73	0.97	0.72	0.94	0.74	0.89	0.43
20	1	0.86	0.98	0.78	0.99	0.65	0.97	0.83	0.95	0.51
21	1	0.8	0.98	0.75	0.98	0.59	0.97	0.85	0.95	0.44
22	0.99	0.81	0.97	0.75	0.97	0.58	0.96	0.86	0.93	0.47
23	0.99	0.79	0.96	0.75	0.99	0.59	0.95	0.84	0.92	0.43
24	1	0.83	1	0.78	1	0.53	1	0.84	0.99	0.45
25	1	0.78	0.99	0.72	0.97	0.58	1	0.82	0.98	0.36
26	0.99	0.86	0.98	0.74	0.98	0.5	0.98	0.85	0.96	0.37
27	1	0.79	0.99	0.71	0.99	0.52	1	0.8	0.99	0.34
28	1	0.77	1	0.73	1	0.43	1	0.86	1	0.34
29	0.99	0.73	0.98	0.67	0.97	0.47	0.99	0.77	0.96	0.23
30	1	0.74	0.99	0.71	1	0.57	0.99	0.8	0.98	0.38

Supplementary Table 4.2 Enrichment analysis of Gene Ontology terms (Biological Process).

GO term ID	Term description	Observed gene count	Background gene count	Percentage	False discovery rate	Matching proteins in your network (IDs)	Matching proteins in your network (labels)
GO:0006954	inflammatory response	6	482	1.24%	0.0024	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000378118,ENSP00000379110,ENSP00000418009	CCL8,CXCL1,IL13,PRKCQ,RARRES2,TNFRSF10B

Appendix II

GO:0002684	positive regulation of immune system process	7	882	0.79%	0.0027	ENSP00000263125,ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000418009	ARNT,CCL8,CXCL1,IL13,PRKCQ,RARRES2,TNFSF13B
GO:0002682	regulation of immune system process	8	1391	0.58%	0.0032	ENSP00000263125,ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000418009,ENSP0000479089	ARNT,CCL8,CXCL1,IL13,OSCAR,PRKCQ,RARRES2,TNFSF13B
GO:0019221	cytokine-mediated signaling pathway	6	655	0.92%	0.0034	ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000365048,ENSP00000378118,ENSP00000379110	CCL8,CXCL1,IL13,MMP1,TNFRSF10B,TNFSF13B
GO:0007166	cell surface receptor signaling pathway	9	2198	0.41%	0.0037	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000355124,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP0000479089	CCL8,CXCL1,IL13,KRT19,MMP1,OSCAR,PRKCQ,TNFRSF10B,TNFSF13B
GO:0040011	locomotion	7	1144	0.61%	0.0037	ENSP00000263125,ENSP00000276431,ENSP00000322788,ENSP00000378118,ENSP00000379110,ENSP0000409007,ENSP00000418009	CCL8,CXCL1,GDNF,MMP1,PRKCQ,RARRES2,TNFRSF10B
GO:0042127	regulation of cell population proliferation	8	1594	0.50%	0.0037	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP0000409007	ARNT,CCL8,CXCL1,GDNF,IL13,PRKCQ,TNFRSF10B,TNFSF13B

Appendix II

GO:0048584	positive regulation of response to stimulus	9	2054	0.44%	0.0037	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP0000409007,ENSP0000418009	ARNT,CCL8,CXCL1,GDNF,IL13,PRKCQ,RARRES2,TNFRSF10B,TNFSF13B
GO:0070663	regulation of leukocyte proliferation	4	213	1.88%	0.0037	ENSP00000263125,ENSP00000304915,ENSP00000365048,ENSP00000378118	CCL8,IL13,PRKCQ,TNFSF13B
GO:0070887	cellular response to chemical stimulus	10	2672	0.37%	0.0037	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP0000409007	ARNT,CCL8,CXCL1,GDNF,IL13,MMP1,PRKCQ,REN,TNFRSF10B,TNFSF13B
GO:0016477	cell migration	6	812	0.74%	0.0042	ENSP00000263125,ENSP00000276431,ENSP00000322788,ENSP00000378118,ENSP00000379110,ENSP0000409007	CCL8,CXCL1,GDNF,MMP1,PRKCQ,TNFRSF10B
GO:0002376	immune system process	9	2370	0.38%	0.0043	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP0000418009,ENSP0000479089	CCL8,CXCL1,IL13,MMP1,OSCAR,PRKCQ,RARRES2,TNFRSF10B,TNFSF13B
GO:0006935	chemotaxis	5	491	1.02%	0.0043	ENSP00000263125,ENSP00000378118,ENSP00000379110,ENSP0000409007,ENSP0000418009	CCL8,CXCL1,GDNF,PRKCQ,RARRES2

Appendix II

GO:0009605	response to external stimulus	8	1857	0.43%	0.0043	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP0000418009	CCL8,CXCL1,GDNF,IL13,PRKCQR,ARRES2,REN,TNFRSF10B
GO:0048583	regulation of response to stimulus	11	3882	0.28%	0.0043	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP0000379110,ENSP0000409007,ENSP0000418009,ENSP0000479089	ARNT,CCL8,CXCL1,GDNF,IL13,OSCAR,PRKCQR,ARRES2,REN,TNFRSF10B,TNFSF13B
GO:0051707	response to other organism	6	835	0.72%	0.0043	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000378118,ENSP00000379110,ENSP0000418009	CCL8,CXCL1,IL13,ARRES2,REN,TNFRSF10B
GO:0009617	response to bacterium	5	555	0.90%	0.0044	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000379110,ENSP00000418009	CXCL1,IL13,ARRES2,REN,TNFRSF10B
GO:0010469	regulation of signaling receptor activity	5	577	0.87%	0.0048	ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000409007	CCL8,CXCL1,GDNF,IL13,TNFSF13B
GO:0032496	response to lipopolysaccharide	4	298	1.34%	0.0048	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000379110	CXCL1,IL13,REN,TNFRSF10B

Appendix II

GO:0042221	response to chemical	11	4153	0.26%	0.0055	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP00000418009	ARNT,CCL8,CXCL1,GDNF,IL13,MMP1,PRKCQ,RARRES2,REN,TNFRSF10B,TNFSF13B
GO:0002687	positive regulation of leukocyte migration	3	127	2.36%	0.0066	ENSP00000378118,ENSP00000379110,ENSP00000418009	CCL8,CXCL1,RARRES2
GO:0006955	immune response	7	1560	0.45%	0.0069	ENSP00000276431,ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000418009,ENSP00000479089	CCL8,CXCL1,IL13,OSCAR,RARRES2,TNFRSF10B,TNFSF13B
GO:0050671	positive regulation of lymphocyte proliferation	3	130	2.31%	0.0069	ENSP00000263125,ENSP00000304915,ENSP00000365048	IL13,PRKCQ,TNFSF13B
GO:0019730	antimicrobial humoral response	3	143	2.10%	0.008	ENSP00000378118,ENSP00000379110,ENSP00000418009	CCL8,CXCL1,RARRES2
GO:0051704	multi-organism process	8	2222	0.36%	0.008	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP0000032788,ENSP00000355124,ENSP00000378118,ENSP00000379110,ENSP00000418009	CCL8,CXCL1,IL13,KRT19,MMP1,RARRES2,REN,TNFRSF10B

Appendix II

GO:0009966	regulation of signal transduction	9	3033	0.30%	0.0113	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP0000379110,ENSP0000409007	ARNT,CCL8,CXCL1,GDNF,IL13,PRKCQ,REN,TNFRSF10B,TNFSF13B
GO:0048568	embryonic organ development	4	417	0.96%	0.0115	ENSP00000351407,ENSP00000355124,ENSP00000409007,ENSP00000418009	ARNT,GDNF,KRT19,RARRES2
GO:0007165	signal transduction	11	4738	0.23%	0.0123	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000355124,ENSP00000365048,ENSP00000378118,ENSP0000379110,ENSP0000409007,ENSP0000479089	CCL8,CXCL1,GDNF,IL13,KRT19,MMP1,OSCAR,PRKCQ,REN,TNFRSF10B,TNFSF13B
GO:0050896	response to stimulus	14	7824	0.18%	0.0123	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000355124,ENSP00000365048,ENSP00000368066,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP00000418009,ENSP00000479089	ARNT,CCL8,CXCL1,GDNF,HAO1,IL13,KRT19,MMP1,OSCAR,PRKCQ,RARRES2,REN,TNFRSF10B,TNFSF13B
GO:0060326	cell chemotaxis	3	183	1.64%	0.0131	ENSP00000263125,ENSP00000378118,ENSP00000379110	CCL8,CXCL1,PRKCQ
GO:0030890	positive regulation of B cell proliferation	2	42	4.76%	0.0158	ENSP00000304915,ENSP00000365048	IL13,TNFSF13B

Appendix II

GO:0071356	cellular response to tumor necrosis factor	3	197	1.52%	0.0158	ENSP00000276431,ENSP00000365048,ENSP00000378118	CCL8,TNFRSF10B,TNFSF13B
GO:006950	response to stress	9	3267	0.28%	0.016	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP000003051407,ENSP00000368066,ENSP00000378118,ENSP00000379110,ENSP0000418009	ARNT,CCL8,CXCL1,HAO1,IL13,PRKCQ,RARRES2,REN,TNFRSF10B
GO:008284	positive regulation of cell population proliferation	5	878	0.57%	0.0163	ENSP00000263125,ENSP00000304915,ENSP00000351407,ENSP0000035048,ENSP00000409007	ARNT,GDNF,IL13,PRKCQ,TNFSF13B
GO:0071622	regulation of granulocyte chemotaxis	2	47	4.26%	0.0169	ENSP00000379110,ENSP00000418009	CXCL1,RARRES2
GO:0032940	secretion by cell	5	959	0.52%	0.0206	ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000418009,ENSP00000479089	CCL8,CXCL1,OSCAR,RARRES2,TNFSF13B
GO:0042325	regulation of phosphorylation	6	1465	0.41%	0.0206	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000378118,ENSP00000418009	ARNT,CCL8,IL13,RARRES2,REN,TNFRSF10B
GO:0051179	localization	11	5233	0.21%	0.0213	ENSP00000263125,ENSP00000276431,ENSP00000322788,ENSP00000365048,ENSP00000368066,ENSP00000376855,ENSP00000378118,ENSP00000379110,ENSP0000409007,ENSP0000418009,ENSP0000479089	CCL8,CXCL1,DPPI0,GDNF,HAO1,MMP1,OSCAR,PRKCQ,RARRES2,TNFRSF10B,TNFSF13B

Appendix II

GO:0051716	cellular response to stimulus	12	6212	0.19%	0.0213	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000355124,ENSP00000365048,ENSP0000378118,ENSP0000379110,ENSP0000409007,ENSP00000479089	ARNT,CCL8,CXCL1,GDNF,IL13,KRT19,MMP1,OSCAR,PRKCQ,REN, TNFRSF10B,TNFSF13B
GO:0042327	positive regulation of phosphorylation	5	984	0.51%	0.0214	ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000378118,ENSP00000418009	ARNT,CCL8,IL13, RARRES2, TNFRSF10B
GO:0030858	positive regulation of epithelial cell differentiation	2	59	3.39%	0.0219	ENSP00000304915,ENSP00000409007	GDNF,IL13
GO:0048518	positive regulation of biological process	11	5459	0.20%	0.0264	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000365048,ENSP00000376855,ENSP0000378118,ENSP0000379110,ENSP0000409007,ENSP00000418009	ARNT,CCL8,CXCL1,DPP10,GDNF, IL13,MMP1,PRKCQ,RARRES2,TNFRSF10B,TNFSF13B
GO:0051247	positive regulation of protein metabolic process	6	1587	0.38%	0.0264	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000378118,ENSP00000418009	ARNT,CCL8,IL13,PRKCQ,RARRES2,TNFRSF10B
GO:0071310	cellular response to organic substance	7	2219	0.32%	0.0264	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000365048,ENSP00000376855,ENSP0000378118,ENSP0000379110,ENSP0000409007,ENSP00000418009	CCL8,CXCL1,IL13,MMP1,REN,TNFRSF10B,TNFSF13B

378118,ENSP0000
0379110

GO:19 00076	regulation of cellular response to insulin stimulus	2	72	2.78%	0.027 6	ENSP0000026312 5,ENSP000004180 09	PRKCQ,RARRES 2
GO:00 50900	leukocyte migration	3	296	1.01%	0.029 4	ENSP0000027643 1,ENSP000003227 88,ENSP00000378 118	CCL8,MMP1,TNF RSF10B
GO:00 70098	chemokine -mediated signaling pathway	2	75	2.67%	0.029 4	ENSP0000037811 8,ENSP000003791 10	CCL8,CXCL1
GO:00 71260	cellular response to mechanical stimulus	2	78	2.56%	0.029 7	ENSP0000027643 1,ENSP000003049 15	IL13,TNFRSF10B
GO:00 01701	in utero embryonic developme nt	3	306	0.98%	0.030 5	ENSP0000035140 7,ENSP000003551 24,ENSP00000418 009	ARNT,KRT19,RA RRES2
GO:00 31401	positive regulation of protein modificati on process	5	1149	0.44%	0.030 5	ENSP0000027643 1,ENSP000003049 15,ENSP00000351 407,ENSP0000037 8118,ENSP000004 18009	ARNT,CCL8,IL13 ,RARRES2,TNFR SF10B
GO:00 33209	tumor necrosis factor- mediated signaling pathway	2	81	2.47%	0.030 5	ENSP0000027643 1,ENSP000003650 48	TNFRSF10B,TNF SF13B
GO:00 01892	embryonic placenta developme nt	2	86	2.33%	0.032 4	ENSP0000035140 7,ENSP000003551 24	ARNT,KRT19

Appendix II

GO:0031399	regulation of protein modification process	6	1747	0.34%	0.0327	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000378118,ENSP00000418009	ARNT,CCL8,IL13,RARRES2,REN,TNFRSF10B
GO:0002824	positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2	89	2.25%	0.0338	ENSP00000263125,ENSP00000365048	PRKCQ,TNFSF13B
GO:0031640	killing of cells of other organism	2	89	2.25%	0.0338	ENSP00000378118,ENSP00000379110	CCL8,CXCL1
GO:0001823	mesonephros development	2	91	2.20%	0.0341	ENSP00000272190,ENSP00000409007	GDNF,REN
GO:0002690	positive regulation of leukocyte chemotaxis	2	91	2.20%	0.0341	ENSP00000379110,ENSP00000418009	CXCL1,RARRES2
GO:0042102	positive regulation of T cell proliferation	2	92	2.17%	0.0341	ENSP00000263125,ENSP00000365048	PRKCQ,TNFSF13B
GO:0044419	interspecies interaction between organisms	4	724	0.55%	0.0344	ENSP00000322788,ENSP00000355124,ENSP00000378118,ENSP00000379110	CCL8,CXCL1,KRT19,MMP1

Appendix II

GO:0051341	regulation of oxidoreductase activity	2	93	2.15%	0.0344	ENSP00000304915,ENSP0000040907	GDNF,IL13
GO:0048522	positive regulation of cellular process	10	4898	0.20%	0.0348	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP00000418009	ARNT,CCL8,CXCL1,GDNF,IL13,MP1,PRKCQ,RARRES2,TNFRSF10B,TNFSF13B
GO:0032879	regulation of localization	7	2524	0.28%	0.0389	ENSP00000272190,ENSP00000304915,ENSP00000376855,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP00000418009	CCL8,CXCL1,DP10,GDNF,IL13,RARRES2,REN
GO:00101237	negative regulation of extrinsic apoptotic signaling pathway	2	104	1.92%	0.0405	ENSP00000276431,ENSP0000040907	GDNF,TNFRSF10B
GO:0006887	exocytosis	4	774	0.52%	0.0416	ENSP00000378118,ENSP00000379110,ENSP00000418009,ENSP00000479089	CCL8,CXCL1,OSCAR,RARRES2
GO:0051094	positive regulation of developmental process	5	1286	0.39%	0.0416	ENSP00000304915,ENSP00000351407,ENSP00000365048,ENSP00000409007,ENSP00000418009	ARNT,GDNF,IL13,RARRES2,TNFSF13B
GO:0061844	antimicrobial humoral immune response mediated by antimicrobial peptide	2	107	1.87%	0.0416	ENSP00000378118,ENSP00000379110	CCL8,CXCL1

Appendix II

GO:0065009	regulation of molecular function	8	3322	0.24%	0.0416	ENSP00000263125,ENSP00000276431,ENSP00000304915,ENSP00000365048,ENSP00000376855,ENSP00000378118,ENSP00000379110,ENSP00000409007	CCL8,CXCL1,DP P10,GDNF,IL13,PRKCQ,TNFRSF10B,TNFSF13B
GO:0001819	positive regulation of cytokine production	3	390	0.77%	0.0451	ENSP00000263125,ENSP00000304915,ENSP00000351407	ARNT,IL13,PRKCQ
GO:0048608	reproductive structure development	3	405	0.74%	0.0483	ENSP00000272190,ENSP00000351407,ENSP00000355124	ARNT,KRT19,REN
GO:0051246	regulation of protein metabolic process	7	2668	0.26%	0.0483	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000378118,ENSP00000418009	ARNT,CCL8,IL13,PRKCQ,RARRES2,REN,TNFRSF10B
GO:0001932	regulation of protein phosphorylation	5	1370	0.36%	0.0486	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000378118,ENSP00000418009	CCL8,IL13,RARRES2,REN,TNFRSF10B
GO:0043085	positive regulation of catalytic activity	5	1381	0.36%	0.0499	ENSP00000263125,ENSP00000276431,ENSP00000378118,ENSP0000039110,ENSP00000409007	CCL8,CXCL1,GDNF,PRKCQ,TNFRSF10B

Supplementary Table 4.3 REVIGO summary analysis of Gene Ontology terms (Biological Process).

GO term ID	Description	Frequency in database	log ₁₀ pvalue	Uniqueness	Dispensability	Representative
------------	-------------	-----------------------	--------------------------	------------	----------------	----------------

Appendix II

GO:0002376	immune system process	0.60%	-2.3665	0.972	0	immune system process
GO:0002682	regulation of immune system process	0.25%	-2.4949	0.6	0	regulation of immune system process
GO:0042127	regulation of cell proliferation	0.31%	-2.4318	0.654	0.225	regulation of immune system process
GO:0048518	positive regulation of biological process	1.74%	-1.5784	0.714	0.281	regulation of immune system process
GO:0043085	positive regulation of catalytic activity	0.82%	-1.3019	0.719	0.611	regulation of immune system process
GO:0031640	killing of cells of other organism	0.02%	-1.4711	0.778	0.173	regulation of immune system process
GO:0051341	regulation of oxidoreductase activity	0.02%	-1.4634	0.755	0.269	regulation of immune system process
GO:0044419	interspecies interaction between organisms	0.26%	-1.4634	0.953	0.669	regulation of immune system process
GO:0051246	regulation of protein metabolic process	1.55%	-1.3161	0.698	0.393	regulation of immune system process
GO:0042325	regulation of phosphorylation	0.47%	-1.6861	0.686	0.243	regulation of immune system process
GO:0065009	regulation of molecular function	1.73%	-1.3809	0.725	0.311	regulation of immune system process
GO:0032879	regulation of localization	0.73%	-1.4101	0.702	0.295	regulation of immune system process
GO:0006954	inflammatory response	0.11%	-2.6198	0.752	0	inflammatory response

Appendix II

GO:0070887	cellular response to chemical stimulus	1.01%	-2.4318	0.625	0.457	inflammatory response
GO:2001237	negative regulation of extrinsic apoptotic signaling pathway	0.02%	-1.3925	0.562	0.638	inflammatory response
GO:0051707	response to other organism	0.30%	-2.3665	0.651	0.352	inflammatory response
GO:0071260	cellular response to mechanical stimulus	0.01%	-1.5272	0.707	0.566	inflammatory response
GO:0001819	positive regulation of cytokine production	0.07%	-1.3458	0.568	0.629	inflammatory response
GO:0009605	response to external stimulus	1.37%	-2.3665	0.702	0.42	inflammatory response
GO:0010469	regulation of receptor activity	0.03%	-2.3188	0.563	0.649	inflammatory response
GO:0048584	positive regulation of response to stimulus	0.46%	-2.4318	0.42	0.367	inflammatory response
GO:0048583	regulation of response to stimulus	1.12%	-2.3665	0.561	0.402	inflammatory response
GO:0006950	response to stress	4.58%	-1.7959	0.671	0.562	inflammatory response
GO:0007166	cell surface receptor signaling pathway	0.92%	-2.4318	0.526	0.321	inflammatory response
GO:0007165	signal transduction	6.62%	-1.9101	0.448	0.679	inflammatory response
GO:0042221	response to chemical	3.07%	-2.2596	0.682	0.475	inflammatory response

GO:1900076	regulation of cellular response to insulin stimulus	0.01%	-1.5591	0.546	0.631	inflammatory response
GO:0030858	positive regulation of epithelial cell differentiation	0.01%	-1.6596	0.592	0.549	inflammatory response
GO:0019221	cytokine-mediated signaling pathway	0.09%	-2.4685	0.499	0.583	inflammatory response
GO:0016477	cell migration	0.29%	-2.3768	0.781	0	cell migration
GO:0032940	secretion by cell	0.76%	-1.6861	0.845	0.257	cell migration
GO:0040011	locomotion	1.00%	-2.4318	0.972	0	locomotion
GO:0050896	response to stimulus	12.21%	-1.9101	0.975	0	response to stimulus
GO:0051179	localization	18.50%	-1.6716	0.977	0	localization
GO:0051704	multi-organism process	0.75%	-2.0969	0.972	0	multi-organism process
GO:0048568	embryonic organ development	0.11%	-1.9393	0.816	0.065	embryonic organ development
GO:0001823	mesonephros development	0.02%	-1.4672	0.839	0.631	embryonic organ development

Supplementary Table 4.4 Enrichment analysis of Gene Ontology terms (Molecular Function).

GO term ID	Term description	Observed gene count	Background gene count	Percentage	False discovery rate	Matching proteins in your network (IDs)	Matching proteins in your network (labels)
GO:0005102	signaling receptor binding	10	1513	0.66%	1.14E-05	ENSP00000272190, ENSP00000304915, ENSP00000351407, ENSP00000357086, ENSP000003	ARNT, CCL8, CXCL1, FCRL6, GDNF, HA O1, IL13, RARRES2, REN, TNFSF13B

Appendix II

						65048,ENSP00000368066,ENSP00000378118,ENSP0000379110,ENSP0000409007,ENSP0000418009	
GO:0005125	cytokine activity	4	216	1.85%	0.002	ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110	CCL8,CXCL1,IL13,TNFSF13B
GO:0005126	cytokine receptor binding	4	272	1.47%	0.002	ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110	CCL8,CXCL1,IL13,TNFSF13B
GO:0048018	receptor ligand activity	5	458	1.09%	0.002	ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000409007	CCL8,CXCL1,GDNF,IL13,TNFSF13B
GO:0008009	chemokine activity	2	48	4.17%	0.0175	ENSP00000378118,ENSP00000379110	CCL8,CXCL1
GO:0005515	protein binding	12	6605	0.18%	0.0331	ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000351407,ENSP00000357086,ENSP00000365048,ENSP00000368066,ENSP00000376855,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP00000418009	ARNT,CCL8,CXCL1,DPP10,FCRL6,GDNF,HAO1,IL13,RARRES2,REN,TNFRSF10B,TNFSF13B
GO:0005488	binding	16	11878	0.13%	0.0343	ENSP00000263125,ENSP00000272190,ENSP00000276431,ENSP00000304915,ENSP00000322788,ENSP00000351407,ENSP00000355124,ENSP00000357086,ENSP00000365048,ENSP00000368066,ENSP00000376855,ENSP00000378118,ENSP00000379110,E	ARNT,CCL8,CXCL1,DPP10,FCRL6,GDNF,HAO1,IL13,KRT19,MMP1,PRKCQ,RARRES2,REN,SPPON1,TNFRSF10B,TNFSF13B

						NSP00000409007, ENSP0000041800 9,ENSP000004602 36	
GO:009 8772	molecular function regulator	6	1793	0.33%	0.037 6	ENSP0000030491 5,ENSP000003650 48,ENSP00000376 855,ENSP0000037 8118,ENSP000003 79110,ENSP00000 409007	CCL8,CXCL1,DPP 10,GDNF,IL13,TNF SF13B

Supplementary Table 4.5 Enrichment analysis of Gene Ontology terms (Cellular Component).

GO term ID	Term description	Observed gene count	Background gene count	Percentage	False discovery rate	Matching proteins in your network (IDs)	Matching proteins in your network (labels)
GO:0005576	extracellular region	9	2505	0.36%	0.0095	ENSP00000272190,ENSP00000304915,ENSP00000322788,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000409007,ENSP0000418009,ENSP0000479089	CCL8,CXCL1,GDNF,IL13,MMP1,OSCAR,RARRES2,REN,TNFSF13B
GO:0044421	extracellular region part	7	1375	0.51%	0.0095	ENSP00000272190,ENSP00000304915,ENSP00000322788,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000418009	CCL8,CXCL1,IL13,MMP1,RARRES2,REN,TNFSF13B
GO:0005615	extracellular space	6	1134	0.53%	0.0101	ENSP00000272190,ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110,ENSP00000418009	CCL8,CXCL1,IL13,RARRES2,REN,TNFSF13B
GO:1904724	tertiary granule lumen	2	55	3.64%	0.0305	ENSP00000379110,ENSP00000479089	CXCL1,OSCAR

Appendix II

GO:0035580	specific granule lumen	2	62	3.23%	0.0308	ENSP00000379110,ENSP00000479089	CXCL1,OSCAR
GO:0034774	secretory granule lumen	3	323	0.93%	0.0484	ENSP00000379110,ENSP00000418009,ENSP00000479089	CXCL1,OSCAR,ARRES2

Supplementary Table 4.6 Enrichment analysis of KEGG Pathways.

GO term ID	Term description	Observed gene count	Background gene count	Percentage	False discovery rate	Matching proteins in your network (IDs)	Matching proteins in your network (labels)
hsa04060	Cytokine-cytokine receptor interaction	5	263	1.90%	0.00012	ENSP00000276431,ENSP00000304915,ENSP00000365048,ENSP00000378118,ENSP00000379110	CCL8,CXCL1,IL13,TNFRSF10B,TNFSF13B
hsa04657	IL-17 signaling pathway	3	92	3.26%	0.0013	ENSP00000304915,ENSP00000322788,ENSP00000379110	CXCL1,IL13,MMP1
hsa05323	Rheumatoid arthritis	3	84	3.57%	0.0013	ENSP00000322788,ENSP00000365048,ENSP00000379110	CXCL1,MMP1,TNFSF13B
hsa05162	Measles	3	133	2.26%	0.0024	ENSP00000263125,ENSP00000276431,ENSP00000304915	IL13,PRKCQ,TNFRSF10B
hsa04064	NF-kappa B signaling pathway	2	93	2.15%	0.0255	ENSP00000263125,ENSP00000365048	PRKCQ,TNFSF13B
hsa04658	Th1 and Th2 cell differentiation	2	88	2.27%	0.0255	ENSP00000263125,ENSP00000304915	IL13,PRKCQ