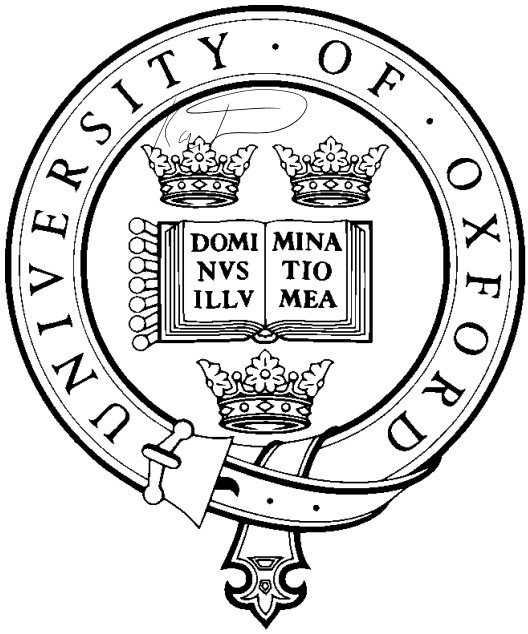
Phenotypic Modelling of Crohn's Disease Severity: A Machine Learning Approach



Katherine Niehaus

Trinity College

University of Oxford

Supervised by
Professor David Clifton
Professor Holm Uhlig
Professor Timothy Peto

Trinity Term 2016

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Declaration

I declare that the work contained in this thesis is entirely my own, except where explictly noted.

Katherine E. Niehaus,
October 5, 2016

Acknowledgements

The road to my DPhil has included many days when I really wasn't sure how it would all come together. I certainly would not have made it without the support of so many people around me.

Thanks goes first to my supervisors. Thank you to Professor David Clifton for taking me on at the very beginning, for trusting me to be able learn what I needed to know, and for giving me the freedom to explore my interests over the past three years. I have really enjoyed seeing the group grow. Thank you to Professor Holm Uhlig for always bringing an unmatched enthusiasm for science to our meetings, for taking the work of an engineer seriously, and for giving me so much of your time. Thank you to Professor Tim Peto for your guidance early on and for continuing to check-in on me, even as my research question shifted from medical microbiology.

Several post-docs have also been invaluable. Thank you to Dr. Stijn Luca for your patience with me as I learned some extreme value theory and for answering my basic questions so willingly. Thank you to Dr. Maria Quaranta for re-explaining immunological pathways to me again every week, for providing interpretation of the biological concepts during our meetings, and for being committed to collaborating on this project.

My DPhil was shaped so positively by my peers around me. Special thanks goes to Rebecca Pullon, for being a calming influence, for engineering advice and discussions, and for comradery as we've journeyed through our DPhils. Thanks also to Dr. Marco Pimentel and Mauro Santos for setting such a high standard for the quality of work expected and for the advice provided during late-evening conversations in the lab when I was still far from the windowed-end of the row. Thank you to my other fellow residents of Time Series Terrace (*the best row*): Dr. Lise Lørup, Farah Colchester, Johanna (Ernst) O'Donnell, Drew Birrenkott, Yanting Shen, and Heloise Greeff Marais. Thanks also to Dr. Tingting Zhu, Dr. Yang Yang, Hamish Tomlinson, Glen Wright Colopy, Peter Gyring, and the growing numbers of the Computational Healthcare Informatics and Biomedical Signal Processing groups. Your neighborly presence made the lab a place where I wanted to be, and I have learned so much both from and with you.

Thank you to my friends from the Stockyard and my fellow members of the Peanut Gallery for such happy homes and wide-ranging conversations. Thanks to my Rhodes class and community, my friends at Trinity, the Oxford cross-country runners, my fellow CDTers, and to the many others, now scattered near and far, whose support and presence have made Oxford the rich environment it has been for me. I have grown so much through your company. Thank you also to my friends "pre-Oxford," whose messages, skype calls, and visits over the past four years have made me feel that I still have a home in the US with friends to come back to.

Of course, I would not have been able to undertake a DPhil without the funding from the Rhodes Trust and the EPSRC, for which I am very grateful.

Finally, deepest thanks goes to my family. To the extended members, I have always felt your love from afar. Emily and Annie, I am lucky to have you as such wise, caring, and inspiring younger sisters (and friends). Mom and Dad, thanks for always encouraging me to challenge myself, for your grounding perspectives, and for believing in me. Your support has enabled me more than I think you will ever quite understand.

Phenotypic modelling of Crohn's disease severity

Katherine E. Niehaus
Thesis submitted for the degree of Doctor of Philosophy
Trinity College
Trinity 2016

Abstract

The growing availability of complex healthcare data holds great promise for improvements in medicine. However, new methodological developments are necessary to realise the potential of these resources. In this thesis we focused upon phenotypic modelling for chronic disease applications, specifically inflammatory bowel disease (IBD). Patients with IBD experience varying clinical trajectories, with the disease course ranging greatly in terms of severity. Our goal was to develop methods to capture IBD patient severity, using data from the electronic health record, and to then use our representations of severity to discover subgroups of patients with similar characteristics. The establishment of patient subgroups and associated genomic factors can enable precision medicine approaches to improve patient care.

Faced with the challenge of unevenly-sampled and sparse clinical time series data, we have proposed a novel approach founded in extreme value theory (EVT) as a means to convert these measurements into interpretable metrics of patient abnormality. We show that our metrics are specifically useful in the modelling of IBD patients, as they provide a condensed representation of a patient's aggregate biochemistry and haematology dynamics. We also found that patient biomarker-based severity is much more representative of classical severity metrics for patients with ulcerative colitis (UC, one of the two primary IBD subtypes) than it is for patients with Crohn's disease (CD, the second subtype).

Thus motivated to combine both our EVT-based and classical phenotypic representations of severity, we implemented a Bayesian clustering model so as to identify latent patient severity profiles. Our model is capable of handling missing data and inferring the number of underlying clusters. We found that consistent patient subgroups were identifiable in our patient cohort, with the majority of CD patients falling into subgroups with severe phenotypic behaviour and the majority of UC patients exhibiting less severe behaviour.

Having identified patient subgroups, we performed a hypothesis-generating association analysis to relate these subgroups (and other clinical features) to genetic loci previously associated with IBD susceptibility. We have presented a number of nominal associations, several of which have plausible biologic mechanisms. Finally, we have illustrated how we can use traditional approaches, machine learning techniques, and our presented EVT methods to answer a practical clinical question regarding the efficacy of two important IBD last-line medications. We examined these drugs in terms of their relative effectiveness, our ability to predict of patient response, and characterisation of this response. We were able to identify distinct ways in which patients respond to these drugs, while also finding that our retrospective data reveals no apparent difference between the two drugs in their effectiveness.

In summary, we have developed a set of methods that can be applied to the challenging problem of finding patterns across patients with heterogeneous disease. Upon using these methods within our specific IBD application area, we have obtained clinically and scientifically useful results.

Publications

Journals and conferences

- **Niehaus, K.E.**, Luca, S., Clifton, D.A., "Point processes for identifying anomalous behaviour in clinical time series data," 2016. *[Under review]*

- Johnson, A.E.W., Ghassemi, M.M., Nemati, S., **Niehaus, K.E.**, Clifton, D.A., Clifford, G.D., "Machine learning and decision support in critical care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444-466, 2016.

- **Niehaus, K.E.**, Uhlig, H.H., Clifton, D.A., "Phenotypic characterization of Crohn's disease severity," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7023-7026, IEEE, 2015.

- **Niehaus, K.E.**, Banda, J.M., Knowles, J.W., Shah, N.H. "FIND FH - A phenotype model to identify patients with familial hypercholesterolemia," in *Proceedings of Data Mining for Medical Informatics*, 2015.

- Clifton, D.A., **Niehaus, K.E.**, Charlton, P., Colopy, G.W. "Health informatics for the clinical management of patients," *Yearbook of Medical Informatics*, International Medical Informatics Association, 2015.

- Walker, T.M., Kohl, T.A., Omar, S.V., Hedge, J., Del Oio Elias, C., Bradley, P., Iqbal, Z., Feuerriegel, S., **Niehaus, K.E.**, Wilson, D.J., Clifton, D.A., Kapatai G., Ip, C., Bowden, R., Drobniewski, F., Allix-Beguec, C., Gaudin, C., Parkhill, J., Diel, R., Supply, P., Crook, D., Smith, E.G., Walker, A.S., Ismail, N., Nieman, S., and Peto, T.E.A., "Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug-susceptibility and resistance: A retrospective cohort study," *Lancet Infectious Diseases*, vol. 15, no. 10, pp. 1193-1202, 2015.

- Clark, I.A., **Niehaus, K.E.**, Duff, E.P., Di Simplicio, M.C., Clifford, G.D., Mackay, C.E., Woolrich, M.W., Holmes, E.A. "First steps in using machine learning on fMRI data to predict intrusive memories of traumatic film footage," *Behaviour Research and Therapy*, vol. 62, pp. 37-46, 2014.

- **Niehaus, K.E.**, Clark, I.A., Bourne, C., Mackay, C.E., Holmes, E.A., Smith, S.M., Woolrich, M.W., Duff, E.P. "Multi-voxel pattern analysis (MVPA) to enhance the study of rare cognitive events: An investigation of experimental post-traumatic stress disorder," in *2014 International Workshop on Pattern Recognition in Neuroimaging*, IEEE, pp.1-4, 2014.

- **Niehaus, K.E.**, Walker, T.M., Crook, D.W., Peto, T.E.A., Clifton, D.A. "Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis," in *2014 IEEE-EMBC International Conference on Biomedical and Health Informatics (BHI)*, IEEE, pp. 618-621, 2014.

Book chapters

1. **Niehaus, K.E.**, Clifton, D.A. "Machine learning for chronic disease" in *Machine Learning in Healthcare*, ed. Clifton, D.A., Institution of Engineering and Technology, 2016. *[In press]*

2. Clifton, D.A., Pimentel, M.A.F., **Niehaus, K.E.**, Clifton, L., Peto, T.E.A., Crook, D.W., Watkinson, P.J. "Intelligent Electronic Health Systems" in *Telemedicine and Electronic Medicine*, eds. Eren, H. and Webster, J.G., CRC Press, pp. 73-97, 2015.

Contents

Abbreviations . . . vi

1 Introduction and Background **1**
- 1.1 IBD Introduction . . . 1
- 1.2 Diagnosis . . . 3
- 1.3 Treatment guidelines . . . 4
- 1.4 Clinically-defined patient groups . . . 6
- 1.5 Genetics introduction . . . 7
  - 1.5.1 Measurement methods . . . 9
  - 1.5.2 GWAS background . . . 10
- 1.6 Immune system introduction . . . 11
- 1.7 Biology of genetic markers . . . 12
- 1.8 Disease severity . . . 14
- 1.9 Prediction and monitoring of disease course . . . 17
- 1.10 Structure of the thesis . . . 19

2 Previous work **21**
- 2.1 Patient subtyping . . . 22
  - 2.1.1 Machine learning for disease subtyping . . . 23
  - 2.1.2 Machine learning for clinical time series . . . 28
  - 2.1.3 Machine learning for computational phenotyping . . . 29
  - 2.1.4 Machine learning for combining -omics data types . . . 31
  - 2.1.5 Summary of clinical and -omics subtyping . . . 32
- 2.2 Previous work in IBD . . . 34
  - 2.2.1 Supervised machine learning prediction . . . 34
  - 2.2.2 Unsupervised machine learning . . . 35
- 2.3 Relation of IBD genetics to phenotype . . . 36
- 2.4 Conclusion . . . 38

3 Introduction to the dataset **39**
- 3.1 Patients . . . 39
- 3.2 Clinical data . . . 39
  - 3.2.1 Laboratory measurements . . . 41
  - 3.2.2 Additional clinical variables . . . 44
- 3.3 Genomic data . . . 46

3.3.1 IBD patient cohort functional data . . . 48
3.4 Conclusions . . . 50

4 Univariate point processes for identifying anomalous behaviour 52
4.1 Classical EVT . . . 53
4.1.1 EVT from a point process perspective . . . 55
4.2 Univariate point process model . . . 56
4.2.1 Model verification . . . 59
4.2.2 Model usage . . . 61
4.3 Model evaluation with simulated data . . . 61
4.3.1 Simulation methods . . . 63
4.3.2 Simulation results . . . 65
4.3.3 Comparison to classical EVT . . . 67
4.4 Application of EVT model to IBD lab data . . . 68
4.4.1 EVT severity metric features . . . 69
4.4.2 Standard lab features . . . 72
4.4.3 Supervised machine learning methods . . . 73
4.4.4 General training formulation . . . 76
4.4.5 Results . . . 77
4.5 Conclusion . . . 87

5 Clustering model for combined data types 89
5.1 General clustering overview . . . 90
5.2 Introduction to mixture models and the EM algorithm . . . 92
5.2.1 Model selection . . . 94
5.3 Extension to a Bayesian model with multiple data types . . . 95
5.3.1 Model specification . . . 95
5.3.2 Model inference . . . 97
5.3.3 Missing data . . . 104
5.4 Clustering coherence . . . 105
5.4.1 Label-switching . . . 105
5.4.2 Cluster stability . . . 107
5.4.3 Cluster similarity . . . 108
5.5 Experiments . . . 109
5.6 Application to IBD patients . . . 113
5.6.1 Functional expression . . . 114
5.6.2 Severity . . . 118
5.7 Discussion . . . 127
5.7.1 Extensions . . . 127
5.7.2 Clinical conclusions . . . 128

6 Associations between genotype and phenotype 130
6.1 Introduction . . . 130
6.2 Data . . . 131
6.2.1 Genetic data . . . 132

6.2.2 Longitudinal and clinical data . . . 133
6.3 Association analysis . . . 134
6.4 Results . . . 138
6.4.1 Tier One: NOD2, IL23R, ATG16L1 . . . 138
6.4.2 Tiers two and three . . . 143
6.4.3 Associations with severity clusters . . . 146
6.5 Survival analysis . . . 147
6.6 Discussion . . . 150
6.6.1 Assumptions and challenges . . . 153

7 Biologic drug comparison **156**
7.1 Introduction . . . 156
7.2 Patient cohort . . . 159
7.3 Survival analysis . . . 160
7.4 Rates of relapse . . . 164
7.4.1 Analysis of relapse event severity . . . 166
7.5 Population-level CRP progression . . . 167
7.5.1 Responders and nonresponders . . . 170
7.5.2 CRP responders . . . 170
7.6 Individual patient trajectories . . . 172
7.7 Predicting patient drug response . . . 177
7.8 Discussion . . . 180
7.8.1 Assumptions and future considerations . . . 183
7.8.2 Conclusion . . . 184

8 Conclusions and future work **185**
8.1 Summary of the thesis . . . 185
8.1.1 Key Contributions . . . 187
8.1.2 Discussion . . . 188
8.2 Future work . . . 189
8.2.1 Laboratory modelling extensions . . . 190
8.2.2 Clustering extensions . . . 191
8.2.3 Additional data sources and questions . . . 192
8.3 Conclusion . . . 193

References **195**

Appendices **215**

A Additional background material **216**
A.1 Additional background from Chapter 1 . . . 216
A.1.1 Reverse-transcriptase quantitative polymerase chain reaction description . . . 216
A.1.2 Enzyme-linked immunosorbent assay description . . . 217
A.2 Additional background from Chapter 2 . . . 217

A.3 Additional background from Chapter 3 . . . 219
A.3.1 ImmunoChip data processing . . . 219
A.3.2 Exome sequencing data processing . . . 220
A.4 Additional background from Chapter 7 . . . 220

B Additional results **222**
B.1 Additional results from Chapter 4 . . . 222
B.1.1 Supporting figures . . . 224
B.2 Additional results from Chapter 5 . . . 225
B.3 Additional results from Chapter 7 . . . 231
B.4 Investigation into localisation subgroups . . . 231

Abbreviations

Genetics	
CNV	Copy number variation
DNA	Deoxyribonucleic acid
ELISA	Enzyme-linked immunosorbent assay
eQTL	Expression quantitative trait loci
GWAS	Genome-wide association study
miRNA	Micro RNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA
PheWAS	Phenome-wide association study
RNA	Ribonucleic acid
RT-qPCR	Reverse-transcriptase quantitative polymerase chain reaction
SNP	Single nucleotide polymorphism
WGS	Whole genome sequencing
IBD	
6-MP	6-mercaptopurine
ADA	Adalimumab
ASA	5-aminosalicylates
ASC	Acute severe colitis
ASCA	Anti-Saccharomyces cerevisiae
ATG16L1	Authophagy-related, 16-like
AZA	Azathioprine
BUD	Budesonide
CD	Crohn's disease
CDAI	Crohn's disease activity index
CRP	C-reactive protein
CS	Classical severity
CYC	Cyclosporine
ESR	Erythrocyte sedimentation rate
FC	Faecal calprotectin
GI	Gastrointestinal
HBI	Harvey Bradshaw Index
IBD	Inflammatory bowel disease
IBDQ	IBD questionaire
IBDu	Inflammatory bowel disease, undetermined
IFN-γ	Interferon γ
IFX	Infliximab
IKZF3	Ikaros family zinc finger 3
IL	Interleukin
IL-23R	IL-23 receptor
L18MDP	Lipidated muramyl dipeptide

LPS	Lipopolysaccharide
MTX	Methotrexate
NF	Nuclear factor
NOD	Nucleotide oligomerization domain 2
PAMPs	Pathogen associated molecular patterns
pANCA	Perinuclear anti-neutrophil antibody
PBMC	Peripheral blood mononuclear cell
PRED	Prednisolone
STAT	Signal transducer and activator of transcription
TF	Transcription factor
Th	T-helper
TLR	Toll-like receptor
TNF	Tumour necrosis factor
UC	Ulcerative colitis
VHI	Van Hees Index
WBC	White blood cell
Methods	
AIC	Aikake information criterion
AMI	Adjusted mutual information
ANOVA	Analysis of variance
AUAAS	Area under the average anomaly score
AUAS	Area under the anomaly score
AUROC	Area under the receivor operator curve
BIC	Bayesian information criterion
BMM	Bernoulli mixture model
CBOW	Continuous bag of words
CDF	Cumulative distribution function
CCCS	Collapsed clustered composite severity
EM	Expectation maximisation
EVT	Extreme value theory
FDR	False discovery rate
GEV	Generalised extreme value
GMM	Gaussian mixture model
GP	Gaussian process
GPD	Generalised Pareto distribution
HMM	Hidden Markov model
ICA	Independent component analysis
i.i.d.	Independent and identically distributed
IOP	Informative observation process
KDE	Kernel density estimation
KW	Krusal-Wallis H-test
LASSO	Least absolute shrinkage and selection operator
LGCP	Log Gaussian Cox process
LR	Logistic regression

MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
ML	Maximum likelihood
MTGP	Multi-task GP
MW	Mann-Whitney U-test
NLP	Natural language processing
NN	Neural network
PCA	Principal component analysis
PDF	Probability density function
PP	Point process
RBF	Radial basis function
RF	Random forest
SMOTE	Synthetic minority over-sampling technique
SVM	Support vector machine
VB	Variational Bayes
General	
COPD	Chronic obstructive pulmonary disease
EHR	Electronic health record
eMERGE	Electronic medical records and genomics
ICD	International Statistical Classification of Diseases and Related Health Problems
ICU	Intensive care unit
IORD	Infections in Oxfordshire Research Database
NHS	National Health Service
PSA	Prostate-specific antigen
RCT	Randomised controlled trials

Chapter 1

Introduction and Background

Many chronic illnesses have heterogeneous disease courses, with both environmental and genetic factors determining their progression. Indeed, this heterogeneity suggests that it is likely that there exist distinct subtypes within many of our current definitions of disease [1–5]. Given the complexity of these diseases, methods for modelling and capturing relevant phenotypic (i.e., non-genetic) aspects of the disease course are necessary in order to understand whether there are subgroups of patients with consistent behaviour. Upon identifying such subgroups, it is then possible to investigate whether there are underlying genetic factors, which may explain subtle differences in disease mechanisms across different patient groups. Such work can then enable "precision medicine" approaches, whereby patients are more appropriately treated for their disease, based upon their personal genomic profile.

Inflammatory bowel disease (IBD) is one such chronic disease characterised by severe gastrointestinal inflammation. Despite being extensively studied, understanding of the aetiology, and particularly the progression, of IBD is still lacking. The focus of this thesis is therefore on the development of modelling techniques that can be used to better describe and understand the phenotypic characteristics of the disease; these approaches are then used to examine the relationship between IBD genotype and phenotype. As this thesis is focused on improving both gaps in clinical IBD care and understanding of the underlying disease, it is worth providing an overview of the relevant medical and physiologic background that will be helpful for understanding the approaches taken.

1.1 IBD Introduction

IBD consists of two primary subdiseases: Crohn's disease (CD) and ulcerative colitis (UC). Cases in which there is inflammation of the colon, but not the small intestine,

and in which there is no further definitive histological evidence for either CD or UC, are termed to be "IBD unclassified" (IBDu) [6].

Incidence of IBD has risen substantially since the 1940s, with the largest increase seen during the 1970s. Incidence varies greatly by geographical regions, with historically higher occurrence in Western countries. However, this is now changing, with increasing rates of diagnosis in countries such as Japan and Korea [7, 8]. It is estimated that 620,000 patients in the United Kingdom (UK) have IBD, which is similar to (though slightly higher than) the prevalence across developed countries of about one in 200 people [9, 10]. Estimates of healthcare costs per patient range from £631 to £3,000 per year. The majority of these costs are due to in-hospital management of episodes of acute flare-up of the disease [9], illustrating the importance of better understanding how to predict and mitigate these events.

IBD is known to be a result of a complex interplay between genetic and environmental factors. Within Caucasian populations, 193 genetic variants (in 163 genetic regions/loci) have been identified through genome wide association study (GWAS)[1] meta-analysis. Some of these loci are involved in pathways of autophagy[2], T-cell differentiation[3], and other inflammatory responses. However, it is still difficult to resolve which genes are involved in many of the identified loci, meaning that there may be many different ways to develop IBD, or that these genetic changes together form a network that results in clinical IBD presentation [11].

To date, these risk alleles[4] are poorly predictive of actual disease, with only 28% of disease inheritance explained in CD through common genetic variants and 22% in UC [12]. Rather, environmental factors seem to play a large role in initiating disease onset within susceptible individuals. For instance, even in monozygous (i.e., identical) twins, the concordance rate for CD is 30-35% and 10-15% for UC [13]. Ananthakrishnan provides a detailed overview of the environmental factors that have been postulated to be associated with IBD [7]. These factors include the "hygiene hypothesis" (i.e., changes in hygiene practice have led to an increase in autoimmune diseases), smoking (which increases risk for CD; but where current smoking is protective for UC), diet, etc. Much attention recently has also focused upon the role of

[1] We provide an introduction to GWAS in Section 1.5.2.

[2] Autophagy is a routine cellular process that involves the destruction of a cell components.

[3] As will be explained in Section 1.6, when triggered by an immune response, T-cells can differentiate into subtypes of cells with specific inflammatory functions.

[4] An allele is a particular genetic variant; this term and other relevant genomic terminology will be described more fully in Section 1.5.

the gut microbiome. Indeed, many of the genetic loci involved with IBD risk are also associated with human responses to mycobacteria in the gut, revealing the complex interplay between genetic, bacterial, and other environmental factors [11]. IBD patients also have different microbiome profiles in comparison to healthy controls [7]. However, the extent to which this is a cause or an effect of the disease remains unclear. It has been hypothesised that changes in the microbiome are contributory towards the increasing incidence of the disease burden worldwide, perhaps due to factors such as alterations in diet, the increasing usage of antibiotics, and general cleanliness [8].

1.2 Diagnosis

Disease onset can occur at any age, though the median age of IBD diagnosis is at 29.5 years, with a second peak around age 65. Approximately 25% of patients are diagnosed by adolescence. The disease symptoms are often more severe for those patients diagnosed in childhood; an earlier disease onset is believed to represent a larger genetic contribution. Patients typically present with abdominal pain (70% of patients), diarrhoea, rectal bleeding, bloody stool, weight loss (60% patients), fever, anaemia, and fatigue [14, 15]. Commonly, there is also inflammation of the joints, skin, and eyes. Diagnosis is determined through a combination of symptoms, endoscopy, radiology, histology, and biochemistry/haematology blood analysis [15]. Possible alternative causes of symptoms include *Clostridium difficile* infection and irritable bowel syndrome. Irritable bowel syndrome lacks visible signs of inflammation in colonoscopy. However, IBD patients may also have comorbid irritable bowel syndrome or *C. difficile* infections.

CD is distinguished from UC in that the extent of ulceration and inflammation is typically confined to the colon and rectum in UC, while it may be anywhere in the gastrointestinal (GI) tract for CD. UC inflammation typically forms a continuous, contiguous region, whereas inflammation may be patchy and segmental in CD. Classically, CD patients tend to exhibit transmural inflammation (i.e., inflammation throughout the entire intestinal wall), while UC is marked by mucosal ulceration (i.e., inflammation confined to the superficial areas of the tissue) [16]. In Figure 1.1 we present an illustration of the various regions of the GI tract. A patient's particular locations of inflammation are frequently used to define subcategories of patients (as will be described in Section 1.4).

In CD patients, inflammation can also cause stricturing (narrowing) of the bowel, fistulas (pathological channels of tissue that form between different regions of the bowel or to other parts of the body), and granulomas (clusterings of macrophages

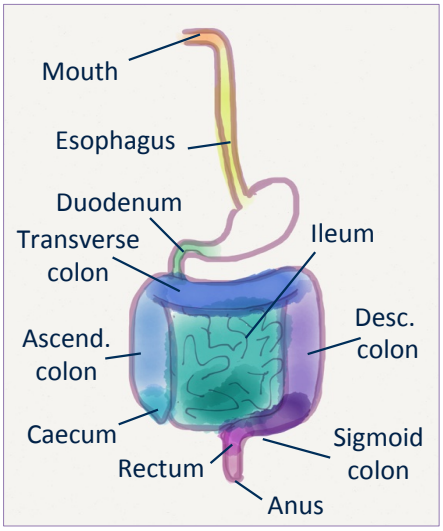


Figure 1.1: *Illustration of the GI tract, with potential areas of inflammation found in IBD patients labelled. Abbreviations: ascend.: ascending, desc.: descending.*

in an attempt to isolate a "foreign" substance in the body). At diagnosis, about 5% to 15% of patients have strictures, and 5% to 15% have fistulas. However, in the 20 years following diagnosis, about half of patients will have a stricturing or fistulating complication; these complications usually require surgery [17].

Once diagnosed, IBD is typically a life-long, chronic condition, with periods of clinical remission and relapse. About half of all patients have at least one relapse every year, 20% of which will require hospitalisation. Periods of relapse are unpredictable, and symptom patterns vary greatly across different patients [9]. For instance, while more than half of CD patients require surgery within 10 years of disease onset, 30% will require neither surgery nor immunosuppressive drugs throughout their lives. Similarly for UC, 20% of patients require colectomy, but half of patients will again require neither surgery nor immunosuppressive drugs [10].

1.3 Treatment guidelines

There is currently no cure for IBD. Medical therapy typically consists of anti-inflammatory and immunosuppressant drug prescription to induce clinical remission; commonly-used drugs are presented in Table 1.1. Traditionally, these drugs are prescribed in a step-up fashion, starting initially with those of lowest strength and fewest side effects. Patients will begin with anti-inflammatory drugs and possibly steroids to induce remission. Those patients who are unable to maintain remission for three months upon steroid removal or decrease of dose are termed "steroid dependent" and

may be started upon immunosuppressants. Patients who do not respond at all to steroid treatment are considered for colectomy or treatment with the biologics infliximab (IFX) or adalimumab (ADA), both of which are anti-tumour necrosis factor (TNF)-α[5] antibodies [18].

UC patients may present with systemic toxicity (elevated pulse rate, temperature, etc.), in which case they will be termed "acute severe colitis" (ASC) and admitted for hospital treatment. Likewise, CD patients may present with fevers and severe weight loss and will likewise be admitted to hospital for treatment. Paediatric CD patients in this state have been found to respond well to nutritional therapy; this approach has not been found to be successful for adults [18].

While the medication practices described above represent conventional clinical practice, recent studies have shown that initial treatment with biologic therapy is actually the most effective option for those patients with severe disease who respond to these drugs [19]. This means that traditional "step-up" therapy is being reconsidered in some cases for "step-down" therapy sequences.

More than half of CD patients have surgery within ten years of diagnosis, with over 70% eventually requiring surgery; requirements for surgery for UC patients is less frequent, with lifetime surgery rates of between $20 - 30\%$ [9]. Most surgeries for CD patients involve removal of the diseased portion of the bowel. While surgery for CD patients often reduces symptoms, it does not cure the condition: clinical symptoms reappear in about a third of surgically-treated patients within about three years. For UC patients, surgery most often involves removal of the entire colon; many patients who receive this procedure remain symptom-free for the rest of their life.

[5] We discuss the relevant inflammatory pathways in IBD, including the role of TNF, in Section 1.6.

Table 1.1: *Common IBD Medications*

Drug	Abbreviation	Drug class	Notes
5-aminosalicylates	ASA	Anti-inflammatory	Includes mesalamine, balasalazide, and olsalazine
Prednisolone	PRED	Steroid	
Budesonide	BUD	Steroid	
Azathioprine, 6-mercaptopurine	AZA, 6-MP	Immunosuppressant/ Immunomodulator	Thiopurines
Cyclosporine	CYC	Immunosuppressant/ Immunomodulator	
Methotrexate	MTX	Immunosuppressant/ Immunomodulator	
Infliximab, Adalimumab	IFX, ADA	Biologics	Anti-TNF-α antibodies

Table 1.2: *Crohn's Disease Montreal Classifications*

Criteria	Code	Description	Notes
Age at diagnosis	A1	Less than 17 years old	
	A2	17 to 40 years old	
	A3	Greater than 40 years old	
Location	L1	Ileal	
	L2	Colonic	
	L3	Ileocolonic	
	L4	Upper alimentary canal	*
Behavior	B1	Non-stricturing, non-penetrating	
	B2	Stricturing	
	B3	Penetrating	
	p	Perianal	*

Starred (*) criteria are used as modifiers for the previous descriptions

1.4 Clinically-defined patient groups

There have been numerous efforts within the medical community to classify IBD patients into more specific subgroups, for the purposes of both guiding medical treatment as well as for more uniform inclusion criteria for clinical studies.

The Montreal Index is widely used to categorise patients based upon descriptive aspects of disease behavior; these were not designed to correspond to disease progression. The first iteration of this index was designed by a working party in 1991 in Rome, which decided upon characteristics of disease location, clinical behavior, and operative history to categorize CD patients. In Vienna in 1998, these classifications for CD were amended: age of onset was included and operative history was removed. Finally in 2005, further amendments were made within these CD categories to produce the currently-used Montreal classifications. The Montreal classifications included a set of criteria to define UC subgroups as well, based upon disease location and severity [6]. The categorisations for CD and UC are outlined in Tables 1.2 and 1.3 (largely drawn from [6]).

The working party that developed the Montreal classifications debated over the incorporation of further blood haematology/biochemistry and genetic factors into the criteria. It was decided that though this would be desirable, the clinical evidence was not sufficiently strong to support further refinement of the classifications [6]. While useful to describe the disease, the Montreal Index does not explicitly represent disease severity in CD patients (though severity is specifically included for UC patients).

Table 1.3: *Ulcerative Colitis Montreal Classifications*

Criteria	Code	Name	Description
Extent of disease	E1	Ulcerative proctitis	Only rectum
	E2	Distal/left-sided UC	Part of colorectum distal to splenic flexure (descending colon in Figure 1.1)
	E3	Pancolitis/extensive UC	Proximal to splenic flexure (ascending colon in Figure 1.1)
Severity	S0	Clinical remission	No symptoms
	S1	Mild UC	Passage of less than five stools/day; no systemic illness; normal inflammatory markers
	S2	Moderate UC	Passage of least five stools/day; limited systemic toxicity
	S3	Severe UC	Passage of at least six bloody stools/day; pulse of $\geq$ 90 bpm; temperature $\geq$ 37.5°C; haemoglobin $\leq$ 10.5g/100mL; ESR $\geq$ 30 mm/h

Abbreviations: bpm: beats per minute, ESR: erythrocyte sedimentation rate

1.5 Genetics introduction

There has been much investigation into the genetic basis of IBD, and we perform our phenotypic modelling in this thesis with the intention of relating patient phenotypic patterns to genetic patterns. Human genetic variation is encoded in long series of deoxyribonucleic acid (DNA) base pairs (adenine, thymine, cytosine, and guanine). Figure 1.2 illustrates the cellular processes involved as DNA is transcribed and then translated into proteins. In the process of transcription, portions of DNA are converted into single-stranded ribonucleic acid (RNA). Codons (triplets of messenger RNA, or mRNA) are then translated into one of the 20 amino acids, which are the building blocks of proteins.

The portions of the genome that encode proteins or other functional products are called genes. Within gene regions, portions are spliced away before producing the mature mRNA - these specified regions are termed introns. The remaining regions, which constitute the portion of the gene that is fully transcribed and then translated, are termed exons. The exome, therefore, is the portion of the genome constituted of the exons. It is estimated that less than two percent of the 3 billion nucleotide bases in the human genome constitute genes; the definition of the exome is not static, as we still do not know all of the protein-coding portions of the genome [20]. Understanding the role of the rest of the genome is also the subject of on-going research, but it is well-established that parts are involved in regulating gene expression (and thereby, protein production). For instance, special types of proteins, called transcription factors (TFs), can bind to the genome to "turn on" the transcription of a certain gene - such a gene is said to be expressed. In addition, non-coding RNAs

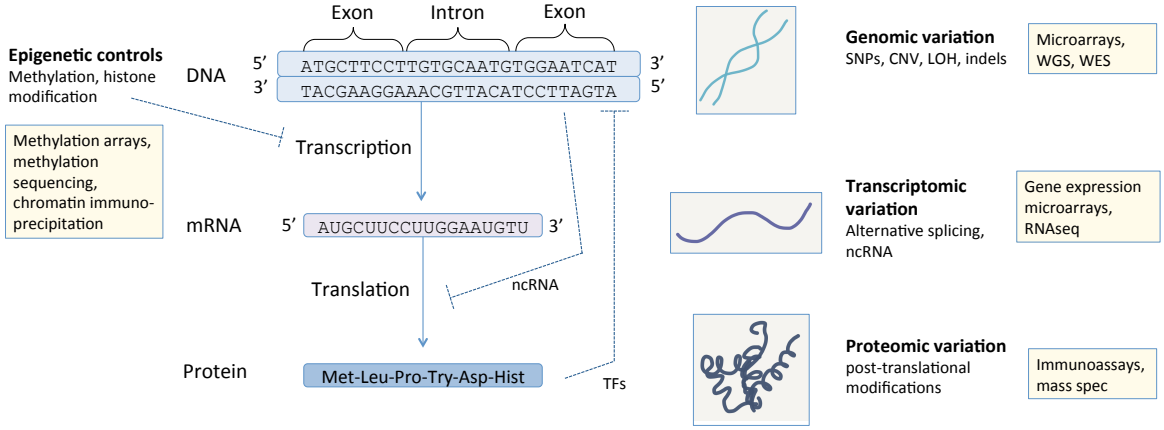


Figure 1.2: *Illustration of the processes involved in cellular protein production, as well as possible sources of variation and regulation. Pale yellow boxes contain methods for measuring the indicated types of variation. Abbreviations: single nucleotide polymorphism (SNP), copy number variation (CNV), loss of heterozygosity (LOH), whole genome sequencing (WGS), whole exome sequencing (WES), non-coding RNA (ncRNA), transcription factors (TFs).*

(ncRNAs) and microRNAs (miRNAs), which are not translated into proteins, act as post-transcriptional modifiers of the final protein or its regulation. Genetic loci that regulate mRNA or protein levels are termed *expression quantitative trait loci* (eQTL). These eQTLs can either act on a gene close to where they are themselves encoded (*cis*-acting) or far from their own genomic location (*trans*-acting) [21].

Heritable (and non-heritable) regulation of gene expression can also be accomplished through changes to the macro-structure of the DNA through chemical alterations such as methylation and histone modification. Both of these processes can make the DNA in the cell less accessible to cellular machinery, meaning that the genes are not easily expressed. Such changes are termed "epigenetic" because they result in changes in gene expression without altering the underlying DNA sequence. These complex regulatory processes mean that different genes are expressed with great precision in different cells throughout the body [22, 23]. Particularly for IBD, there is evidence that epigenetic and regulatory factors are associated with the disease, based upon the location of various risk loci in miRNA and ncRNA coding sequences, as well as the abundance of eQTL loci [13].

There are many ways in which the genome itself can vary across individuals. A single nucleotide polymorphism (SNP) is a single-base change in the DNA found in at least one percent of the population (i.e., these are relatively common polymorphisms). SNPs can result in proteins with altered functionality, or, if within a regulatory region, changes in protein production. Across populations, there may be a few different

common variants of a gene, termed alleles. Insertions and deletions of nucleotide bases (indels) can result in a shift in how mRNA bases are translated into a protein (because mRNA bases are read together in triplets); such variants can therefore be referred to as "frameshift." Copy-number variants (CNVs) are also a source of inter-human variation. CNVs are large portions of the genome that have been deleted or duplicated; they constitute approximately 12% of human DNA [24]. Most multicellular organisms, including humans, have two copies of each chromosome. If the alleles on both chromosomes are the same, the allele is said to be homozygous; if the alleles differ, then it is said to be heterozygous. Sometimes portions of the genome can be lost, resulting in only a single parental copy of a region of the genome. Such a situation is termed a "loss of heterozygosity" (LOH). While some diseases are monogenic, meaning that a single mutation is sufficient to cause disease, most common diseases involve the combinatorial effect of many variants across the genome. These effects can be additive or have more complex nonlinear interactions; such non-additive effects are often termed "epistatic" [25]. A visual representation of where these sources of genomic variation act is shown in Figure 1.2.

1.5.1 Measurement methods

Patient genetic information may be measured in a variety of ways. Microarrays (also known as DNA chips) are perhaps the most common source of data in large databanks of phenotypic and genetic information. Microarrays are small chips that contain thousands of DNA probes attached to their surface. When a fluorescently-labelled sample is introduced, any pieces of DNA that are complementary to the attached DNA probes will bind to the probe, indicating which sequences were present in the sample. Specialised microarrays can be created for specific purposes; for instance, the Immunochip is a microarray that has been designed to provide coverage of SNPs suspected to be involved in immune-related diseases (e.g. IBD, rheumatoid arthritis, etc) [26].

Microarrays can be very cost-effective for wide coverage of common SNPs, but they do not provide information about every mutation in a patient's genome. Alternatively, whole-genome sequencing (WGS) methods can be used to obtain the full genome sequence, which identifies SNPs and indels throughout the genome; CNVs are still sometimes difficult to identify. Illumina (United States of America, USA), the current market-leader in sequencing platforms, announced the sale of the first machines capable of the long-heralded "one-thousand dollar" human genome in early

2014[6]. Some newer sequencing companies are offering longer read lengths[7], allowing for perhaps easier interpretation of CNVs. Full exome sequencing is another sequencing alternative, in which all exonic regions of a patient's genome are sequenced (where, as mentioned, the exome represents only about 1% of the human genome). Exome sequencing can be useful for identifying rare (<1% frequency in the population) variants that are not covered in standard microarrays. WGS is often necessary, however, to identify regulatory mechanisms involved in the disease [13].

While sequencing data can provide information about a patient's genetic background, it does not identify which proteins are actually being produced in particular cells in the body at any given time (due to the advanced regulatory mechanisms involved). For this, mRNA levels or protein levels must be obtained. mRNA can be measured through microarrays, a real time reverse-transcriptase quantitative polymerase chain reaction (RT-qPCR), or RNA sequencing. Gene expression microarrays are again most useful for known and common gene transcripts, and as a less accurate technology than RT-qPCR, they are best-suited for large-scale initial analyses [27].

Proteins levels can be assessed through antibody-based methods or mass spectometry. Enzyme-linked immunosorbent assays (ELISAs) remain the gold-standard method for measurement of the relative abundance of proteins. Mass spectometry is most useful for large-scale screening studies in which measurement sensitivity is less important, as it is relatively cost-effective and the proteins of interest do not have to be pre-specified. However, especially for studies involving cytokine (cytokines are small cell signaling proteins) measurements, mass spectometry is not sufficiently sensitive [28].

These various measurement methods are also noted in the yellow boxes in Figure 1.2. More details on the RT-qPCR and ELISA methodologies, which were used to obtain the functional measurements in this thesis, are presented in Appendix A.1.

1.5.2 GWAS background

The dominant paradigm for identifying genomic associations with human phenotypes is through GWAS. In such studies, the variants present for several hundred thousand SNPs are assessed for a group of individuals with some phenotypic characteristic (e.g., heart disease, diabetes, etc.) and a "control" group without the characteristic. Genomic variants are typically assessed through microarrays that probe for SNPs that

[6]This $1,000 genome is practically only achievable for large scale research labs: Illumina only sells the machines in batches of ten for $10 million.

[7]Read lengths refer to the length of the raw DNA sequence that can be read at a time by a sequencing technique or machine.

are present in about 5% of the population[8] [29]. In standard analyses, chi-squared tests are used to identify SNPs that are significantly associated with the phenotype. Corrections must be made for multiple testing, usually through Bonferroni correction, using a false discovery rate (FDR) procedure, or permutation testing [30]. In addition, corrections are often made for population structure through adjustment by a genomic inflation constant[9] or using principal component analysis (PCA) [31].

To date, more than 1,600 papers have published GWAS results, identifying 2,000 genotype-phenotype associations [32]. However, these associations typically together account for less than 10% of the phenotype's heritability[10], as estimated through family studies, and are often associated with only 1.1 to 1.5 fold increases in risk. This "missing heritability" could be due to many factors, such as the possibility of rare variants, CNVs, interactions across SNPs that are not currently captured, epigenetic factors, and environmental influences [29,32]. The definition of phenotypes is another limiting challenge with GWAS. Medical diagnosis standards differ across institutions and countries, and human illness phenotypes exist on a spectrum, meaning that it is difficult to obtain clean groups of "cases" and "controls" (e.g. [33]). Indeed, the heterogeneity of phenotypes is one of the motivating factors for clinical subtyping.

1.6 Immune system introduction

IBD is an inflammatory disease, meaning that its relevant biology is closely tied to the immune system. There are two primary arms to the immune system: innate and adaptive immunity. The innate response in the gut involves dendritic cells (DCs), macrophages (a type of white blood cell, WBC), and epithelial cells. These cells recognise pathogen associated molecular patterns (PAMPs) on bacteria and other microorganisms through pattern recognition receptors. Toll-like receptors (TLRs) and nucleotide-binding oligomerisation domain (NOD)-like receptors are two such examples. When PAMPs are identified, nuclear factor (NF)-$\kappa\beta$ is activated, which

[8] The identification of causal SNPs is hampered by linkage disequilibrium. Linkage disequilibrium is the degree to which a certain genetic variant is correlated with another genetic variant, due to the fact that chromosomal recombination physically occurs in discrete chunks. This means that SNPs that are found to be strongly associated with the phenotype of interest can only be viewed as markers for the "true" causal SNPs.

[9] Genomic inflation is a statistical method for dealing with the large number of false positives expected in GWAS due to linkage disequilibrium. It is calculated by dividing the median χ^2 value by the median expected χ^2 value. An inflation factor of greater than 1 indicates that there is underlying population structure.

[10] Heritability is "the proportion of the total phenotypic variation in a trait that can be attributed to genetic effects" [32].

then triggers the response of additional inflammatory cytokines. One of these responses is the activation of T-cells (a type of lymphocyte, which itself is another type of WBC; these cells play a role in adaptive immunity as described below). In addition to responding to external stimuli, the epithelial cells form a physical line of defense against invading pathogens and also produce anti-bacterial molecules such as α- and β-defensins [34].

While the innate immune cells respond broadly to pathogens, the adaptive arm serves to provide targeted responses that are "learned" and "remembered" based upon previous exposures. In response to different cytokines, naïve T-cells will differentiate into various types of T-helper (Th) cells: Th1, Th2, or Th17 cells. These in turn produce additional cytokines, which again contribute to the inflammatory response. Regulatory T-cells act to inhibit T-cell differentiation, thereby decreasing inflammation [34].

1.7 Biology of genetic markers

In general, there appears to be a stronger genetic component in CD than in UC, though both are believed to share many of the same genetic mechanisms [8, 11, 13]. GWA studies have consistently found loci associated with IBD, and these findings have also held in a large meta-analysis of 15 trials (14,763 CD patients, 10,920 UC patients, and 15,977 controls) [11]. While early studies have focused almost entirely upon individuals of European ancestry, more recent analyses have begun to examine differences across individuals of different genetic backgrounds [35]. The most well-understood mechanisms that have arisen from these studies involve polymorphisms in the *NOD2* genes, genes from the interleukin-23-type (IL-23)/Th17 pathway, and autophagy genes. In addition, additional gut homeostatic processes, such as barrier function, endoplasmic reticulum stress, and metabolic pathways have also been implicated [8, 13].

The NOD2 protein, as mentioned in Section 1.6, is an intracellular protein expressed in many cell types in the gut (e.g., epithelial cells, macrophages, etc.) that detects bacterial peptidoglycean (found within bacterial cell walls) and initiates a cellular response. The genomic region containing *NOD2* was first identified in 1996 through genomic linkage studies[11]; this finding was one of the early successes in ge-

[11] Genomic linkage studies are based upon the fact that certain regions of the genome will be inherited together, or linked. These links can be used to identify the portion of the genome likely associated with a disease by examining the probability of traits in a family pedigree [36].

netic epidemiology and indicates the strength of the *NOD2* association with disease susceptibility [37]. Three different polymorphisms in the *NOD2* gene have been associated with risk of CD. There are differing hypotheses about the role of *NOD2* variants in CD [38]. Early evidence suggested that a defective protein results in lower levels of Paneth[12] cell defensin production, which reduces the patient's ability to mount proper inflammatory responses [39]. More recent evidence suggests that *NOD2* variants also influence the microbiome (again perhaps modulated by disease localisation) [40]. However, there have also been numerous other potential hypotheses for the role of NOD2 in IBD; it is likely that many mechanistic factors are at play, rather than a single explanation [34]. Approximately 30% of CD patients with European ancestry have one or more of these *NOD2* polymorphisms, but these polymorphisms are neither necessary nor sufficient for causing IBD. Phenotypically, patients with these polymorphisms are more likely to have inflammation within the ileum than those without [8, 13].

As previously described, autophagy is the process of intracellular component destruction. Autophagy processes are important in cell maintenance, and they also provide a mechanism for destroying bacterial pathogens within the cell. Polymorphisms in the "authophagy-related, 16-like" gene (*ATG16L1*) have been associated with CD, but not UC. The genetic region containing the "immunity-related guanosine triphosphate (GTP)ase M" gene, another gene involved in autophagy, contains three loci associated with CD and possibly also with UC. Patients with polymorphisms in these genes have abnormal Paneth cells [8].

IL-23 is a cytokine secreted by macrophages and dendritic cells that induces the transcription of proteins involved in inflammatory responses. Th17 cells produce cytokines such as IL-17, which induce epithelial cells to produce anti-microbial proteins. Polymorphisms in many genes across the IL-23 signaling pathway, such as IL-23 receptor (IL23R), IL-12B p40 subunit, and "signal transducer and activator of transcription 3" (STAT3), have been associated with both CD and UC. In addition, SNPs in chemokine C-C motif receptor 6 has also been associated with CD. These genes produce subunits of IL-23 cytokines, are receptors for IL-23, are produced in response to cytokines, and are involved in inflammatory cell recruitment, respectively [8].

In short, most of the genetic risk factors for IBD point towards irregular inflammatory responses within the gut. The polymorphisms with the greatest relative risk are found in *NOD2*: individuals who are heterozygous for a *NOD2* risk allele have a

[12]Paneth cells are epithelial cells found largely in the ileum and the jejunum (the jejunum is the area between the ileum and the duodenum).

two- to four-times greater risk of developing CD, while homozygous individuals carry a 20- to 40-times greater risk [37]. However, these loci still have a low positive predictive value. As such, genetic screening is not yet recommended for standard clinical practice [15].

As mentioned, it is estimated that less than 30% of the heritability in IBD can be explained through identified genetic loci [8,12]. This points towards the possibility of complex epistatic interactions between genomic regions, and the potential for many different rare mutations that produce similar cellular phenotypic outcomes. Improved understanding of these genetic mechanisms in IBD would allow for the development of new therapeutic agents, as well as acting as a more useful tool for clinical practice. It would be very desirable to be able to provide more specific prescription of medical therapy given an individual's combination of pathogenic genetic polymorphisms.

Interestingly, more than half of the loci associated with IBD have also been associated with other inflammatory or autoimmune diseases (such as rheumatoid arthritis or type 1 diabetes). However, the risk-associated variants sometimes have opposing effects in different diseases [13], pointing to either cases of strong linkage disequilibrium or key proteins for understanding differences in the pathobiology of these diseases.

1.8 Disease severity

IBD disease severity does not have a single simple clinical representation; severity can be defined in terms of intestinal malformations (e.g. the formation of fistulas), clinical outcomes (e.g. requiring intestinal resection or hospitalisation), medication management (e.g. requiring immunosuppressant or biologic therapy), etc [41]. Walsh et al. recommend evaluating disease severity in terms of the distinct axes of inflammation (which can be assessed via biomarkers, endoscopy, etc.), clinical symptoms, and quality of life [42]. Severity is also not a static concept, as the probability of requiring surgery increases with increasing time since diagnosis. Nevertheless, there are numerous approaches and metrics that have been applied for the purposes of representing disease "severity" so as to study factors associated with severity and to assess the efficacy of treatment options.

One widely-accepted metric for assessing CD patient disease severity is the CD Activity Index (CDAI). This index was created in 1976 by the National Cooperative Crohn's Disease Study group. The investigators examined 18 different variables for 112 CD patients (e.g., albumin and haematocrit laboratory values, degree of abdominal pain, number of days with bloody stools, nausea) and fit these variables

Table 1.4: *Crohn's Disease Activity Index Criteria*

Item	Value Range	Weighting
Number of liquid stools	(Number)	2
Abdominal pain score	0 to 3	5
General well-being	1 to 4	7
Summed presence of the following:	(Sum)	20
Arthritis		
Mucocutaneous lesions		
Iritis		
Anal fissure/fistula		
External fistula		
Fever $>$ 37.8°C		
Antidiarrhoeal use	(1/0)	30
Abdominal mass	0 to 5	10
47 - haematocrit (males)		6
42 - haematocrit (females)		
100 x (1 - [bodyweight / standard weight])		1

to physician assessment of patient well-being using linear regression. They removed variables using a backwards step-wise elimination procedure to produce a set of eight items [43]. The final CDAI was created by combining these eight variables together using a simplified weighting in order to make the index more easily-used in clinical practice, as shown in Table 1.4. A CDAI of less than 150 indicates clinical remission, whereas a CDAI of greater than 450 indicate very severe disease. The CDAI was subsequently verified in a larger study of 1058 patients [44].

While the CDAI is frequently used in clinical trials and as a clinical measure, some physicians believe that it is too cumbersome (requiring patients to keep a week-long diary) and that it relies on too many subjective measures. The Harvey Bradshaw Index (HBI) was created to simplify the CDAI: it uses symptoms over only the previous 24 hours; removes antidiarrhoeal usage, haematocrit level, and body weight from the formula; and uses no weighting factor. The Van Hees Index (VHI) was created to reduce subjectivity. Using a similar methodology as the CDAI, the VHI was created using multiple regression and step-wise deletion of variables to fit to physician assessment. The final variables included were: serum albumin, erythrocyte sedimentation rate (ESR), patient sex, weight:height index, abdominal mass presence, temperature, stool consistency, previous intestinal resection, and fissures/fistulas [44].

From an opposing point of view, other physicians have felt that the most important severity indicator for the patient is quality of life, and therefore an alternative metric based only upon quality of life has been created: the IBD Questionnaire (IBDQ). The IBDQ was created with patient input in the design process and consists of 32

questions that cover bowel and systemic symptoms, as well as emotional and social functioning. The IBDQ has been used in several studies as an outcome assessment for therapy efficacy [44]. Walsh et al. provide a comprehensive review of various criterion that have been developed for disease activity monitoring for both CD and UC, with recommendations as to those that are most useful clinically [42].

The choice of severity definition in a research context typically depends upon the clinical question and the available data. In clinical trials, for instance, comprehensive severity metrics can be calculated, such as the CDAI. In retrospective studies, however, only the information in the patient's record is available. As such, outcomes such as the progression to GI surgery are commonly employed [45]. Surrogate markers, such as the Vienna or Montreal criteria, have also been used, since age of diagnosis (Montreal [A] criteria) and ileal location (Montreal [L] criteria) tend to correlate with severe disease, and the disease behaviour (Montreal [B] criteria) represents one interpretation of severity [10, 46, 47]. Table 1.5 shows a number of different severity metrics that have been employed in the literature, for both prospective and retrospective studies.

Table 1.5: *Severity criteria employed in previous studies*

Study Authors	Year	Ref.	Diag.	Severity definition
Mow et al.	2004	48	CD	Disease behaviour and resections
Solem et al.	2005	49	CD	Any of: failed to respond to 5-ASA, moderate abdominal pain, vomiting, fistulas, fever, weight loss >10%, haemoglobin <10g/dL
Solem et al.	2005	49	UC	Any of: $\geq$ 7 stools daily, acute systemic response (fever, $\geq$ 10% weight loss, orthostasis, haemoglobin $\leq$ 10g/dL), hospitalisation
Russell et al.	2005	46	CD	Vienna classification
Beaugerie, et al.	2006	50	CD	Any of: 2 steroid courses, steroid dependence, hospitalisation for flare, $\geq$ 12 months of disabling symptoms, immunosuppressive therapy, intestinal resection
Wolters et al.	2006	51	CD	Surgical recurrence, medication escalation
Solberg et al.	2007	52	CD	Requirement for surgery; patient self-identified disease pattern
Henriksen et al.	2007	53	CD, UC	Requirement of surgery, corticosteroid or AZA therapy, and frequency of relapse
Hoie et al.	2007	54	UC	Surgery or escalation of medication
Henriksen, et al.	2008	45	CD	Intestinal resection
Loly et al.	2008	55	CD	Any of: 2 steroid courses, steroid dependence, hospitalisation for flare, $\geq$ 12 months of disabling symptoms, immunosuppressive therapy, intestinal resection
Weersma et al.	2009	47	CD	Vienna classification
Weersma et al.	2009	47	UC	Age of onset, location, colectomy, extra-intestinal
Solberg et al.	2009	56	UC	Colectomy; patient self-identified disease pattern
Schoepfer et al.	2010	57	CD	Endoscopic disease activity; created own score cutoffs based on Simple Endoscopic Score
Cleynen et al.	2012	58	CD, UC	Any of: stenosing/penetrating behavior, perianal disease, bowel resection
Cleynen et al.	2015	10	CD, UC	Montreal classification
Siegel et al.	2016	59	CD	Any of: surgery; stricturing or penetrating disease behaviour

1.9 Prediction and monitoring of disease course

It has proven very difficult to predict the course of IBD. This is partially due to the difficulty associated with defining severe and nonsevere disease, as discussed. Nevertheless, some factors have been found consistent, regardless of the specific severity criteria employed. In terms of disease phenotypes and patient demographics, previous studies of CD patients have found that a young age at diagnosis (less than 40 years), initially requiring steroids, perianal disease, ileal disease localisation, stricturing or penetrating disease behaviour, upper GI inflammation, smoking, and male sex have all been associated with more severe disease. UC has been less well-studied, but extensive colitis and more severe disease at diagnosis (e.g., ASC, which typically involves fever and weight loss) have been associated with more severe disease [19, 60, 61].

IBD care could also be much improved if it were possible to use common biochemistry/haematology laboratory tests (which we hereafter refer to as "labs") to monitor disease activity and severity over time. From a lab measurement perspective, C-reactive protein (CRP) has been the most-studied biomarker for IBD severity. CRP is an acute-phase inflammatory marker; it is primarily produced by hepatocytes (liver cells) in response to inflammatory stimulus and its production is mediated through the cytokines IL-6 and IL-1β. Notably, there are genetic variants that can result in low basal (i.e., baseline) levels of CRP [62]. CRP has been found to correlate well with disease activity [63]. For instance, Solem et al. studied 104 CD patients and found CRP $>$ 8mg/L to be associated with histologic, colonoscopic, and clinical active disease using univariate logistic regression (LR) [49]. Schoepfer et al. examined the ability of CRP (as well as faecal calprotectin, discussed below) to represent endscopic disease activity in 122 CD patients [57]. The authors found that CRP $>$ 5 mg/L had a positive predictive value of 88% for endoscopically-active disease (accuracy of 66%), while use of the CDAI had a positive predictive value of 80% and accuracy of only 40%. Henriksen et al. found, in a 2008 study on 454 CD and 200 UC patients, that CRP levels at diagnosis are predictive of increased risk of surgery for CD patients with ileal disease and UC patients with extensive colitis [45]. The correlation also has been found to depend upon the disease localisation; severity for UC patients with inflammation confined to the rectum and CD patients with inflammation confined to the upper ileum did not have good correlation with CRP in a previous study [63].

The utility of CRP and other laboratory tests for predicting disease relapse has also been examined. ESR and CRP have been found to be predictive of relapse in CD, though these findings have not been consistent across all studies [17, 63]. Never-

theless, the European consensus guidelines recommend the use of CRP as a predictive indicator for relapse (with high values corresponding to risk of relapse) [15]. Generally, levels of ESR, serum albumin, CRP, haematocrit, and other common serological measures are more extreme in patients with more severe disease. Monitoring of activity depends not only on whether values are extreme for a population, but also on their comparison to the patient's typical range of values [63].

Notably, recent work has shown the utility of lab measurements for the prediction of ASC [64]. Cesarini et al. were able to predict ASC within three years through a point-based score assessed at diagnosis/first colonoscopy, which consisted of extensive disease localisation, CRP $>$ 10 mg/L, and haemoglobin $>$ 12 g/dL (females) or $<$ 14 g/dL (males). Their scoring system was validated in two independent patient cohorts.

Another important biomarker that has been associated with IBD severity and disease course is faecal calprotectin (FC), a protein that can be measured in the faeces. High levels of FC indicate that high levels of neutrophils are present in the GI mucosa. Several studies have shown that high levels of FC are reflective of high disease activity and relapse, and that this biomarker is more reflective of endoscopic disease activity than CRP levels or the CDAI, with an 87% accuracy in classifying endoscopically-active disease [17, 57, 63, 65]. Elevated levels of FC have also been consistently found to predict the need for surgery in ASC patients and relapse within 12 months, with estimates of predictive value ranging 69% to 90% sensitivity and 69% to 82% specificity [17, 66, 67].

In terms of serologic metrics of disease severity, it has been found that high levels of antibodies to various microbes have also been associated with disease course. High reactivity to anti-*Saccharomyces cerevisiae* (ASCA), outer membrane protein, flagellin, anti-I2, and antiglycan have been associated with more severe disease. On the other hand, higher reactivity to perinuclear anti-neutrophil antibody (pANCA) has been associated with less severe disease and more UC-related symptoms than CD symptoms [19, 48, 60].

As far as genetic markers, variants in *NOD2* have been most strongly associated with severe disease [60]. For instance, a study focusing on young-onset Scottish patients (less than 16 years old at diagnosis) found the *NOD2* p.L1007fsX1008 polymorphism to be associated with an increased requirement for surgery, as well as increased CRP and decreased albumin levels at diagnosis [46]. Ryan et al. conducted a study on 127 Canadian CD patients to examine whether common IBD-risk associated SNPs in *NOD2*, *ATG16L1*, *IL23R*, or blood levels of seven antibodies were related to severe CD [68]. The authors found that ASCA reactivity was predictive

of stricturing/penetrating disease; no other associations were significant using multivariate LR. However, this may be due to the study's small sample size; there was a trend towards significance for the *NOD2* p.L1007fsX1008 polymorphism.

However, it is quite likely, given the variety of inflammatory pathways involved, that the genetic load could be cumulative. Weersma et al. found, in their study of 2400 Dutch IBD patients, that patients with an increased number of IBD risk alleles have more severe disease [47]. The authors included risk-associated variants in *NOD2*, *IBD5*, *DLG5*, *ATG16L1*, and *IL23R*. Similarly, Cleynen et al. studied 1528 European CD patients and constructed a risk score consisting of the total count of risk alleles across those previously (as of [69]) associated with CD risk and also which were univariately associated with severe disease in their study [58]. As in [47], they found that patients with increased numbers of risk-associated SNPs progressed more quickly to severe disease. However, family history of disease does not appear to be associated with disease severity, indicating that perhaps additional loci beyond those associated with disease onset risk may be involved in mechanisms associated with severity [60].

Seigel et al. recently developed a web-based tool to predict disease complications (defined as either stricturing or penetrating disease behaviour, or GI surgery) based upon a model they developed for this purpose [59]. The relevant variables found to be predictive included ileal and perianal disease localisation, serological indicators, and *NOD2* variants.

1.10 Structure of the thesis

The structure of the thesis is organised as follows: in Chapter 2, we provide an overview of previous work in IBD severity modelling and disease prediction. We preface this with a review of relevant methodological approaches that have been employed in other diseases for modelling disease phenotypes and combining these with genetic data. In Chapter 3, we introduce the dataset that will be examined through the rest of the thesis.

The contribution of this thesis is both a set of methods that may be used for probing relationships between clinical and genomic data types generally, as well as specific findings for our chosen application area of IBD. We therefore start in Chapter 4 with an introduction to extreme value theory and present our extensions, which are relevant for representing longitudinal lab measurements. We show how our methods capture standard features very concisely. To better probe the underlying latent subtypes of IBD, we present in Chapter 5 an unsupervised clustering model for combining

multiple data types. We show how this model can uncover relationships between the various features of interest in our patient population.

We then begin to relate our methods to genomics and clinical outcomes. In chapter 6, we perform a large hypothesis-generating analysis to identify associations between genotype and disease phenotype. We also clarify previous associations and identify potential new relationships. In Chapter 7 we tackle an important clinical question of comparing medications and predicting efficacy; we use our developed methods to aid in this analysis. Finally, in Chapter 8 we summarise and conclude, with ideas for future improvements upon these methodologies.

Chapter 2

Previous work

There is growing excitement about the potential of combining genomic and clinical data to uncover new understandings of disease mechanisms, which can be used to build the foundations for "precision medicine" [1, 70–73]. Precision medicine (a term that has now largely superseded the previously used "personalised medicine") describes an approach to medicine that aims to take into account patient-specific factors. Such an approach is not new; doctors almost always try to take individual factors into account when treating patients, and many subtypes of disease have already been identified that are associated with differing treatments (e.g., type-1 and type-2 diabetes). However, the growing ability to capture more in-depth and descriptive information about individuals through electronic health records (EHRs) and genomics means that insights drawn from such information have great potential to improve medicine.

EHRs have been developed for the purposes of aiding physicians in clinical practice, but their nature as a repository of patient information, accumulated during the routine care of patients, also makes them an attractive source of data for research. With legislation in the USA and elsewhere incentivising the implementation of EHR systems, there has been significant growth in EHR uptake, and as a result, much of patient information collected in hospitals is now accessible in electronic formats.

Simultaneously, the declining cost of genome sequencing has allowed for increased availability of genomic information for research. One resulting trend is the creation of large biobanks of linked clinical and genomic data sources. Examples of such EHR/biobank databases include the Electronic Medical Records and Genomics (eMERGE) network, the Kaiser Permanente Research Program on Genes, Environment, and Health, the Informatics for Integrating Biology and the Bedside service created by the National Center for Biomedical Computing in the USA, the China

Kadoorie Biobank, and the UK Biobank [74–76]. However, genome-wide association studies (GWAS) have illustrated that links between genotype and phenotype are more complex than originally thought [29,32]. Indeed, it is likely that important discriminatory factors between patient phenotypes may lie in transcriptional differences (for instance regulated through epigenetics), rather than purely genotypical differences. For this reason, new efforts to use a wider range of "-omic" data sources, such as gene expression and transcription, methylation patterns, and CNVs have also been underway.

One key way in which these data can be used is to define subtypes of disease. The growing availability of EHR data has motivated the development of subtyping based upon clinical features. The goal of such work is to identify groups of patients with consistent patterns in their phenotypic behaviour, which can be used to define distinct patient groups. These patient groups may have differences in the underlying biology of their disease. Likewise, subtyping based upon the growing availability of molecular -omic data types has also been a growing field. In either case, identification of consistent patient subgroups can allow such patients to be treated differently if appropriate and can also help guide future experiments. The hypothesis underlying such work is that many diseases, as currently defined, are actually composed of many, mechanistically-distinct subtypes. However, from an analysis perspective, the optimal way to combine these various data sources to extract clinically-meaningful input is still an active area of research. Here, we discuss such previous work, followed by an overview of machine learning techniques that have been applied to IBD.

2.1 Patient subtyping

Disease subtyping is a developing field, with varying definitions and perspectives of the term "subtyping." Saria and Goldenberg discuss the notion of patient subtyping [71], defining it as "identifying subpopulations of similar patients that can guide treatment decisions for a given individual." They distinguish subtypes from endotypes, which they define as patient subgroups that have a known biologic, mechanistic distinction. On the other hand, Howard et al. use the term endotypes to describe data-derived patient subgroups, despite an absence of (currently known) mechanistic differences [77]. Boland et al. introduce the term "verotype," which they define as a group of individuals that all have a particular and identical genotype, phenotype, and disease subtype [78]. As an example in our application, a group of IBD patients (phenotype) with a deleterious NOD2 variant (genotype) who share the same CRP response to a

particular drug (disease subtype) may be termed a verotype[1]. Here, as in [71], we will use the term subtyping to describe techniques that attempt to identify coherent patient sub-classes. We use the term "phenotypic" to encompass all non-genomic characteristics of an individual; this typically refers to data collected from the EHR.

Our goal in this thesis is to develop tools for patient subtyping and to take the next step of relating these patient groups to underlying biological differences. In our evaluation of the current literature, we find that there are several categories of work that are relevant background material. An overview of the primary papers we discuss is shown in Table 2.1, with further details for each presented in Appendix A.2. We discuss direct modelling of patient subgroups, based upon clinical features, in Section 2.1.1. Early work in this field has focused upon static features, with recent research focused on capturing the time-varying nature of many disease conditions. We are interested in chronic disease, which indeed progresses over time, and so we also describe recent work on time series methods for medical data in Section 2.1.2. A set of methods that fall under the category of "computational phenotyping" is described in Section 2.1.3. These approaches are designed with a slightly different purpose, but they can also be applied to the question of patient subtyping. In addition, because we are also interested in combining our clinical findings with genomic and other -omic data sources, we describe a selection of subtyping approaches that have been undertaken for the purposes of combining many -omic data types in Section 2.1.4. Finally, in Section 2.1.5 we discuss this previous work in aggreggate - we find that cohesive studies that combine all of these aspects are largely lacking in the literature.

2.1.1 Machine learning for disease subtyping

Data-driven patient subtyping has been undertaken for many disease conditions. As an early example, Lewis et al. used k-means clustering to identify subtypes of the early progression of Parkinson's disease [91]. Cho et al. and Castaldi et al. similarly used factor analysis followed by k-means to find subgroups of patients with chronic obstructive pulmonary disease (COPD) [94, 102]. Within cardiovascular medicine, Shah et al. used Gaussian mixture models (GMMs) to model latent groups of patients with heart failure [101]. Howard et al. outline several examples of subtyping of patients with asthma [77].

Newer methodological developments in machine learning have also been applied to the question of patient subtyping. For example, Li et al. focused on diabetes in their

[1]Verotype grows from the term "biotype," which is defined as a group of individuals with a particular genotype together with a particular phenotype [78].

Table 2.1: *Relevant previous work*

Authors	Ref.	Year	Data types	Model	Inference procedure	Number of classes
Phenotyping and time series studies						
Roque et al.	79	2011	Binary	HC over cosine distances	Presume: recursive merges	Inspection
Lasko et al.	80	2013	Time series	Two autoencoders over GP	GPs: ML; NN: stochastic grad. desc.	Set number of features
Ho et al.	81	2014	Counts	Non-negative Poisson tensor decomposition	Alternating minimisation approach	Set
Kale et al.	82	2015	Time series	Three-layer feed-forward NN + causal models	Stochastic gradient descent for NN	n/a
Wang et al.	83	2015	Counts	Constrained non-negative tensor factorisation	ADMM	Set
Pivovarov et al.	84	2015	Counts	Latent Dirichlet allocation	Collapsed Gibbs sampling	Presume: emp. Bayes
Ghassemi et al.	85	2015	Time series	Multi-task Gaussian process	Marginal likelihood (gradient descent)	n/a
Durichen et al.	86	2015	Time series	Multi-task Gaussian process	Marginal likelihood (gradient descent)	n/a
Lipton et al.	87	2015	Time series	LSTM RNN	Presume: stochastic gradient descent	n/a
Lee et al	88	2016	Counts	BN-CTPF	Variational inference	Nonparametric
Gligorijevic et al.	89	2016	Counts	Multi modal NN embedding + K-means	Negative sampling	Prediction performance
Liu et al.	90	2016	Time series	Population-level LDS + residual MTGP	EM	n/a
Subtyping studies						
Lewis et al.	91	2005	Continuus	K-means	Presume: iterative updates	Set
Chen et al.	92	2007	Continous	Hierarchical clustering	Presume: recursive merges	Inspection
Saria et al.	93	2010	Time series	Hierarchical DP over AR processes	Block Gibbs sampling	Set four topics
Cho et al.	94	2010	Continous	Factor analysis + K-means	Presume: iterative updates	Silhouette method
Marlin et al.	95	2012	Time series	GMM	Maximum a posteriori	Set
Ross et al.	96	2013	Time series	DP mixture of GPs with MRF for constraints	Variational inference	Automatic
Pimentel et al.	97	2013	Time series	HC over GPs	Presume: recursive merges	Inspection
Belgrave et al	98	2014	Time series	HMMs	Infer.NET	Model evidence
Ross et al.	99	2014	Time series	Dual BP over GP	Variational inference	Automatic
Doshi-Velez et al.	100	2014	Time series	HC	Presume: recursive merges	Inspection
Shah et al.	101	2014	Continuous	GMM	Presume: EM	BIC
Castaldi et al.	102	2014	Continous	Factor analysis + K-means	Presume: iterative updates	Normalised MI
Schulam et al.	103	2015	Time series	Hierarchical model over GP	EM	BIC
Li et al.	104	2015	Cont., binary	Topological analysis	Ayasdi commercial software	Ayasdi-defined “flares”
Schulam et al.	105	2016	Time series	GP latent variable model / LMM	Variational inference	Continuous latent variable
Coley et al.	106	2016	Time series	Bayesian hierarchical multilevel latent class	Gibbs sampling (JAGS)	Defined as two
Omics studies						
Shen et al.	107	2009	Continuous	Latent Gaussian variable model	EM	Cluster separability plots
Yuan et al.	108	2011	Multinomial	DP with feature selection	Gibbs sampling	Automatic
Kirk et al.	109	2012	Many	DMA with dependencies between datasets	Gibbs sampling	Automatic
Lock et al.	110	2013	Continuous	Matrix factorisation	Iterative updates	Permutation procedure
Ray et al.	111	2014	Continuous	Beta-Bernoulli process	Gibbs sampling	Automatic
Wang et al.	112	2014	Continuous	Network coalescence + spectral clustering	Iterative updates	Presume: eigengap

Abbreviations: ADMM: alternating direction method of multipliers, AR: autoregressive, BIC: Bayesian information criterion, BN-CTPF: Bayesian non-parametric collaborative topic Poisson factorization, BP: Beta process, Cont: continuous, DMA: Dirichlet multinomial allocation, DP: Dirichlet process, EM: expectation maximisation, emp: empirical, GMM: Gaussian mixture model, GP: Gaussian process, HC: hierarchical clustering, HMM: hidden Markov model, JAGS: just another Gibbs sampler, LDS: linear dynamical system, LMM: linear mixed model, LSTM: long short-term memory, MI: mutual information, ML: maximum likelihood, MRF: Markov random field, MTGP: multi-task GP, NN: neural network, RNN: recurrent neural network

2015 study, using recently-commericalised techniques from the field of topological analysis to identify patient subtypes [104]. The authors used topological clustering (Ayasdi, Menlo Park, USA) to identify three distinct groupings of patients, based upon blood test measurements, diagnostic codes, etc. extracted from the EHR at a single time point. In topological clustering, data points are transformed, here via singular value decomposition. The data are then split into overlapping groups according to this output variable. Using a metric of similarity, defined here as cosine distance, each group of points is clustered independently to form sets of nodes. Then, nodes with shared data points (since the groups were overlapping) are connected by edges to create a network. This series of steps produces networks that allow for easy visualisation of points that are similar [113, 114].

In a different approach, Gligorijevic et al. applied the recently-developed word2vec framework[2] in a "disease2vec" approach for the purpose of subtype analysis of sepsis [89]. They compared the common continuous bag of words (CBOW) and SkipGram architectures for a neural network (NN), but as in recent language developments [117], they allowed for multiple "senses" of a word. In other words, they allowed for multiple types of meaning embeddings for the medical concept of sepsis; they termed these models t-CBOW and t-SkipGram. Using diagnostic codes from a large inpatient sample, they were able to use their embeddings as features with improved predictive accuracy than other methods and were also able to identify subgroups of sepsis with distinct characteristics.

In all of these noted cases, the input variables were well-defined disease-specific clinical variables, taken at a static point. Such analysis can be informative, particularly with large sample sizes, but many aspects of disease are time-varying. Clinical time series data are often characterised by intermittent sampling, differing lengths of follow-up, and missingness; these motivate new approaches for handling such challenges within clustering models.

In one of the first examples of patient subtyping with longitudinal data, Chen et al. framed clinical EHR data in terms of a "clinarray" (analogous to a microarray, where substantial informatics development had already been undertaken) to guide their analysis. To summarise lab measurements taken over time for CD patients, they simply took the average over all time, and they then performed hierarchical clustering. They were able to differentiate between severe and less-severe patients (with severity measured by a surrogate marker: the frequency at which lab measures

[2]Word2vec consists of language embeddings that carry semantic meaning [115]. An introductory description of this approach can be found, for instance, in [116].

were taken) [92].

Subsequent analyses have focused on the time series nature of the data more directly. Doshi-Velez, et al. looked at trajectories of patients with autism spectrum disorder by examining International Statistical Classification of Diseases and Related Health Problems (ICD) 9 diagnostic codes [100]. They looked at counts of diagnostic codes over time, where the counts were binned over six-month intervals. Within a collection of 13,000 patients, they identified three clusters of patients, using hierarchical clustering, each of which was found to have distinct clinical signatures based upon these input diagnostic codes (e.g. one group characterised by seizures, etc). Belgrave et al. used a set of hidden Markov models (HMMs) to model differing hypotheses regarding progression of asthma symptoms in children [98]. They found that their cohort of 9,801 children did not support the conventional understanding of asthma symptom progression; they instead found several latent subclasses of symptom progression.

In an example of subtyping using inpatient clinical time series data, Marlin, et al. proposed the use of GMMs with an empirically-defined mean-based prior (parameters estimated through maximum a posteriori, MAP) to deal with the abundance of missing values that are encountered in EHR physiological data [95]. As a pre-processing step, they discretised their time series into hour-long windows by taking the mean over each hour period. They then showed how patient clustering can be informative for secondary predictive tasks by training per-cluster mortality classifiers (a Bernoulli model and LR) and evaluating the predictive ability on new test data: they first calculated the new test patient's probability of belonging to each cluster, and then took a weighted average of the predictions associated with each cluster to come to a final mortality prediction. In a similar scenario, Pimentel et al. used GPs to identify common trajectories of vital signs (e.g., heart rate, breathing rate) following surgery [97]. After fitting a GP to each patient's vital signs, the authors computed a likelihood-based similarity metric between each patient-specific GP (essentially determining the likelihood that one patient's GP accurately models the second patient's time series data). Hierarchical clustering was then used on the values of the inter-GP similarity metric to group these trajectories. Previously-unseen test patient trajectories were compared to the time series clusters to determine if the test patients were similar to "normal" or "abnormal" clusters.

There are also a few examples of involved graphical modelling approaches that have been developed to capture changing patient characteristics over time. Saria, et al. developed a generative hierarchical Dirichlet process model they termed the Time

Series Topic Model to examine common features within their neonatal intensive care unit (ICU) patient heart rate data [118]. Their model drew upon natural language processing (NLP) approaches for clustering documents by topic, with the topics defining distributions over words. By analogy, they assumed that there were an underlying set of "topics" within the neonatal vital-sign signals, and that any given time series was formed through a transition matrix over these topics; any topic was associated with a probability of "words," which were features of the signal. These words were also learned during model fitting. Schulam et al. took a different approach to a time series clustering model, in which they defined a set of generative linear prototype functions to describe the behaviour of individual clinical features over time for patients with sclerodoma (a connective tissue disease) [103]. They incorporated patient covariates and individual variation into their model, the parameters of which they fit using expectation maximisation (EM). Their approach can be described as a latent clustering over GPs. Schulam and Arora have since developed further extensions to this original approach for modelling sclerodoma subtypes, deriving a model that is similar to Lawrence's Bayesian Gaussian process latent variable model or can alternatively be viewed as a linear mixed model extension [105, 119].

Ross and Dy presented a nonparametric model for clustering times-series of COPD patient clinical variables that used a Dirichlet mixture of GPs [96]. The authors performed inference using variational Bayes (VB) methods. They also allowed for domain knowledge through the inclusion of "must-link" and "cannot-link" constraints through a Markov random field, which they used to encourage individual patient's measurements to all be considered as part of a single latent group. Ross et al. further extended this method to allow for certain clinical time series variables to be important only in certain groups, and to allow individuals to belong to multiple groups [99]. They formulated this as a dual beta process over GPs, again performing VB inference to determine the most probable clusters.

Coley et al. described a Bayesian hierarchical latent class model for the purposes of predicting underlying prostate cancer status as aggressive or not, based upon prostate-specific antigen (PSA) measurements over time and other clinical factors [106]. The fact that measurements are being made can be informative, and so they incorporated this aspect directly into their model. They used a multi-level model to represent individual PSA measurements over time; a Bernoulli partially-latent variable to describe cancer state (partially latent because this information is available for the subset of patients who actually undergo surgery); and LR functions at each time step to represent the occurrence of biopsies, their results, and decisions for surgery. (They termed

these latter components of the model an "informative observation process" or IOP.) Relevant clinical covariates were also incorporated. They used Gibbs sampling to estimate the parameters of their model and found that their approach yielded slight improvements upon both a simple LR approach and a version of their model that did not use the IOP components.

2.1.2 Machine learning for clinical time series

While the above methods have been designed specifically with patient subtyping as the primary objective, it is also relevant to discuss recent advances in clinical times series modelling. Such approaches can be used within standard clustering models to identify patient subgroups, much as in Pimentel's clustering over GPs [97]. GPs have also been used to identify patient deterioration much earlier than would be possible using traditional methods [120]. Building upon previous work on single-task GPs, Dürichen et al. used multi-task Gaussian processes (MTGPs) to model patient vital signs, meaning that correlations between vital signs could be captured and used to inform interpolation during, for example, sensor dropout [86]. Ghassemi et al. used MTGPs to model changes in textual ICU patient records over time [85]. The authors first used topic modelling over free-text notes to extract the proportion of topics in each note, and they then modelled the changing proportion of nine of these topics over time per patient using MTGPs. They finally used the MTGP hyperparameters for each patient as a set of features to predict mortality, finding that their approach performed better than standard baselines.

Liu et al. presented a two-step approach that is able to incorporate both individual patient- and population-based information into forecasting predictions [90]. With patient blood counts for inpatients as their features over time, they used linear dynamical systems to model the population-level characteristics. They then built patient-specific MTGPs fit to individual patient residuals. They found that their approach yielded lower forecasting errors than a MTGP fit on individuals without the population component. Liu et al.'s work builds upon the work of Quinn et al., who illustrated the use of switching linear dynamical systems to model neonatal physiology in the ICU [121]. Their approach allowed the model to switch between modes of "artefact" (e.g. opening of the incubator), known abnormalities (e.g. bradycardia), and unknown abnormalities.

In a different approach, Wiens, et al. created a day-by-day patient risk score for becoming infected by *Clostridium difficile* [122]. This risk score was derived from a support vector machine (SVM) classifier with >10,000 features from the patient

EHR (e.g., reason for admission, demographics, lab results, room location, vital sign measurements) as input. They then modelled this risk score as a time series with an HMM. Their methods were able to predict patient risk more successfully than traditional approaches of taking aggregate or daily features.

2.1.3 Machine learning for computational phenotyping

Patient subtyping is closely related to the similar field of computational patient phenotyping, which involves the use of patient records to ascribe disease diagnoses automatically. This is useful for purposes such as identifying cohorts of patients for secondary study (such as phenome-wide association studies or PheWAS[3]). Phenotyping differs from sub-typing in that phenotyping is typically attempting to learn known disease states from medical representations of data. Sub-typing, on the other hand, attempts to uncover new, latent states within a single disease entity; this is hypothesis-generating by nature. However, the input data and methodologies for phenotyping and subtyping are often similar, and in many cases, phenotyping technologies could be used for subtyping even if not explicitly designed as such.

Computational phenotyping has grown recently as a field, in large part due to efforts such as the eMERGE network. This group has created a repository of handcrafted algorithms for defining phenotypes from EHRs, termed Phenotype KnowledgeBase [76]. The manual nature of such work has motivated automatic phenotyping, growing first from supervised learning problems for individual conditions (e.g., [123]), but now also expanding to unsupervised analyses of many disparate diagnoses at once. Such unsupervised analyses are often also applicable for patient subtyping problems.

In one straightforward approach, Roque et al. used hierarchical clustering to define patient subgroups across a set of patients with various psychiatric conditions [79]. The input data consisted of patient diagnosis codes, with distance between patients defined in terms of cosine similarity. Pivovarov et al. took an unsupervised probabilistic approach, using topic-modelling techniques, to identify phenotypes represented in EHR data [84]. They modelled free text, lab tests, medications, and diagnostic codes separately and learned the latent phenotypes present in both an inpatient and outpatient dataset; they termed their latent dirichlet allocation model "UPhenome." Lee et al. further extended the approach of using topic modelling with their Bayesian nonparametric collaborative topic Poisson factorisation model [88]. Their approach

[3]PheWAS are essentially the reverse of GWAS: rather than looking for SNPs that are associated with a given disease, the question is which phenotypes may be associated with a given SNP.

allowed them to not only infer groups of related topics, but also to relate these topics to patient demographics (e.g., age, gender). They were then able to perform patient prediction conditioned upon demographics.

Tensor factorisation methods have been used as well. Building upon their initial tensor factorisation method (Limestone) [124], Ho et al. developed Marble, a sparse non-negative tensor factorisation approach for patient phenotyping [81]. With input data consisting of counts of diagnosis and procedure codes, the authors were able to retrieve some examples of subcategories of disease, such as heart failure with different comorbidities. Rubik is another such tensor factorisation method. It is a scalable non-negative tensor factorisation approach that is able to incorporate known medical knowledge and allows for missing data; the number of groups must be pre-specified [83]. By introducing particular constraints on the tensor optimisation, Rubik is able to investigate a limited subset of diagnoses, which, for example, can enable subtyping within a single condition. As such, the authors presented an example of using their technique to find subtypes of patients with hypertension.

The recent success of NNs in computer vision and natural language processing has motivated their application to patient phenotyping [125]. In the first such application, Lasko et al. developed a methodology for creating features in an unsupervised manner from clinical time series measurements [80]. They fit GPs to sequential measurements of serum uric acid, where the time series were first warped to improve stationarity (the time between close measurements was increased and the time between sparse measurements was decreased). They then provided these probability distributions (in random 30-day sections) to a NN composed of two layered autoencoders to produce a sparse set of time series trajectories. When they projected these features into two dimensions, patients with gout could be distinguished from those with acute leukemia (both of these types of patients exhibit unusual and repeated uric acid measurements), though the entire feature-generation process was unsupervised.

Following on from this, Lipton et al. used long short-term memory NNs to train a multi-class classifier to learn diagnostic codes based upon multivariate clinical measurements from the pediatric ICU [87]. They were able to achieve slightly better performance with their model (with raw time series) than with hand-generated features and a multi-layer perceptron. In an interesting approach, Kale et al. took advantage of the features that can be learned through feed-forward NNs as input into causality analysis to better understand which aspects of the patient trajectory may be driving disease classification [82]. Large sample sizes are needed for success with NNs; even with $> 8,000$ patients, Kale et al. reported a suspected underfitting of

their models due to small sample size.

2.1.4 Machine learning for combining -omics data types

Alongside efforts to develop methods for subtyping based upon clinical features, there has been significant work within the bioinformatics community to develop tools that allow subtyping across various types of -omics data. Such data can include raw genetics, mRNA expression levels, CNVs, protein levels, and DNA methylation patterns, all of which constitute different aspects of a patient's molecular phenotype and genotype. Gligorijevic et al. provide an overview of these various data types and examples of machine learning methods that have been developed for the purposes of disease subtyping [72]. Much of this work has focused on cancer applications, where there is now a history of success in identifying distinct cancer subtypes that respond differently to treatment. One of these early examples involved work by van 't Veer et al. and Van De Vijver et al., who used supervised and unsupervised machine learning techniques on gene expression data from breast cancer cells to provide large improvement over existing clinical prognostic markers [126, 127]. These techniques have now been developed into a commercial tool (MammaPrint, Agendia, Netherlands) that is used for breast cancer prognosis in the clinic.

We do not attempt an exhaustive literature review of this topic here, but rather provide an overview of methods that show the current state of the field. In an early instance of data fusion using kernels, Lanckriet et al. employed a set of kernels (e.g., fast Fourier transform, radial basis function, sequence alignment algorithms), each specific to the data type at hand [128]. They then defined a final fused kernel as a weighted sum of each individual data type's kernel. They used SVMs as their method for classification, but with the modification that the weights over each of the feature kernels must be optimised in addition to the standard SVM parameters. They found that their method generated improved prediction results in comparison to weighting each of the data types equally.

However, most methods relevant to suptyping are unsupervised. Lock et al. presented a matrix factorisation approach, which was able to assess the amount of variation explained by each individual input dataset versus the entire set together, thus motivating their model name Joint and Individual Variation Explained [110]. They employed their model to find common structure across miRNA and mRNA expression for gliobastoma patients. Taking a network-based approach, Wang et al. designed a method they termed Similarity Network Fusion to construct patient-similarity networks that could identify clusters based upon both individual and shared patterns

within datasets [112]. Patient data were clustered together if either their correlation was strong within one of more datasets or if there were consistent weaker correlations across many datasets.

In terms of probabilistic models, Shen et al. presented iCluster, a Gaussian latent variable model to combine data from multiple datasets [107]. They modelled underlying disease subtypes with a continuous latent variable, which they assumed to be identical across datasets. This model has since been used as a comparison method in more recently-developed approaches. Yuan et al. constructed a Dirichlet process clustering model capable of feature selection to manage combining both CNV and gene expression data [108]. However, their model was only capable of handling multinomial input data, requiring them to discretise their continuous features into bins. In another nonparametric model, Ray et al. used a Beta-Bernoulli process to allow for feature selection across their input gene expression, CNV, and methylation data [111]. Zhao et al. presented their Bayesian group factor analysis model, which was an extension of factor models to the case of multiple observation matrices [129]. Their goal was to capture the covariance structure of a low-rank approximation to their original data. They therefore placed a sparsity-inducing prior on the loading matrix: the three-parameter beta prior. Feature selection is a desirable quality when working with very high-dimensional data, as is often the case with -omic data.

These methods all take different assumptions as to how various datasets should be combined. Such assumptions may be valid or not in various applications; Ritchie et al. discuss various ways in which multi-scale -omics data can be combined [22]. However, all of these presented methods are only capable of handling a single distributional type of data (i.e., only continuous or only multinomial). Methods are also being developed for integrating heterogeneous genomic data. Kirk et al. used a Dirichlet-multinomial allocation mixture model (a finite approximation to a Dirichlet process mixture model) to integrate multiple datasets [109]. The underlying idea was that the clustering within one dataset informs the clustering in other datasets, which the authors refer to as "correlated clustering." Their model was capable of handling continuous, binary, multinomial, and time series data.

2.1.5 Summary of clinical and -omics subtyping

In our review of previous work, we found (as in a beta-binomial process) that we could categorise studies as belonging to multiple relevant categorisations. Table 2.2 shows this topic membership. An overview of previous methods in this way is informative, as it shows the growing focus on using time series approaches in patient clustering.

Table 2.2: *Overview of relevant studies*

Authors	Ref.	PhenoT	SubT	Time series	Genetic	EHR/clinical	Multi -omic
Phenotyping and Time series studies							
Roque et al.	79	x	x		x	x	
Lasko et al.	80	x	x	x		x	
Ho et al.	81	x	x			x	
Kale et al.	82	x		x		x	
Wang et al.	83	x	x			x	
Pivovarov et al.	84	x				x	
Ghassemi et al.	85			x		x	
Durichen et al.	86			x		x	
Lipton et al.	87	x		x		x	
Lee et al	88	x				x	
Liu et al.	90			x		x	
Subtyping studies							
Lewis et al.	91		x			x	
Chen et al.	92		x			x	
Saria et al.	118		x	x		x	
Cho et al.	94		x		x	x	
Marlin et al.	95		x	x		x	
Ross et al.	96		x	x		x	
Pimentel et al.	97		x	x		x	
Belgrave et al	98		x	x		x	
Ross et al.	99		x	x	x	x	
Doshi-Velez et al.	100		x	x		x	
Shah, et al.	101		x			x	
Castaldi et al.	102		x			x	
Schulam et al.	103		x	x		x	
Li et al.	104		x		x	x	
Schulam et al.	105		x	x		x	
Coley et al.	106		x	x			
Gligorijevic et al.	89		x			x	
Omics studies							
Shen et al.	107		x				x
Yuan et al.	108		x				x
Kirk et al.	109		x	x			x
Lock et al.	110		x				x
Ray et al.	111		x				x
Wang et al.	112		x				x

Abbreviations: PhenoT: phenotyping, SubT: subtyping.

At the same time, as is evident, there is a divide between subtyping studies that specifically model both clinical and -omic data. This may be because large datasets that combine clinical and multiple -omic data are still not yet available. Indeed, the presence of The Cancer Genome Atlas has undoubtedly spurred interest in combining these various modalities within cancer applications (as was the case for the majority of the -omic studies covered here). In terms of model complexity, cancer is a natural place to begin for -omics subtyping because cancer cells are defined by their abnormal genomic characteristics. The genomic aspects of other disorders are often less well-defined than with cancer, making it less clear which -omic measurements are relevant, and also often resulting in fewer available datasets in the public domain.

While the majority of the current clinical-data subtyping studies focused entirely on the identification of phenotypic subtypes, some took additional steps to relate

these subtypes to underlying biologic differences. For instance, Roque et al. examined common comorbidities within their identified clusters and, using a systems biology approach, constructed protein networks based on possible co-involved proteins [79]. Ross et al. presented odds ratios of three genetic mutations known to be associated with certain forms of COPD for being present in each of their identified clusters; their patient data was derived from a large cohort of genotyped patients [99]. Li et al. assessed the genetic enrichment of SNPs (obtained via Illumina microarrays) in each of their three patient sub-clusters (using hypergeometric tests), which they mapped to gene regions and compared to previous findings in the literature (via the VarDi database). They also performed a pathway analysis for the SNPs associated with each subgroup using Qiagen's Ingenuity Pathway Analysis programme. The three groups were found to differ in terms of genetic enrichment.

While challenging, the relation of latent disease clinical phenotypes to additional -omic modalities and biological underpinnings holds great potential both for disease understanding and improved clinical care. Jensen et al. and Denny et al. provide high-level overviews of the challenges and opportunities involved with combining EHR and genomic data sources [70, 130]. Johnson et al. discuss the many data challenges inherent in retrospective EHR analysis [131].

2.2 Previous work in IBD

We now turn our attention to machine learning as applied to IBD. In general, there is very little previous work in this area, though statistical association approaches have been widely employed to ascertain clinical and genomic relationships.

2.2.1 Supervised machine learning prediction

In the literature, the most prevalent application of machine learning techniques to IBD has involved the diagnostic question of identifying non-IBD versus IBD patients, and within the latter category, the classification of CD versus UC; e.g. [132–137]. Input data for such studies have included features extracted from a wide range of sources, including blood-based antibodies, circulating miRNA, the microbiome, the proteome, and common inflammatory-associated SNPs. In addition, the medical imaging community has examined this problem (and the question of severity) from a magnetic resonance imaging perspective [138–140].

However, while diagnostic tools have great clinical potential, particularly for the question of IBD versus non-IBD, this question is not the focus of our work in this thesis. Indeed, there is increasing genetic evidence that the traditionally-defined

notions of CD and UC are likely to represent a spectrum of disease. Cleynen et al. showed that perhaps a better description, based upon a genetic perspective, is the classification of ileal CD, non-ileal CD, and UC [10]. Future medical treatment may therefore shift according to these changing definitions.

Prediction of disease course and drug response

Several studies have used statistical techniques such as univariate analysis, repeated measures analysis of variance (ANOVA), and multivariate LR to relate serum and faecal biomarkers to severity; we presented the results of this work in Chapter 1.

Wei et al. reported on their ability to predict IBD risk based upon Immunochip genotyping [137]. Using a sample size of over $17,000$ CD cases and $13,000$ UC cases, together with over $22,000$ control cases, they performed a three-staged machine learning pipeline. In the first step, they performed a univariate screening of the $178,822$ SNPs in the dataset by removing all SNPs with a p-value of $> 10^{-4}$ in chi-squared testing and with a minor allele frequency of < 0.01. They next used $L1$-penalised LR to further select a subset of predictive SNPs, and they tested the performance of the best-performing regularised classifier on the test set. The final sensitivity and specificity were 71% and 83%, respectively, for CD.

Machine learning techniques have also been used to predict drug response for IBD patients using laboratory markers. Walgee et al. used a random forest (RF) algorithm to predict patient response to thiopurines based upon patient age and lab tests [141]. Lab markers included levels of neutrophils, WBC count, total bilirubin, sodium, and haemoglobin, and derived quantities such as red cell distribution width and mean corpuscular volume. The group was able to obtain 72% sensitivity and 77% specificity in predicting clinical response, which was better than that possible with the only reproducible alternative method.

In general, previous studies that have examined IBD severity have relied upon traditional statistics such as LR; newer machine learning approaches have not been widely employed.

2.2.2 Unsupervised machine learning

There has been very little data-driven investigation of IBD subtypes. Chen et al. used their clinarray approach, as described in Section 2.1.1, to identify groups of severe and nonsevere CD patients based upon average laboratory test results. Essentially their work showed that patients with many lab measurements (which was used as their indicator of severity) tended to have distinguishable mean lab measurements from less severe patients. While a useful first step, this work was based entirely on

lab measurements, rather than any external representation of disease severity.

The CD Vienna classifications, which were the precursor for the current Montreal classifications, were informed to some degree by large patient cohorts [142]. The Working Party for the World Congress of Gastroenterology, which constructed the guidelines, met multiple times to decide upon the variables to include in the classification scheme. They then amended these variables based upon redundancies or lack of available data as assessed from approximately 5,000 patients from North America and Europe. Thus, though informed by patient data, the classifications were not set based upon a specific clustering or subtyping methodology and were ultimately decided based upon expert clinical opinion. The Montreal classifications were then constructed through further amendments to the Vienna classifications based upon clinical opinion and literature review [6, 143].

Besides these initial studies, to our knowledge there has been no machine learning investigation into IBD subtypes.

2.3 Relation of IBD genetics to phenotype

While there is little previous work on modelling of IBD clinical phenotypes, progress has been made in relating IBD patient genetics to already-defined phenotypic characteristics. A large study in 2016 by Cleynen et al. examined how genetic loci are associated with different classically-defined IBD phenotypes [10]. These phenotypes included the location of inflammation, age of onset, behaviour of the disease, and the requirement of surgery (essentially, those phenotypes captured in the Montreal classification). The authors examined a cohort of nearly 30,000 IBD patients, each of whom had Immunochip genotyping performed. They investigated how the genomic loci were associated with IBD phenotypes by first fitting various model representations of the phenotypes (e.g., grouping levels of the Montreal categories together in several permutations) and then choosing the model with the highest Bayesian information criterion (BIC). For instance, they found that grouping UC disease extent into the two categories of extensive disease (E3) versus proctitis and left-sided disease (E1 + E2) was the best phenotypic representation, based upon the fit to the genetic data. Using these "best-fit" models, they then conducted univariate associations (corrected for the population structure via the first five principal components as well as correlated phenotypes) for each of the genotyped variants. From this analysis, they found that only SNPs in *NOD2*, the major histocompatibility complex, and macrophage stimulating 1 (*MST1*) were significantly associated with age of disease onset and disease location.

They also combined the 193 known risk-associated IBD loci together into a single risk score to examine whether this provided discriminatory power between phenotypes. The authors constructed these risk scores for their phenotypes of interest by performing LR using the 193 risk-associated SNPs (together with the five population-based principal components) onto their binary phenotypes of interest. Using the resulting p-values and odds ratios, they included all SNPs with $p < 0.05$ in their final model, which they designed as a linear model with the following form [144]:

$$r_n = \left[1 + e^{-(\mu_0 + \boldsymbol{\beta}^T \boldsymbol{g}_n - \bar{\beta})}\right]^{-1} \tag{2.1}$$

where r_n is the risk score for a given patient n, μ_0 is the prevalence of the condition of interest, $\boldsymbol{\beta}$ is the vector of log odds-ratios associated with each included SNP, $\boldsymbol{g}_n$ is the vector of a patient's risk-associated SNP counts ($g_{n,i} \in [0, 1, 2]$), $\boldsymbol{f}$ is the vector of allele frequencies derived from a control population, and $\bar{\beta} = \sum_i 2\beta_i f_i$ (essentially, the baseline population risk). The authors found that while their UC versus CD risk score was unable to provide predictive discriminatory power between various phenotypes, the subgroup means did follow a progressive pattern, suggesting that colonic CD is perhaps better described as intermediate disease between UC and CD (as already discussed).

Ananthakrishnan et al. conducted a similar study in 2013 to assess whether patients' cumulative risk for IBD was associated with disease phenotype [145]. The authors assessed phenotype in terms of age of disease onset, disease locations, requirement for surgery, biologics, or immunomodulators, and disease behaviour (again, essentially the Montreal criteria, but with the addition of drug usage). To assess genetic contribution to the overall risk, they summed the log-odds ratio of each of the 163 IBD risk loci, weighted by each patient's number of alleles in each location (i.e., homozygous for a risk allele would be given a weight of 2; heterozygous given a weight of 1, and wild type given a weight of 0). This scoring system therefore differs from that employed in the Cleynen study described earlier, in that it does not incorporate any metrics of baseline allele frequency. In addition, Ananthakrishnan et al. asked a slightly different question than Cleynen et al.: their risk score is a reflection of the risk of IBD incidence, while Cleynen et al. retrained their model onto their phenotypes of interest. Ananthakrishnan et al. studied a cohort consisting of 1,105 patients. They found only a modest effect of increased genetic risk for ileal disease location and early age of onset in CD, but no associations between genetic risk on UC phenotypes. This could indicate that genetic loci associated with IBD-risk are not the same as the

loci involved in the phenotypes that the authors examined, or it could be that larger patient sample sizes are required because the effect sizes may be small.

2.4 Conclusion

As we have discussed, there have been exciting new developments in the patient subtyping field in recent years. However, subtyping methods that are capable of incorporating phenotypic features of many different distributional forms, as well as longitudinal clinical data, are still needed. In addition, few studies have connected their modelled subtypes to disease biology. While in many disease areas, such as asthma, patient subtypes have been well-studied [77], this analysis is lacking in the clinical area of IBD. As such, in our next several chapters, we concentrate on addressing some of these challenges in the context of IBD.

Chapter 3

Introduction to the dataset

In this chapter we discuss the many types of data available for the patients in our cohort. We separate our data into two categories: clinical and genomic data.

3.1 Patients

Our patient cohort consisted of patients who have been diagnosed with CD, UC, or IBDu. All IBD patients who were treated at the Oxford University Hospitals National Health Service (NHS) Foundation Trust, UK, starting from 2010, were approached for consent to participate in this cohort study. The cohort is continuing to grow at the time of writing, as new patients are diagnosed. Appropriate ethical approval was received before cohort collection began (REC 09/H2104/30). This consent model follows that of the UK Biobank, in which patients grant consent for long-term access to their health care records and agree that researchers involved in future projects can have access to their de-identified data and tissue samples [146].

The patients collected in this sample are broadly representative of IBD patient populations in the UK, with characteristics as shown in the summary distributions in Figures 3.1 and 3.2. The gender ratio is approximately equally split, and approximately one-third of patients have required surgery by the time they enter our cohort. The largest peak in age of diagnosis is in the mid-twenties. Unlike the general population, however, the number of CD patients outnumbers UC patients in our cohort.

3.2 Clinical data

While some information is now available in electronic form at the hospital, not all is easily extractable or in a machine-interpretable format. Therefore, various clinical

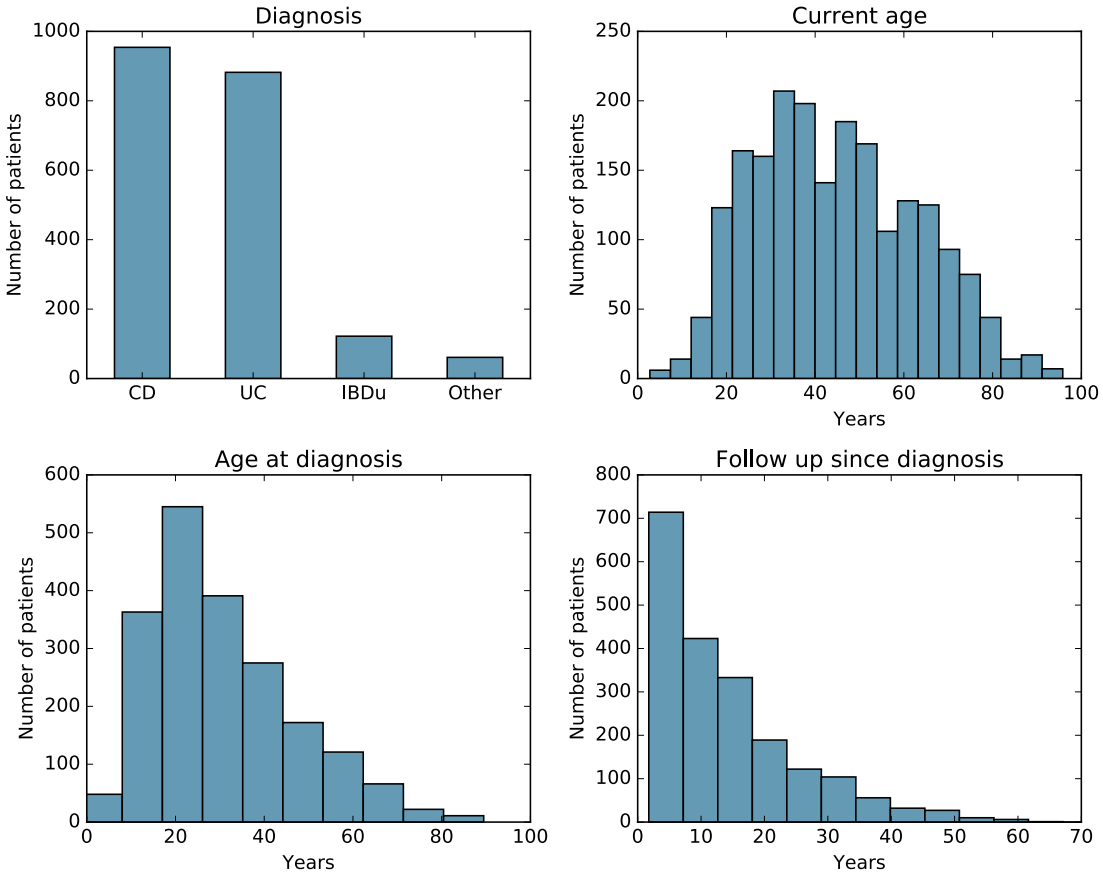
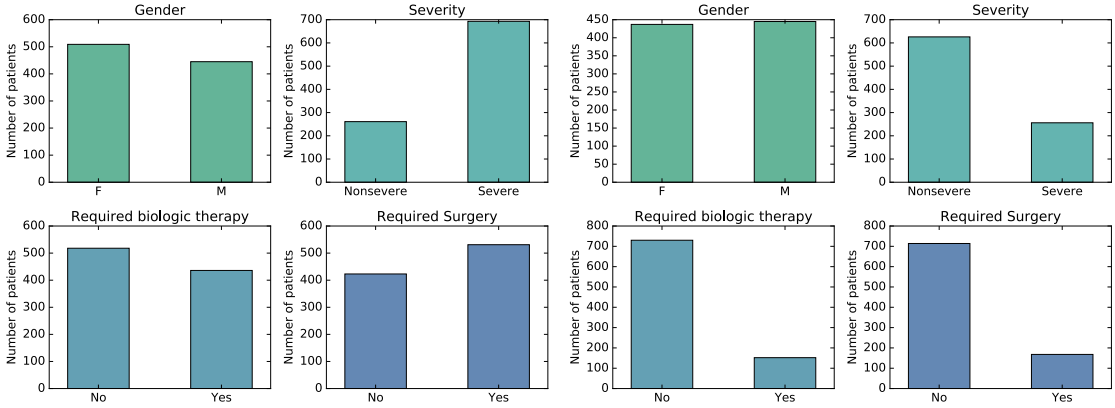


Figure 3.1: *Distribution summaries of the patients recruited into the patient cohort*



(a) CD (b) UC

Figure 3.2: *Summaries of binary characteristics describing the (a) CD and (b) UC patients recruited into the patient cohort. "Severity" defined as ever requiring GI surgery or biologic therapy.*

parameters were manually extracted from the patient notes[1]. The extracted variables were: date of diagnosis, diagnosis (CD/UC/IBDu), year of birth, gender, locations of inflammation in the GI tract (12 binary features), previous usage of IBD medications (11 primary treatments), co-morbidities associated with IBD (e.g. arthritis, anaemia), disease behaviour (inflammatory/stricturing/penetrating disease).

For many patients, large portions of information were missing or difficult to interpret from the available notes (even for a medical professional). For instance, while a stated mention of the use of a drug indicates that, with high probability, the specific drug was prescribed, the absence of a clear mention does not guarantee that the drug was not prescribed from another doctor when the patient lived elsewhere, etc. To indicate the scope of missingness, of the 2439 consented patients, 412 (16%) had a missing date of diagnosis.

Both patient age at diagnosis and the date of patient diagnosis were collected. In cases in which these two ages disagreed by two or more years, we examined the available extracted patient information (e.g., dates of previous colonoscopies; history of laboratory measurements and hospital visits) to make a manual assessment of which date was more accurate.

Due to the manually-extracted nature of this data, there were a variety of formats for encoding information (e.g., misspellings, both numerical and text indicators to indicate the presence of a variable, various date formats, etc). The data were therefore standardised into a consistent format before analysis. Aggregate summary statistics for the subset of our patient cohort with genetic data available (these patients constitute the majority of the analyses undertaken in this thesis) are presented in Table 3.1.

3.2.1 Laboratory measurements

Laboratory haematology and biochemistry blood measurements, hereafter referred to as "labs," were extracted from the Infections in Oxfordshire Research Database (IORD). The latter comprises data from the Oxford University Hospitals Patient Safety System. The lab information within this database includes the date of measurement, the units of measurement, the reference range, the name of the test performed, the study ID of the person whose measurements were taken, and the actual measurement value. Often the reference range was missing, so a consistent reference range was assumed throughout. The earliest labs available were from the year 1987,

[1]This task was performed by medical students, who have access to NHS records. Both paper and electronic notes were used.

Table 3.1: *Summary statistics for clinical variables of interest*

		IBD	CD	UC
	Number of patients	646	346	266
			Percentage	
	Male	46.6	44.2	48.5
Treatments				
	ASA	67.3	52	83.8
	Steroids	65.8	64.7	67.3
	Immunosuppressants	68.1	79.5	53.8
	Biologics	40.7	57.5	20.3
	Surgery	40.9	57.5	22.6
Montreal Extent				
	Ileal		13.3	
	Colonic		23.4	
	Ileocolonic		30.6	
	Upper GI		0	
	Ileal + upper GI		4	
	Colonic + upper GI		4.6	
	Ileocolonic + upper GI		8.4	
	No data		15.6	
Montreal Age				
	<16 years old		29.8	
	17-40 years		57.8	
	>40 years		10.4	
	No data		2	
Montreal Behaviour				
	Non-stenosing/penetrating		48	
	Stenosing		17.3	
	Penetrating		7.5	
	Perianal disease alone		12.4	
	Stenosing + perianal disease		4	
	Penetrating + perianal disease		9.8	
	No data		0.9	
			Mean (std)	
	Age of diagnosis (years)	28.1 (16.1)	24.4 (15.4)	32.4 (15.0)
	Follow-up time (years)	16.5 (11.6)	16.5 (12.0)	17.0 (11.4)

with the final extract taken in November 2015, for a possible maximum of 28 years of lab data.

There were many lab types that were only measured once in the database (e.g. "SERUM CHROMIUM"). Indeed, as Figure **??** shows, the distribution of the number of measurements per lab name decreased sharply after the most-commonly ordered labs. The same lab tests were also represented by many names in the hospital database (e.g. "T.BILIRUBIN", vs. "T. BILIRUBIN", vs. "TOT. BILIRUBIN" and "CRP", vs. "C REACT PROT"). We examined all lab names in the database and manually combined those that represented the same concept (as confirmed by a physician). We included the lab measurements represented in Table 3.2 as candidates in our analyses; these were the relevant IBD labs that were taken for the majority of patients in the

Table 3.2: *Lab Tests*

Test name	Type	IBD behaviour	Normal range
Alanine aminotransferase (ALT)	Enzyme	Increased	$10-45$ IU/L
Albumin	Protein	Decreased	$35-50$ g/L
Alkaline Phosphatase	Enzyme	Increased	$95-320$ IU/L
Amylase	Enzyme	Increased	$25-125$IU/L
Basophils	WBC	Increased	0 - 0.19 10^9 pts/L; $0-2$% WBCs
Bilirubin	Pigment	Increased	$3-16$ μmol/L
Creatinine	Waste molecule	Depends	$54/70-140/150$ μmol/L
C-reactive protein (CRP)	Inflammatory protein	Increased	$0-8$ mg/L
Eosinophils	WBC	Increased	$0-0.55\times10^9$ pts/L; $0-5$% WBCs
Erythrocyte Sedimentation Rate (ESR)	RBC characteristic	Increased	2 - 29 mm/hr
Ferritin	Protein	Decreased	$17/20-230/300\mu$g/L
Haematocrit	Volume of RBCs in blood	Decreased	$0.4-0.5$ L/L
Haemoglobin	O_2-carrying protein	Decreased	$13-17$ g/dL
Lactate dehydrogenase (LDH)	Enzyme	Increased	$90/110-240/250$ IU/L
Lipase	Enzyme	Increased	$6-60$ IU/L
Lymphocytes	WBC	Decreased	$0.72-5.3\times10^9$ pts/L; $18-48$% WBCs
Mean corpuscular/cell volume (MCV)	Average size of RBCs	Decreased	$83-105$ fl
Monocytes	WBC	Increased	$0-0.99\times10^9$ pts/L; $0-9$% WBCs
Neutrophils	WBC	Increased	$1.92-8.03\times10^9$ pts/L; $48-73$% WBCs
Platelets	Cell type	Increased	$150-400\times10^9$ pts/L
Red Cell Count (RC Count)	Count of RBCs per volume	Decreased	$4.5-5.5\times10^{12}$ pts/L
White blood cells (WBCs)	Count of WBCs per volume	Increased	$4-11\times10^9$ pts/L

hospital database.

For many labs, the measured value was recorded as "$< x$", with x being some numerical value. We encoded these measurements as x, and treated $> x$ variables in the same way. We discarded any measurements with non-numerical values, and we converted various entry mistakes (e.g., using commas for decimals and Os for 0s) into the corrected format.

WBCs consist of a variety of subtypes (i.e., basophils, eosinophils, lymphocytes, monocytes, neutrophils), the measurements of which were reported as percentages of the total WBC count prior to 2015. These percentages were converted to scalar values by multiplying the percentage by the number of WBCs. However, for lab measurements taken before 2002, sometimes the timing of the neutrophil, etc. measurements

did not align exactly with the WBC measurements. This was often because multiple WBC measurements would be performed in a day, and the temporal resolution for these earlier measurements was only specified in terms of days (as compared to more recent measurements, which have time-stamps that are exact to the nearest minute). In these cases, we took the average of the WBC measurements collected at ostensibly the "same" time point and multiplied the percentages by this number to yield the final scalar. In the case that WBC or the WBC component measurement dates simply did not match, we only included those measurements where the timing was consistent.

Based upon some lab measurements, it is clear that the type of lab equipment in use has changed over time. For instance, Figure 3.3 shows the strata in CRP measurements as the equipment has changed. There are numerous distinct change points in maximum and minimum measurements. Of the labs examined, this behaviour is most prominent for CRP. The changing maximum cutoff measurement must be dealt with directly during the modelling process, as will be discussed in Chapter 4.

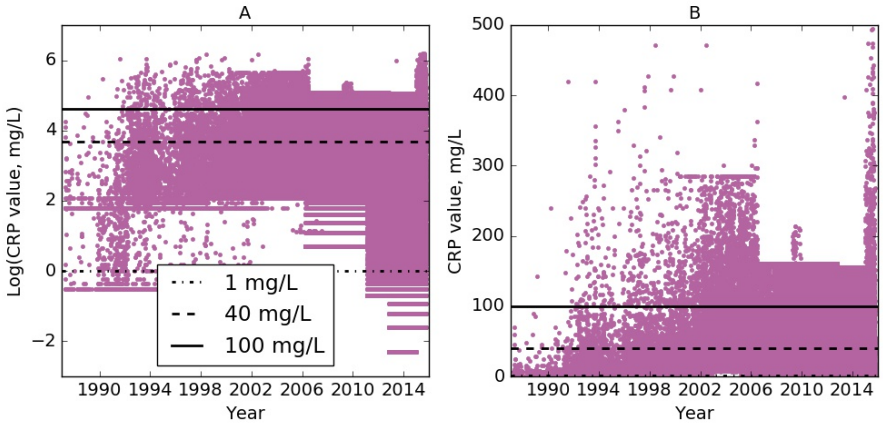


Figure 3.3: *CRP measurements taken over time to illustrate the presence of maximum/minimum cutoffs in the measurements recorded and the changes in the way measurements are recorded over time. (A) shows the cutoffs apparent at low thresholds, with the log(CRP) measurements shown, while (B) shows the cutoffs at high thresholds, with the raw CRP measurements shown.*

3.2.2 Additional clinical variables

We extracted a variety of additional clinical variables from various electronic sources throughout the hospital. We obtained diagnostic codes and admissions, again from IORD; these were available from the year 1997 onwards. We also extracted surgical procedures from Janus, the hospital surgery registry; this information has been available in electronic format since 2007. We obtained records for the fulfillment of the

prescription of biologic drugs from the hospital pharmacy, where the range of data spanned the interval from 2007 to 2014. We extracted notes concerning endoscopy procedures and findings from the endoscopy department, but this information has only been electronically filed since 2013. All of these extracts were from data sources not originally intended for research, and the data contained were of varying formats and interpretability. For instance, endoscopy findings were recorded in free text, but different physicians recorded this free text in different locations in the endoscopy record, some of which were extractable, while others were not.

Preliminary analysis of the diagnostic codes revealed that the vast majority of diagnostic codes that were applied to IBD patients were IBD-related. Table 3.3 presents the 25 most commonly-applied ICD10 codes, together with the proportion of patients who received at least one of these codes. As is evident, co-morbidities (at least as indicated through presence in the diagnostic codes) were rare. The most common co-morbidity was asthma, with 14% of patients having an asthma diagnosis code. Asthma is also an inflammatory disease, and previous studies have identified genomic loci that are associated with both disorders [69, 147]. Arthritis, psoriasis, and other inflammatory disorders are also common comorbidities with IBD, though generally these are seen at a rate of less than six percent in IBD patients, likely explaining their absence here [148].

Table 3.3: *The most commonly-applied diagnostic codes for our patient cohort, together with the proportion of patients who received each code at least once*

Proportion of patients	Code	IBD-related
0.48	Crohn's disease, unspecified	x
0.45	Ulcerative colitis, unspecified	x
0.36	Other chemotherapy	x
0.31	Noninfective gastroenteritis and colitis, unspecified	x
0.29	Acquired absence of other parts of digestive tract	x
0.26	Intestinal bypass and anastomosis status	x
0.25	Personal history of diseases of the digestive system	x
0.24	Crohn's disease of large intestine	x
0.18	Iron deficiency anemia, unspecified	x
0.17	Crohn's disease of small intestine	x
0.17	Other and unspecified abdominal pain	x
0.16	Other Crohn's disease	x
0.15	Polyp of colon	x
0.14	Asthma, unspecified	
0.12	Essential (primary) hypertension	x
0.12	Single live birth	
0.12	Ileostomy status	x
0.12	Mental and behavioural disorders due to use of tobacco, harmful use	
0.12	Hemorrhage of anus and rectum	x
0.11	Anemia, unspecified	x
0.11	Other and unspecified intestinal obstruction	x
0.1	Ulcer of intestine	x
0.09	Nausea and vomiting	x
0.08	Personal history of psychoactive substance abuse	
0.08	Other specified diseases of anus and rectum	x

We used these supplemental clinical variables to create a composite dataset for our IBD patients, with information regarding timing of important clinical events (e.g. surgeries, IBD-related hospital admissions, drug usage, etc.).

3.3 Genomic data

For our patient population, we obtained information about genetic variability through the use of the Immunochip array. Patient genotyping was funded in part by Eli Lilly and Company (Indianapolis, Indiana, USA). As mentioned, the Immunochip was designed specifically for immune-related diseases by a consortium instigated by the Wellcome Trust Case-Control Consortium [26]. The chip is less costly than GWAS chips and includes immune-related variants. It contains probes for 196,524 polymorphisms, of which 0.4% are indels, and SNPs constitute the remainder. For our analyses, we predominantly focused upon the 193 genomic loci already identified to be associated with IBD [11], as presented in Chapter 1. While of course this narrow focus means that we may be missing loci from the remainder of the genome, it also increases the biological plausibility of any findings in exploratory analyses (as is the nature of many of the analyses in this thesis).

Our patients' genetic data was processed as described in Appendix A.3. This resulted in a set of 0s, 1s, and 2s, indicating that the patient was homozygous for the non-Immunochip-detected variant, heterozygous for the Immunochip-detected variant, or homozygous for the Immunochip-detected variant. Notably, the Immunochip-detected allele is not always the one associated with IBD, and some of the IBD-associated alleles are defined in such a way that the dominant allele is protective against IBD (meaning that a patient with a "2" has lower risk for the disease). We undertook a procedure to standarise all of these measurements, as will be explained in Chapter 6.

Table 3.4: *Cytokines involved in IBD inflammatory pathways*

mRNA	Purpose
IL-17A	Th17 pathway
IL-23	Th17 pathway
IL-10	Modulator in Th17 pathway
IL-6	Th1 signature; responds to LPS
IL-22	NOD2 signature; responds to L18MDP
IFN-γ	Th1 signature

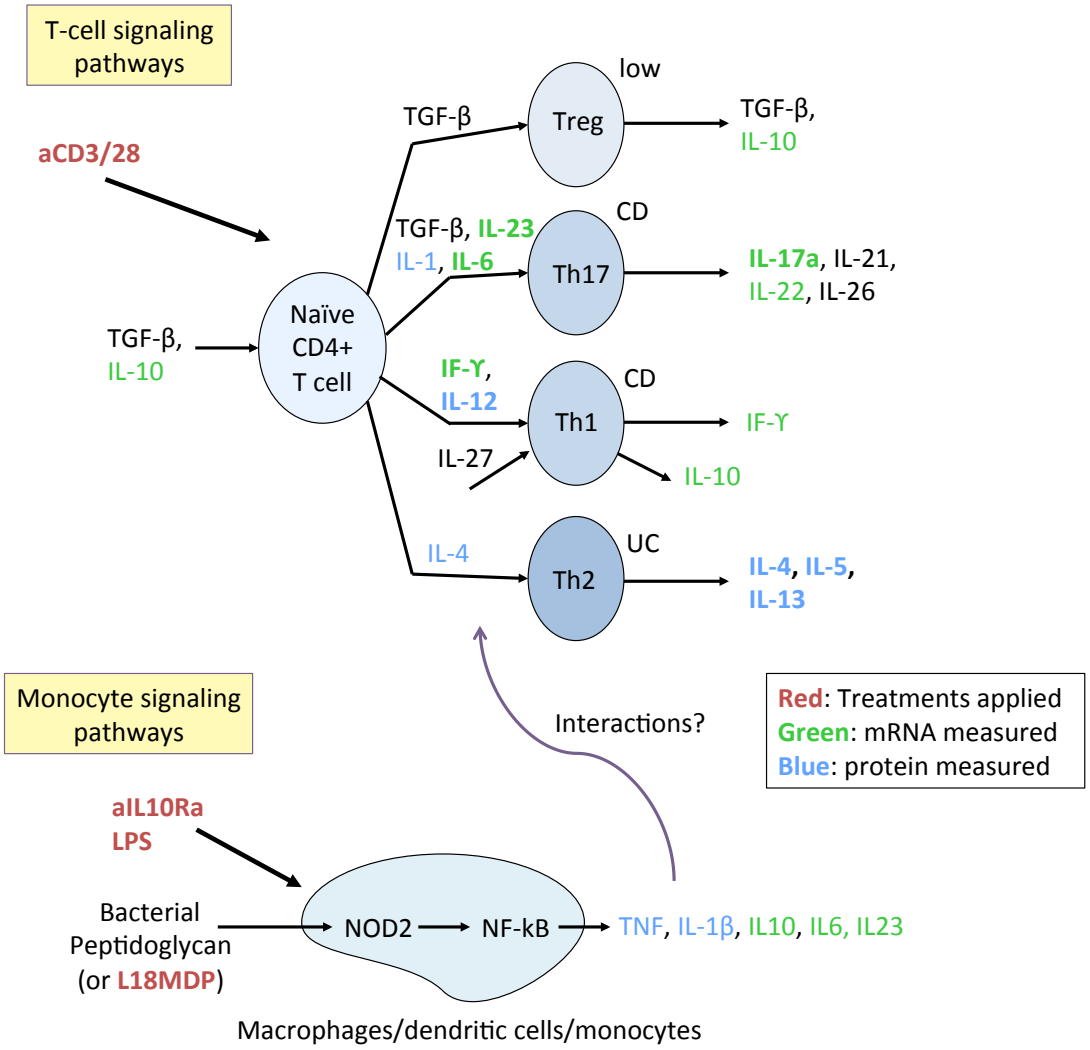


Figure 3.4: *Pathways upon which we focused for our functional analysis of IBD patients. Compounds in red indicate stimulations that were added for our experimental treatments; mRNA and protein levels of the compounds in green were measured (except only mRNA for IL-22 was measured); protein-only levels were measured for the compounds in blue. Abbreviations: Treg (regulatory T cells)*

3.3.1 IBD patient cohort functional data

In addition to genetic variation, we assessed functional (transcriptome and proteome level) variation. As has been mentioned, it is thought that a significant reason for the modest success of research to date in relating genetic variation to complex disease is because transcriptomic and proteomic data have not been adequately considered [22]. Here, our focus was upon several inflammatory pathways thought to be disregulated in IBD patients; these were presented in Chapter 1 and are shown in more detail here in Figure 3.4. The primary cytokines involved in these pathways are presented in Figure 3.4, together with their primary stimuli. As mentioned in Chapter 1, the immune system is stimulated by bacterial PAMPs; lipidated muramyl dipeptide (L18MDP) and lipopolyssaccharide (LPS) are two such stimuli commonly used in experimental settings to stimulate monocytes. Likewise, anti-CD3/CD28 (aCD3/28) mimics the behaviour of antigen-presenting cells to stimulate T-cells. Meanwhile, treatment with anti-IL-10Ra (aIL10Ra) inhibits monocyte response to bacterial stimuli by interfering with IL-10 production. We therefore assessed both levels of the associated mRNA and the protein for these six cytokines under various combinations of these stimulation conditions[2], as shown in Table 3.5. RT-qPCR was used to measure the levels of mRNA, and multiplexed ELISAs (Luminex, Austin, Texas, USA) were used to measure protein levels. Both mRNA levels and proteins were measured in replicate.

Table 3.5: *Experiments performed*

	Treatment	Purpose
1	(Baseline)	No stimulation (control)
2	LPS	Stimulates general inflammatory pathway
3	L18MDP	Stimulates NOD2 pathway
4	aCD3/28	Stimulates T-cells
5	aIL10Ra	Stimulates monocytes mainly, + T-cells
6	LPS + aCD3/28	T-cells + general inflam.
7	L18MDP + aCD3/28	T-cells + NOD2
8	LPS + aIL10Ra	Monocytes + general inflammation
9	L18MDP + aIL10Ra	Monocytes + NOD2
10	aCD3/28 + aIL10Ra	T-cells + monocytes
11	aCD3/28 +aIL10Ra + LPS	T-cells + monocytes + general inflam.
12	aCD3/28 + aIL10Ra + L18MDP	T-cells + monocytes + NOD2

The standard approach for normalising mRNA measurements from RT-qPCR experiments is the "$2^{-\Delta\Delta C_T}$" method, which normalises both by a housekeeping gene

[2]These experiments were conducted by Dr. Maria Quaranta, Nuffield Department of Medicine, and by collaborators at Eli Lilly.

and by a control condition in the experiment [149]. However, given that we were interested in not just the change from baseline control conditions but also each sample's relative expression at baseline, we used only a single normalisation by the housekeeping gene. Our normalisation was therefore as follows: $C_q = 2^{-(\mu_s - \mu_h)}$, where μ_s was the mean cycle point at which the mRNA expression of interest crossed the measurement threshold line and μ_h was the mean cycle point at which housekeeping gene expression crossed the measurement threshold line. The base of two arises because the amount of cDNA in the sample doubles with every cycle.

Each experiment conducted by our collaborators involved the stimulation of patient peripheral blood mononuclear cells (PBMCs) with these treatments. Mononuclear cells are defined as those cells with a single round nucleus, such as lymphocytes and monocytes. The sample roughly is composed of about 50% T-cells, 20% B-cells, 20% lymphocytes, and about 10% monocytes. The PBMCs were obtained from fresh samples of patients' blood when they came into the hospital for an appointment; experiments performed on frozen samples found that monocyte responses were dampened in frozen versus fresh samples. The experiments were run for 16 hours and then RT-qPCR was conducted. These experiments were conducted for 45 patients randomly selected from our patient cohort who also had Immunochip microarray analysis performed.

Despite being a controlled procedure, RT-qPCR and immunoassay experiments sometimes fail due to machine malfunction or poor data quality, meaning that we had instances of missing data (5.4% and 4.1% for the mRNA and protein levels, respectively, across all conducted experiments). We have presented a heatmap (with missing values filled with the column mean) to illustrate the patterns present in this data in Figure 3.5). As is apparent, there were clear effects on the levels of mRNA expression in response to various experimental stimulations. For instance, IL-10 expression was largest during stimulation with LPS and aIL-10Ra (Experiments 2, 5, 8, and 11), which fits with our understanding of the cell signaling pathways (as shown in Figure 3.4). As another example, the baseline conditions (Experiment 1) resulted in lower expression levels in comparison to the remaining experimental treatments (Experiments 2-12). Though these effects were generally consistent across patients, there was some variability. We will revisit these per-patient differences in Chapter 5.

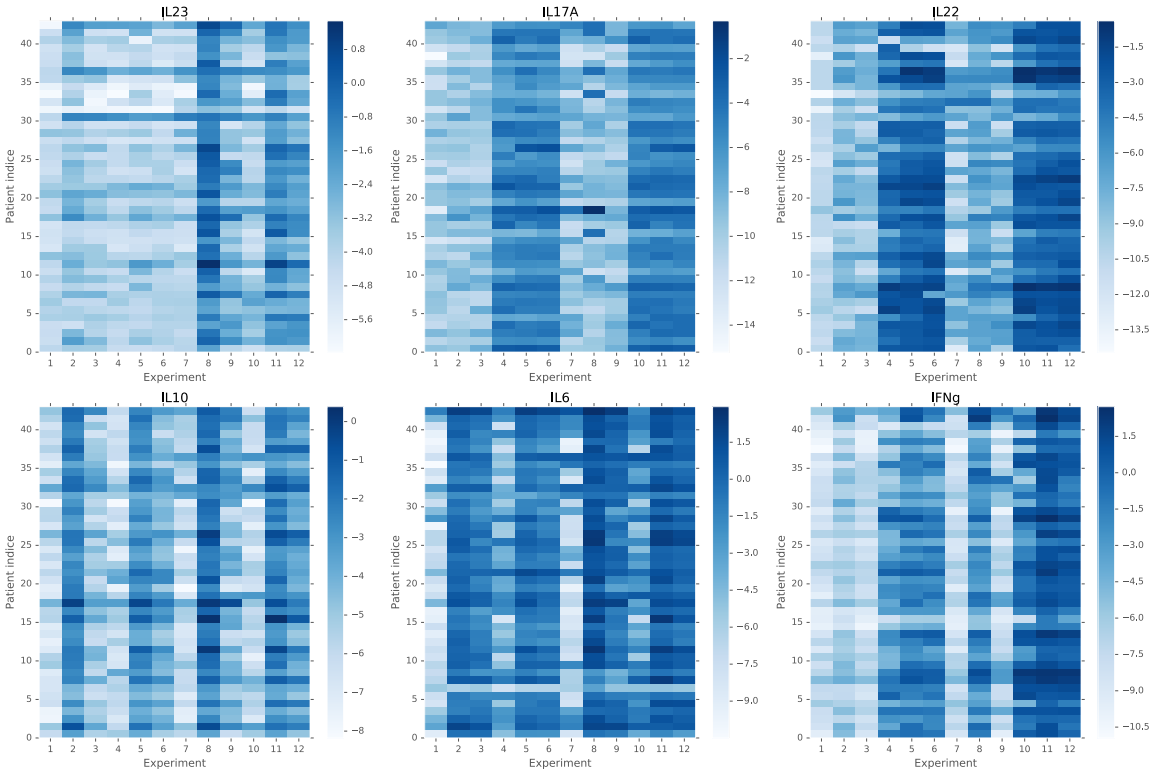


Figure 3.5: *Heatmaps showing the level of mRNA produced in each of the 12 experiments, shown for each mRNA, by patient and experiment (color indicates log(mRNA), where mRNA levels were determined as the number of qPCR cycles relative to housekeeping gene expression).*

3.4 Conclusions

As summarised in Table 3.6, we have assembled a large compository of information related to our IBD patient cohort. In all cases, after merging the various data sources using the necessary hospital identifiers, the patients were assigned anonymised identifiers for use during analysis.

As this introduction to our dataset has indicated, there are many challenges associated with working from clinically-derived data. Information is incompletely recorded, electronically-recorded values may represent meaningless quantities, hospital equipment will change over time, etc. While the nature of our data must be kept in mind through analysis, we uphold that a great deal of insight can still be gained (as will be discussed in the remainder of this thesis). Indeed, a dataset with such a wide range of information concerning a large cohort of IBD patients does not yet exist, to our knowledge, elsewhere in the world.

Nevertheless, there are still many aspects of IBD that have not been captured

in the data to which we have access. For instance, Rogler and Vavricka provide an overview of the concept of the "exposome," a composite of all environmental factors that a given person has been exposed to in his or her life [150]. Many environmental factors have previously been found to be associated with IBD, and twin studies, together with the increasing incidence in countries as they become more "Westernised," suggest that the environment plays a significant large role in the disease [7,13]. Another relevant set of data is the microbiome, which consists of all microorganisms in the body. The microbiome has been shown to be particularly relevant for IBD pathogenesis [7,8]. We lack access to such information, meaning that we can only hope to examine the portion of the disease that is attributable to factors within our dataset.

Table 3.6: *The various data sources that have been gathered for our cohort of* ≈ 2900 *IBD patients*

Phenotypic data	Time period	Type
Clinical data	n/a	Manual extracts
Laboratory measurements	1987-2015	Electronic extract
Endoscopy results	2013-2015	Electronic extract
Pharmacy prescriptions	2007-2014	Electronic extract
Surgeries performed	2007-2014	Electronic extract
Admissions	1997-2015	Electronic extract
Clinical codes	1997-2015	Electronic extract
Genomic data	N	Type
Immunochip data	≈ 900	n/a
Exome data	≈ 650	n/a
Gene expression data	45	n/a

Chapter 4

Univariate point processes for identifying anomalous behaviour

Patients with chronic illness are often monitored throughout the course of their disease. In the case of IBD, one form of this monitoring is through laboratory measurements. In our efforts to model differing patient trajectories throughout the disease, we therefore are presented with the challenge of summarising the critical information from long time-periods of longitudinal lab data.

These laboratory measurements are numerous, sampled at non-uniform intervals, and span time periods of up to twenty years. Different labs capture distinct axes of information regarding patient well-being. Most lab measurements for IBD patients are typically characterised by fluctuations within the clinically-defined normal range, punctuated by occasional above-normal transients or "spikes." Measurements within the "normal" range of human physiology are typically not concerning; rather it is measurements beyond this normal range that are of clinical interest. Both long periods of low-level inflammation and acute severe relapse are important clinical events, meaning that both the number of extreme measurements and their magnitudes are informative. With these motivating factors in mind, we have developed a model to describe the probability of encountering an exact pattern of extreme lab measurements within a given time window. Our model builds upon extreme value theory (EVT), which is a branch of statistics that seeks to model the behaviour of the tails of distributions. Classical EVT has shown potential for being useful in modelling extreme lab measurements [151].

In this chapter we begin with an introduction to classical EVT, which we follow with our model derivation. We then illustrate the benefit of our model, in the form of an anomaly score, in an example with simulated data. We finally apply our model directly to IBD-patient data to show how it concisely captures patterns in longitudinal

lab data. We use our EVT modelling approach to provide lab-derived features in Chapters 5, 6, and 7.

4.1 Classical EVT

There are two primary formulations for classical EVT models: a block maxima approach and an exceedances-over-threshold approach. Our formulation follows that presented by Coles [152]. We start with a sequence of independently and identically distributed (i.i.d.) random variables, $X_1, ..., X_n \sim F$, collected at time points $1...n$. See Figure 4.1 for a visual overview of these models. The block maxima approach attempts to model the maximum of these random variables obtained within a given time frame: $M_n = \max(X_1, ..., X_n)$. The probability of M_n being less than some value z_m is therefore $F(z_m)^n$. However, since we do not know the distribution F, and since estimates of F^n will vary greatly depending upon the estimate of F, we instead focus on modelling the extremes themselves. In the limit of infinite data ($n \to \infty$), M_n will approach a Dirac delta function on the maximum possible value of F. To avoid this, we re-normalise M_n as $M_n^* = \frac{M_n - \mu}{\sigma}$ for two constants, the scale $\sigma > 0$ and the location μ.

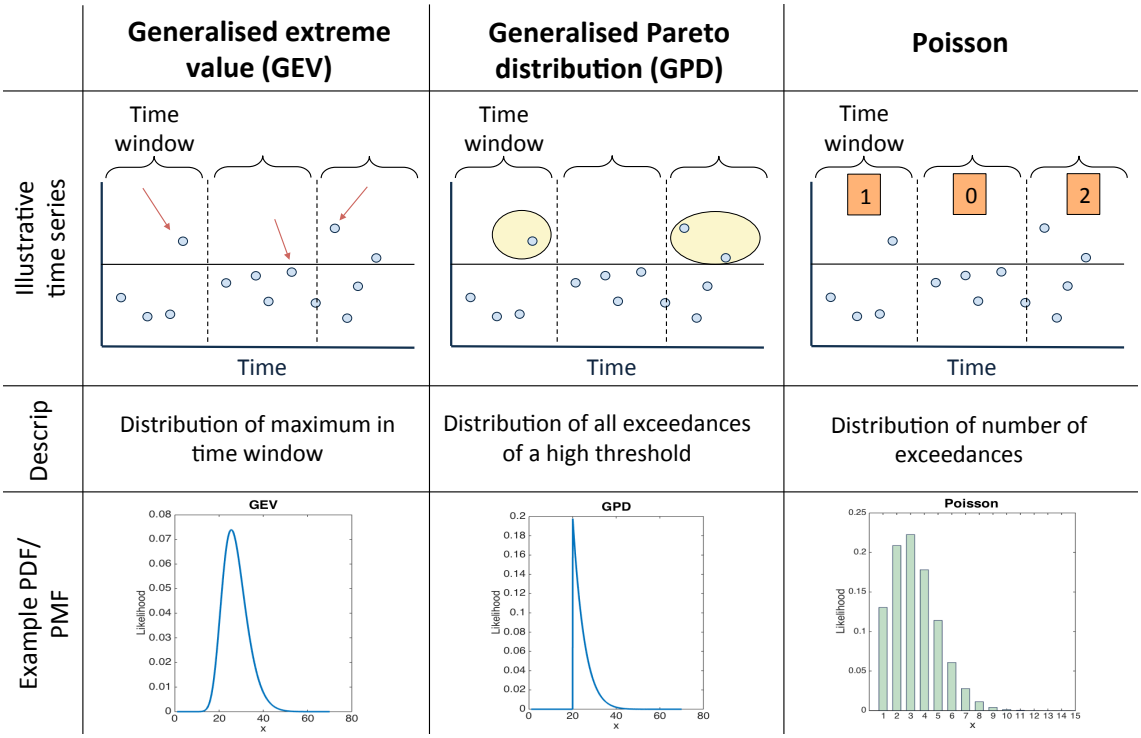


Figure 4.1: *Illustration of EVT distributions.*

It can then be shown that, regardless of the distribution F, in the limit as $n \to \infty$, the distribution of M_n^* approaches the generalised extreme value (GEV) distribution

[152]. The GEV cumulative distribution function (CDF) is as follows:

$$P(M_n^* \leq z) \rightarrow G(x) = \exp\Big(- \Big[1 + \xi M_n^*\Big)\Big]^{-1/\xi}\Big) \tag{4.1}$$

where ξ is termed the shape parameter. $G(x)$ is defined for $1 + \frac{\xi(x-\mu)}{\sigma} > 0$. When $\xi = 0$, the limit is taken to obtain $G(x) = \exp\big(- \exp\big(- \frac{x-\mu}{\sigma}\big)\big)$. The values of ξ correspond to three special cases of the GEV: when $\xi > 0$, the distribution is called the Fréchet type; when $\xi < 0$, the distribution is called the Weibull, and when $\xi = 0$, the distribution is called the Gumbel[1]. This tells us that the maximum value of a set of i.i.d. variables collected over a period of time will tend to follow a GEV distribution.

As mentioned, an alternative perspective is to model points exceeding some "high" threshold. This allows us to use more of the data, rather than simply the maximum values within each time window. Here, we are interested in a slightly different question: the probability of seeing a point above the threshold, u. Formally, we are interested in $P(X > u + y \mid X > u)$. We can obtain this by building upon our findings for the GEV above. Because $F(x)$ is a CDF, it tends towards a value of one for large values of x. Therefore, the Taylor expansion of $\log F(x) \approx F(x) - 1$ for large x. Using this fact to rewrite our formulation of the GEV above, we can obtain [152]:

$$P(X > u + y \mid X > u) \approx H(y) = 1 - \Big(1 + \frac{\xi y}{\sigma_p}\Big)^{-1/\xi} \tag{4.2}$$

$$\sigma_p = \sigma + \xi(u - \mu) \tag{4.3}$$

$H(y)$ is the form of a family of distributions called the generalised Pareto distribution (GPD). As is evident, the GPD and GEV models are closely related; the parameter ξ is shared between the two models, and σ_p is a function of σ and μ. It makes intuitive sense that the distribution governing the probability of high values above some threshold u will be related to the maximum value observed within a given time window.

These two models form the cornerstone of classical EVT. As has been emphasised, these rely upon the assumption of i.i.d. data. In the case of a stationary time series, however, there are temporal dependencies in the data. Nevertheless, given a time-dependency that decreases with increasing time difference, the GEV model is still an appropriate model for block maxima [152]. In the case of the exceedances-over-

[1] These distributions are often presented with slight reparameterisations in different references and textbooks.

threshold model, the GPD model is no longer as appropriate because exceedances will tend to occur in clusters. For instance, a high value is more likely to be followed by another high value. This issue is often managed by "declustering" the data, which we refer to as "deduplicating" so as to reduce confusion with unsupervised clustering methods. This involves using a (typically rule-based) approach to identify clusters of exceedances, where only the maximum value in the cluster is retained in the dataset. The resulting exceedances are thus assumed to still follow the GPD.

4.1.1 EVT from a point process perspective

We will now take a slightly different approach, which is that of modelling a time series as a point process (PP). We will define a set $\mathcal{M}$ (e.g. a defined period of time), upon which there is some stochastic process that generates events. The PP approach is also valid for high dimensional spaces, though we will focus on the two-dimensional case (value vs. time) here. The expected number of points in any subset of the time series is termed the intensity measure of the process: $\Lambda = \mathrm{E}\{N(M)\}$, for every $M \subset \mathcal{M}$. The prototypical example of a PP is the Poisson process, in which we model a homogeneous PP with the number of points according to the Poisson distribution: $P(N(M) = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, where λ is the Poisson parameter (and which is its mean and variance). We can also allow λ to vary as a function of time to allow for a inhomogeneous Poisson process. The general form for the likelihood of the Poisson process is $L(\theta) = \exp\big[-\Lambda(\mathcal{M})\prod_{i=1}^{k}\lambda(x_i)\big]$, with $\Lambda(\mathcal{M}) = \int_{\mathcal{M}}\lambda(x)dx$.

We take this approach because it turns out that the block maxima and exceedances-over-threshold models (the GEV and GPD models just presented, respectively) are special cases of the PP approach. This makes the PP formulation often more useful when modelling in practice because all parameters of both models can be fitted simultaneously. Given a sequence of i.i.d. random variables, the points that appear above a high threshold u converge to a homogeneous Poisson process as the number of points $k \to \infty$ [152]. The intensity measure for a region $[t_1, t_2]$ can be written

$$\Lambda = (t_2 - t_1)\ \lambda(x) = (t_2 - t_1)\Big[1 + \xi\Big(\frac{u-\mu}{\sigma}\Big)\Big]^{-1/\xi} \tag{4.4}$$

The Poisson process likelihood with this intensity measure is then:

$$L \propto \exp\left(-n_y\Big[1 + \xi(\frac{u-\mu}{\sigma})\Big]^{\frac{-1}{\xi}}\right)\ \prod_{i=1}^{N(A)}\sigma^{-1}\left[1 + \xi\frac{(x_i-\mu)}{\sigma}\right]^{\frac{-1}{\xi}-1} \tag{4.5}$$

where n_y is a scaling parameter (the number of years of data) and $N(A)$ is the number

of points exceeding the threshold in region A. The parameters for the Poisson process model for EVT can then be determined via, for example, maximum likelihood.

The PP approach is therefore very powerful because, simply by fitting this single model, we obtain the parameters for the GEV, GPD, and Poisson: $\mu, \sigma, \xi, \sigma_p, \lambda$. We can then use the corresponding models to answer questions about the likelihood of observing a given value as the maximum in a time window, etc.

In choosing a value for threshold u, there is a trade-off between model bias and variance. If u is too low, it will lead to bias in the model because the assumption of a high threshold, which was needed to derive the limiting approximation, will be violated. If u is too high, it will mean that there is very little data above u with which to estimate the model, leading to large variance in parameter estimation. Taking the threshold u as being the 95th percentile of the data is a common pragmatic approach. Another approach is to vary the threshold and assess how the maximum likelihood estimates of the parameters change. If the GPD is an appropriate model, the parameters chosen should be valid for any subset of the data above the high threshold. Therefore, the threshold can be chosen as the point at which the fitted model parameters begin to stabilise. In our clinical application, a "high" threshold has typically already been established through medical experience: lab measurements are accompanied by a reference range that defines the normal physiological range for the population.

4.2 Univariate point process model

We now present our univariate PP model, which provides an extension to the theory of EVT [153]. We are interested in evaluating the probability of seeing sets of abnormally large values in sequences of patient laboratory measurements. We can view our data as a PP, where we say that our PP is a random element Z, with a distribution over all point patterns in a given subspace D. Given a PP, we would like to be able to determine the probability of an example point pattern belonging to a set A, where $A \subset D$. In more detail, we will consider a PP Z_X, consisting of point patterns $\tilde{\mathbf{x}}$, each of which consists of a finite number of i.i.d. points. This model is completely specified by the total number of points in $\tilde{\mathbf{x}}$, defined as the random variable N, and the magnitudes of the points in $\tilde{\mathbf{x}}$, described as the random variable X with probability density function (PDF) $f_X(x)$. We will define $\mathcal{X}_a$ as all point patterns falling within A. Making the assumption that the number of points is independent of the magnitudes of those points, we can then write the probability of a new example

Z_x being an element of $\mathcal{X}_a$ as

$$P(Z_x \in \mathcal{X}_a) = \sum_{k=1}^{\infty} P(N = k) P(\{X_i\}_{i=1}^{k} \in A|_{N=k}) \tag{4.6}$$

where k is the number of points in $\tilde{\mathbf{x}}$.

In our case, we are interested in extremes. As we have discussed, as k approaches infinity, and given a large enough threshold, u, the number of exceedances above this threshold converges to the Poisson distribution and the distribution of the exceedances themselves follows a GPD. Thus $N \sim \text{Poisson}(\lambda)$ and $f(x) \sim \text{GPD}(u, \sigma_p, \xi)$.

Following the derivation framework of [154], we can therefore write

$$P(Z_{x^e} \in \mathcal{X}_a) = \sum_{k=1}^{\infty} \int_{A_k} p_k \prod_{i=1}^{k} f(x_i) \mathrm{d}x_i \tag{4.7}$$

where $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$ (the Poisson probability mass function, as before), $f(x_i) = p_\xi(x_i) = \frac{1}{\sigma_p}\left[1+\xi(\frac{x_i-u}{\sigma_p})\right]^{\frac{-(\xi+1)}{\xi}}$ $(\xi \neq 0)$ (which is the pdf of the GPD), and $A_k = \{\mathbf{x}^e \subset A|N = k\}$ with $\tilde{\mathbf{x}}^e$ being a non-empty point pattern of exceedances. This is the probability that a non-empty point pattern Z_{x^e} of exceedances falls within A.

In the PP literature, the term $\tilde{f}(\tilde{\mathbf{x}}) = k! p_k \prod_{i=1}^{k} f(x_i)$ is called the Janossy density function of the PP. The Janossy density is essentially a representation of the probability of seeing an exact pattern of points $\tilde{\mathbf{x}}$; the $k!$ must be included as the points have no intrinsic order (because they are i.i.d.). From this definition of the Janossy density, we still require a means by which to evaluate the novelty of a given pattern of points - i.e., how likely is a given Janossy density? In order to answer this, we must integrate over all of the point patterns, as is given in Eq. 4.7. This integral has an analytical solution, as will be presented below.

We only are interested in time windows in which there is at least one exceedance (i.e., values above some high threshold u). We can therefore define $G(v)$ as the probability of obtaining $\tilde{\mathbf{x}}^e$ of length k with Janossy density less than v. This corresponds to identifying the patterns $\mathbf{x^e}$ that are "unlikely" with respect to the model, where we condition on the space of non-empty point patterns:

$$G(v) = \frac{P(\tilde{f}(\tilde{\mathbf{x}}^e) \leq v, k \geq 1)}{P(k \geq 1)} = \frac{P(\tilde{f}(\mathbf{x^e}) \leq v, k \geq 1)}{1 - e^{-\lambda}} \tag{4.8}$$

We can use Eq. 4.7 to determine $\overline{G}(v)$, the numerator in Eq. 4.8, where we define the region A_v as consisting of those regions in which $k \geq 1$ and where the Janossy

density is less than v:

$$A_v = \mathbf{x}^e \mid \tilde{f}(\tilde{\mathbf{x}} \leq v\} \;=\; \bigcup_{k\geq 1} \{\tilde{\mathbf{x}}^e \mid k! p_k p_\xi(x_1)...p_\xi(x_k) \leq v\} \tag{4.9}$$

Defining $c_k(v,\xi)^2 = \frac{-2}{1+\xi}\log\Big(\frac{(\sigma_p)^k v}{k! p_k}\Big)$ and
$\frac{u_i^2}{2} = \frac{1}{\xi}\log\Big(1+\xi\big(\frac{x_i-u}{\sigma_p}\big)\Big)$, we can re-formulate Eq. 4.9 as

$$A_v = \bigcup_{k\geq 1}\Big\{\tilde{\mathbf{x}}^e \mid \sum_{i=1}^{k}\big[u_i^2 \geq c_k(v,\xi)^2\big]\Big\} \tag{4.10}$$

Our numerator of Eq. 4.8 is now

$$\overline{G}(v) = \sum_{k\geq 1}^{\infty} p_k \int_{A_v} \frac{1}{(\sigma_p)^k}\prod_{i=1}^{k}\Big(1+\xi\Big(\frac{x_i-u}{\sigma_p}\Big)\Big)^{\frac{-1}{\xi}-1}\mathrm{d}\tilde{\mathbf{x}}^e \tag{4.11}$$

which, after manipulation, we can write as

$$\overline{G}(v) = \sum_{k\geq 1}^{\infty} p_k \int_{\sum_{i=1}^{k} u_i^2 \geq c_k(v,\xi)^2} e^{-\sum_{i=1}^{k}\frac{u_i^2}{2}} u_1...u_k \mathrm{d}u_1...\mathrm{d}u_k \tag{4.12}$$

We can then introduce a k-dimensional spherical coordinate system, with
$u_i = \rho_k \sin(\varphi_i)\prod_{j=1}^{i-1}\cos(\varphi_j)$, $u_k = \rho_k\prod_{j=1}^{k-1}\cos(\varphi_j)$, $\rho_k^2 = \sum_{i=1}^{k} u_i^2$, and
$J_k = \rho_k^{k-1}\prod_{i=1}^{k-2}(\cos(\varphi_i))^{k-1-i}$, which allows us to write

$$\overline{G}(v) = \sum_{k\geq 1} p_k \int_{c_k^2}^{\infty} \frac{1}{2^k\Gamma(k)}\overline{\rho}^{(k-1)} e^{\frac{-\overline{\rho}}{2}}\mathrm{d}\overline{\rho} \tag{4.13}$$

with $\overline{\rho} = \rho^2$. If we make the substitution $w = \frac{-2}{1+\xi}\log v$, we can write:

$$\overline{G}_W(w) = \sum_{k\geq 1} p_k \int_{0}^{c_k^2} \frac{1}{2^k\Gamma(k)}\overline{\rho}^{(k-1)} e^{\frac{-\overline{\rho}}{2}}\mathrm{d}\overline{\rho} \tag{4.14}$$

which gives us our final desired distribution over Janossy densities, where we define $\overline{p}_k = \frac{p_k}{1-e^{-\lambda}}$:

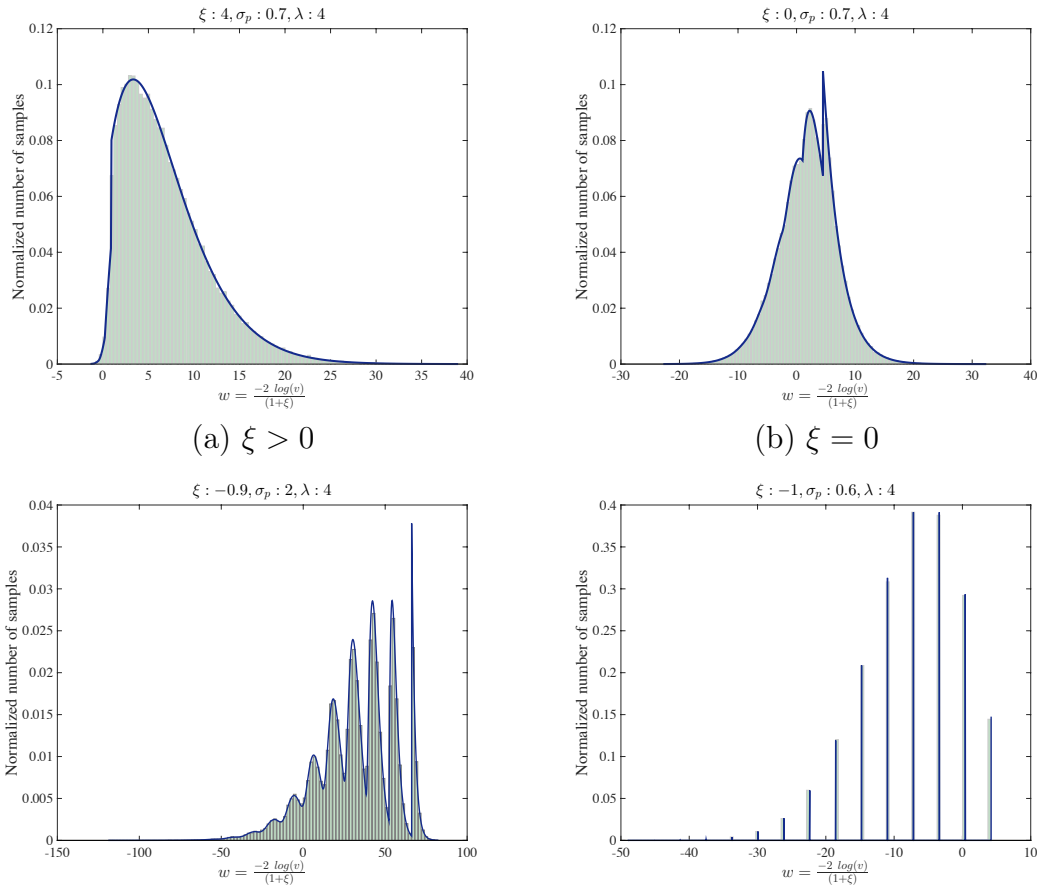
$$G_W(w) = \sum_{k\geq 1} \overline{p}_k \int_{0}^{c_k^2} \frac{1}{2^k\Gamma(k)}\overline{\rho}^{(k-1)} e^{\frac{-\overline{\rho}}{2}}\mathrm{d}\overline{\rho} \tag{4.15}$$

This is a mixture of translated chi-squared distributions, which provides us with an

analytical form for the CDF over our Janossy densities. Due to the form of this analytical solution, special cases occur when $\xi < -1$, $\xi = 0$, and $\xi = -1$. The derivation with these special cases is presented in Appendix **??**.

4.2.1 Model verification

To verify the accuracy of these calculations, we produced various simulations of Janossy densities and compared them with the expected analytical formulations. The simulated densities were obtained by selecting a set of ξ, σ_p, and λ parameters. $100,000$ samples were obtained from a Poisson distribution parameterised by λ to determine the number of exceedances observed. The resulting number of exceedances was drawn from a GPD parameterised by $\mu = 0$, σ_p, and ξ to produce the final set of simulated patterns. The distribution of Janossy densities of these patterns was then calculated.



(a) $\xi > 0$ (b) $\xi = 0$

(c) $\xi < 0$ (d) $\xi = -1$

Figure 4.2: *PDFs over the Janossy density for example ξ parameter values, with transformed $\left(w = \frac{-2}{1+\xi} log\ v\right)$ Janossy scores calculated from simulated point patterns (green histogram) and the analytical solution (blue line).*

Figure 4.2 illustrates several example PDFs that were obtained over the Janossy densities. As is evident, the shapes of these densities are heavily dependent upon the shape parameter of the GPD, ξ. As they are a sum of translated chi-squared densities, the multi-modal shape for some parameter values is expected. It is also worth making a few notes regarding the interpretation of these likelihood functions. Often in alarm-triggering applications, one may be interested in the low-likelihood regions of a likelihood function. For instance, an alarm may be triggered if a new instance (e.g. w_i) has only a 5% change of occuring, i.e., a threshold u is set such that $P(w_i \leq u) = 0.95$. However, with a likelihood function that has such a starkly multimodal shape, as for instance in Figure 4.2 (c), this CDF-based approach may give misleading results.

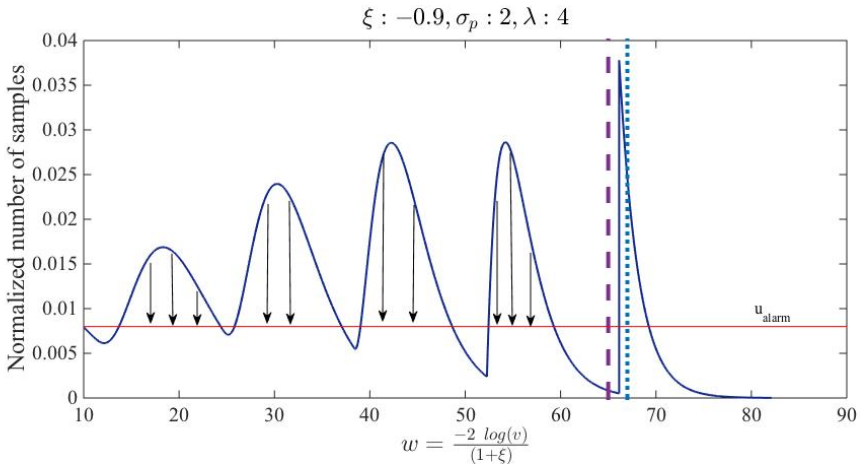


Figure 4.3: *A CDF-based alarm-triggering approach would yield misleading results in a multimodal PDF. An integration from the PDF downwards may be more appropriate. Dashed vertical lines illustrate two example values of w that have differing likelihoods but would have very similar CDF values.*

We use a zoomed-in version of this distribution, shown in Figure 4.3, to illustrate why this is problematic. The $P(w_i > 67)$ is very small in this instance, but the actual likelihood of $w_i = 67$ is very high. Triggering an alarm for $w_i = 67$ is inappropriate, but it would be appropriate for nearby values of w, such as $w = 65$. In such a situation, a more appropriate approach for setting an alarm-triggering threshold u_{alarm} (or several, as may be required with a stongly multi-modal shape) may be to integrate from the top of the density downwards to u_{alarm} such that the resulting area constitutes, e.g., 95% of the PDF: $\int_a^b g_W(w) - u_{alarm} dw$ with a, b being the minimum and maximum solutions of $g_W(w) = u_{alarm}$ (rather than across with increasing values of w, as in a CDF: $G_W(w) = P(W \leq w) = \int_{-\infty}^{w} g_W(t)dt$). Such an operation would then more appropriately capture the regions of w that correspond to unlikely point

patterns and should therefore trigger an alarm.

However, it is also worth understanding what sort of GPD distributions would give rise to these multi-modal distributional forms. Figure 4.4 illustrates GPD distributions for a range of shape parameters ξ. A shape parameter of $\xi = -1$ corresponds to an exact boxcar function, which means that every possible observed value of given variable x is equally likely. As a result, the corresponding likelihood over the transformed Janossy density w is discrete because there are only a finite set of potential point patterns that can be obtained if the value of each point is equally likely (the Poisson parameter is then dominating the Janossy density likelihood function). A shape parameter of $\xi = -0.5$ corresponds to a triangle-shaped GPD likelihood function. This means that the likelihood of extreme points above a high threshold follows a linear function. It is situations between these two instances (i.e., $-1 < \xi < -0.5$) in which the extreme multi-modal behaviour occurs. In most physiologic situations, the tail of the variables of interest follows a more typical smoothly decreasing shape, as in the case of $\xi > -0.5$. Indeed, the range of parameters encountered in our lab data was $0 > \xi > 0.5$, where the likelihood function of our EVT model is more well-behaved.

4.2.2 Model usage

We have presented a model with which it is possible to evaluate the probability of a particular pattern of points appearing within a pre-defined time interval. We can employ this model in a variety of ways to probe our longitudinal laboratory data. Our primary means will be through an "anomaly score" - a score to indicate how "anomalous" a particular pattern of lab measurements is. Our model provides us with likelihoods and/or probabilities associated with a particular pattern of points. We can calculate these likelihoods over consecutive (or overlapping) time intervals, transform this likelihood by taking the negative logarithm, and thereby create a score over time. This score will reflect how unusual a particular pattern of points is with respect to the model, with high values indicating highly unusal patterns and low values indicating patterns of high likelihood (i.e., which match the characteristics of the training data to which the model was fitted). We can then use these anomaly scores to compare the trajectories of different patients, as we will discuss below and in subsequent chapters.

4.3 Model evaluation with simulated data

One task for which our model can be applied is that of identifying patients with similar trajectories over time. We will consider the situation of a time series characterised by

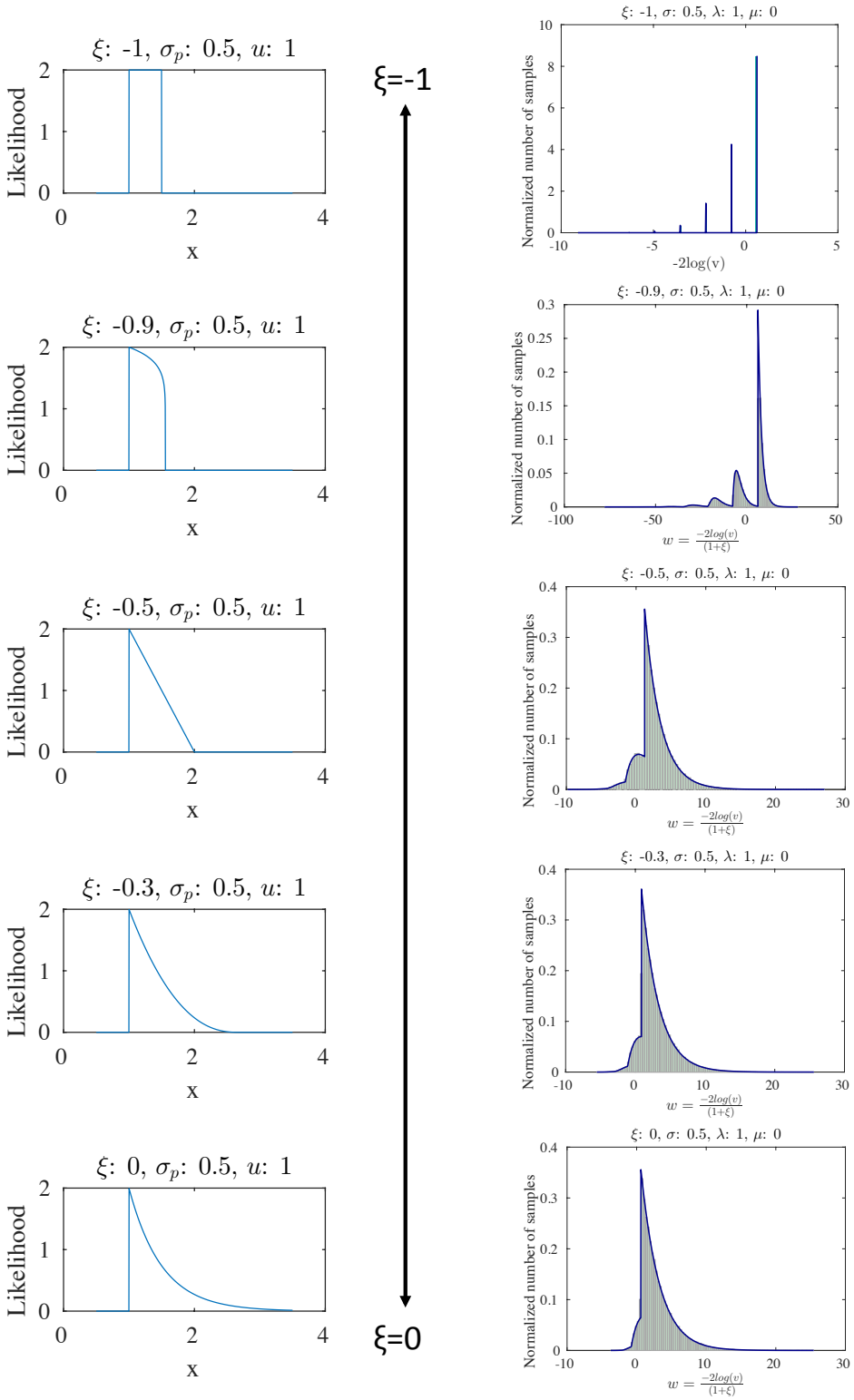


Figure 4.4: *Illustration of GPD distributions (left) and likelihoods over Janossy densities (right) as ξ varies.*

many low-level measurements and occasional large exceedances. CRP is characterised by such behaviour, with four illustrative examples of patient time series shown in Figure 4.5. Here, we will use simulated data so that we can evaluate whether we can detect underlying latent behaviours when we know them to exist in the data. This is a clustering problem: from a clinical perspective, we are investigating whether there are patients who exhibit similar relapse patterns over time. For example, do some patients have large spikes (which we can view as relapses) early in their clinical trajectory, while others only relapse later? Are these patterns shared across the patient population?

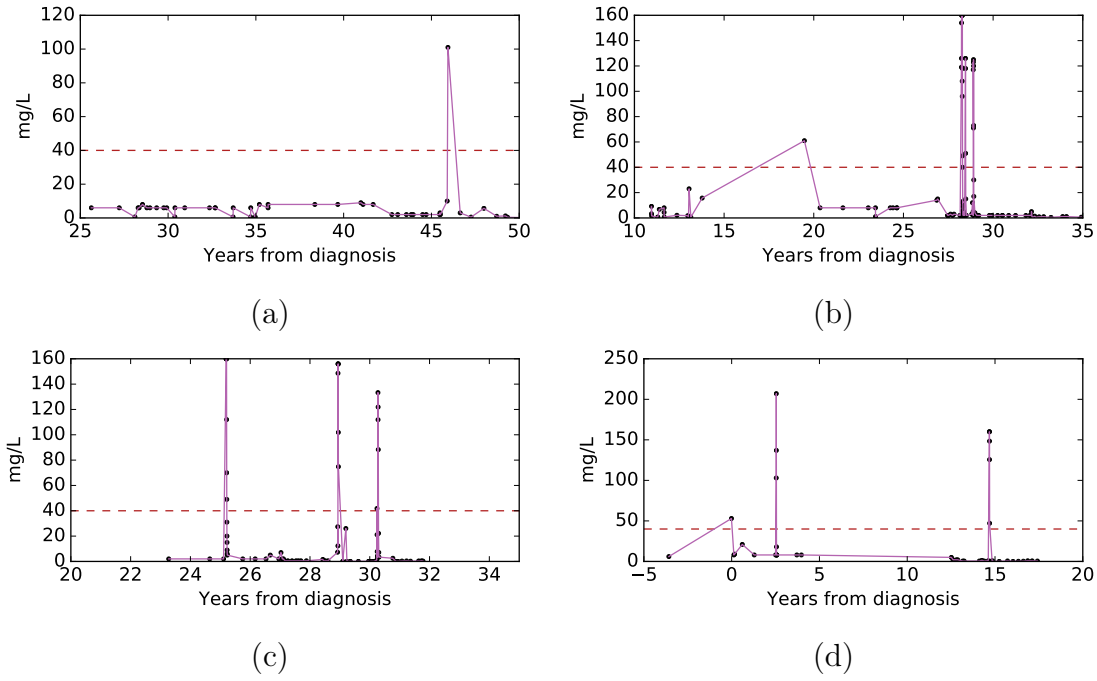


Figure 4.5: *Example CRP time series*

4.3.1 Simulation methods

We designed four prototype generating functions, P_i, $i \in [1, 4]$, to reflect lab measurement time series that could be encountered clinically. Each of these prototype functions was examined over an arbitrary "time period" of 60 months; we defined a threshold of $u = 40$. For all P_i, data points at any time point were drawn from a log-normal distribution f_n with a specified mean ($\mu_n = 5$) and standard deviation ($\sigma_n = 0.3$). This completed the specification of prototype function P_1. For prototype function P_2, we introduced three above-threshold data points in the time period between $0 - 30$ months. Prototype function P_3 was defined by a single exceedance

at a time point between $20 - 50$ months. Prototype function P_4 consisted of three exceedances in the time period between 30 and 60 months. Each exceedance was drawn from a log-normal with mean $\mu_{ne} = 100$ and standard deviation $\sigma_{ne} = 0.075$. From each of these prototype functions, we drew 30 simulated time series, each consisting of 20 data points unevenly sampled over time. We will call these time series our simulated evaluation dataset. Our goal was to cluster these $4 \times 30 = 120$ time series into their original generating prototype functions; we chose to use hierarchical clustering (our objective was not to compare various clustering methods).

As comparison methods for producing features as input for our clustering, we used (a) kernel density estimation; (b) classical EVT; (c) our new univariate PP model of Janossy densities. As training data for all methods, we generated another set of prototype functions, with fewer anomaly events to simulate "normal" patient data.

For (a), we applied a kernel density estimation (KDE) to the training dataset, which we used to evaluate the log likelihood of every point x_i in our evaluation set, $S_{k,i} = -\log p(x_i|\theta)$ for KDE p with bandwidth θ. We set θ according to Scott's Rule, which is of the form $\theta = C\sigma N^{\frac{-1}{d+4}}$, where N is the number of training data points, $d = 1$ is the number of dimensions, C is a constant, and σ is the training data standard deviation. Scott's Rule is an asymptotic estimation of the optimal kernel width so as to minimise the mean squared error between the KDE and the true distribution [155]. In order to provide a fair comparison to results obtained fom our PP method, which evaluates data over a time window (thereby providing smoothing over time), we used a sliding time window and took the average likelihood score across all points in the time window. This produced a trajectory of anomaly scores, S_k.

As previously discussed, classical EVT evaluates tail distributions primarily through two approaches: time-window maxima and exceedances-over-threshold. We remind our readers that the GEV distribution is a distribution over the maximum value observed in a given time window, and the GPD is a distribution over all exceedances above a high threshold. For both classical EVT (b) and our Janossy model (c), we fit the GPD, Poisson, and GEV distributions to our training dataset using the likelihood for windows of extremes in a Poisson process as presented in Equation 4.5. We fit the model parameters using maximum likelihood with five-fold cross validation within the training dataset.

With the classical EVT approach (b), we summarised the information obtained from these three distributions (GEV, GPD, Poisson) by making an assumption of independence between the three. We defined an anomaly score S_c for a given time window as the sum of the log probabilities of (i) seeing a given number of exceedances

(Poisson), (ii) of observing the maximum value or greater in the window (GEV), and (iii) the mean probability of seeing each exceedance value or greater (GPD). With a set of observed points $\mathbf{x_l}$ in a time window, with n_e exceedances $\mathbf{x_e}$, where the ith exceedance is written as x_{ei}, our score is therefore:

$$S_c = -\log\left[P_p(n_e)\left[1 - P_{gev}(X \leq \max(\mathbf{x_l}))\right]\left(1 - \frac{\sum_{i=1}^{n_e} P_{gpd}(X \leq x_{ei})}{n_e}\right)\right] \tag{4.16}$$

where P_p represents the Poisson pmf, P_{gpd} represents the GPD cdf, and P_{gev} represents the GEV cdf. We again used a sliding time window to evaluate the time series in our evaluation set, allowing all extreme points in the time window to contribute towards the score.

For the Janossy model (c), all of the information from these different distributions is already incorporated in the likelihood. Therefore, after fitting the model parameters on the training data, we defined the negative log-likelihood over the Janossy densities (i.e., the derivative of Eq. 4.15) as our anomaly score S_j for the set of points in a given time window. Again, to calculate S_j for the evaluation set data, we used a moving time window.

We then used these scores from methods (a), (b), and (c) (S_k, S_c, and S_j, respectively), at every time point as our feature set for hierarchical clustering. We used a sliding "one-year" window, so the number of features n_f for every trajectory (and for each of these methods) was consistent. We took our initial window such that month "zero" was at the median of the time window and moved the window forward in monthly intervals for a final $n_f = 61$. We evaluated the ability of these various feature-generation methods to resolve the evaluation dataset into its generating functions.

4.3.2 Simulation results

Fig. 4.6 shows the results of our model comparisons with simulated data. The actual simulated data, separated into the true underlying clusters, is plotted in the lower panel ("True cluster 1" - "True cluster 4"). For these simulated data, the Janossy method (S_j) and the classical EVT approach (S_c) were similarly able to resolve the data into its true groups, while the KDE (S_k) approach was less successful: the Janossy method correctly classified every evaluation example into its correct grouping. The classical EVT approach mislabelled one example, while the KDE approach could not distinguish time series from P_1 from those of P_3.

The KDE method's poor performance was likely to be due to the fact that it

was incorporating *too much* information into its summary of the patient trajectory. Since, in this application, it is extreme departures from normality in which we are interested, the KDE approach of modelling the entire data density could actually be detrimental. We note that the separation did not resolve with further branches of the dendrogram tree (i.e., the poor performance is not simply because of an enforcement of four clusters).

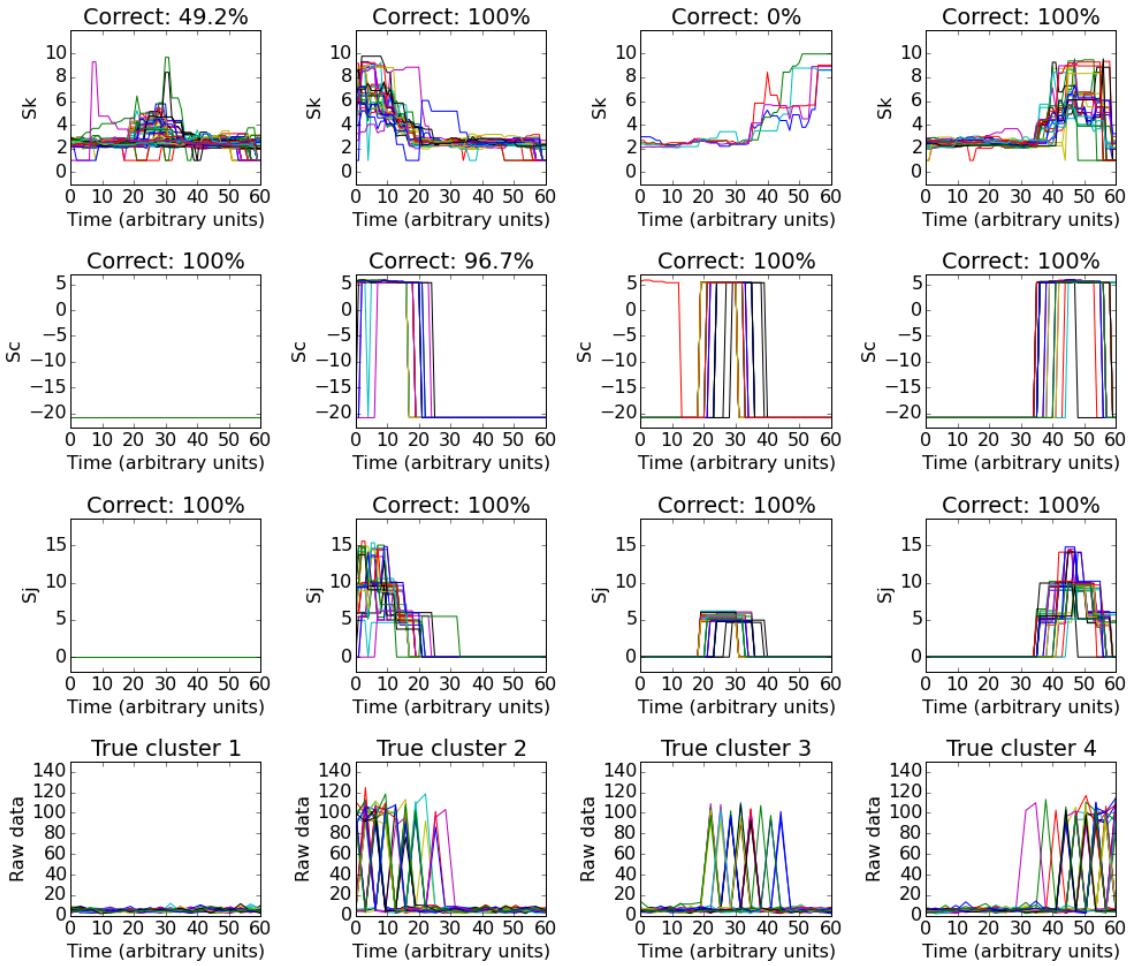


Figure 4.6: *The first $m = 4$ clusters taken from hierarchical clustering of the anomaly-scores S_k, S_c, S_j. The bottom panel illustrates the raw data, grouped into the four true clusters. The clusters generated from the KDE approach (S_k) do not align well with the true clusters, while the classical EVT (S_c) and Janossy (S_j) model anomaly scores better reflect the generating functions. The percentage of correctly-assigned time series are shown above each subplot.*

As is also clear from Fig. 4.6, in comparing the classical S_c and Janossy S_j model features, the Janossy approach was better able to reflect the accumulation of multiple extreme measurements (e.g., higher scores when there are multiple spikes in clusters 2 and 4). This is because it incorporates all of the information about the above-

threshold measurements jointly, rather than summarising the information from each of the classic extreme value distributions somewhat arbitrarily.

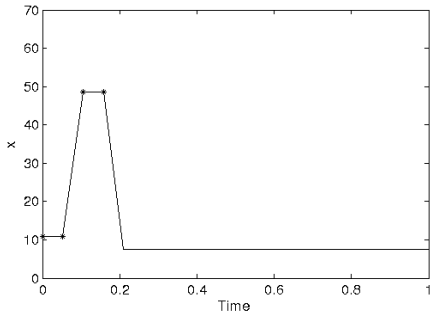
4.3.3 Comparison to classical EVT

We have explained that our model (c)'s approach of calculating the Janossy density and evaluating the likelihood of this density score provides a more principled way in which to capture the behaviours of point exceedances, in comparison to the classical EVT approach (b). This is because model (c) integrates over all possible scores to obtain the likelihood of a particular point configuration. It could be possible that in practice, however, the two methods behave very similarly. In our simulated clustering example above, we have seen that model (c) offers advantages in that it better represents accumulated measurements in a short time frame (i.e., a greater resolution of lab-based severity is shown for the S_j scores in Figure 4.6 than for the S_c scores). In addition, the model (b) requires us to make decisions about how to incorporate the various EVT models. As presented in Equation 4.16, we made the choice to take the mean of the GPD scores here, but other approaches (median, maximum, etc) are possible.

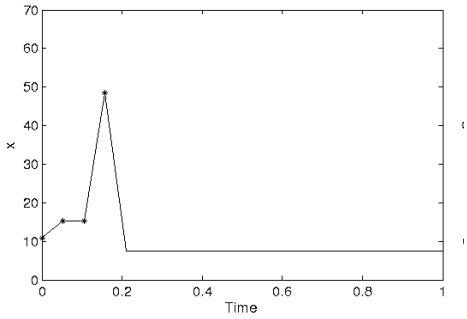
To provide further examples of how the two models (b) and (c) compare, we will now present examples of how sets of data points result in different representations between the two. We note that the "anomality" of a particular point pattern is dependent upon the training data - the score is a representation of how probable a particular set of points is to be generated, given the model fitted to the training data. Here, we train both (b) and (c) models on the same training data and obtain the parameters in the same way. The difference is that the score S_c obtained through model (b) is a combination of the scores obtained from the independent EVT models (GEV, GPD, PP).

Figure 4.7 shows two sets of point patterns. Both result in the same anomaly score S_c according to model (b), but the left panel (1) obtains a higher anomaly score S_j from model (c). This fits with our intution that the additional high-magnitude points that are present in pattern (1) should be considered to be more abnormal than those in (2). The mean of the exceedance points is the same, however, and thus model (b) scores S_c are the same in both cases.

Similarly, Figure 4.8 shows another two sets of point patterns, but here both obtain the same score S_j according to model (c). This is a reflection that, based upon the training data, the set of points with so many exceedances (i.e., as shown in (1)) is just as anomalous as a single large exceedance as shown in (2). This is true even

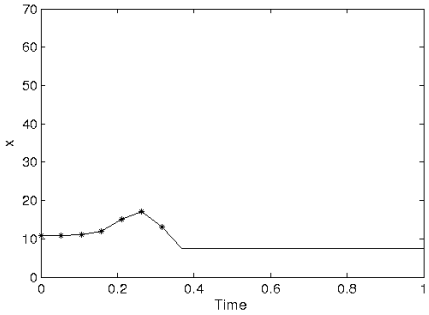


(1) $S_j = 0.76$, $S_c = 1.10$

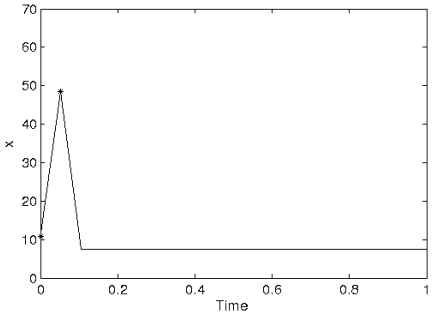


(2) $S_j = 0.69$, $S_c = 1.10$

Figure 4.7: *Illustration of how two sets of points can produce the same score according to classical EVT, model (c) (S_c), but different, and arguably more appropriate, scores with our model (b) (S_j).*



(1) $S_j = 0.68$, $S_c = 0.75$



(2) $S_j = 0.68$, $S_c = 1.32$

Figure 4.8: *Illustration of how two sets of points can produce different scores according to classical EVT, model (c) (S_c), but the same scores with our model (b) (S_j).*

though the magnitudes are not particularly large. For our application to IBD, this is behaviour that we would like to capture: a long period spent in an above-normal range is clinically relevant, as is a single, acute event. Both are representative of anomolous behaviour.

4.4 Application of EVT model to IBD lab data

We have shown that our EVT model can capture relevant lab dynamics in a simulated example. We now illustrate how we can use a severity metric derived from our model to consisely represent standard features that can be extracted from patient laboratory time series.

Our patient population is that described in Chapter 3. For our purposes here, we defined "classically severe" (CS) patients as those requiring surgery or prescription of biologic drugs. The remainder were termed "nonsevere." We recognise that defining

severity in this way results in a bias because patients who have had longer follow-up tend to have accumulated more severe outcomes. To address this partially, we limited our patient population to those with at least three years of follow-up. Another problem with defining CS in this way is that clinical management of IBD has changed over time, with some drug options such as anti-TNF-α not even becoming available until 1999. A more finely-grained outcome, such as time-to-surgery or biologic therapy would be preferred, but such information is not available for our patient cohort before 2007[2]. Nevertheless, this rough CS definition reflects broad categorisations that have been employed previously, as discussed in Chapter 1.

We used those nonsevere patients who had not been genotyped ($N_{ng} = 599$) as our training population from which to fit our EVT model parameters. We used the laboratory measurements of the remaining patients ($N_g = 646$) within a supervised learning framework to classify patients according to our CS definition as severe or nonsevere.

4.4.1 EVT severity metric features

Our purpose in this study was to compare the performance of (i) models built using only our EVT-based approach described in this chapter with (ii) those built with a set of standard time series features (described below). We therefore derived a single "severity metric" for each lab for each patient using our EVT model.

We focused upon five labs for this analysis: CRP, haemoglobin, WBCs, neutrophils, monocytes, and eosinophils. We focused upon these labs in particular because they reflect differing mechanisms of inflammatory behaviour [34], they are commonly employed in the clinic to guide treatment decisions, and they may reflect different characteristics across patients. As illustrated in Figure 4.9, we then defined features for each lab. We chose a threshold u as the maximum or minimum of the clinically-normal range. The choice between maximum and minimum was taken based upon the direction that the lab typically takes during IBD-related events. For CRP, WBCs, neutrophils, and monocytes, this was the maximum (i.e., exceedances above threshold are clinically the most common in IBD), and for haemoglobin, this was the minimum. In order to satisfy the i.i.d. assumptions of our EVT model, we deduplicated all above-threshold measurements taken within ten days of each other, leaving only the maximum (or, for haemoglobin, the minimum) of these values. This is the "deduplicating" step described previously in Section 4.1.

[2]We have taken this approach when possible for specific clinical questions, as will be presented in Chapters 6 and 7.

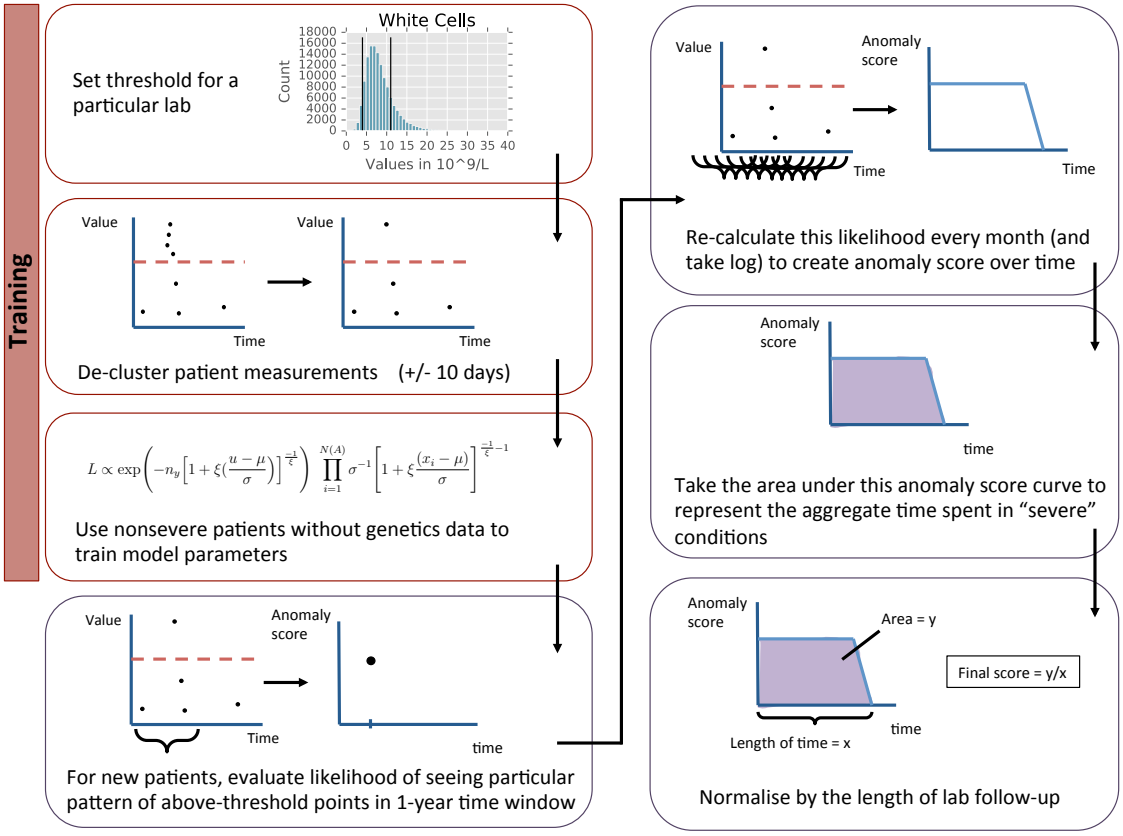
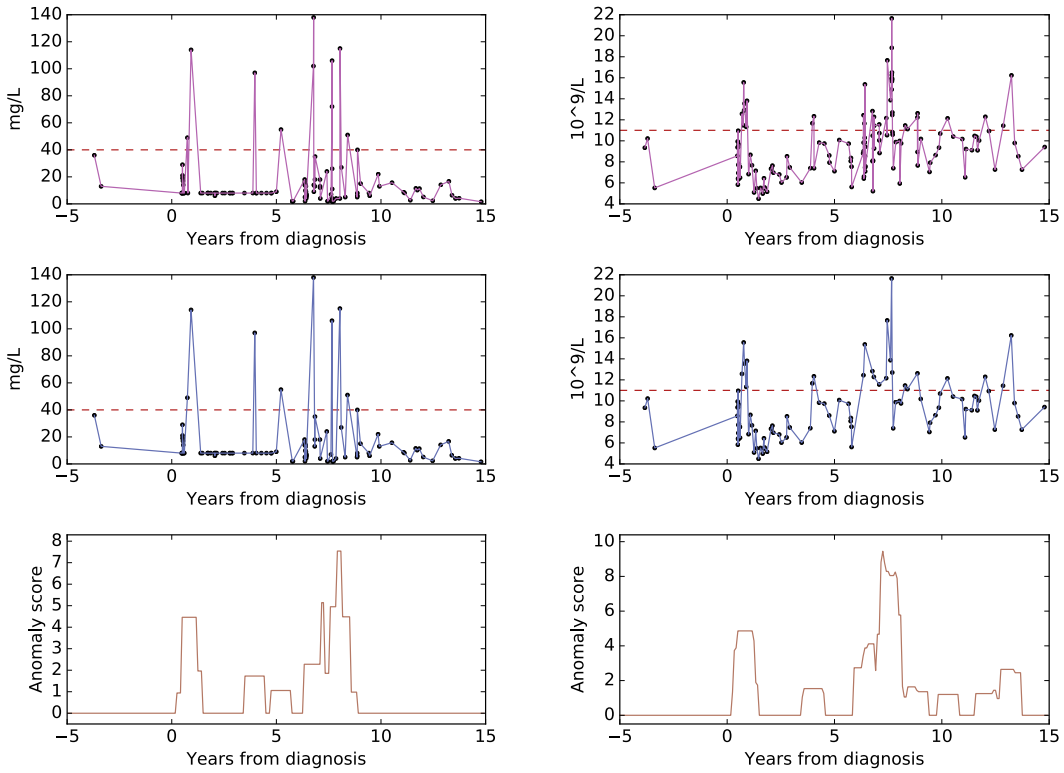


Figure 4.9: *Illustration of steps involved in creating anomaly metric*

We note that for CRP, the clinically "abnormal" cutoff is 8 mg/L, a threshold that would include 45.8% of the lab measurements. We therefore instead used a cutoff of 40 mg/L, which is just above the level of CRP increase typically expected during a viral infection (19.5% of pre-deduplicated measurements were above 40 mgL , corresponding to 5.7% of deduplicated nonsevere measurements) [156, 157]. Similarly, for haemoglobin, we used a cutoff of 11 g/dL because 56.2% measurements were below the clinical abnormal cutoff of 13 g/dL (and where 19.1% of pre-deduplicated measurements were below 11 mgL, corresponding to 7.4% of deduplicated nonsevere measurements).

We then used the lab measurements from all nonsevere patients without genetic data to fit the EVT PP model for each lab type. We used five-fold cross validation to obtain parameter estimates for λ, σ_p, and ξ using maximum likelihood (see Equation 4.5) based upon the deduplicated labs. We used a window size of $w = 1$ (one year).

CRP is again problematic because, as presented in Chapter 3, the maximum reported cutoff has changed over time, resulting in an aggregation of points at the levels of 156, 160, and 285 mg/L. Measurements of values equal to 156, 160, or 285 mg/L



(a) CRP (b) WBC count

Figure 4.10: *Illustration of the EVT trajectories obtained for an example patient. Top: original lab measurements. Middle: deduplicated measurements. Bottom: corresponding anomaly score. (a) CRP; (b) WBC count.*

accounted for 13.8% of the total set of exceedances in the training dataset (0.10% of the entire training set with within-threshold values included). Such an aggregation can result in parameters that do not actually capture the "true" representation of the data. Therefore, we took a simple approach to address this: for all points measured at 156, 160, and 285 mg/dL, we replaced these values with a sample from the empirical distribution of points above these values and added random, normally-distributed noise (with zero mean and 10mg/dL standard deviation). We used these adjusted lab measurements to fit our model parameters.

Upon obtaining our PP model parameters, we then employed our Janossy density model to calculate the likelihood of point patterns for the labs of all patients with genetic data, as shown in Figure 4.9: for a given patient's lab measurement time series, we calculated the log-likelihood of the lab measurements that exceeded the threshold within the first time period w (see Equation 4.15). We created a new time series of these log-likelihood anomaly scores, $s_{n,i}$, for each patient n and lab i by moving

the window forward by amount $\delta_w = 0.0833$ years (one month) and calculating the log-likelihood in this new window. Having obtained $s_{n,i}$, we then created a single metric by taking the area under this curve, normalised by the length of follow-up. We will call this final metric for a given patient n and lab i the $\text{AUAS}_{n,i}$ (area under the anomaly score). We did not include any measurements taken three months prior to diagnosis or sooner so as to only capture events following IBD diagnosis.

An example with real lab measurements is shown in Figure 4.10. We see that the general shape of the trajectory is preserved when moving from the original lab measurements (top) to the deduplicated measurements (middle). However, only a single above-threshold measurement for periods of high measurement frequency remains. Then, in both the CRP and WBC example lab trajectories, we see that the anomaly score captures periods of abnormal behaviour. Time periods with above-threshold lab measurements in the original and deduplicated data correspond to periods of increased anomaly scores. Measurements with values that are more extreme (i.e., further above the threshold) are captured with higher-magnitude anomaly scores, as are periods with a high frequency of measurements.

Having obtained these severity metrics for each lab for each patient, we thus obtained a set of five features per patient; we defined this to be the feature set $F_{sevMetric}$.

4.4.2 Standard lab features

We also extracted the following features from each patient's time course for each of CRP, haemoglobin, WBC count, neutrophils, monocytes, and eosinophils[3]: maximum, minimum, median of all points, median of points within normal threshold, standard deviation, proportion of measurements that occurred outside normal thresholds, number of above-threshold (or below-threshold, if relevant) measurements (after de-clustering) normalised by length of follow-up, and value of the measurement at diagnosis (using the closest measurement within one year of diagnosis). This gave us a set of $6 \times 8 = 48$ features for each patient; we will call this feature set $F_{standard}$. Missing data (e.g., if lab measurements started more than one year following diagnosis) were replaced with the mean of that variable across all patients in the training set (training/testing scheme described below).

[3]We found that eosinophils were maintained within the "normal" clinical range, so we did not consider these lab measurements in the EVT-derived features.

4.4.3 Supervised machine learning methods

We employed several different supervised classification algorithms for our classification task. This comparison allowed us to understand how well the assumptions of each method (e.g., linear combinations of features; independent features) were substantiated in the data. We examined LR, SVM, and RF classifiers.

For the purposes of introducing these classifiers, we will consider a subset of N patients to represent a training set of examples $\mathbf{x}_1 ... \mathbf{x}_N$ with labels $\ell_1 ... \ell_N$, $\ell \in \{0, 1\}$, with 1 indicating severity and 0 indicating nonseverity. Each example $\mathbf{x}_i$ is composed of a vector of D features.

Logistic regression

LR is a linear classification method that optimises a set of weights $\mathbf{w}$ assigned to each input feature to provide the best classification performance using a training dataset. LR can be formulated by considering the sigmoidal hypothesis function:

$$P(\ell_n = 1|\mathbf{x}_n, \mathbf{w}) = h(\mathbf{x}_n) = \frac{1}{1 + \mathrm{e}^{-\mathbf{w}^\mathbf{T}\mathbf{x}_n}} \tag{4.17}$$

which is the probability that the given example is of class 1. An example is assigned to class 1 based upon whether the hypothesis function $h(\mathbf{x}_n)$ is greater than or less than a decision threshold T. We define a cost function that includes a penalty when the hypothesis is incorrect:

$$f(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^{N} \Big[\ell_n \log(h(\mathbf{x}_n)) + (1 - \ell_n) \log(1 - h(\mathbf{x}_n)) \Big] \tag{4.18}$$

Adding an $L2$ regularisation term to discourage the weights from overfitting the data by penalising large values in $\mathbf{w}$, the final equation to be minimised is:

$$f(\mathbf{w})_R = f(\mathbf{w}) + \frac{\lambda}{2N} \sum_{i=1}^{D} \mathbf{w}_j^2 \tag{4.19}$$

where λ is a hyperparameter that governs the degree of regularisation. We also examined LR with the "least absolute shrinkage and selection operator" (LASSO) regularisation method, which imposes the constraint that the $L1$ norm $||\mathbf{w}|| = \sum_i |w_i|$ does not exceed some threshold value. From a Bayesian perspective, $L2$ regularisation is equivalent to placing a zero-mean Gaussian prior on the feature weightings, whereas $L1$ regularisation is equivalent a zero-mean Laplace prior. The prior assumption is

that the feature is not important until the training data shows otherwise.

Support vector machine

The SVM is a classification algorithm that attempts to separate two groups by the widest margin possible in some feature space. A hyperplane defining this separation is determined by maximising the distance between it and the closest training points from each class, which are termed the support vectors. Here we will consider a set of labels $\ell_1 ... \ell_N$, $\ell \in \{-1, 1\}$, in keeping with the SVM literature.

The formulation of an (SVM) begins by considering the distance of each training example $\mathbf{x}_i$ from the hyperplane $y(\mathbf{x}_n) = \mathbf{w}^T\mathbf{x}_n + b$, where b is a bias parameter and $\mathbf{w}$ is again a vector of weights. This distance is written as $\frac{|y(\mathbf{x}_n)|}{||\mathbf{w}||}$. This is subject to the constraint that $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$ because the goal is to classify all examples correctly. The data are not always linearly separable, however, which is taken into account through the introduction of a "slack variable" for every training example, ξ_n, and a cost hyperparameter, C. The slack variable $\xi_n = 0$ if the example datapoint $\mathbf{x}_n$ lies on or within its correct boundary, and $\xi_n = |\ell_n - y(\mathbf{x}_n)|$ otherwise. The parameter C penalises misclassified examples. C is analogous to a regularisation parameter in that lower values of C correspond to decision boundaries that are more slowly-varying in the feature space, with less curvature (because misclassifications are not penalised heavily), and vice-versa. The constraint is therefore $\ell_n y(\mathbf{x}_n) \geq 1 - \xi_n$. As the goal is to maximise the distance between the hyperplane and the closest training example, which requires maximising $||\mathbf{w}||^{-1}$, this is equivalent to minimising $||\mathbf{w}||^2$, where the square is introduced to avoid taking the root in $||\mathbf{w}||$. This therefore requires the minimisation of $f(\mathbf{w}) = C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$, which is referred to as the primal form of the classifier. The primal form can be re-written in terms of the feature vectors themselves in the dual form, which requires maximisation of:

$$f(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{j,k=1}^{N}\alpha_j\alpha_k\ell_j\ell_k\mathbf{x}_j^T\mathbf{x}_k, \quad 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^{N}\alpha_i\ell_i = 0 \tag{4.20}$$

where α is another vector of weights, with $\mathbf{w} = \sum_{i=1}^{N}\alpha_i\ell_i\mathbf{x}_i$[4]. The dual form allows for the use of the "kernel trick" to project data into a high-dimensional space, in which the two classes may be linearly separable. The kernel trick is a method by which, rather than using the actual vector of features that define each $\mathbf{x}_n$, a kernel function is used that describes the features of each example in relation to each other.

[4]In deriving the dual form from the Lagrangian, α is the vector of Lagrange multipliers.

Through Mercer's theorem, any positive semi-definite kernel function corresponds to a high-dimensional space, for which $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})\dot{\phi}(\mathbf{x}')$ and where $\phi(\mathbf{x})$ is some mapping from our original data space to the higher-dimensional space. That is, we can avoid operating in the high-dimensional space because we require only the dot product in Eq. 4.20, and our kernel function gives the scalar product in that space. The Gaussian radial basis function (RBF) kernel is one of the most-commonly used kernels because of its straightforward interpretation as a similarity metric between two points:

$$\mathrm{k}(\mathbf{x}, \mathbf{x}') = \exp\big(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\big) \tag{4.21}$$

Random forest

RFs are ensemble learners, which means that the RF prediction is based upon the votes of a committee of many weak "base learners." The base learner for a RF is a decision tree, which is a branching composite of binary decisions so as to separate the data into the classes of interest. At each node of the tree, a single feature d and threshold t are chosen to separate the input examples $\boldsymbol{x}$ such that the resulting split has minimal impurity (in terms of the class labels). Impurity can be defined in multiple ways; here we have used the entropy H, which in a two-class learning situation is:

$$H = -\sum_{c=1}^{2} \hat{\pi}_c \log \hat{\pi}_c \tag{4.22}$$

where $\hat{\pi}_c$ is the proportion of examples in class c. (Minimising entropy is equivalent to maximising information gain.) The tree is constructed in a greedy manner, where the locally optimal d and t are chosen at each split. Decision trees tend to overfit, thus motivating the RF approach.

In a RF, each decision tree is formed from a random subset of the available features and a random subset of the available training set. The RF is built upon the concepts of boosting and bagging: boosting is the idea that a useful classifier can be constructed from many weak classifiers (i.e., the classifications produced are correlated with the true label, but not by a margin much larger than chance). Bagging (bootstrap aggregating) is the technique of repeatedly sampling from a larger set of training examples to construct a model (bootstrapping), and then aggregating the results through a majority-vote (aggregating). The number of trees built and the number of features selected in each tree are both controlled by tuning hyperparameters within the RF algorithm.

Features can be assigned an importance metric based upon their impact on de-

creasing the node impurity, weighted by how important a given node is in classification. This is formulated as

$$I_{DT} = \left(H_{pre} - \sum_{s=1}^{2} H_{post,s}\right) \times P_n \tag{4.23}$$

where P_n is the proportion of samples that reach the given node, H_{pre} is the entropy of the node before the split, and $H_{post,s}$ is the entropy of the child node s after the split. These I_{DT} scores are then averaged over all of the built decision trees to produce a final I_{RF} importance weighting.

4.4.4 General training formulation

In our classification task we have imbalanced classes, with a ratio of approximately 1:4 for nonsevere:severe for CD. (The opposite trend is the case for UC, with a 2:1 ratio for nonsevere:severe.) While such an imbalance is relatively mild and does not motivate the use of novelty detection or one-class classification, it could result in bias for our discriminative classifiers. The preferred approach for handling imbalanced classes is debated in the literature, with various methods having examples of success [158, 159]. The optimal choice likely depends upon the specific problem at hand. General approaches can be divided into two categories: sampling-based adjustments and algorithmic adjustments. Sampling options include over-sampling the smaller class, under-sampling the larger class, or a combination. Algorithms such as the synthetic minority over-sampling technique (SMOTE) involve the addition of new, synthetic samples to the smaller class, constructed by interpolating between existing examples. Algorithmic adjustments include placing a larger penalty upon misclassifications of the smaller class. In addition, the metric by which the algorithm is evaluated can be adjusted to better capture the performance of interest. For example, the F-score is a commonly used metric to capture a better representation of model performance than accuracy or area-under-the-receivor-operator-curve (AUROC). It combines the positive predictive value (i.e., precision) and sensitivity (i.e., recall) into a single metric, which can be weighted according to which of these is viewed as most important for the specific application.

In our application, we do not have a clear cost associated with a misclassification, meaning that we would have to introduce arbitrary costs in both the F-score and algorithmic adjustment approaches. In addition, our classes are typically highly overlapped, meaning that the introduction of synthetic examples is a dangerous procedure. Therefore, to address the class imbalance in our data, we randomly subsampled from

the class with a larger number of patients to produce a dataset with equal numbers of each class. We then used this balanced dataset for training and testing our models. However, similar to a bootstrapping approach, we repeated this process many times with different samplings of the larger class to provide robust estimates of model performance.

We trained each algorithm, A, on 80% of our balanced dataset and tested on the held-out 20%. During training, we used internal five-fold cross-validation to determine the SVM RBF width σ (see Eq. 4.21), the SVM soft margin parameter C (see Eq. 4.20), the LR regularisation parameter λ (see Eq. 4.19), and the number of RF trees and proportion of features selected. These optimised parameters were then used to train a final algorithm using all the training data (i.e., all of the 80% of the data that were previously used for cross-validation). This final algorithm was then used for prediction on the "held-out" 20% data in the test set. We repeated this process N_I times for our $N_A = 4$ algorithms, as presented in Algorithm 1.

The mean and standard deviation of the area under the receiver-operator curve (AUROC) produced from each method was subsequently calculated across the N_I iterations, allowing an assessment of the variation in the process due to the stochastic selection of the training and test data.

Algorithm 1 Learning framework

for $i = 1 : N_I$ **do** % For N_I iterations
  Randomly subsample from larger class to create balanced dataset
  Split dataset into 80%:20% training:test set
  for $A = 1 : N_A$ **do** % For every classification algorithm
    Five-fold cross-validation within training set to optimise hyperparameters
    Use hyperparameters to train model using entire training set
    Evaluate performance on test set
Take average performance across N_I iterations, for each algorithm A separately

4.4.5 Results

We will present results related to each feature set individually, followed by a comparison of the feature sets in terms of their performance in the classification task of determining traditionally-defined disease severity.

Standard features

Within our feature set $F_{standard}$, many of our input features captured related measurements. Correlations[5] between the $F_{standard}$ features (as well as the $F_{sevMetric}$ features) for CRP and neutrophils are shown in Figure 4.11. These correlations help to reveal some of the underlying structure of the labs. For instance, the number of greater-than-threshold events correlated with the proportion of beyond-normal measurements and the AUAS (Spearman correlation coefficient $\rho > 0.7$ for both CRP and neutrophils). The minimum of the variable correlated only weakly with the other features ($\rho < 0.61$ for the correlation between the minimum and all other features). The maximum correlated with the standard deviation ($\rho > 0.9$), which makes intuitive sense for labs in which the IBD-related events are defined by exceedances of the threshold. Interestingly, however, diagnosis values did not strongly correlate with any of our other measured features ($\rho < 0.4$ for the correlation between the diagnosis value and all other features). (Haemoglobin exhibited the opposite characteristic for some features, noting that it is shortfalls that are relevant for this variable.)

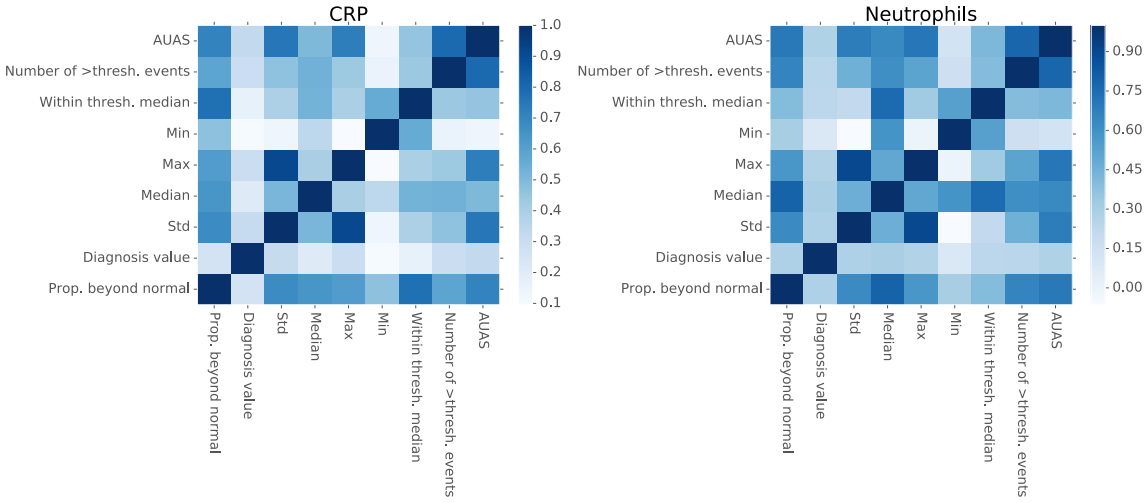
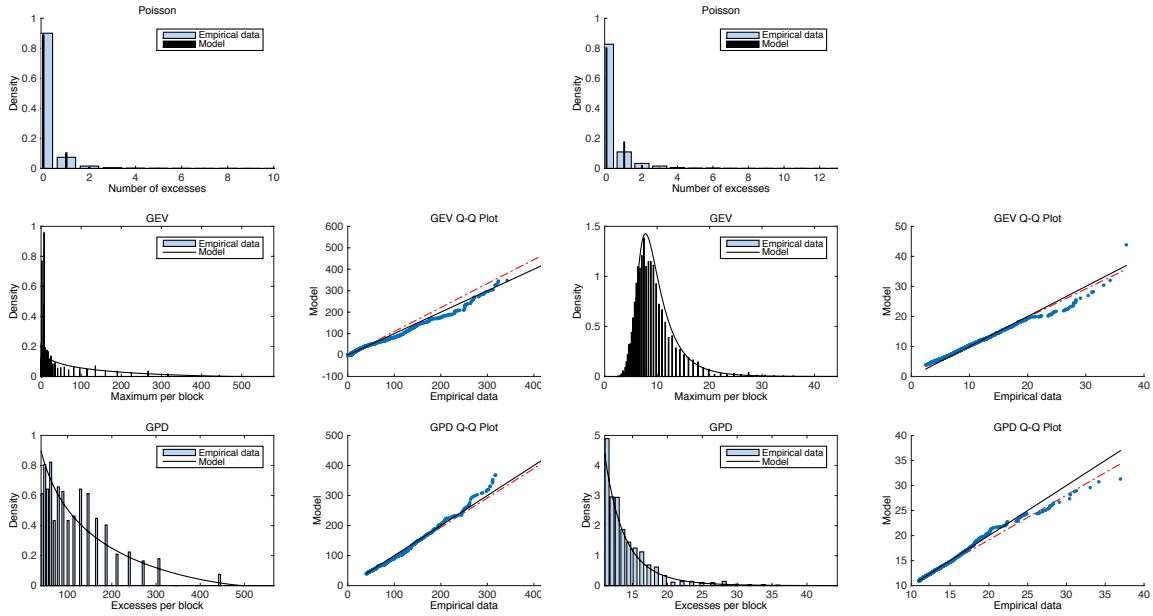


Figure 4.11: *Covariance matrices for CRP and neutrophils.*

EVT model fits

We found that the exceedances of our lab time series were well-fit by the univariate PP model. The model fits for the Poisson, GEV, and GPD distributions are shown in Figure 4.12 for CRP and WBCs and and in Figure 4.13 for neutrophils and

[5]Here, we took the covariance of the standardised feature values, meaning that correlation was equivalent to covariance.



(a) CRP (b) WBCs

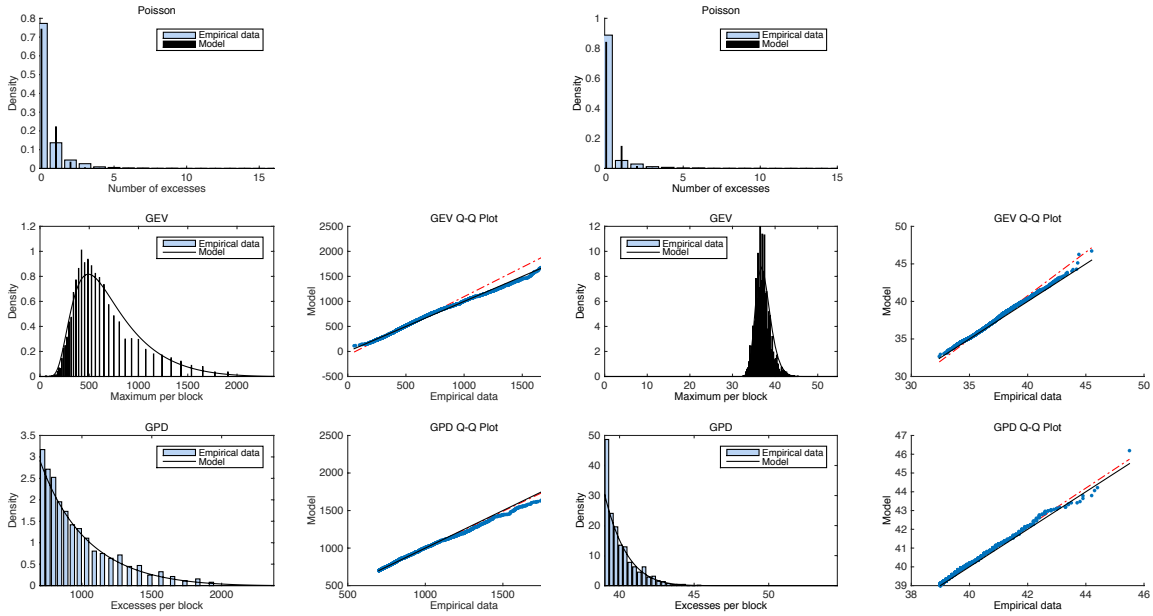
Figure 4.12: *CRP and WBC model fits. In Q-Q plots, the black line illustrates the theoretical perfect fit between model and data; the red dashed line illustrates the projection between the first and third quartile of the data.*

haemoglobin. Histograms of the empirical data and the fitted distribution, as well as Q-Q plots, are shown.

Histograms of the distribution of AUAS anomaly metrics, $s_{n,i}$, for each patient n and lab i including haemoglobin, CRP, and neutrophils are shown in Figure 4.14. As is apparent, many patients had lab trajectories that never exceeded the threshold u (i.e., $s_{n,i} = 0$).

Classifier and feature set comparisons

We found three predominant patterns across our substudies of disease subcategory (IBD, CD, UC) and feature type ($F_{standard}$, $F_{sevMetric}$). First, as shown in Table 4.1, for feature set $F_{standard}$ we mostly found better performance for more complex classifiers than simple linear classifiers. The RF classifier performed the best across CD, UC, and IBD, with significantly better performance in comparison to all other classifiers within IBD and UC ($p < 0.05$ for all comparisons, Wilcoxon signed-rank test). Within CD, there was a nonsignificant difference in comparison to the LR classifiers. In contrast, linear classifiers performed quite well for feature set $F_{sevMetric}$, with either LR-ridge or LR-lasso significantly outperforming the SVM and RF in



(a) Neutrophils (b) Haemoglobin

Figure 4.13: *Neutrophil and haemoglobin laboratory meaurements model fits. In Q-Q plots, the black line illustrates the theoretical perfect fit between model and data; the red dashed line illustrates the projection between the first and third quartile of the data.*

every case ($p < 0.05$ for all comparisons, Wilcoxon signed-rank test).

Secondly, when we compared across the two feature sets, we found that similar and slightly higher (nonsignificant) classification performance was obtained when using the $F_{sevMetric}$ feature set as compared to the full $F_{standard}$ feature set. However, as noted, the highest performance was obtained with a RF classifier when training with $F_{standard}$, while the simple LR had highest performance with the $F_{sevMetric}$ feature set. This is a useful finding, suggesting that similar classification performance is possible with either (a) a complex classifier that captures nonlinear interactions between the 48 features in $F_{standard}$ or (b) a simple classifier and the reduced set of 5 features in $F_{sevMetric}$. We conclude that our severity metric captured the available information in the laboratory dataset, as represented in terms of our CS measure of patient severity (i.e., requiring surgery and/or biologic therapy).

Finally, as shown in Table 4.1, across all feature sets, the performance of UC patient classification (up to 81.7% AUROC) was much better than the performance for CD patients (up to 68.6% AUROC). This suggests that patient lab trajectory patterns were closely aligned with higher-level severity outcomes for UC patients.

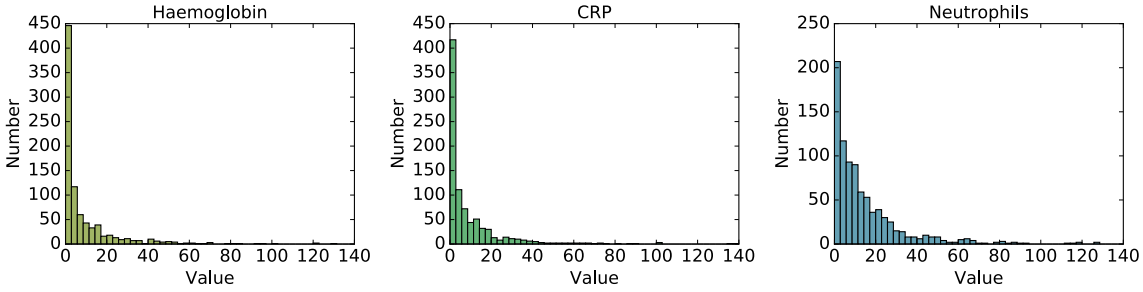


Figure 4.14: *Distribution of AUAS metrics for a representative selection of labs.*

Table 4.1: *Results for lab-based severity classification*

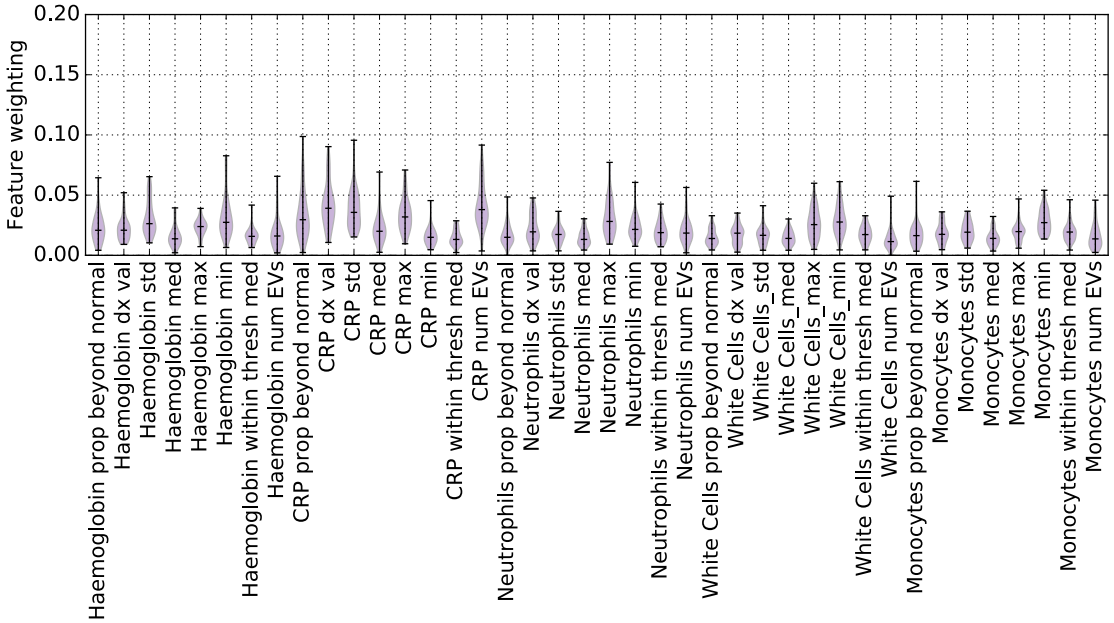
		CD	UC	IBD
$F_{standard}$	LR, ridge	65.1 (9.7)	78.3 (7.6)	77.5 (4.5)
	LR, lasso	64.8 (8.9)	79.8 (8.2)	78.0 (4.4)
	RF	**65.8 (8.8)**	**80.6 (6.8)**	**80.0 (4.2)**
	SVM	57.8 (14.5)	79.8 (7.8)	78.8 (4.2)
$F_{sevMetric}$	LR, ridge	67.4 (8.8)	**81.7 (6.4)**	**78.7 (3.1)**
	LR, lasso	**68.6 (9.5)**	81.3 (6.6)	78.7 (3.9)
	RF	65.8 (10.7)	77.8 (7.0)	76.7 (4.0)
	SVM	64.2 (10.8)	79.9 (6.8)	78.1 (3.9)

CD patients, on the other hand, apparently had more heterogeneity in their lab-based phenotypic behaviour. When combining both patient groups together, classification performance improved for CD patients and slightly decreased for UC patients.

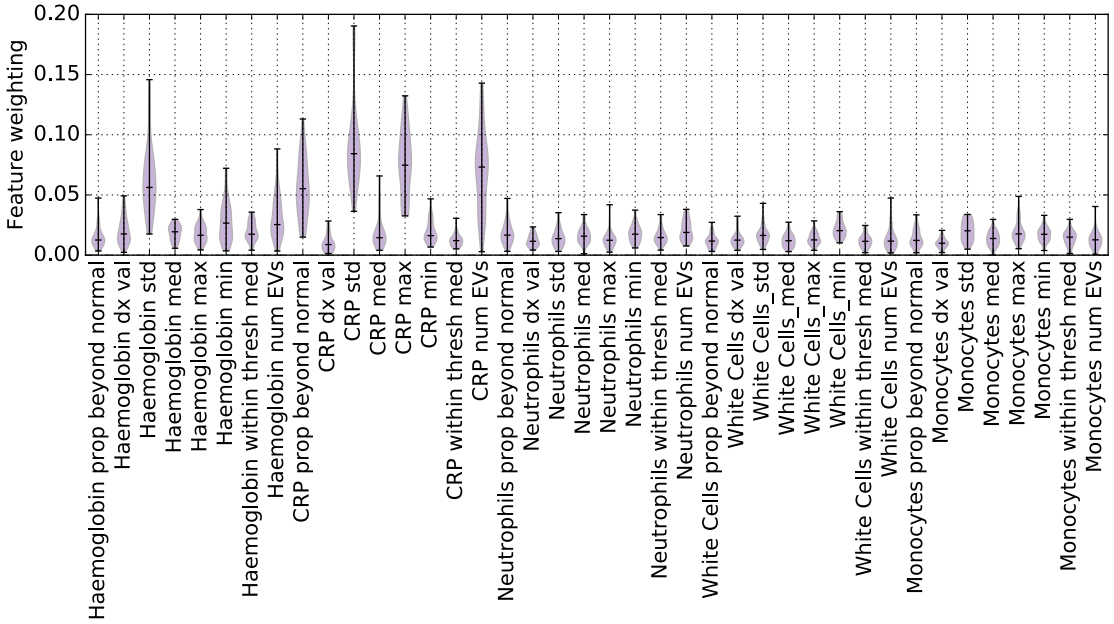
Individual feature importance

We next examined the feature weightings of those features involved in these classification tasks. We will discuss the feature weightings for each feature set, $F_{standard}$ and $F_{sevMetric}$, according to the classifier with the best performance. For $F_{standard}$, this was the RF classifier, and we therefore use feature importance I_{RF} discussed previously. For $F_{sevMetric}$, we present the values of the feature coefficients $\boldsymbol{w}$ from logistic regression. These weightings across feature sets should therefore not be compared with each other, but rather used to compare the various features within the set.

Figure 4.15 shows the feature weightings obtained from the RF classifier with $F_{standard}$ features. CRP-derived features (e.g., diagnosis value and number of above-threshold events, both with feature weightings of 0.042) were most important for CD patient classification, though all features appeared to play some role. The high variance across training iterations was somewhat to be expected, given the apparently poor correspondence between CD lab measurements and CS (i.e., poor classification



(a) CD



(b) UC

Figure 4.15: *RF feature weightings across N_I subsampling iterations for (a) CD and (b) UC patients with feature set $F_{standard}$*

performance based upon these lab metrics).

For UC, again, we found CRP-derived features to be very important for classification performance, with haemoglobin-derived metrics (e.g., standard deviation, with a feature weighting of 0.062) being of secondary importance. However, the highly-weighted CRP-derived features slightly differed from classification within CD patients, with the standard deviation (weighting of 0.089) and maximum (weighting of 0.076) having a larger role. Neutrophil and WBC levels were less representative of high-level severity.

Figure 4.15 shows the feature weightings obtained from the LR lasso classifier with $F_{sevMetric}$ features. Again, the CRP-derived feature was the most important for both IBD patient subtypes, with weighting 0.50 for CD patients and 1.1 for UC patients. For CD patients, this metric seemed to be the primary driver in classification. For UC, on the other hand, haemoglobin and, to a lesser degree, monocyte levels, also aided in classification, as shown by their feature weightings of 0.33 and 0.20, respectively. This was likely because blood loss from the colon is sometimes higher in UC patients (as disease localisation is generally concentrated in the colon), meaning that their haemoglobin levels would capture this dimension of patient severity. Interestingly, when all IBD patients were considered, neutrophil levels were also predictive (weighting of 0.24), though CRP was still the most important feature (feature weighting of 0.93).

Dimensionality reduction

The relationship between lab-based severity and clinical severe outcomes such as biologic therapy and surgery can also be visualised through a projection into 2-D space. We used PCA[6] to identify the axes of largest variation within our patient dataset. As input features for this visualisation, we used patient diagnosis, drug usage, disease location, surgery requirement (seven binary features), and also the AUAS for each of CRP, haemoglobin, WBCs, neutrophils, and monocytes (five continuous features).

As we can see in Figure 4.17, there were three primary axes: one separating the patients by diagnosis (top to bottom across the locus of projected data), one

[6]We note that standard PCA is not designed for binary features, as was the case for several of our input features here. We focus upon this challenge of combining various data types in Chapter 5. We also note that the patterns we have obtained were very similar between PCA and independent component analysis (ICA) (As will also be explained in Chapter 5, PCA and ICA can be formulated in the same way within a Bayesian framework.) Thus we have presented results obtained using PCA for simplicity.

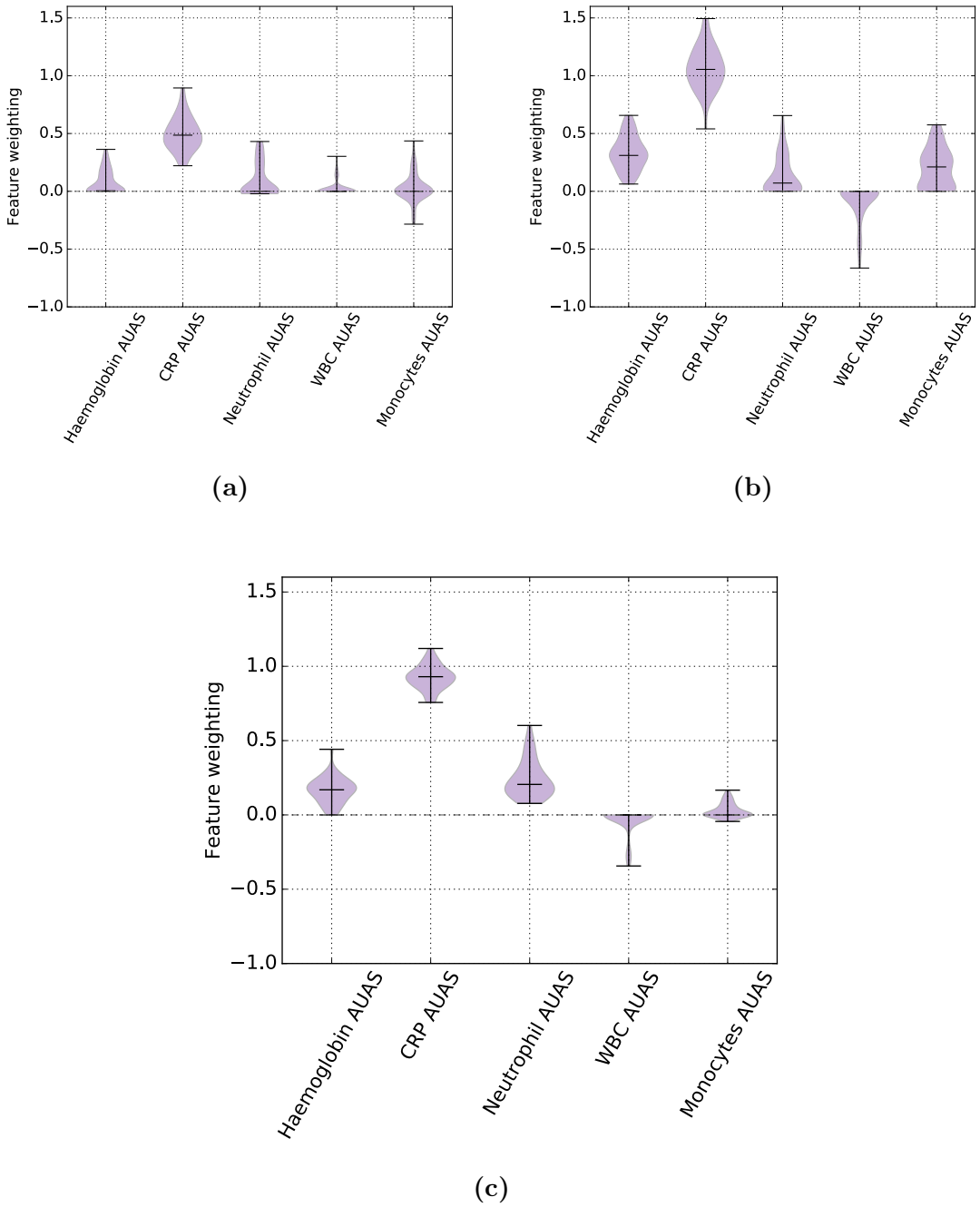


Figure 4.16: *LR-lasso feature weightings across N_I subsampling iterations for (a) CD, (b) UC, and (c) all IBD patients with feature set $F_{sevMetric}$*

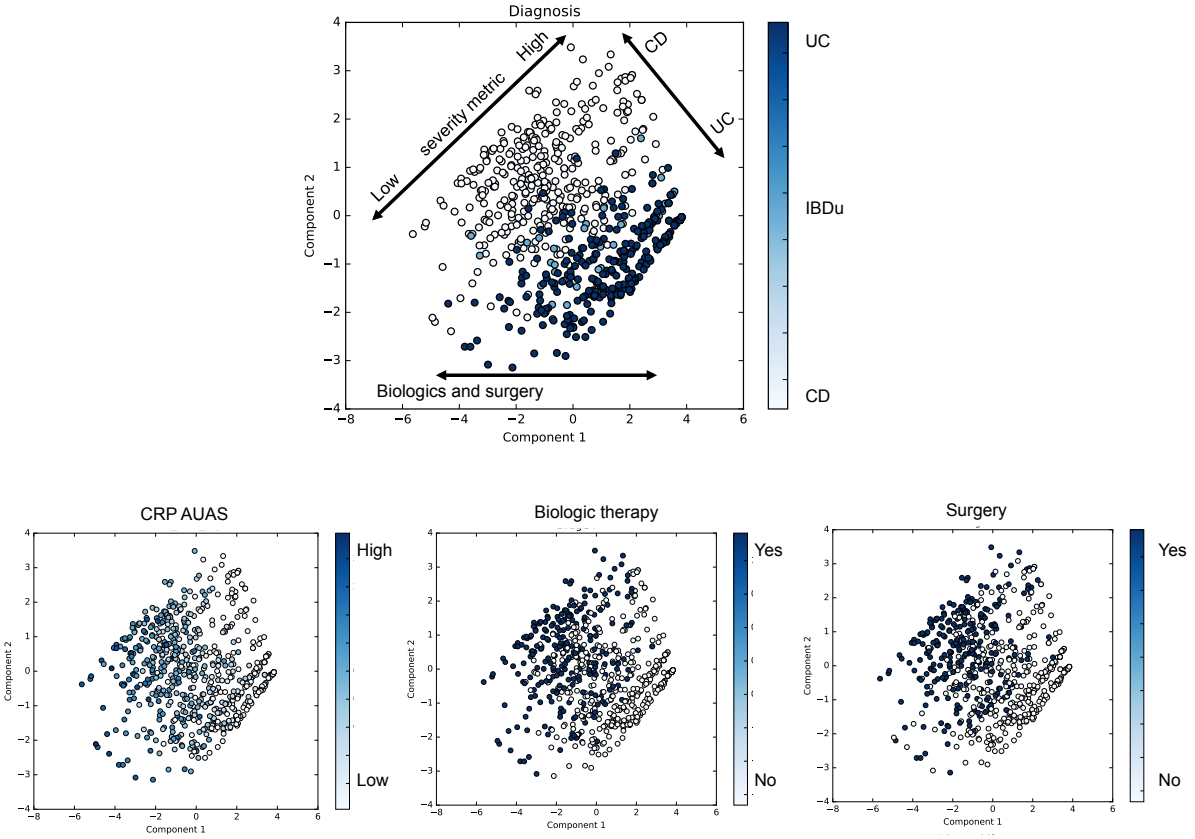


Figure 4.17: *PCA projection of data from IBD patients. First two components are shown. Four selected features (from the 12 used as input in the PCA decomposition) are highlighted in a subplot.*

separating the patients by the lab AUAS metrics (diagonal across the locus), and one separating the patients by severe drug prescription and surgery (left to right across the locus). This suggests that across IBD as a whole, the lab-based AUAS metrics overlap to some degree with CS, but that these two aspects of the disease also capture distinct information. However, as noted above, it is also helpful to view the breakdown of these projections by patient sub-type.

Figure 4.18 shows the results for a projection performed just with CD patients. Here, four illustrative features are shown for ease of interpretation; figures with full subplots for each of the input features are shown in Appendix B.1.1. We can see that there was an axis of lab-based severity from left to right across the locus of the projected data (i.e., severe lab measurements on the left, less severe on the right). In addition, many of the patients requiring surgery also had severe lab measurements (e.g., patients in top left of the locus). However, patients requiring surgery or biologics

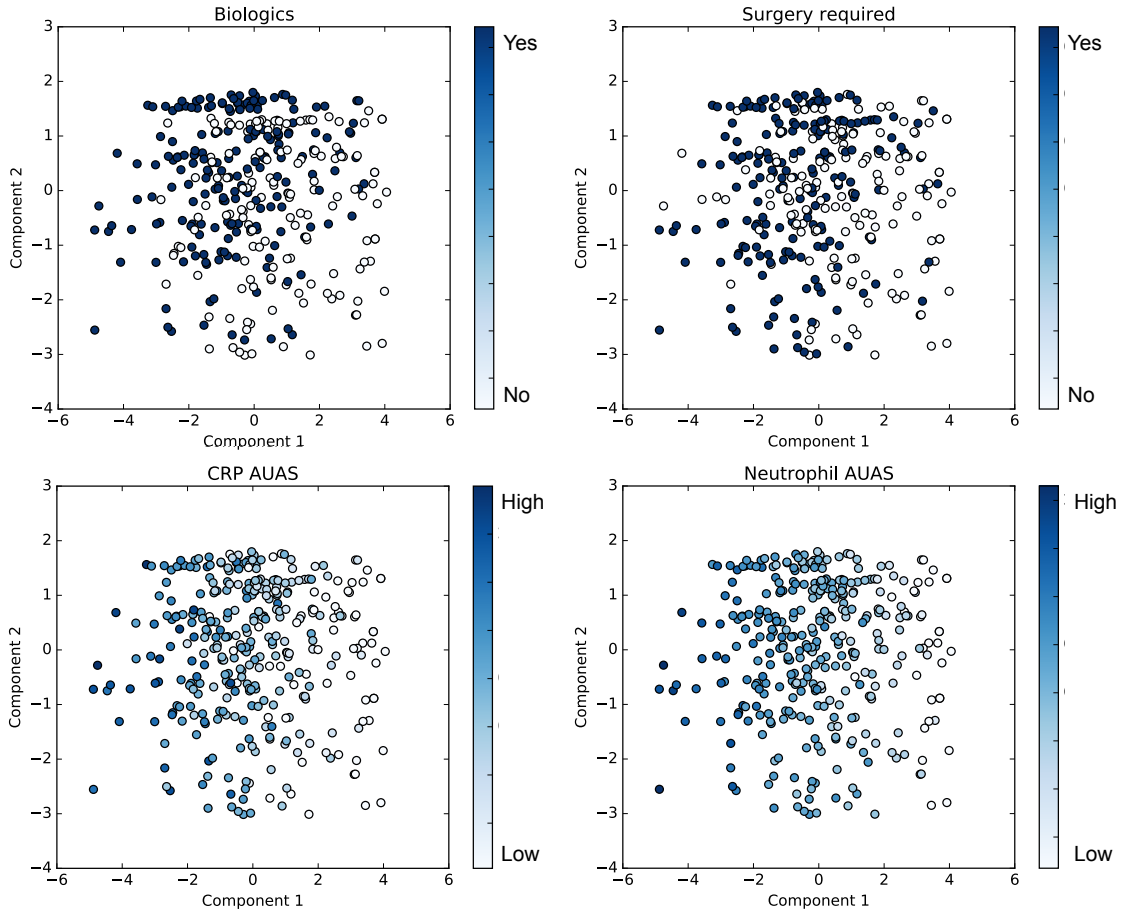


Figure 4.18: *PCA projection of data from CD patients. First two components are shown. Four selected features (from the 11 used as input in the PCA decomposition) are highlighted in a subplot.*

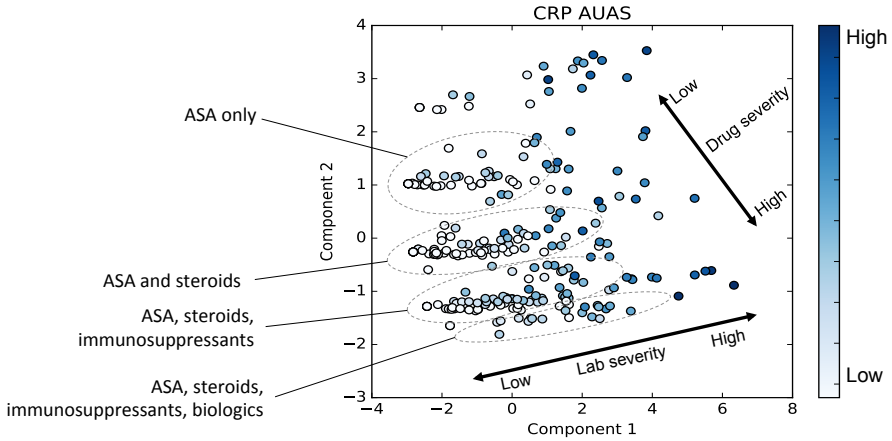


Figure 4.19: *PCA projection of data from UC patients. First two components are shown. Single feature shown, with labels to describe characteristics of other features.*

were also spread across the entire locus, without a strong pattern. For UC patients, as shown in Figure 4.19, a consistency between lab-based and outcome-based severity indicators was more clear. Here, most of the variability across the input set of patients was due to medication prescription, which corresponded to the three primary clusters apparent in the projected data. Patients that did not fall into these clusters exhibited more severe lab measurements and also tended to require more extensive medications or surgery. Again, figures with full subplots for each of the input features are shown in Appendix B.1.1.

These findings match with our findings from our supervised classification reported earlier in Table 4.1, in which lab-based metrics were found to be more representative of CS for UC patients than for CD patients (81.3% versus 69.4% for classification AUROCs, respectively). Nevertheless, in both cases, we found that a single metric descriptor was not sufficient to capture a full representation of patient severity. In other words, it is necessary to combine metrics across various different "axes" in order to fully capture patient severity. We delve further into this question in Chapter 5.

Our results have shown that in the context of IBD, our anomaly score metric very concisely captured the relevant CS information (i.e., progression to surgery or biologic therapy) available in lab-based metrics. Our model was well-suited to the problem of capturing anomalous lab dynamics, with very good model fits. We found also that UC CS was much more easily described by lab-based severity than for CD, but that in both major disease subtypes, long-term inflammation represents an alternative axis of severity in comparison to classical clinical outcome metrics.

4.5 Conclusion

In this chapter, we have presented a method for evaluating the likelihood of obtaining an exact pattern of measurements during a given time-window. The ability to calculate such a likelihood is desirable because it enables us to compare the behaviour of IBD patient lab measurements over time and between patients. The analytical formulation makes it quick and easy to calculate, without relying upon time-consuming sampling methods.

We have shown that our derived model is able to capture the important dynamics of patient trajectories over time in a simulated dataset that is representative of CRP lab measurements. Alternative approaches that attempt to model the entire patient time series actually suffer in comparison due to the high degree of noise within "normal-range" measurements. We have also presented case examples of how our model's approach compares to a more simplified approach of naïvely combin-

ing aspects of classical EVT models: our model offers a consistent and appropriate framework for capturing information about several measurements within a particular window of time.

Finally, we have shown that a single metric (the area under our anomaly score trajectory, or AUAS) can capture the relevant lab dynamics to represent high-level patient CS as successfully as a suite of standard time series features. This suggests the potential of this feature as a tool for summarising patient severity information into a single value, which will be useful for high-dimensional modelling of patient subtypes, as described later in this thesis.

Having developed this EVT-based model, we can use it to probe various aspects of IBD patient trajectories. As in the simulated example presented in this chapter, we can use the model as a means of modelling patient trajectories in order to cluster similarly-progressing patients together. We can also use it as a summary metric for the accumulated "severity" that a patient experiences during the course of their disease, for instance by taking the area under the anomaly score curve. Alternatively, we can directly calculate the likelihoods of particular lab measurements obtained within a given time window, which can be used to indicate the severity of a new patient's clinical presentation.

Having developed our model and applied it to our clinical application, we have also gained a better understanding of long-term inflammatory dynamics within IBD patients. We have found that CD patients have a much more heterogeneous disease progression than UC patients and that UC patients can be quite accurately classified as severe or not based upon a time series of their lab-based severity.

Chapter 5

Clustering model for combined data types

In the Chapter 4, we developed a lab-based severity metric. However, we have also seen that IBD patients (particularly CD patients) exhibit great heterogeneity in terms of their lab-based severity and other clinical severity markers (such as progressing to GI surgery). To investigate this heterogeneity and to better understand potential latent IBD subtypes, we turn to unsupervised analysis. Here, our motivation is the development of a model that is able to incorporate many aspects of the IBD patient characteristics: clinical, genetic, and functional.

Our goal in this chapter is to describe our approach in developing a clustering model that is capable of handling multiple different data types because our features of interest span a variety of forms. Several of our clinical indicators (e.g., the presence of inflammation in various bodily locations) are binary, while lab and functional measurements are continuous, and genetic SNP data can be described as binomial (whereby each copy of the chromosome can be viewed as a trial with a certain probability of being a variant). We also require a model that is able to handle missing data, which can infer the number of clusters present in the dataset, and which is capable of assigning new patients to previously-defined clusters. For these reasons, we have focused upon mixture modelling, a class of unsupervised learning that is capable of handling these requirements in a principled manner.

First, we provide a general introduction to clustering analysis. We then introduce mixture modelling through the example of a Bernoulli mixture model (BMM). We explain how parameter inference may be performed through EM. We then extend this to the case of a fully-Bayesian model capable of handling continuous, binary, and

binomial data types.

5.1 General clustering overview

There are a number of ways in which one can begin to assess groups of patients with similar characteristics. In this context any patient must first be represented by a set of features, which may include continuous, binary, categorical, or one of many other data types (for example, as presented in Chapter 3). The input into a clustering model can either be the basic features (i.e., an $N \times D$ matrix, where N is the number of patients and D is the number of features) or a distance matrix specifying the distance between each patient in terms of some specified metric (i.e., a $D \times D$ matrix).

A common first approach for clustering patient data is hierarchical clustering. Hierarchical clustering involves the iterative joining (or separating, if using divisive clustering) of the "most similar" groups of patients. Distance between data points must be defined; some example metrics are shown in Table 5.1. There are additionally a number of ways in which the clusters can be conjoined. In many real-world applications, Ward's criterion, which minimises the covariance within clusters, often results in distinctive cluster identification when other methods reported in the literature (e.g. average linkage, maximum linkage) do not. The clustering of time-series data, such as that available for many chronic disease and critical care medical applications, is often more challenging than the clustering of static (i.i.d.) data because the distance metric between two time-series is less well-defined. Relevant distance metric options for time series include Euclidean distance, Pearson's correlation factor, and dynamic time warping methods [160].

However, hierarchical clustering is heuristic in nature. Another approach is to model the underlying distributions generating the data, rather than defining a distance metric. This has the benefit that new patients can be assigned a probability of belonging to any of the clusters, and the clusters are defined by a generative model. The parameters can also be used as identifying descriptions for the given cluster.

Table 5.1: *Commonly-used distance metrics*

Metric	Type of data	Distance formula
Euclidean	Continuous	$(x_1 - x_2)^2$
City block	Continuous	$\lvert x_1 - x_2 \rvert$
Correlation coefficient	Time series	$\sum_{d=1}^{D} x_{1,d} x_{2,d}$
Hamming	Categorical	$\sum_{d=1}^{D} \mathbf{1}(x_{1,d} \neq x_{2,d})$

For two data points, x_1, x_2, with feature dimension D

Particularly within a Bayesian framework, generative mixture models can also handle missing data through marginalisation. Mixture modelling essentially makes the assumption that the underlying data can be separated into discrete latent variables, defined by the number of clusters, K, where we define a set of parameters to describe the underlying parametric cluster distributions. Mixture models are therefore sometimes classified as being "semi-parametric" methods.

Another approach, particularly when confronted with a large number of variables, is to assume that the "true" number of variables is much smaller than the number of available features, and that the dimensionality of the problem can be greatly reduced. This is essentially assuming that there is a continuous latent space upon which the data can be projected (in contrast to the discrete latent space of mixture models). These models (termed "latent factor models"), attempt to find a low-dimensional representation, L, of some D-dimensional data, X onto a subspace of dimension M. We can say $L = SX + \epsilon$, where S is termed the loading matrix (of size M by N), and ϵ is (often Gaussian) noise. We will write $W = S^{-1}$. In other words, this method involves a matrix factorisation of one matrix (X) into two low-rank matrices (W and L). PCA, ICA, and canonical correlation analysis are all special cases of latent factor models.

PCA attempts to find an orthogonal projection of some input data that maximises the variance of the projected data. We can derive PCA in a probabilistic way from our framework above ($X = WL + \epsilon$), if we assume that the latent variables are normally distributed with isotropic covariance: $L \sim N(\mu, \mathbf{I})$, and that the noise is also Gaussian: $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. This means that $p(X|L) \sim N(WL, \sigma^2\mathbf{I})$. When a maximum likelihood method is used to estimate the parameters of this model, we recover the PCA result. The full probabilistic formulation is sometimes termed probabilistic PCA. When the constraints of the model are relaxed and the latent variables are no longer assumed to have a strictly Gaussian distribution, this is equivalent to ICA [161–163].

As the above overview suggests, all of these clustering models can be approached from first principles from a Bayesian graphical modelling framework, with priors specifying the required constraints. Engelhardt and Stephens explain how mixture models can also be formulated as latent factor models, with modelling choices such as Bayesian priors acting as constraints on the matrix factorisation [164]. Bouveyron et al. explain how various model-based mixture methods build upon each other and are related [163].

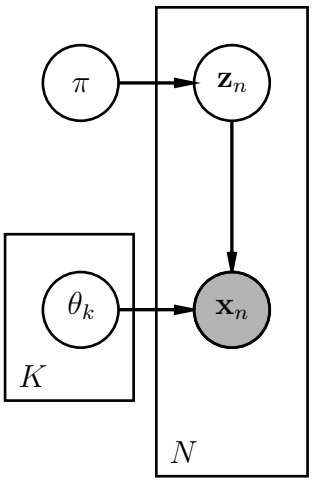


Figure 5.1: *Probabilistic graphical model for a BMM. As is standard in the literature, shaded variables indicate observed variables; plates indicate repeated variables; small closed dots indicate predefined hyperparameters.*

5.2 Introduction to mixture models and the EM algorithm

We now provide more detail on a specific parameterised form of mixture modelling, the BMM. As mentioned, a mixture model is a type of unsupervised learning model that assumes that a set of data is generated from a mixture of underlying parametric distributions. These underlying distributions each define a cluster, allowing the model to be used to probe for latent patterns in the data. For instance, given a set number of K latent groups, a BMM assumes that all data are derived from a mixture of these K underlying multi-dimensional Bernoulli distributions[1].

We will define $\mathbf{x}_n$ as a vector of binary variables of dimension D for patient n, where the entire set of N patients is represented as $\mathbf{X} = \mathbf{x}_1 ... \mathbf{x}_N$. We can model each feature according to a Bernoulli distribution with parameter θ_d for the d^{th} feature; the value of θ_d is the probability of the i^{th} feature being "on" $(= 1)$. Taking an assumption of independence across variables, we can then say $p(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{d=1}^{D} \theta_d^{x_d}(1-\theta_d)^{(1-x_d)}$.

If we say that the entire set of N patients can be represented as a mixture of K Bernoulli trials, then it is helpful to introduce a latent variable $\mathbf{Z}$, where the value of z_{nk} represents the assignment of patient x_n into group k, $z_{nk} \in [0, 1]$. We also include

[1]Here we a considering feature vectors where each feature is defined according to a Bernoulli trial; we are not considering a multivariate Bernoulli, which would be a joint distribution over all binary features. Equivalently, we considered a multivariate Bernoulli where we assume that the various dimensions are independent.

a mixing coefficient, $\boldsymbol{\pi} = \pi_1 ... \pi_k, k = 1...K$, where π_k is the probability of belonging to group k. The likelihood of the patient data, given $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, is then

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \; p(\mathbf{x_n}|\boldsymbol{\theta_k}) \tag{5.1}$$

In a generative interpretation, this means that the data is drawn from a mixture of K multivariate Bernoulli distributions, where each Bernoulli is drawn according to the mixing coefficient π_k, and the Bernoulli-distributed features are independent. This mixture model is represented in probabilistic graphical model form in Figure 5.1.

We would like to optimise this likelihood function to obtain estimates of π_k and $\boldsymbol{\theta}_k$. The maximum likelihood solution is not solvable in closed form, so we instead use the EM algorithm. The latter alternates between (1) calculating the expectation of the latent variable $\boldsymbol{Z}$ and (2) optimising the parameters based upon the resulting data allocations. We write the "complete-data" log-likelihood, which is the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Z}$, as:

$$\log \; p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \; \pi_k + \sum_{d=1}^{D} x_{nd} \log \; \theta_{kd} + (1 - x_{nd}) \log(1 - \theta_{kd}) \right] \tag{5.2}$$

Our goal is to then take the expectation with respect to the latent variable $\mathbf{Z}$'s posterior distribution; this is the expectation step (1). The expectation of z_{nk}, which we define as $\gamma(z_{nk})$, is also termed the "responsibility;" it is the mean posterior probability of component k, given $\boldsymbol{x}_n$.

$$\gamma(z_{nk}) = E[z_{nk}] = \frac{\pi_k p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)}{\sum_j^k \pi_j p(\boldsymbol{x}_n|\boldsymbol{\theta}_j)} \tag{5.3}$$

Intuitively, this represents the fact that the expectation of a particular patient n being assigned to cluster k is the probability of the patient's data, given the parameters defining cluster k, normalised by the sum of probabilities of the patient's data for all clusters.

We then perform the maximisation step (2): we use the newly-calculated expectation of z_{nk} to maximise the complete-data log likelihood, in order to obtain the optimal values for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. We therefore take the derivative of the complete-data log-likelihood in terms of these parameters. For $\boldsymbol{\pi}$, we also must enforce the constraint that all π_k values must sum to one, and so we use Lagrange multipliers. After

setting the derivatives to zero and simplifying, we obtain:

$$\boldsymbol{\theta}_k = \frac{\sum_N \gamma(z_{nk})\boldsymbol{x}_n}{\sum_N \gamma(z_{nk})} \qquad \pi_k = \frac{\sum_N \gamma(z_{nk})}{N} \tag{5.4}$$

We then iteratively alternate between the expectation (1) and maximisation (2) steps with the updated parameters until we reach convergence. Convergence can be defined to be a negligible change in the log-likelihood. EM produces a final set of parameters for each group: $\boldsymbol{\theta}$ (size $K \times D$) describing the probability of each feature being "on" ($= 1$) in group k, and $\boldsymbol{\pi}$ (size $1 \times K$) describing the probability of belonging to each group.

5.2.1 Model selection

EM provides a maximum likelihood (ML) estimate of the model parameters. This is a useful approach, but it suffers from the characteristic challenges of ML estimators. For instance, ML tends to overfit. In a clustering application, this means that singlarities can occur (e.g., a Gaussian mixture model can result in a single Gaussian representing a single data point, with infinitely small covariance). In addition, ML provides only single-point estimates for our parameters, rather than distributions over their values. Furthermore, the model likelihood will continue to increase as additional model components (i.e., additional clusters) are added. In other words, there is no clear way to determine the correct number of clusters. Therefore to estimate cluster numbers, approximations such as the BIC or Akaike Information criterion (AIC) are often employed. The BIC and AIC provide an estimation of the model posterior probability by implementing a penalty term for increasing the number of clusters.

The BIC is derived from several approximating steps of the model marginal likelihood. The BIC penalises large models (i.e., those with more parameters and hence with greater complexity) more heavily than the AIC. The AIC estimates the expected Kullbeck-Leibler discrepancy between the true model and the approximating model; it asymptotically selects that model from the available candidates that minimises the prediction error. The BIC $= K[\log\ N + \log\ 2\pi] - 2\log\ L$, and the AIC $= 2K - 2\log\ L$, with K being the number of groups, L being the log-likelihood, and N being the number of examples. The "best" model is that which minimises the chosen criterion [165].

If we wish to have a more powerful model capable of overcoming these limitations of the ML approach, then we must extend our inference procedure to a Bayesian formulation, as we will present in the next section. A Bayesian approach has the

advantage over ML that the incorporation of priors mitigates singularities and over-fitting, and that it also provides a built-in penalty for increasing model complexity. This allows us to use the marginal likelihood to more easily compare models and to determine the "best" number of clusters for a given dataset. In a fully-Bayesian formulation, models could be compared by examining the marginal likelihood at the level of the model hyperparameters, which is proportional to the posterior probability of a model (assuming all models are equally likely). Here, we will use an empirical Bayes approach (termed "maximum likelihood type II" in the literature), which provides us with a means by which to compare various clustering models, given our chosen hyperparameters.

5.3 Extension to a Bayesian model with multiple data types

We will now extend our formulation of a mixture model into a Bayesian construction by placing priors over the model parameters. We will also expand to the situation in which we have not only binary data, but also binomial-distributed and Gaussian-distributed data types for each patient. For a given patient n, we will define the portion of data that is Gaussian-distributed as $\mathbf{x}_{1,n}$, the portion that is Bernoulli as $\mathbf{x}_{2,n}$, and the portion that is binomial as $\mathbf{x}_{3,n}$. The dimensionality of each of these datasets is d_1, d_2, d_3, respectively. We will assume a naïve Bayesian model in regards to the different data types, meaning that we will treat each of the three dataset types as being independent from one another.

5.3.1 Model specification

As before, we will take a latent variable approach, modelling each patient's entire set of data, $\mathbf{x}_n$, as being drawn from a latent class, defined through the variable $\boldsymbol{z}_n$. We also again model the probability of being in any of the K possible classes as $\boldsymbol{\pi}$. However, we here introduce a Dirichlet prior over $\boldsymbol{\pi}$, governed by hyperparameter α.

We will parameterise the multivariate Gaussian distributions underlying data X_1 with mean $\boldsymbol{\mu}_k$ and precision Λ_k, where $\boldsymbol{\mu}_k$ is drawn from a zero-mean multivariate normal with precision $\beta\Lambda_k$. Λ_k is drawn from a Wishart distribution with scale matrix Λ_0 and ν degrees of freedom. As described previously, we will model the binary features X_2 as being drawn from independent Bernoulli distributions, where the vector of parameters will be defined as $\boldsymbol{p}_{1,k}$ (noting that each $\boldsymbol{p}_{1,k,d}$ represents the probability of a feature d being "on"). In turn, we place a Beta prior over this vector of parameters, with shape parameters a_1 and b_1. Similarly for the binomial data X_3,

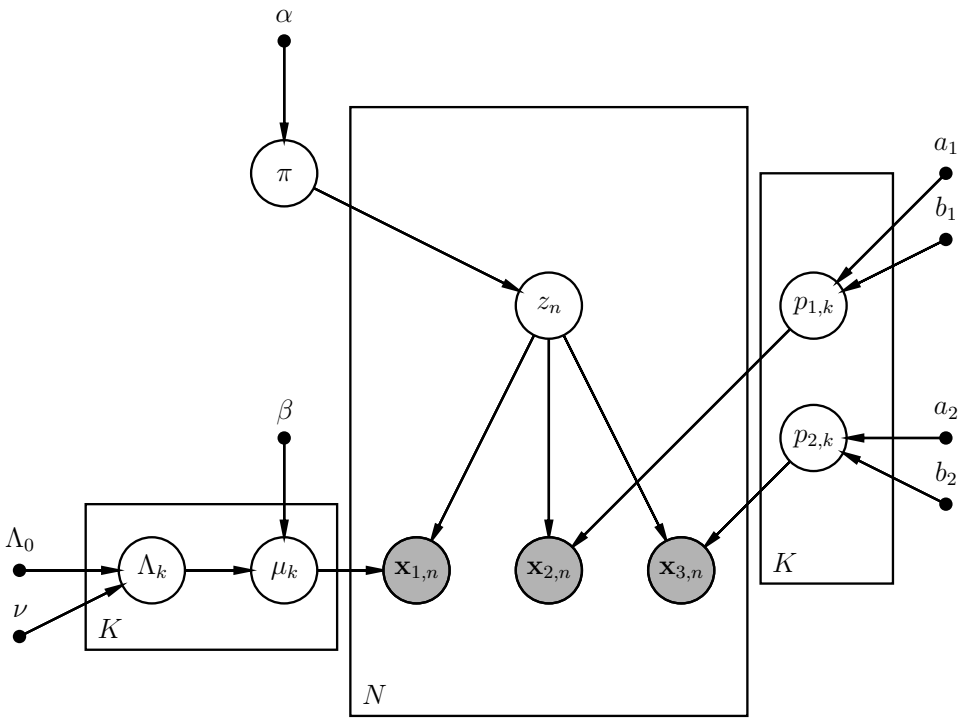


Figure 5.2: *Probabilistic graphical model for our mixture model over multiple data types.*

we define the vector of parameters p_2. We are designing the binomial portion of the model to best represent human genomic data, and thus we set $n = 2$ to represent the potential two possibilities of obtaining a mutation at every genomic locus (with one possibility for each chromosome). Of course, the Bernoulli and binomial distibutions are very similar, with the binomial corresponding to multiple Bernoulli trials. We have separated these two variable types here, though they could be combined by setting the binomial parameter $n = 1$ for the binary variables.

The above procedure gives us a model as represented in Figure 5.2. Our task is then to find the optimal values of $\boldsymbol{\mu}_k$, Λ_k, $\boldsymbol{p}_{1,k}$, $\boldsymbol{p}_{2,k}$, and $\boldsymbol{z}_n$, given our defined hyperparameters α, β, Λ_0, ν, a_1, b_1, a_2, and b_2. We present the distributions involved in our model in Figure 5.3. We note that though we are modelling each type of data independently through distinct parametric forms, these are all linked through the latent variable $\boldsymbol{Z}$. It is through $\boldsymbol{Z}$ that the clustering is operating, thus all data types are considered together in obtaining the optimal patient clustering assignments.

$\mathbf{x}_{n,1}|k \sim \text{Normal}(\boldsymbol{\mu}_k, \Lambda_k^{-1})$

$\mathbf{x}_{n,2}|k \sim \text{Bernoulli}(\mathbf{p}_{1,k})$

$\mathbf{x}_{3,n}|k \sim \text{Binomial}(\mathbf{p}_{2,k})$

$\mathbf{z_n}|\pi, \alpha \sim \prod_{k=1}^{K} \boldsymbol{\pi}_k^{z_{nk}}$

$\boldsymbol{\pi}|\alpha \sim \text{Dirichlet}(\boldsymbol{\alpha})$

$\boldsymbol{\mu}_k|\Lambda_k^{-1} \sim \text{Normal}(\mathbf{0}, (\beta\Lambda_k)^{-1})$

$\Lambda_k|\Lambda_0, \nu \sim \text{Wishart}(\Lambda_0, \nu)$

$\mathbf{p}_{1,k}|a_1, b_1 \sim \text{Beta}(a_1, b_1)$

$\mathbf{p}_{2,k}|a_2, b_2 \sim \text{Beta}(a_2, b_2)$

$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_k, \Lambda_k, \mathbf{p}_{1,k}, \mathbf{p}_{2,k}]$ (parameters)

$\boldsymbol{h} = [\beta, \Lambda_0, \nu, a_1, b_1, a_2, b_2]$ (hyperparameters)

Figure 5.3: *Specification of mixture model with multiple data types*

5.3.2 Model inference

We have chosen to use Gibbs sampling to perform inference in our model to obtain a posterior distribution over our parameters. Numerical sampling techniques (also termed "Monte Carlo" techniques) hold an important role in machine learning; we refer to [161,162,166,167] in our introduction to the methodology. We often encounter situations in Bayesian statistics in which we need to approximate integrals with no analytical solution. For example, we may have a function $\tilde{g}(\mathbf{x})$ that we can evaluate, but we actually wish to evaluate some function $g(\mathbf{x})$, which is equivalent to $\tilde{g}(\mathbf{x})$ up to an (unknown) normalising constant Z_g: $g(\mathbf{x}) = \frac{\tilde{g}(\mathbf{x})}{Z_g}$. This is often the situation when seeking to obtain a posterior distribution: we have functional forms to calculate the likelihood and the prior of a model $\tilde{g}(\mathbf{x})$, but in order to obtain a probability distribution $g(\mathbf{x})$, we must normalise by the intractable marginal likelihood Z_g.

A straightforward way to approximate $g(\mathbf{x})$ is through rejection sampling, which illustrates the basic idea of sampling. We introduce a proposal distribution $q(\mathbf{x})$, which is a simpler distribution from which it is possible to sample; $q(\mathbf{x})$ may be, for instance, a Gaussian. In order to employ rejection sampling, we need a function that

is greater than our unnormalised function $\tilde{g}(\mathbf{x})$ at every point, so we multiply $q(\mathbf{x})$ by some constant c such that $cq(\mathbf{x}) \geq \tilde{g}(\mathbf{x})$ for all $\mathbf{x}$. We then sample $\mathbf{x}_0 \sim q(\mathbf{x})$ and $u_0 \sim U[0, cq(\mathbf{x_0})]$ where U is the uniform distribution. If $u_0 > \tilde{g}(\mathbf{x_0})$, then $(\mathbf{x_0}, u_0)$ is rejected. This is equivalent to accepting samples $\mathbf{x}_0$ with probability $\frac{\tilde{g}(\mathbf{x})}{cq(\mathbf{x})}$. The retained samples will be distributed according to $g(\mathbf{x})$. Rejection sampling has several challenges, such as deciding how to choose c and $q(\mathbf{x})$, that can be addressed through further modifications of the method (e.g., adaptive rejection sampling). However, its inability to scale well to high-dimensional spaces (because the acceptance ratio becomes very small) motivates the use of Markov chain Monte Carlo (MCMC) methods.

As in rejection sampling, MCMC methods sample from a proposal distribution and reject these samples with a certain probability. However, this proposal distribution is now made to be dependent upon the current sample: $\mathbf{x}^\star \sim q_t(\mathbf{x}|\mathbf{x_{t-1}})$, where $\mathbf{x}^\star$ is the newly-drawn sample and $\mathbf{x}_{t-1}$ is the previous sample. If $\mathbf{x}^\star$ is rejected, $\mathbf{x_t} = \mathbf{x}_{t-1}$; otherwise $\mathbf{x_t} = \mathbf{x}^\star$. The samples $\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, ...$ follow the Markov property of conditional independence and form a Markov chain. The chain is designed such that, with the availability of infinite samples, it will reach an equilibrium distribution (namely, $g(\mathbf{x})$). In the widely-used Metropolis-Hastings (MH) algorithm, samples are accepted with probability

$$A_t = \min\left(1, \frac{\tilde{g}(\mathbf{x}^\star) q_t(\mathbf{x_{t-1}}|\mathbf{x}^\star)}{\tilde{g}(\mathbf{x_{t-1}}) q_t(\mathbf{x}^\star|\mathbf{x_{t-1}})}\right) \tag{5.5}$$

If we consider a chain of samples as exploring the space of a parameter's posterior distribution, then a criterion of this sort will push the exploration towards regions of high likelihood. However, because samples are accepted with probability A_t, there is always the chance to step in "unfavorable" directions, which allows for the exploration of, for instance, multi-modal distributions. As in rejection sampling, we must still specify the form of the proposal distribution. In practice when performing MCMC, the initial I_b samples must usually be discarded as the "burn-in" period before the chain reaches stationarity.

Gibbs sampling is a specific type of MCMC in which the proposal distribution $q_t(\mathbf{x}^\star|\mathbf{x}_{t-1}) = g(x_k^\star|\mathbf{x}_{\backslash \mathbf{k}})$, where for the k^{th} parameter contained in $\mathbf{x}$, the term $\mathbf{x}_{\backslash \mathbf{k}}$ corresponds to all parameters except k. In other words, at each step all parameters are held fixed except one, which is updated based upon sampling from the conditional distribution of that parameter, given the rest. If we insert this proposal distribution into our MH criterion, we find that $A_t = 1$, so all samples are accepted. Bayesian network models without loops (i.e., directed acyclic graphs) often lend themselves well

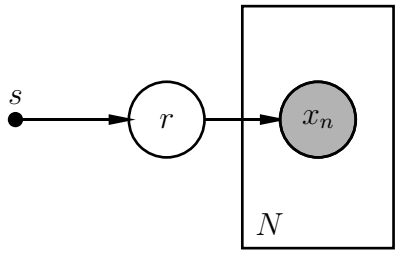


Figure 5.4: *Example probabilistic graphical model, as illustration in explanation of collapsed Gibbs sampling*

to Gibbs sampling because we have clear definitions of the conditional distributions[2].

Returning to our model, we need to find the optimal values of $\boldsymbol{\mu}_k$, Λ_k, $\boldsymbol{p}_{1,k}$, $\boldsymbol{p}_{2,k}$, and $\boldsymbol{z}_n$. As shown in Figure 5.2, our conditional independencies are encoded in our graphical model. For instance, μ_k can be sampled as $p(\mu_k^\star|\beta, \Lambda_k, \mathbf{X_1}, z_n)$. We could therefore draw a sample for each parameter in turn according to the posterior distribution corresponding to the functional forms presented in Figure 5.3.

In our case, however, we can take advantage of the conjugacy of our model to use a collapsed Gibbs sampler. We provide Figure 5.4 as a simple sample model for reference during our explanation of this process. In this model, the data X are modelled according to a distribution with parameter r. This parameter r is in turn modelled according to a distribution with parameter s. Thus s is a hyperparameter of the model. The data likelihood is $p(X|r, s)$ with a prior $p(r|s)$; the posterior can be written as $p(r|X, s)$:

$$p(r|X, s) = \frac{p(X|r, s)\ p(r|s)}{p(X)} \tag{5.6}$$

A prior is said to be conjugate to a particular likelihood if the prior and the corresponding posterior have the same functional form. Thus, if $p(r|s)$ is of the same form as $p(r|X, s)$, then it is conjugate to $p(X|r, s)$. The normal-Wishart distribution is a conjugate prior to the normal; the Beta is a conjugate prior for the binomial and Bernoulli.

A collapsed Gibbs sampler marginalises over some portion of the model parameters so that only a smaller subset are involved in the sampling update steps. Returning again to Figure 5.4, a marginalisation over r, written as $\int_r p(x_n|r, s)p(r|s)$, results in a

[2]The conditional distributions present in a graphical model are represented by a given node's Markov blanket. For a directed acyclic graph, the Markov blanket corresponds to a given node's parents, children, and children's parents.

compound probability distribution. If $p(r|s)$ is a conjugate prior, this marginalisation can be performed in closed form. In our case, this means that we can marginalise over $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, Λ_k, $\mathbf{p}_{1,k}$, and $\mathbf{p}_{2,k}$. This means that we only need to sample the posterior $\boldsymbol{z}_n$ variables in our sampling procedure. We will now present the details of the compound distributions that result from our marginalisations.

When we marginalise over the parameters of the Gaussian $\boldsymbol{\mu}_k$ and Λ_k, this involves the following computation:

$$\mathrm{p}(\mathbf{x_{1,n}}|\beta, \Lambda_0, \nu) = \int_{\boldsymbol{\mu}_k} \int_{\Lambda_k} \prod_{n|z_n=k} \mathrm{p}(\mathbf{x_{1,n}}|\boldsymbol{\mu_k}, \Lambda_k)\ \mathrm{p}(\boldsymbol{\mu_k}, \Lambda_k|\beta, \Lambda_0, \nu)\ \mathrm{d}\boldsymbol{\mu_k}\mathrm{d}\Lambda_k \tag{5.7}$$

Our normal-Wishart prior is conjugate to the multivariate Gaussian, and so the multiplication inside the integral produces another normal-Wishart distribution. After integrating over $\boldsymbol{\mu}_k$ and Λ_k for this normal-Wishart distribution, the resulting compound distribution is multivariate t-distributed with location $\boldsymbol{\mu}_t$, scale Σ_t, and ν_t degrees of freedom [168]. The formula for each of these parameters is presented below:

$$\nu_t = \nu + N_k - D + 1 \tag{5.8}$$

$$\boldsymbol{\mu}_t = \frac{\beta\boldsymbol{\mu}_0 + N_k\bar{\mathbf{x}}}{\beta + N_k} \tag{5.9}$$

$$\Lambda_t = \frac{T_n(\beta + N_k + 1)}{(\beta + N_k)(\nu_t)} \tag{5.10}$$

where:

$$T_n = \Lambda_0 + S + M \tag{5.11}$$

$$S = \sum_{i=1}^{N_k} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{5.12}$$

$$M = \frac{\beta N_k}{\beta + N_k}(\boldsymbol{\mu}_0 - \bar{\mathbf{x}})(\boldsymbol{\mu}_0 - \bar{\mathbf{x}})^T \tag{5.13}$$

where $\bar{\mathbf{x}}$ is the empirical mean of the data, $\boldsymbol{\mu}_0 = \mathbf{0}$ for our defined model, and, as before, D is the number of features.

We perform the same process for our Bernoulli and binomial variables. In the Bernoulli case, we must calculate

$$\mathrm{p}(\mathbf{x_{2,n}}|a_1, b_1) = \int_{\mathbf{p_{1,k}}} \prod_{n|z_n=k} \mathrm{p}(\mathbf{x_{2,n}}|\mathbf{p_{1,k}})\mathrm{p}(\mathbf{p_{1,k}}|a_1, b_1)\ \mathrm{d}\mathbf{p_{1,k}} \tag{5.14}$$

where, to remind the reader, a_1 and b_1 are the hyperparameters for the Beta distribu-

tion that governs the Bernoulli parameter $\mathbf{p_{1,k}}$. The compound distribution for the Bernoulli-beta conjugacy is again a Bernoulli distribution, with $\mathbf{p}'_{1,k} = \frac{\mathbf{a}'_{1,k}}{\mathbf{a}'_{1,k}+\mathbf{b}'_{1,k}}$. The updates are as follows:

$$\mathbf{a}'_{1,k} = a_1 + \sum_{n=1}^{N_k} \mathbf{x}_{2,n} \tag{5.15}$$

$$\mathbf{b}'_{1,k} = b_1 + N_k - \sum_{n=1}^{N_k} \mathbf{x}_{2,n} \tag{5.16}$$

For the binomial-beta portion of the model, when we marginalise over $\mathbf{p}_{2,k}$, we obtain a beta-binomial distribution. The integral is

$$\mathrm{p}(\mathbf{x_{3,n}}|a_2, b_2) = \int_{\mathbf{p_{2,k}}} \prod_{n|z_n=k} \mathrm{p}(\mathbf{x_{3,n}}|\mathbf{p_{2,k}})\mathrm{p}(\mathbf{p_{2,k}}|a_2, b_2)\ \mathrm{d}\mathbf{p_{2,k}} \tag{5.17}$$

where, as we recall, $\mathbf{p_{2,k}}$, together with $n_b = 2$, are the parameters of the binomial model. The hyperparameters a_2 and b_2 parameterise our Beta prior over the binomial. The parameters defining the resulting beta-binomial compound distribution are $\mathbf{a}'_2$, $\mathbf{b}'_2$, and $n_{bb} = 2$, the number of trials. Our updates are very similar to the Bernoulli case (the Bernoulli is a special case of the beta-binomial distribution when $n_{bb} = 1$). Intuitively, $\mathbf{a}'_{2,k}$ represents the number of successes (feature value of ≥ 1) and $\mathbf{b}'_{2,k}$ is the number of failures (feature value of 0). These updates are therefore as follows:

$$\mathbf{a}'_{2,k} = a_2 + \sum_{n=1}^{N_k} \mathbf{x}_{2,n} \tag{5.18}$$

$$\mathbf{b}'_{2,k} = b_2 + N_k n_{bb} - \sum_{n=1}^{N_k} \mathbf{x}_{2,n} \tag{5.19}$$

The Dirichlet-multinomial conjugacy is again a further generalisation of the beta-binomial conjugacy presented above, this time to the case of more than two categories. The updated $\boldsymbol{\alpha}'$ is essentially a re-weighting by the number of examples found in each class: $\alpha'_k = N_k + \frac{\alpha}{K}$.

Having now obtained our update equations for each data type, we will explain how these come together within the Gibbs sampling framework. We have collapsed our model by marginalising over several of our model parameters ($\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, Λ_k, $\mathbf{p}_{1,k}$, and $\mathbf{p}_{2,k}$), and so our only remaining latent variable is $\boldsymbol{Z}$. We can therefore perform iterative updates of the conditional distribution $p(\boldsymbol{z}_n = k|\boldsymbol{z}_{\setminus n}, \boldsymbol{x}, \boldsymbol{h}, \alpha) = p(\boldsymbol{z}_n =$

Algorithm 2 Gibbs sampling for multiple data types mixture model

```
Randomly initialise Z
for i=1:number of iterations do
    for n=1:N do
        Remove x_n from data before calculating updates
        for k=1:K do
            Compute t-distribution parameter updates: ν_t, μ_t, Σ_t
            L1 = T(x_{1,n}|ν_t, μ_t, Σ_t)            % Compute likelihood of x_{1,n}
            Compute Bernoulli parameter update: p'_{1,k}
            L2 = Bernoulli(x_{2,n}|p'_{1,k})         % Compute likelihood of x_{2,n}
            Compute beta parameter updates: a'_{2,k}, b'_{2,k}
            L3 = Beta(x_{3,n}|a'_{2,k}, b'_{2,k})    % Compute likelihood of x_{3,n}
            Compute Dirichlet-multinomial update: α'_k
            L_{n,k} = L1 × L2 × L3 × α'_k
        Sample from: p(z_n = a) = L_{n,a} / Σ_k L_{n,k}
        Store sampled z_n                            % New sample for z_n
Disregard initial I_b samples (burn-in)
```

$k|\boldsymbol{z}_{\backslash n}, \alpha)$ $p(\boldsymbol{x}_n|\boldsymbol{x}_{\backslash n}, \boldsymbol{z}_{\backslash n}, \boldsymbol{z}_n = k, \boldsymbol{h})$, where we define $\boldsymbol{z}_{\backslash n} = \boldsymbol{z}_1, ...\boldsymbol{z}_{i-n}, \boldsymbol{z}_{n+1}, ...\boldsymbol{z}_N$. We remind the reader that n refers to the n^{th} patient. The first term in $p(\boldsymbol{z}_n = k|\boldsymbol{z}_{\backslash n}, \boldsymbol{x}, \boldsymbol{h}, \alpha)$ is our prior for the k^{th} cluster, and the second term is the likelihood of the data given that cluster. Due to our independence assumption between data types, we can write this update as:

$$p(\boldsymbol{z}_n|\boldsymbol{z}_{-n}, \boldsymbol{x}, \boldsymbol{h}, \alpha) \propto \alpha'_k \text{ T}(x_{1,n}|\nu_t, \mu_t, \Sigma_t) \text{ Bernoulli}(x_{2,n}|p'_{1,k}) \text{ Beta}(x_{3,n}|a'_{2,k}, b'_{2,k}) \tag{5.20}$$

where T indicates the multivariate t distribution. This means that the product of our Dirichlet-multinomial update α'_k and the likelihood of each of our data partitions is proportional to the likelihood of that latent variable taking the value $z_n = k$. However, because this is a marginal distribution, we must calculate our update equations with our current sample $\boldsymbol{x}_n$ removed. We write the remaining data as $X_{\backslash n}$.

To obtain a probability distribution, we must normalise the likelihood L_n by the sum over all values of k: $p(z_n = k) = \frac{L_n}{\sum_k L_{n,k}}$. Having obtained this distribution, we can then sample from it to obtain our update for z_n. We perform this sampling step by sampling from the uniform distribution, projecting this onto the CDF associated with $p(z_n = k)$, and selecting the corresponding k cluster as the new cluster assignment for patient n. The pseudocode for our Gibbs sampler is presented in Algorithm 2.

Throughout this procedure, we also keep track of the log likelihood and marginal

likelihood. We recall the formula for our collapsed model,

$$\frac{p(\boldsymbol{X}|\boldsymbol{z},\alpha,\boldsymbol{h})\ p(\boldsymbol{z}|\alpha)}{\sum_z p(\boldsymbol{X}|\boldsymbol{z},\alpha,\boldsymbol{h})\ p(\boldsymbol{z}|\alpha)} = p(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{h},\alpha) \tag{5.21}$$

where $\boldsymbol{h}$ represents the model hyperparmaters defined previously. The joint log-likelihood is $\log\ p(\boldsymbol{X},\boldsymbol{z}|\boldsymbol{h},\alpha) = \sum_{k=1}^{K} p(\boldsymbol{X}_k|\boldsymbol{h})p(\boldsymbol{z}|\alpha)$, where $\boldsymbol{X}_k$ is the portion of $\boldsymbol{X}$ assigned to cluster k. The summation over $\boldsymbol{z}$ in Equation 5.21 means that the full log-marginal likelihood is essentially a summation over the number of clusters. This can be written as:

$$\begin{aligned}\log\ p(\boldsymbol{X}|\boldsymbol{h},\alpha) = K\log\ p(\boldsymbol{z}|\alpha)\ +\ \sum_{k=1}^{K}\Bigg[\sum_{i=1}^{N}\log\ \mathrm{T}(x_{i,1}|z_i,\mu_{t,k},\Sigma_{t,k},\nu_{t,k})+\\ \log\ \mathrm{Bernoulli}\Big(x_{i,2}|z_i,\frac{a'_{1,k}}{a'_{1,k}+b'_{1,k}}\Big)+\log\ \mathrm{Beta}\Big(x_{i,3}|z_i,\frac{a'_{2,k}}{a'_{2,k}+b'_{2,k}}\Big)\Bigg]\end{aligned} \tag{5.22}$$

where

$$\begin{aligned}\log P(\boldsymbol{z}|\alpha) = \log\ \mathrm{Gamma}(K\alpha) - \log\ K\mathrm{Gamma}(\alpha) - \log\ \mathrm{Gamma}(K\alpha+N)+\\ \sum_{k=1}^{K}\log\ \mathrm{Gamma}(\alpha+N_k)\end{aligned} \tag{5.23}$$

As described earlier, a non-Bayesian setting results in the likelihood continuing to increase as additional clusters are added (because each data point will have a higher likelihood if described by a distribution centred upon itself). In our Bayesian setting, the addition of the prior provides a penalty term in the marginal likelihood for adding increased numbers of clusters. For instance, here α offers control over the number of clusters, where a larger value of α provides a prior over more clusters. This takes effect in the equations through the term $p(\boldsymbol{z}|\alpha)$. A larger value of α means that the "penalty" for increased numbers of clusters is relatively smaller than for smaller values of α. As a result, the marginal likelihood provides a good metric by which different models (e.g. different numbers of clusters) can be compared, given a particular set of hyperparameters. The best model (again, given the specified hyperparameters) will be that with the largest log-marginal likelihood - this is an "empirical Bayes" approach.

5.3.3 Missing data

In our clinical application, missing data is common as a reality of using actual patient data acquired from the hospital. A Bayesian mixture model provides a principled approach for dealing with missing data. During inference, missing data must be dealt with in two places: (1) calculating summary statistics of the data (minus x_n for a given step) for the Gibbs sampling updates; (2) when calculating the log likelihood of x_n, given a particular cluster assignment.

First we consider the case when there is missing data from a sample $\boldsymbol{x}_{1,n}$ (i.e., the portion of $\boldsymbol{x}$ modelled as a multivariate Gaussian). To address (1) above, we sample from the multivariate Gaussian that describes the data from the previous iteration step, conditioned on the non-missing values of $\boldsymbol{x}_{1,n}$. We will define the indices of the missing features as i_m, and those that are not missing as $-i_m$. We are therefore sampling from the conditional Gaussian

$$x_{1,n,i_m} \sim \text{Normal}(\mu_k, \Sigma_k | x_{1,n,-i_m}) \tag{5.24}$$

This can be written as

$$p(x_{1,n,i_m} | x_{1,n,-i_m}) \sim \text{Normal}(\mu_M, \Sigma_M) \tag{5.25}$$

where

$$\mu_M = \mu_{i_m} + \Sigma_{12}\Sigma_{22}(x_{1,n,-i_m} - \mu_{-i_m}) \tag{5.26}$$

$$\Sigma_M = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{5.27}$$

Here, μ_{i_m} is the portion of the cluster k mean vector that corresponds to the missing feature indices and μ_{-i_m} is the remainder. Similarly, Σ_{11} is the portion of the covariance matrix that corresponds to the missing indices; Σ_{12} is the portion that corresponds to the missing indice rows and the non-missing indice columns, Σ_{21}, corresponds to the non-missing indice rows and missing indice columns, and Σ_{22} is the remainder of the covariance matrix.

This operation provides us with a draw from the distribution describing the missing values, conditioned upon the other feature values for a given patient. This is the preferred approach for inferring missing data, rather than sampling from the missing dimension's marginal distribution, for instance, because it allows the patient's available information to be taken into account.

We then obtain a matrix without any missing data, from which we can compute the

Gibbs sampler updates. Then, when evaluating the log-likelihood for patients with missing data, we marginalise over the missing values. This simply means that we remove the missing dimensions from the calculation (noting that the marginal distribution of a subset of dimensions of a multivariate t distribution is also t-distributed).

We employ a slightly different approach for handling missingness with our binary and binomial data because we have designed our model with the assumption that the features within these data types are independent. We therefore sample from the marginal distribution for the missing feature values for these data types because this is the same as the conditional distribution when the variables are independent. We thus sample from the Bernoulli or binomial distribution obtained in the previous iteration step. In the same manner as employed in the Gaussian case, we then remove any dimensions with missing data during the calculation of the log-likelihood (i.e., we marginalise over the missing values).

If we are then interested in inferring the value of a missing data point, we can examine the distribution over the missing samples across the sampling runs (post burn-in).

5.4 Clustering coherence

We now present several practical challenges associated with clustering and our approaches for addressing them.

5.4.1 Label-switching

In repeated runs of a clustering algorithm, the cluster assignments are not constant - the labeling of a particular cluster is arbitrary. This means that, for example, a group of patients assigned to a cluster with the label "5" in one run of an algorithm may all be assigned to a cluster labelled "2" in a subsequent run. This is the so-called "label-switching" problem. To make the situation less trivial, the patients will often not be assigned to the exact same group each time (i.e., only a proportion of those assigned to "5" will be in "2" in the next run). In addition, when taking an empirical Bayesian approach, the number of clusters may vary across runs of the algorithm (meaning that clusters may be split). The challenge, then, in an MCMC scenario, is to match clusters across runs. We note that the typical approach of averaging across MCMC iterations to take an expectation is inappropriate for a multinomial variable such as cluster assignment [161, 162].

In simple univariate settings, one approach to manage this problem is to apply identifiability constraints [169]. Such constraints order the groups so that, for in-

stance, if there are groups 1, 2, and 3, the means $\mu_1 < \mu_2 < \mu_3$. While this approach is useful in some scenarios, such constraints quickly become difficult to define and apply in any meaningful way with data of higher dimensionality. For instance, there is no certainty that a single cluster will have "lower" mean values in every dimension; similarly, two clusters may have identical mean values in every dimension except for one critical dimension.

Another approach that has been taken is that of "relabelling algorithms," which seek to minimise some loss function associated with transforming labels. For instance, some early approaches matched parameters between successive runs of the sampler [170]. More recently, this relabelling has been applied directly to the latent variable Z. Cron et al. present an approach that uses what is essentially a label-mismatch matrix between a new labelling and a reference labelling as a cost matrix [171]. They define a $K \times K$ matrix C that contains components defined as $C_{hj} = \sum_{i=1}^{N} \mathbb{1}(z_i^R = h, z_i^w \neq j)$, where Z^R is the reference labelling and Z^w is the new labelling. They then find the optimal permutation of the new labelling to best-match the reference, which is equivalent to minimising $\mathrm{tr}(C)$. This can be performed using the Hungarian algorithm (i.e., Munkres' algorithm) to find the optimal permutation. Such an approach was also used by, e.g., [172]. The approach of relabelling, as Jasra et al. describe, is essentially an automatic way to choose identifiability constraints [169], where the identifiability constraints do not need to be explicitly defined.

Our approach

Here, to provide consistent cluster assignments, we employed a methodology that is similar to Cron et al.'s relabelling approach based upon the example allocations. However, the problem with the use of Cron et al.'s method in our context is that the number of our clusters can change between iterations. This means that we do not always wish to find an exact re-ordering to match cluster assignments, but rather that sometimes we wish to allow multiple cluster assignments in one iteration to be mapped to a single cluster assignment in another iteration. Therefore, we reframe the "minimisation of the non-overlap" between two labellings as the maximisation of the overlap between the two. We choose one clustering (typically that with the largest marginal likelihood) as the reference cluster, Z^R, and we then create a confusion matrix between subsequent clusters and the reference cluster. We map the cluster assignment from a new assignment, Z^w, to Z^R by defining a class within Z^w as the class within Z^R that has the largest overlap in patient assignments. In the case of a tie, one of the reference clusters is chosen randomly. This procedure is shown in

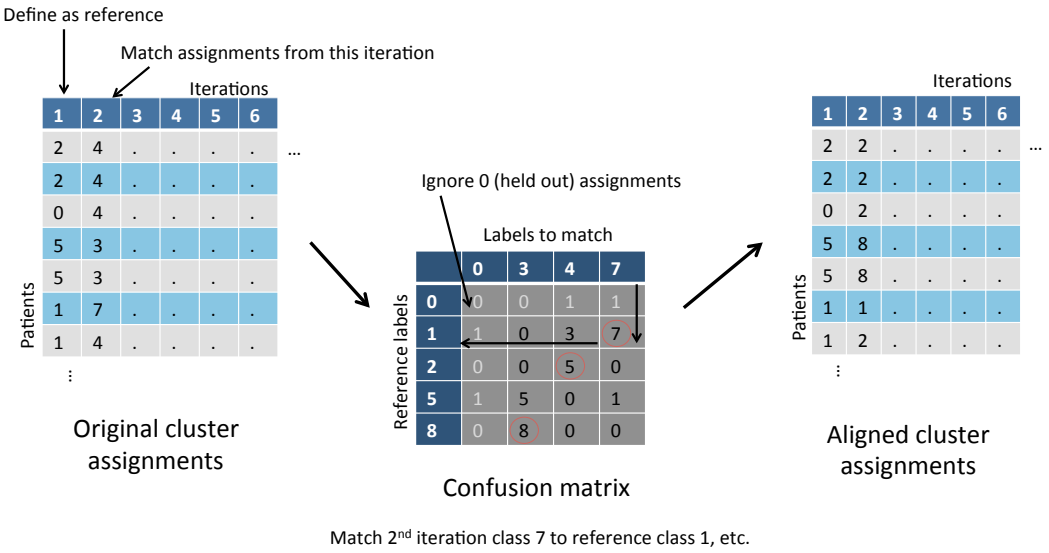


Figure 5.5: *Illustration of the steps involved in the label-matching procedure. Only the first several rows of z assignments are shown to provide illustration.*

Figure 5.5, and which is further described in the next section. Our approach has a straightforward interpretation and is quick to calculate, making it a desirable choice for our application.

5.4.2 Cluster stability

Another aspect of clustering that is of practical importance is how to assess the consistency of the clustering across runs and the ability of the clustering to generalise to previously-unseen data. Within supervised machine learning frameworks, techniques such as cross-validation are used to assess the ability of an algorithm's performance to generalise to new data. However, frequently clustering algorithms are run only once, and it is these results that are presented. We feel that this provides inadequate representation of how sensitive the clustering results are to the exact set of patients available.

In order to provide results that are robust, we perform several rounds of clusterings, with a proportion of patients held out each time. Our approach follows that of, for example, Li et al. [104]. In other words, we assign a "training" set and perform the clustering based upon only the training set. After repeating this subsampling many times, we can, using the same relabelling approach presented above, assess cluster-assignment stability. Here, the only modification required is that we ignore patients who have been held out from the clustering, which is accomplished by simply not

Table 5.2: *Commonly-used cluster comparison metrics*

Metric	Formula	Intuition
Purity	$\frac{\sum_i(\sum_{j=1}^{K} N_{ij})(\max_j p_{ij})}{N}$	Consistency of group assignments
Rand index (RI)	$\frac{TP+TN}{TP+FP+FN+TN}$	Proportion of "correct" clustering decisions
Mutual information (MI)	$\sum_{i=1}^{R}\sum_{j=1}^{C} P(i,j)\ \log\frac{P(i,j)}{P_A(i)P_B(j)}$	Overlap between two clusters
Jaccard index (JI)	$\frac{TP}{TP+FP+FN}$	Number shared in both groups divided by total number across both groups

Purity: N_{ij} is the number of class i in class j; $p_{ij} = N_{ij}/(\sum_{j=1}^{K} N_{ij})$; K is the number of classes
RI, JI: TP= true positive, TN= true negative, FP= false positive, FN= false negative
MI: R, C=number of clusters in two partitions of the dataset, A and B, respectively
$P(i,j)$=probability of patient being in class A_i and B_j; $P_A(i)$=probability of being in class A_i
$P_B(i)$=probability of being in class B_i

performing a mapping from the held-out patients (given a cluster label of "0") to the reference cluster (also shown in Figure 5.5).

5.4.3 Cluster similarity

A related, but distinct, problem is that of assessing the similarity between various clustering results. Unlike the label switching problem, where we needed to re-define labels for each individual example (i.e., patient), such similarity metrics assess the similarity between two entire sets of cluster assignments. For instance, if we wish to compare the similarity between successive runs of our stochastic sampler or if we wish to compare results of various clustering methods (e.g., hierarchical versus mixture model), we can use these similarity metrics to obtain a global view of similarity.

There are many metrics that can be used to compare the similarity of clusters that are generated from the same underlying dataset, some of which are presented in Table 5.2 [162]. The overall consistency across subsamplings can then be compared. While some of these indices, such as the Rand index, rely on a "true" label, as above, a single instance Z^R of the subsampling procedure can be identified as the "true" label, and other clusterings can be compared to Z^R to assess consistency.

Some of these metrics have been further developed to account for chance in cluster similarity. It is particularly important to take chance similarities into account when the number of clusters within the dataset is large. The adjusted Rand index is formulated as $ARI = \frac{RI-I_{exp}}{I_{max}-I_{exp}}$, where RI is the Rand index as above, I_{exp} is the expected index, and I_{max} is the maximum index. The adjusted mutual information (AMI) follows the same formulation.

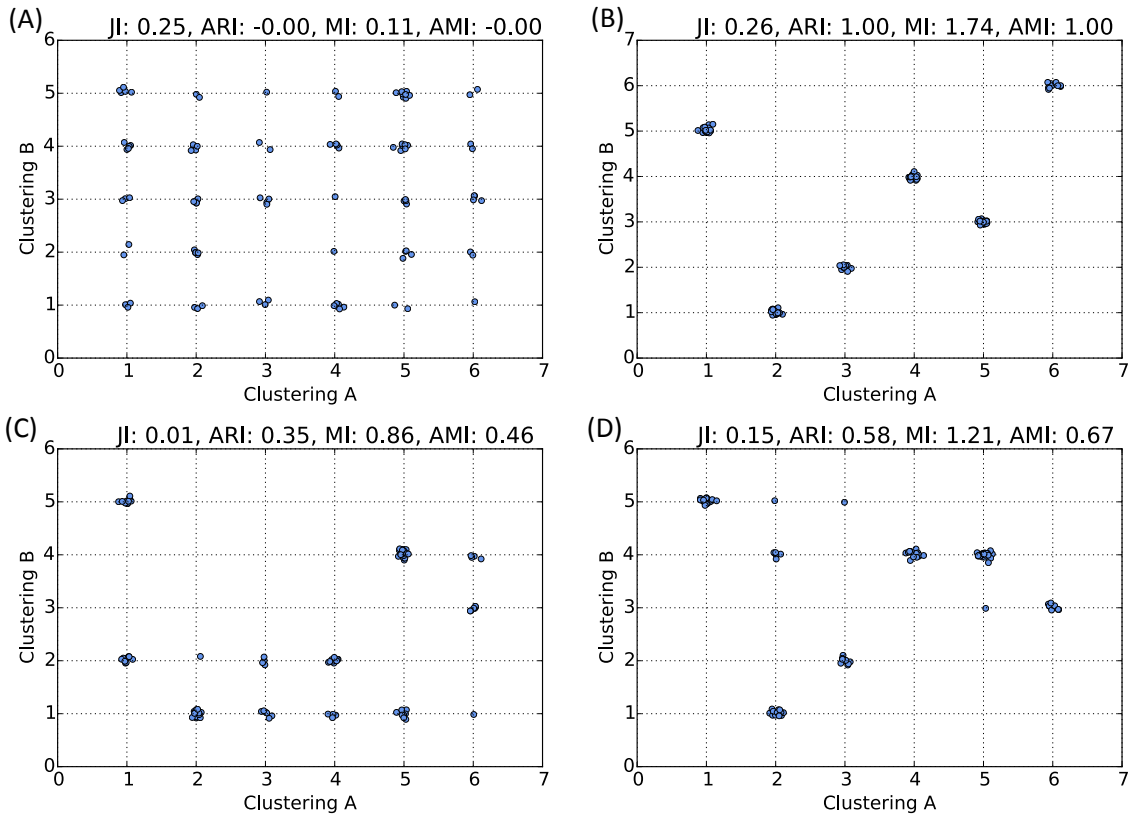


Figure 5.6: *Comparison of various cluster evaluation metrics. Panel A illustrates a random clustering, Panel B illustrates perfect agreement between clusters, and Panels C and D illustrate situations between these two extremes. JI=Jaccard Index, ARI=Adjusted Rand Index, MI=Mutual Information, AMI=Adjusted Mutual Information.*

Figure 5.6 illustrates a comparison of different clustering outcomes, ranging from random class assignments (Panel A) to perfect agreement (Panel B). As is evident, some metrics require first the matching up of corresponding groups. For instance, the Jaccard Index is only able to match groups with the same label, yielding a low score even in the case of perfect agreement. We also see the benefit of the adjusted indices, where the MI score during random assignment (panel A) is 0.11, but the adjusted score is 0.

5.5 Experiments

We now present some simulations illustrating our mixture model's functionality. We first used Fishers' classic iris flower dataset, which consists of fifty examples each of three types of iris flowers, each of which is defined by four continuous-valued features [173]. We used the Gaussian portion of our mixture model only in this simulation,

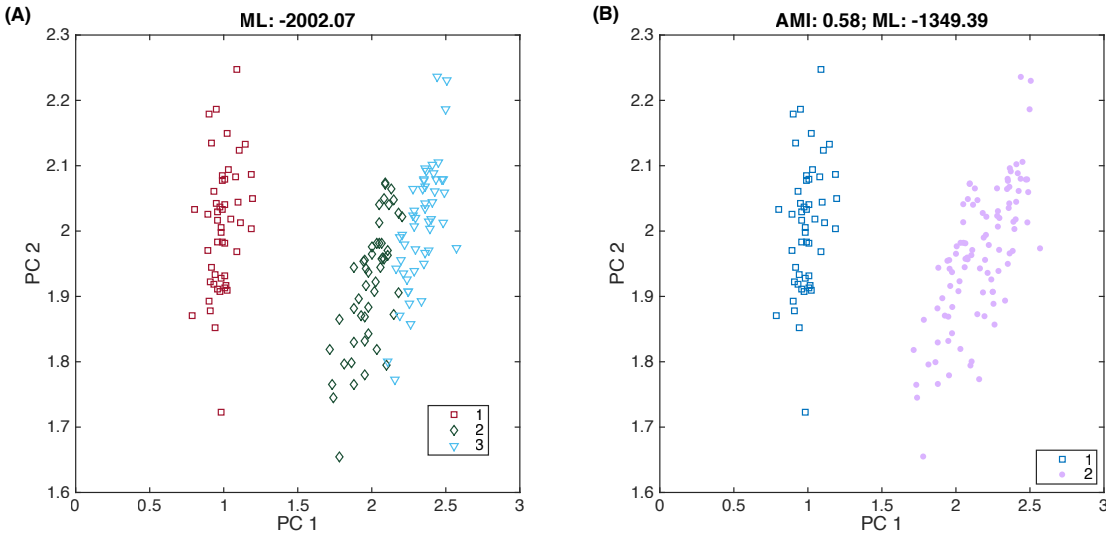


Figure 5.7: *Comparison of true flower types (A) and model clustering results (B) for the iris flower dataset. Presented as a projection onto first two principal components. AMI presented between the two clusterings.*

as is appropriate given the scalar type of data considered. In Figure 5.7 we show a comparison of the true flower types (A) and the mixture model results (B). The marginal likelihood of the three flower types is actually lower (according to our model) than the marginal likelihood for two types, which suggests that though the model has combined two groups of flowers togther, based upon the data, this is actually a preferred solution. Indeed, when examining the clustering apparent in the projection onto the two primary principal components, this may be considered to be a reasonable clustering result. We could influence the results further by changing the model priors.

We then randomly selected features and replaced them as being missing (ensuring that no example had more than three missing features out of the total four features per example) to evaluate the ability of the model to manage missing data. We compared our clustering results with the clustering structure apparent through PCA, with the missing values replaced with the mean value of the variable for PCA. Even with up to 200 missing features (i.e., an average of 1.3 missing values for each flower example), we obtained reasonable results that offered improvement upon PCA-based clustering. Figure 5.8 (A) shows the data projected onto the first two PCs, with the "true labels" coloured. As is apparent, several of the class 1 (red) examples tended to cluster spatially with the class 2 and class 3 examples (green and blue). Panel (B) shows the projected data, now coloured according to the clustering results from our model. As in the case with no missing data, our model merged examples from the "true" labels 2 and 3 together into a single cluster (both in the purple cluster 2 in panel (B)).

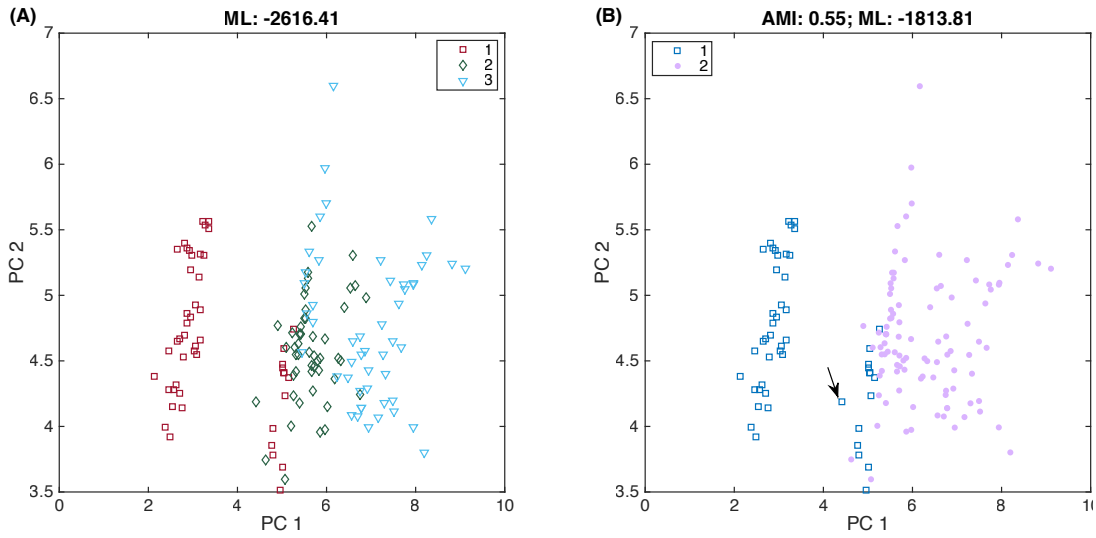


Figure 5.8: *Iris flower clustering results with 200 missing data features. Data projected onto first two PCs and colored in (A) according to the true classes and in (B) according to our model's clustering results. Arrow indicates a data point belonging to the "true" class 2, but mistakenly assigned to the wrong cluster by our model.*

Table 5.3: *Iris flower example: Comparison of cluster assignments between hierarchical clustering and our mixture model.*

		Hierarchical			Mixture model		
		1	2	3	1	2	3
True labels	1	14	0	36	0	50	0
	2	42	8	0	49	1	0
	3	13	37	0	50	0	0

Our model, however, was still able to correctly separate examples from the "true" clusters 2 and 3 from the "true" cluster 1, with all of the "true" cluster 1 examples being classified together. Only one example from the "true" cluster 2 was mistakenly classified with the other examples from the "true" cluster 1 (marked with an arrow). This improved performance was because our model is able to more appropriately infer over the missing feature values by conditioning upon the available data.

When we examined the results from hierarchical clustering (using Ward's criterion and Euclidean distance, with missing data imputed by the population mean of that variable), we found that the resultant clusters were similar to the PCA results in that they did not well-represent the true underlying labels. As is shown in the confusion matrix in Table 5.3, hierarchical clustering was able to identify three underlying clusters (through manual inspection of the dendrogram), while, as mentioned, our model was only able to identify two clusters. However, the true flower types did not

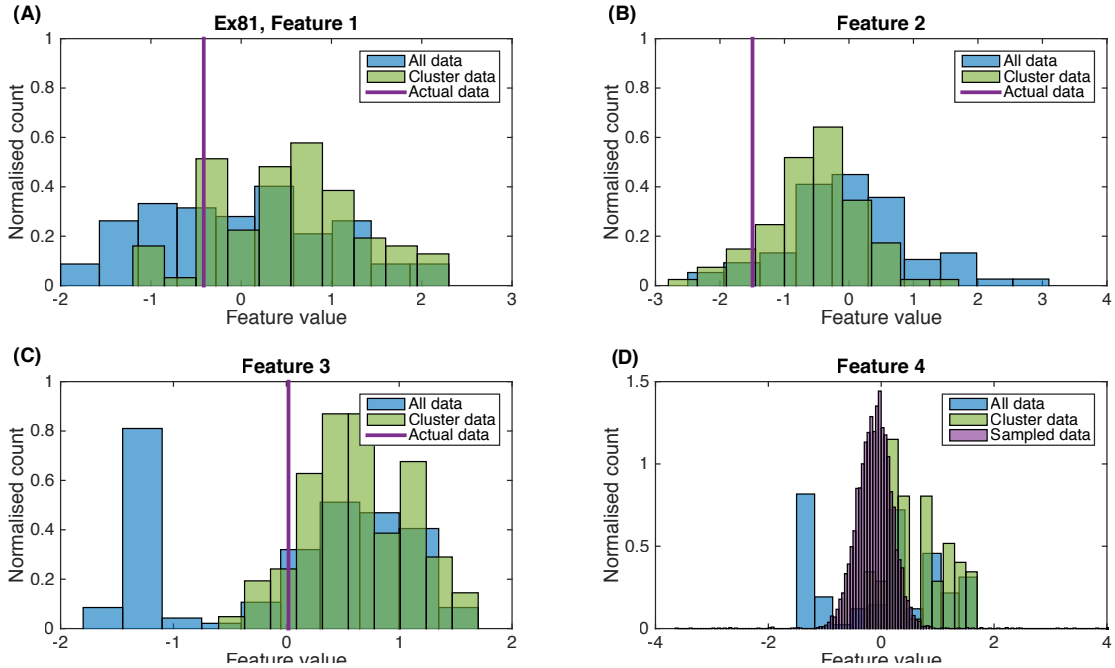


Figure 5.9: *Example of the sampled distribution over a feature artificially induced as being missing within the iris flower dataset.*

overlap well with the identified clusters in hierarchical clustering, largely because of the challenge of the missing data. As in our model, the "true" label 2 and 3 examples were largely overlapped, with $8 + 13 = 21$ of these examples incorrectly assigned. However, while our model was able to correctly distinguish "true" label 1 from "true" labels 2 and 3 (with only 1 misclassified example), 14 "true" label 1 examples were mistakenly grouped with the "true" label 2 cluster by hierarchical clustering.

An example of the distribution obtained over a missing feature value through our model is shown in Figure 5.9. For each of the four features, the distribution of the feature for the dataset as a whole is shown in blue. The within-cluster distribution of the feature for the particular cluster in which this example was assigned is shown in green. As is apparent, this particular example tended to have feature values that lie to the left side of the within-cluster feature distributions, as indicated by the vertical purple lines in plots (A)-(D). We would therefore expect that the sampled values for the missing feature in panel (D) would tend towards the left side of the within-cluster feature distribution. Indeed, the sampled distribution is centred on the left side of the within-cluster distribution, as illustrated with the purple histogram.

We then simulated a dataset with 50 Gaussian-distributed variables and 10 binary features. We sampled according to our model with the following hyperparameters: $\alpha = 1$, $\Lambda_0 = \text{diag}(1)$, $\nu = d_{normal}$, $\beta = 1$, $a = 0.5$, $b = 0.5$. We simulated a set of

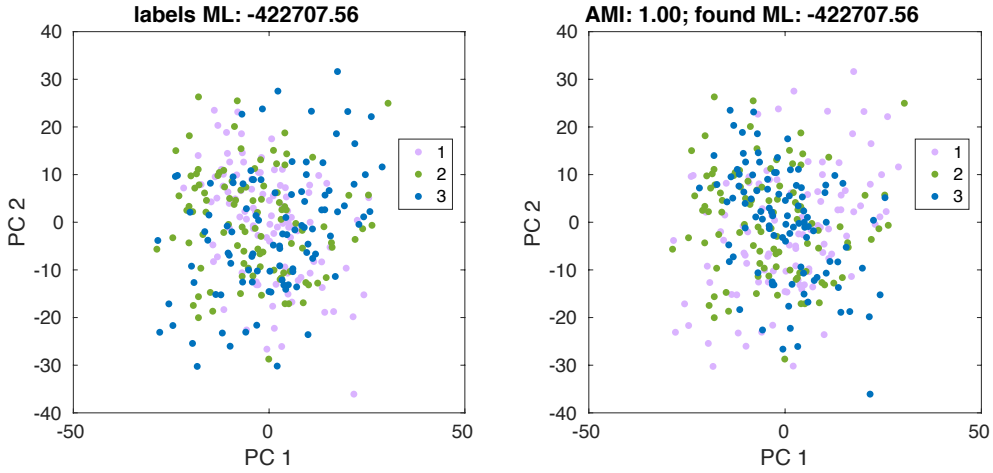


Figure 5.10: *Comparison of true clusters from simulated dataset (left) and model clustering results (right). Presented as a projection onto first two principal components. All examples were correctly assigned (pertinent to our label-switching discussion previously, Cluster 1, 2, and 3 on the left correspond to Cluster 3, 2, 1 on the right, respectively). The marginal likelihood (ML) and adjusted mutual information (AMI) between the two clusterings are also presented.*

Table 5.4: *Simulated data example: Comparison of cluster assignments between hierarchical clustering and our mixture model for multiple data types.*

		Hierarchical			Mixture model		
		1	2	3	1	2	3
True labels	1	37	11	22	0	0	100
	2	22	58	16	100	0	0
	3	41	31	62	0	100	0

$N = 100$ example vectors for $k = 3$ clusters. Unlike the iris flower dataset, when this simulated dataset was projected onto its first two principal components, there was no apparent structure to the data. Hierarchical clustering also failed to recover the true cluster labels, as shown in the confusion matrix in Table 5.4. However, the mixture model was able to identify the underlying clusters, as illustrated in Figure 5.10. The assignment of group labels was different between the two graphs, but the actual assignments matched exactly, as emphasised in Table 5.4.

5.6 Application to IBD patients

Having described our clustering algorithm, we now apply it to our application of clustering data from IBD patients. We examined two experimental questions: (1) whether there are consistent subtypes of patients in terms of their functional gene-expression behaviour; and (2) whether there are groups of patients with similar characteristics

in terms of phenotypic severity.

5.6.1 Functional expression

For experimental question (1), we were restricted to the subset of $N_1 = 45$ patients with gene expression data collected through stimulation experiments[3]. As described in Chapter 3, we recall that the stimulation experiments, which were designed to mimic bacterial activation of monocytes and T-cells, consisted of 12 experimental treatments. These treatments included different combinations of LPS, L18MDP, aIL10Ra, and aCD3/28. For each treatment, five mRNA levels and 50 protein levels were measured. Our goal was to identify patients with similar expression profiles. However, we lack a sufficient number of patients to cluster across all of these $55 \times 12 = 660$ measurements.

Therefore, to reduce the dimensionality of this problem, we focused upon those five cytokines that are known to be heavily involved in our pathways of interest (i.e., those measured in the mRNA experiments). We could cluster across all of the mRNA and protein levels for these cytokines (which would yield $10 \times 12 = 120$ features), or use only the mRNA or the protein levels (60 features). However, as several of the experimental treatments are redundant (i.e., distinct treatments that stimulated the same pathway were applied), and as some of the treatments do not result in stimulation of all of the cytokines, we performed two filtering steps.

First, we limited our experimental treatment set to those in which the correlation between mRNA and protein levels was $\rho \geq 0.6$ (for Pearson correlation coefficient ρ). When mRNA and protein levels are highly correlated, it is indicative that the stimulation is acting upon the cytokine of interest - a coordinated response provides a validation that the patterns across patients are not simply due to noise. This filtering step results in a loss of information concerning post-transcriptional modifications (i.e., cell-mediated differences between mRNA and protein levels), but given the small patient size, the filtering provides a way to reduce the effect of noise on the individual experimental data while also capturing the major expression patterns across patients.

Second, having identified all experiments with correlated findings (e.g., those boxed in Figure 5.11), we then selected a subset so as to reduce the number of duplicated treatments. We did this so that a single treatment would not dominate the clustering results. This effect is shown in Figure 5.11. Here, all treatments with aCD3/28 as a component resulted in a correlated mRNA/protein response. The fact that this response is present with aCD3/28 alone and that it persists with other

[3]These experiments were time-intensive and were conducted by Dr. Maria Quaranta, Nuffield Department of Medicine, as described in Chapter 3.

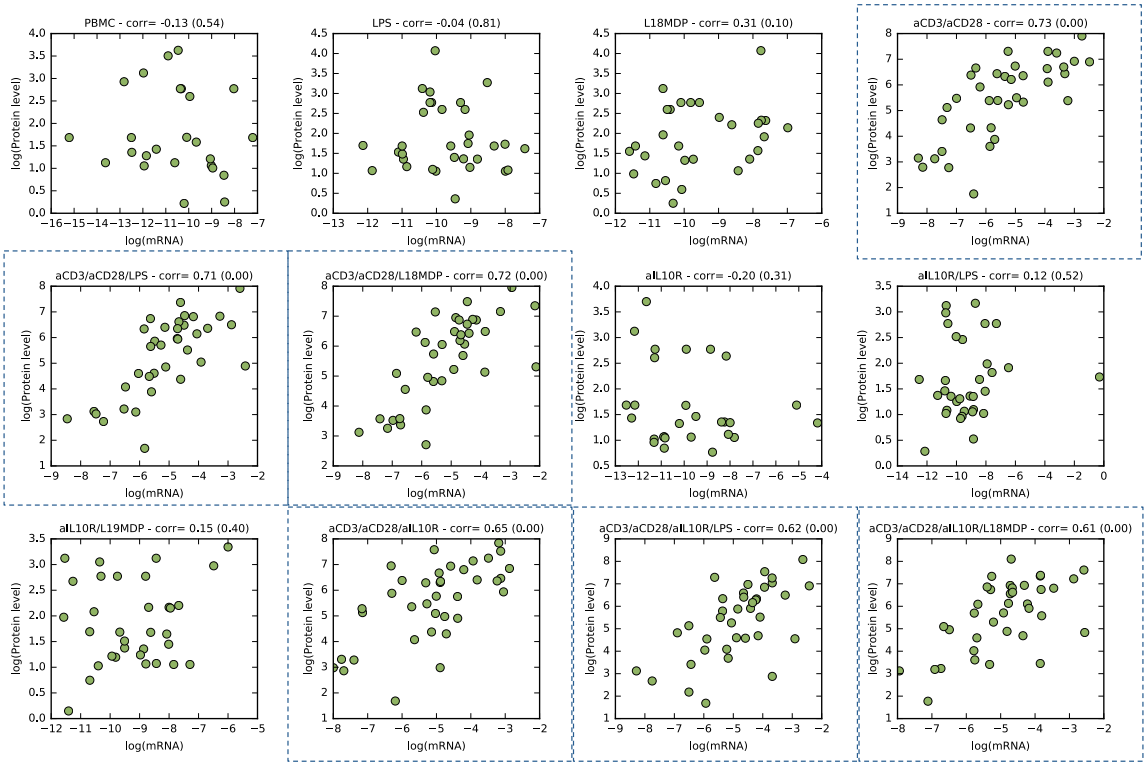


Figure 5.11: *Correlations in mRNA and protein levels for cytokine IL17A. Pearson correlation coefficients, with p-values in parethenses, are shown in titles. Stimulations that resulted in* $\rho \geq 0.6$ *are boxed.*

treatments suggests that it was aCD3/28 driving the response. To include further treatments is redundant and likely is only adding noise, as indicated by the reduction in the correlation coefficient r from 0.72 for aCD3/28 alone to 0.63 when IL17A, aIL10R, and L18MDP were added as well (final subplot in Figure 5.11). To identify the strongest such treatment in each cytokine, we manually examined the correlated experimental treatments in this manner and chose the single "most representative" treatment. This was either the simplest decomposition of the treatment (as above in the aCD3/28 example) or that with the largest correlation in response.

Following these procedures, we obtained a set of five features to include in our clustering algorithm: those shown in Table 5.5. We took the log transform of all features, as the range of variation following perturbation followed a log scale (as is typical in many cellular biology scenarios).

Clustering

Having defined our feature set, we then used our clustering model described previously. Since all features were Gaussian, this was a Gaussian mixture model (we will

Table 5.5: *Experimental treatments inducing strongest mRNA/protein correlation*

mRNA/protein	Selected treatment	ρ	Response
IL17A	aCD3/28	0.72	T-cell
IFNγ	aCD3/28	0.83	T-cell
IL10	aIL10R + L18MDP	0.80	Monocyte
IL23	aIL10R + L18MDP	0.65	Monocyte
IL6	aIL10R	0.79	Monocyte

use the Bernoulli portion of the model as well in the next section). We set the values of the hyperparameters as follows: the Dirichlet concentration parameter $\alpha = 10$ (corresponding to a small prior belief that we have multiple clusters); the normal-Wishart scale matrix $\Sigma_0 = 50000 \times \mathrm{I}_{5x5}$ (where we believed that the covariance within clusters was small), the normal-Wishart parameter $\beta = 1$ (no strong prior for our mean estimate), and the normal-Wishart degrees of freedom $\nu = 5$ (with five features, ν must be ≥ 5).

We performed five rounds of clustering, where we held out 10% of our patients in each round. We then found the coherent clusters across these clustering rounds by matching clusters based upon patient overlap (as presented in Section 5.4.2). We ran 1500 iterations of the Gibbs sampler for each run, and we disregarded the initial 500 runs as burn-in. We then used our relabelling algorithm to match-up cluster assignments across iterations. To find a patient's final clustering assignment z_i based upon the assignments z_{ij} in each iteration j, we took a marginal likelihood-adjusted mode. This is

$$z_i = \arg\max_k(\boldsymbol{S_i}) \tag{5.28}$$

where

$$S_{ik} = \sum_j \phi_{nj} \mathbb{1}(z_{ij} = k) \tag{5.29}$$

with the normalised marginal likelihoods, $\boldsymbol{\phi_{norm}}$ as

$$\boldsymbol{\phi_{norm}} = \frac{\boldsymbol{\phi_r}}{\max(\boldsymbol{\phi_r})} \tag{5.30}$$

where $\boldsymbol{\phi_r} = \boldsymbol{\phi} + |\min(\boldsymbol{\phi})|$, with $\boldsymbol{\phi}$ being the marginal likelihoods for each iteration. In other words, we weighted the assignment from each round by the marginal likelihood (relative to the maximum marginal likelihood obtained across all 10 rounds) of that round, summed over all rounds, and then took the assignment with the maximum value as our final assignment.

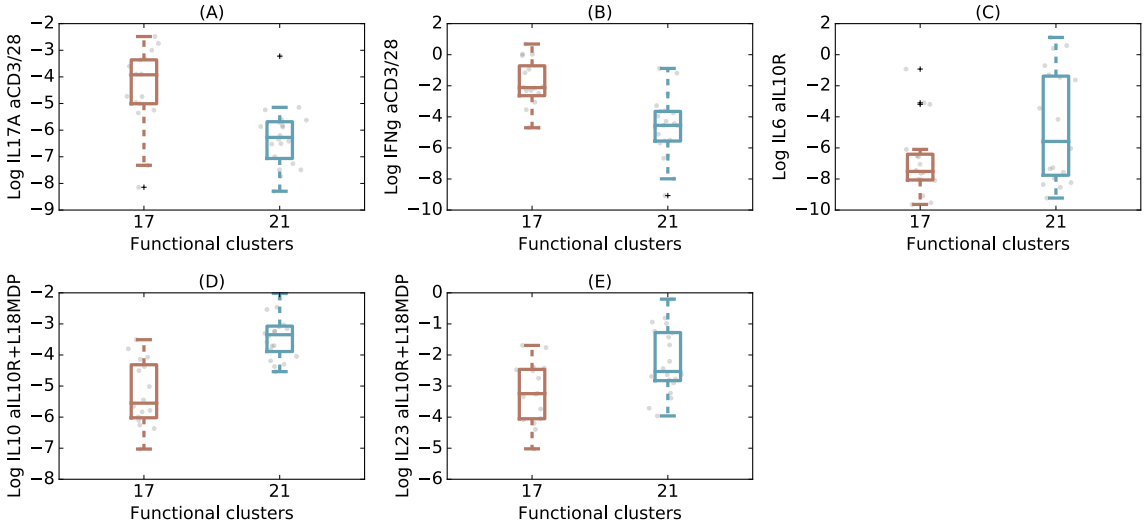


Figure 5.12: *Group distributions for the two clusters identified (left: high T-cell, low monocyte response; right: low T-cell, high monocyte response), for the input features. The number of patients in each cluster are shown below on the x-axis (i.e., 17 patients and 21 patients, respectively).*

Results

We found two clusters within our dataset (diagnostic plots are shown in Appendix B.2). The first cluster can be described as a group with high T-cell responses but low monocyte responses ($N = 17$), while the second group had high monocyte responses but low T-cell responses ($N = 21$). We will refer to these groups as $FC1$ and $FC2$, respectively. Summary graphs of the input continuous features for these two clusters are shown in Figure 5.12, with $FC1$ (high T-cell/low monocyte group) in brown (left) and $FC2$ (high monocyte/low T-cell group) in blue (right).

As is apparent in subplots (A) and (B), the $FC2$ group had very little IL17A or IFNγ production in response to stimulation with aCD3/28. These stimulations were designed to ellicit T-cell response, indicating a dampened response in the $FC2$ group of patients in comparison to $FC1$. For instance, the median level of IFNγ upon stimulation with αCD3/28 was 0.12 for $FC1$, while it was 0.011 for $FC2$. In contrast, in subplots (C), (D), and (E), we observe the opposite pattern. When stimulated with aIL10R and L18MDP, both of which stimulate monocytes, levels of IL10, IL23, and IL6 were high for $FC2$ but low for $FC1$. This indicates that monocytes in $FC1$ patients were responding to these stimulations, while monocytes in $FC2$ patients had a much lower level of response. For instance, stimulation with aIL10R and L18MDP resulted in median IL10 levels of 0.0040 for $FC1$, in comparison to levels of 0.035 for $FC2$. These stark contrasts suggest that these two groups of patients have some

Table 5.6: *Input feature set for severity phenotype clustering*

Binary	Continuous
Prescription of ASA	Age of diagnosis
Prescription of steroids (PRED, BUD)	CRP AUAS
Prescription of immunosuppressants (AZA, 6MP, CYC, MTX, MYC)	Haemoglobin AUAS
Prescription of biologics (IFX, ADA)	White blood cell AUAS
Requirement for surgery	Neutrophil AUAS
Ileum inflammation or not	Monocyte AUAS

difference in underlying physiology.

5.6.2 Severity

We have seen that both lab metrics and outcomes such as medication prescriptions are inadequate, alone, to represent patient severity. For instance, there may be patients with very mild systemic inflammation, but this may simply be because they are on intensive medication treatment or have had extensive surgery. We need to take into account all of these factors in order to understand the severity patterns present across our available variables. We therefore have included information contained in both classical clinical severity outcome measures (e.g. whether surgery is required, various medication prescriptions) as well as our quantitative, novel lab-measurement-based metrics. In addition, we have included age of diagnosis and ileal disease localisation, as these are both correlated with disease "severity." As a result, our input feature set consisted of both binary and continuous features: these are presented in Table 5.6. Our AUAS (area-under-anomaly-score) metrics for each lab were as derived in Chapter 4. Again, as previously, we used nonsevere patients who have not been genotyped to train these EVT model parameters ($N_{tr} = 599$).

We then applied our clustering model, with both binary and continuous features, across those patients with at least three years of follow-up and who had been genotyped ($N_2 = 646$). We clustered (i) across all IBD patients, (ii) within CD patients, and (iii) within UC patients. Similarly to our functional clustering model, we set $\alpha = 1$, $\Sigma_k = 500 \times \mathrm{I}_{6x6}$, $\beta = 1$, and $\nu = 6$. Again, we performed five clustering rounds, holding out 10% of the patients in each round, with 1500 Gibbs sampler iterations (with the first 500 considered as burn-in).

Results

Our clustering results provide a representation of the complex array of patient subtypes within IBD (diagnostic plots presented in Appendix B.2). For IBD as a whole, we obtained eight clusters, which can each be described with the identifying char-

Table 5.7: *Defining characteristics of IBD clusters*

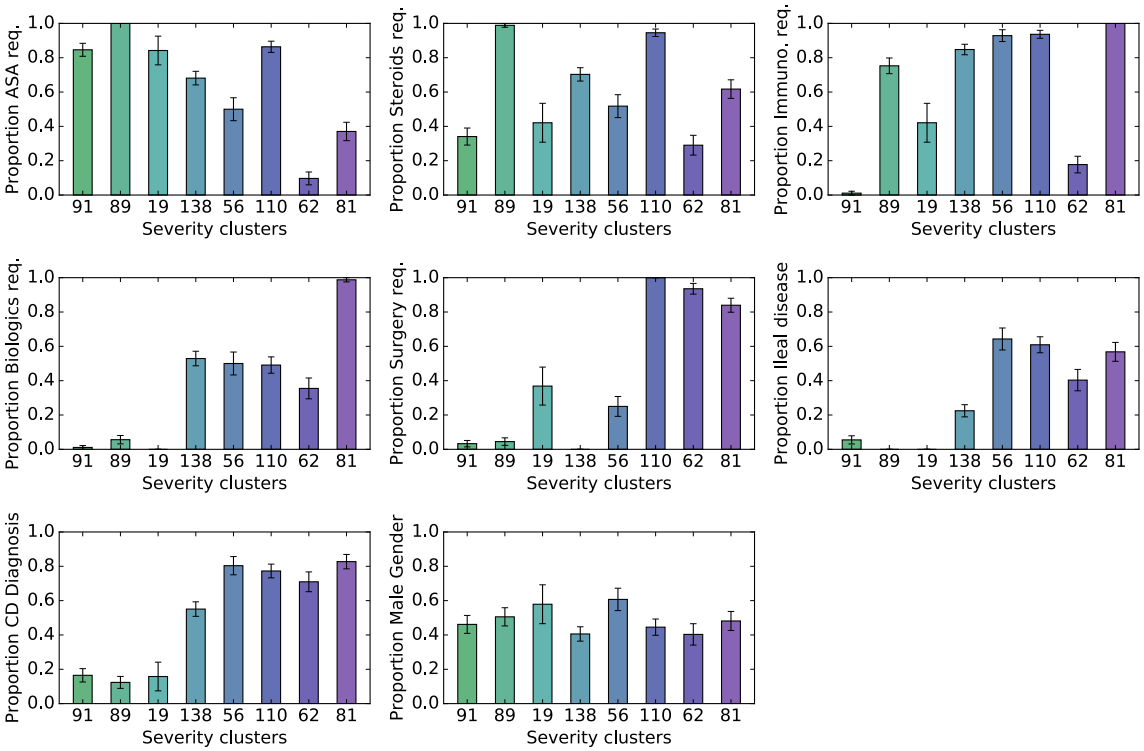
N patients	Description	CCCS
91	Very mild disease	Nonsevere
89	Mild, but require immunosupp.	Nonsevere
19	Late diagnosis age	Severe
138	Moderate disease; no surgery	Severe
56	Mild labs; severe clinical outcomes	Severe
110	Severe disease	Severe
62	High surgery rate; low drug prescription	Severe
81	Very severe disease	Severe

acteristics presented in Table 5.7. We will refer to these clusters with a subscript to indicate the number of patients in the cluster (e.g., C_{91} refers to the first "very mild disease" cluster in Table 5.7, with 91 patients). As shown in Figure 5.13 (a), cluster C_{91} patients generally were prescribed ASA (a first-line medication), but they less commonly were prescribed steroids (34.1%) and almost never were prescribed immunosuppressants or biologics or require surgery (1.0%, 1.0%, 3.3%, respectively). As shown in Figure 5.13 (b), their AUAS severity metrics were low in comparison to the other clusters, with a median AUAS of 0 for CRP, monocytes, and haemoglobin. Their age of diagnosis was also generally late, at a median of 37 years old. Due to these characteristics, we have described C_{69} patients as having "very mild disease."

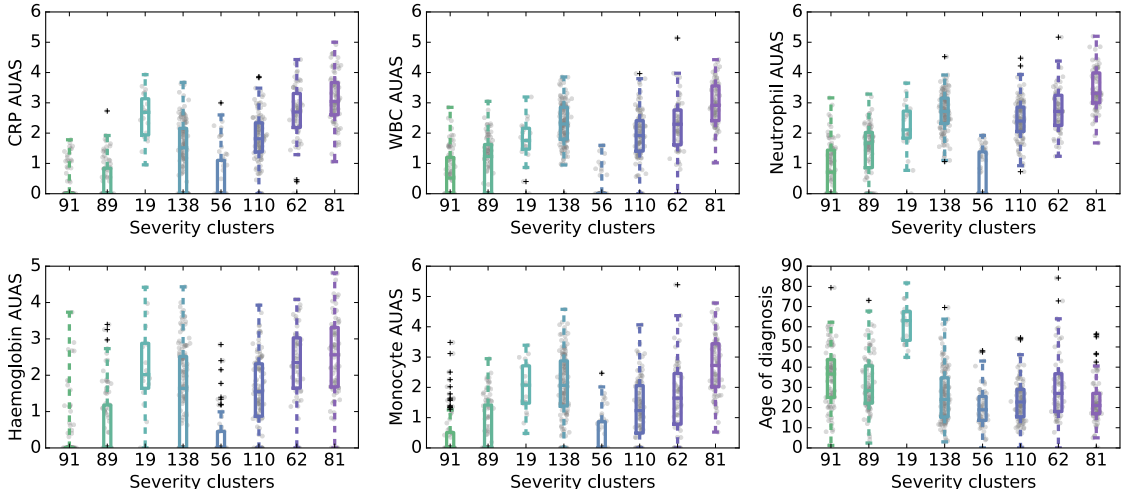
In contrast, as shown in Figure 5.13 (A), patients in cluster C_{81} were nearly all prescribed immunosuppressants and biologics and required surgery (100%, 98.8%, 84.0%, respectively). Their AUAS metrics, shown in Figure 5.13 (B), were extremely elevated, with a median of 3.0, 2.9, 3.3, 2.7, and 2.6 for CRP, WBCs, neutrophils, monocytes, and haemoglobin, respectively. Their median age of diagnosis was low as well, at 21 years old. These patients have had very severe disease (see Table 5.7).

The remaining clusters exhibited characteristics between these two extremes, but with specific distinguishing factors. For instance, again as illustrated in Figure 5.13, cluster C_{89} was similar to cluster C_{91} in that these patients also did not require biologics or surgery and also had relatively mild lab AUAS levels. However, they can be distinguished in that these patients did require steroids (98.9%) and often required immunosuppressants (75.3%). As labelled in Table 5.7, this group of patients can also be considered to have relatively mild disease.

Cluster C_{19} represented a subgroup of patients with a very late diagnosis age (median of 63 years old). These patients had high AUAS metrics (≥ 1.8), but low rates of biologic therapy or surgery (0.0% and 36.8%, respectively). It could be that these patients had mild disease but other comorbidities (due to their older age) that



(A) Binary features



(B) Continuous features

Figure 5.13: *Summary characteristics of IBD clusters, for both input features for clustering algorithm and additional informative features (diagnosis, gender). Numbers on x-axis beneath clusters indicate the number of patients assigned to that cluster.*

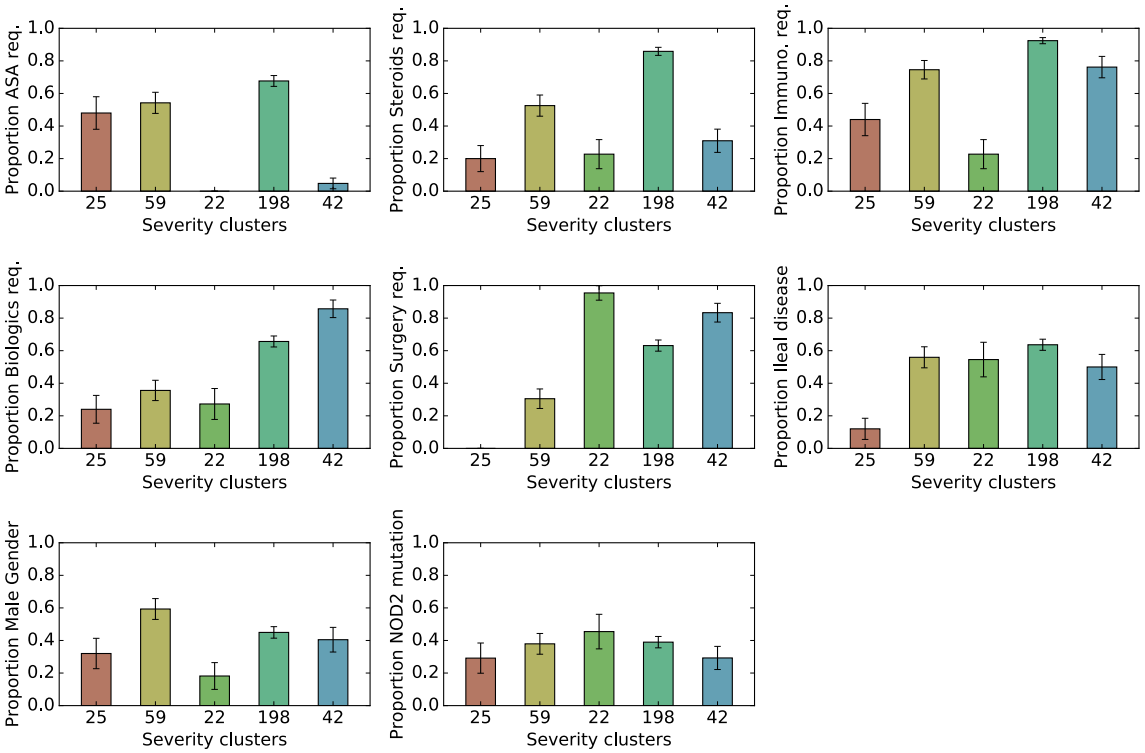
resulted in high numbers of abnormal lab measurements. Cluster C_{138} represents a group of patients with rather severe AUAS metrics ($\geq$ 1.5) and rates of drug prescription (e.g. 52.9% requiring biologics), but they were notable in that these patients have not required surgery (0.0%).

Cluster C_{56} patients did progress to severe outcomes, with 25% patients requiring surgery, 92.9% requiring immunosuppressants, and 50% requiring biologics, but as visually apparent in Figure 5.13, their lab AUAS metrics were very low, with a median of 0.0 for all measured labs. Those patients in cluster C_{110} seemed to represent a set of patients with severe disease, but of a slightly less severe degree than those in cluster C_{81}: cluster C_{110} patients had higher rates of prescription of the first-line drugs ASA and steroids (86.3% and 94.5%, respectively) and also lower lab AUAS metrics (median of $\leq$ 2.4 for all labs). Cluster C_{62} patients were notable in that they had very high rates of surgery (and high lab AUAS metrics), but as shown in Figure 5.13 (A), very low rates of other drug prescriptions (9.7% requiring ASA and 17.7% requiring immunosuppressants, for instance).

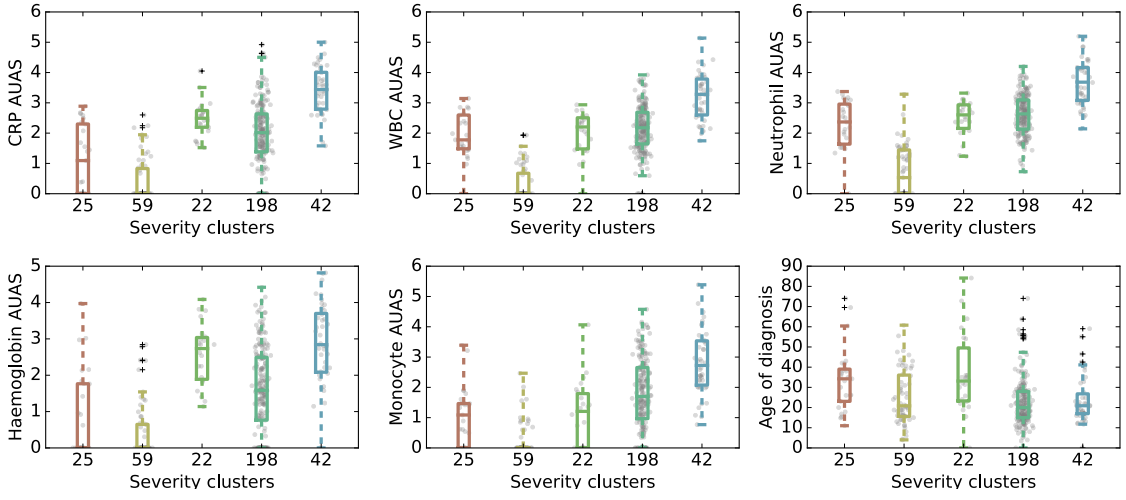
Our analysis therefore reveals the presence of these differing "types" of patients within the cohort. Some of these subgroups may represent differing progressions through disease treatment. For instance, C_{62} likely represents severe patients who quickly progress to surgery because their prescription of all drugs is low, while the cluster has a very high rate of surgery. Cluster C_{138}, on the other hand, likely represents those patients with slightly less severe disease who have been managed through medications alone. Cluster C_{56} represents a patient group with relatively severe outcomes, but with very mild lab measurements. These patients may not mount a CRP response or have some other notable underlying factor that diminishes their peripheral blood-based response to inflammation.

We also see that the severity of our clusters was related to disease diagnosis. As shown in Figure 5.13 (A), the proportion of patients with a CD diagnosis was higher in those more-severe clusters. Within the "'very mild disease" cluster C_{91}, for example, 16.5% of patients had CD, whereas 82.7% of the patients in the "very severe disease" cluster C_{81} had CD. Notably, however, this overlap was not one-to-one. There were still 12 UC patients in the very-severe cluster C_{81} (15%), and there were 15 CD patients in the very-mild cluster C_{91} (16%). This indicates that, in terms of phenotypic severity, the diagnoses of CD and UC overlap to some degree.

This correlation between severity and diagnosis motivates our clustering within each of these diagnoses individually. Figure 5.14 illustrates the summary characteristics of each identified cluster. Descriptions of these identified clusters are shown in



(A) Binary features



(B) Continuous features

Figure 5.14: *Summary characteristics of CD clusters, for both input features for clustering algorithm and additional informative features (gender, NOD2 status). Numbers on x-axis beneath clusters indicate the number of patients assigned to that cluster.*

Table 5.8. We see that the majority of CD patients fell into a single "severe" category, $C_{cd,198}$, in which 65.7% of patients required biologics, 63.1% required surgery, and the median AUAS metrics were 2.0, 2.2, 2.6, 1.6, and 1.7 for CRP, WBCs, neutrophils, haemoglobin, and monocytes, respectively.

A small subgroup, $C_{cd,25}$ had more mild disease, with 24% requiring biologics and 0% requiring surgery, as shown in Figure 5.14 (A). As shown in Figure 5.14 (B), however, these patients did have elevated lab measurements, with median AUAS metrics of ≥ 1 for CRP, WBCs, neutrophils, and monocytes. Another small group, cluster $C_{cd,59}$, had the opposite behaviour, with very low AUAS metrics (median of 0.0 for CRP, WBCs, monocytes, and haemoglobin), but with a higher proportion of patients requiring steroids, immunosuppressants, biologics, and surgery than C_{25}, as shown in Figure 5.14 (A).

Cluster $C_{cd,22}$ was mostly similar to the large cluster $C_{cd,198}$ in terms of AUAS metrics (as shown in Figure 5.14 (b); median of 2.5, 2.2, 2.6, 2.7, and 1.2 for CRP, WBCs, neutrophils, haemoglobin, and monocytes, respectively). However, this group had low rates of all drug prescriptions (fewer than 22.7% of patients in this group), combined with very high rates of surgery (95%), as shown in Figure 5.14 (A).

Cluster $C_{cd,42}$ appeared to phenotypically be the most severe subgroup. These patients had a very low rate of ASA prescription (4.8%) and high rates of biologics (85.7%) and surgery (83.3%). They also had very high AUAS metrics (median of ≥ 2.7 for all labs). These patients seem to have quickly progressed to the most severe drug classes. Their high AUAS metrics, however, suggest that they have remained in a severe state. These high AUAS metrics do not seem to be simply because of single severe episodes and short time of follow-up; median follow-up for $C_{cd,42}$ patients was 13 years, while it was 14 years for $C_{cd,198}$ patients.

Generally, all identified clusters within CD exhibited more severe characteristics than within UC, as illustrated by contrasting Figure 5.14 with the summary characteristics for UC patients in Figure 5.15. Rates of biologic drug usage and surgery were much lower for the UC patients.

The majority of UC patients fell into two mild categories: $C_{uc,116}$ and $C_{uc,101}$, as also described in Table 5.8. Cluster $C_{uc,116}$ patients had very mild disease, with 0.0% requiring biologics, 1.7% requiring surgery, and median AUAS metrics of ≤ 1.0 for all examined labs. As illustrated in Figure 5.15, cluster $C_{uc,101}$ was similar to cluster $C_{uc,116}$, but cluster $C_{uc,101}$ exhibited slightly higher rates of biologic therapy and surgery (29.7% and 23.7%, respectively) and higher AUAS metrics (median of 1.3, 1.9, 2.4, 1.3, and 1.5 for CRP, WBCs, neutrophils, haemoglobin, and monocytes,

Table 5.8: *Defining characteristics of CD and UC clusters*

CD			UC		
N	Description	C	*N*	Description	C
25	Mild disease	N.	116	Very mild disease	N.
59	Mild; higher steroid/imm.	N.	101	Mild disease; more severe labs	N.
22	High rates drug prescrip.	S.	25	Severe disease	S.
198	Severe	S.	24	Severe disease; higher drug prescrip.	S.
42	Very severe	S.			

Abbreviations: C: CCCS designations; imm. : imunosuppressive drug prescription; N: nonsevere; S: severe

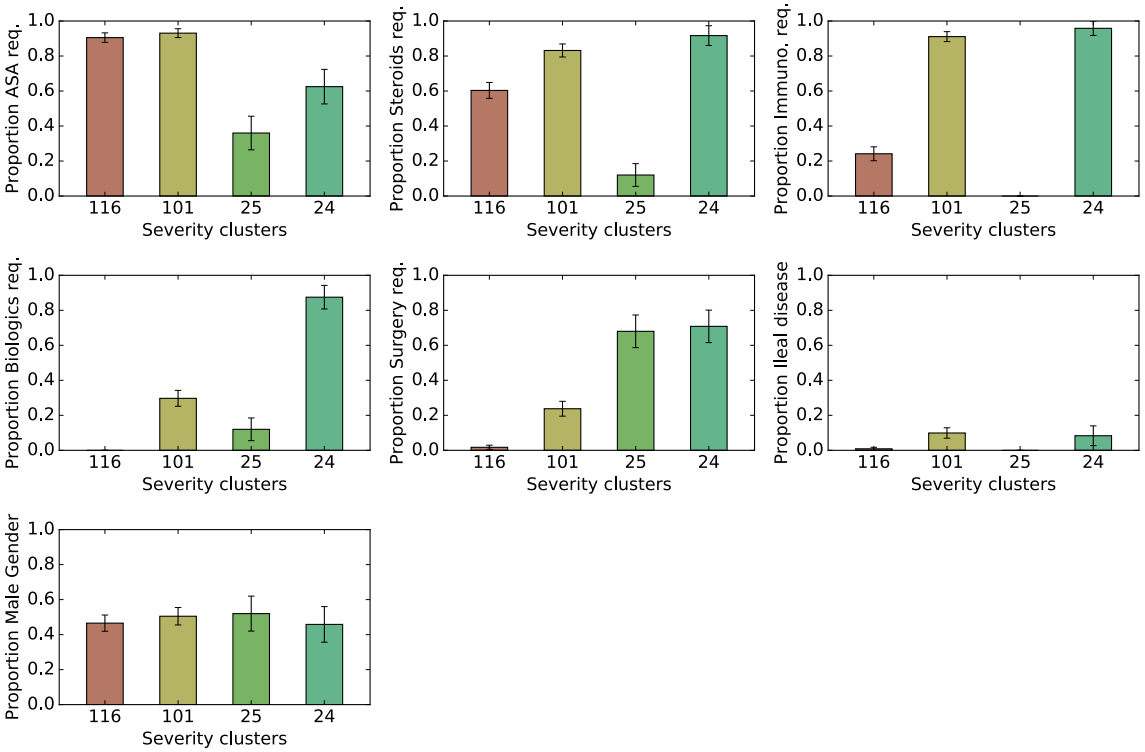
respectively).

Cluster $C_{uc,25}$ patients had more severe disease, with 68.0% requiring surgery and AUAS metrics of ≥ 2.3 for CRP, WBCs, and neutrophils. Cluster $C_{uc,24}$, as shown in Figure 5.15, exhibited the most severe characteristics of all the other UC clusters. Most of these patients (87.5%) required biologics, and 8% required surgery. Their AUAS metrics were also very elevated, with medians of ≥ 2.6 for all labs.

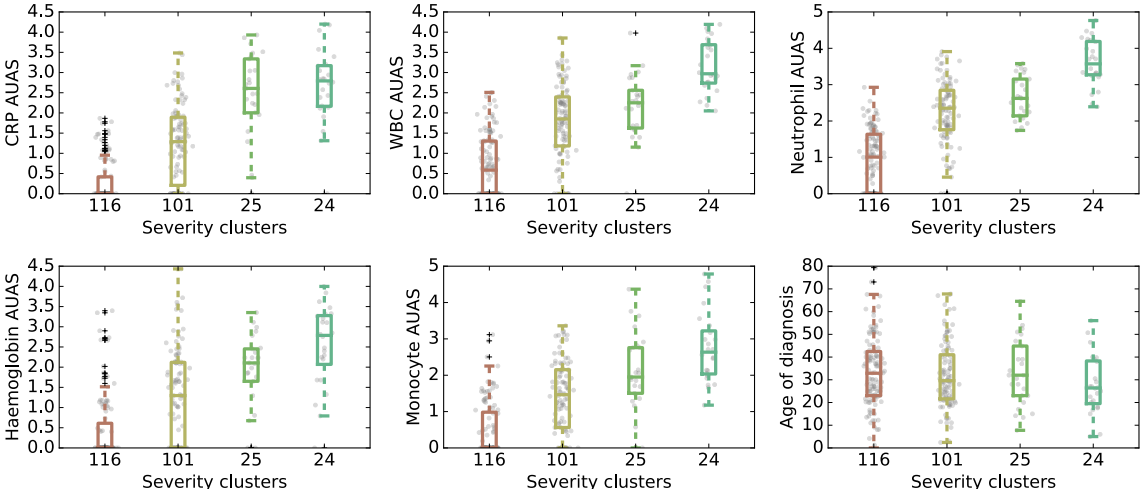
Comparison between severity metrics We conclude that our approach provides us with a more holistic representation of patient severity than a rule-based representation as is commonly employed by clinical studies. We have obtained several patient subgroups with distinct phenotypic patterns, which shows the many different ways in which an IBD patient can be "severe." Some of these patient groups have elevated labs combined with mild treatment; others have the reverse pattern. We also see a spectrum of severity within each disease subcategory (i.e., within UC and CD).

To compare our subgroups with CS classifications (i.e., requiring biologics or surgery), we have manually collapsed our subclusters into two groups. We refer to these two-class clusters as collapsed composite clustered severity (CCCS), both here and in future chapters. For IBD, we combined C_{91} and C_{89} into a "nonsevere" class, with the remainder of patients being "severe." Similarly, for CD, we combined $C_{cd,25}$ and $C_{cd,59}$ into a "nonsevere" class; for UC we combined $C_{uc,116}$ and $C_{uc,101}$ into a "nonsevere" class. The summary distributions defining these CCCS groups are presented in Appendix B.2.

When we looked at the correspondence between the CCCS and CS groups, there was a great deal of overlap, as shown in Table 5.9. Within IBD patients, 362 patients were classified as severe through both methods. However, 18% of patients were classified differently between the two methods. Our CCCS groups provided a conservative



(a) Binary features



(b) Continuous features

Figure 5.15: *Summary characteristics of UC clusters, for both input features for clustering algorithm and additional informative feature (gender). Numbers on x-axis beneath clusters indicate the number of patients assigned to that cluster.*

Table 5.9: *Confusion matrix between our CCCS and classical (CS) definitions*

	CS - IBD		CS - CD		CS - UC	
	Nonsevere	Severe	Nonsevere	Severe	Nonsevere	Severe
CCCS nonsevere	167	13	47	37	165	52
CCCS Severe	104	362	28	234	9	40

representation of severe patients, with many more (104) patients being considered severe through our approach than through the CS definition alone (while only 13 patients requiring surgery/biologics were considered nonsevere in CCCS groups). In other words, additionally considering lab relapse events allowed us to capture an additional set of patients who exhibit severe disease trajectories, despite not (yet) progressing to surgery or biologic therapy. Thus, this supports our conclusion that our metric may better capture severity than conventional metrics. Only 13 patients who were classified as severe according to the CS definition were classified as nonsevere according to the CCCS definition. These patients may have had a very mild lab-based and drug-based disease course despite having a surgery or requiring biologics (an additional possibility is that there might have been an error in the manual transcription of their clinical notes into the clinical database).

Within CD patients, both methods classified the majority (234) of patients as severe, while the majority of UC patients (165) were classified as nonsevere by both methods. This fits with our finding that CD patients generally exhibit more severe behaviour than UC patients. Similarly to the case within IBD as a whole, 18.8% of CD patients and 22.3% of UC were classified differently between the traditional and collapsed clustering methods.

Within CD and UC, we see a substantial number of patients who were severe using CS (37 patients for CD and 52 for UC) who, according to CCCS, fell into a nonsevere category. This was partially because CCCS combined patient subgroups with varying levels of "severity"[4]. However, this finding also points to the fact that these patients had more characteristics in common with nonsevere patients than severe across the various severity metrics under consideration (e.g., medication prescriptions, age of diagnosis, lab measurements). This might mean that, despite requiring these last-line treatments, the patients would actually be better considered as nonsevere. Depending upon the question at hand, a conservative approach for identifying associations related to severity could consider all patients classified by either method as severe

[4]We note that by applying different labels (particularly within UC) to collapse the clusters, we could obtain different results in terms of sensitivity and specificity in relation to CS. For instance, we could use only $C_{uc,116}$ as our nonsevere class within UC.

or, alternatively, consider only those patients who are classified by both methods as severe.

5.7 Discussion

IBD is a highly heterogeneous disease, with many potential inflammatory mechanisms that might be disregulated differently across patients. This leads us to hypothesise that there are likely to be latent subgroups of patients with consistent phenotypic presentations. Just as patients have heterogeneous presentations, these presentations are described by heterogeneous data types. We have therefore created a clustering model that can combine information across these disparate data types, particularly given that we often are missing features for particular patients and that we have no prior information concerning how many clusters there should be.

Our model employs a Bayesian framework, which allows us to incorporate prior information (e.g., prior belief concerning the number of clusters, how much covariance we expect within these clusters, etc.) and address the various challenges just mentioned. Given our pre-specified prior parameters, we have then used an empirical Bayes approach to compare models. We found that in simulated scenarios, we were able to recapitulate complex clustering structure that is not evident through conventional methods, such as hierarchical clustering.

5.7.1 Extensions

While our current model provides the functionality we require for our problem, there are numerous extensions and alternatives that could be pursued in future research. One area is adjustment of our sampling approach. Rather than using MCMC, we could employ a VB inference procedure. VB provides a fully Bayesian alternative for estimating model parameters that is typically faster than MCMC approaches. VB is similar to EM in that it is an iterative procedure, but it produces estimates of distributions over parameters rather than single point estimates. It is attempting to minimise a metric (the Kullbeck-Leibler divergence) of similarity between the "true" and estimated distributions.

Regarding our model specification, we currently have implemented our model with the assumption of independence between our binary and binomial variables. We could extend the model to the situation of non-independent forms of these distributions. For instance, the multivariate Bernoulli distribution is described by Dai et al. [174]. In addition, we are currently working with relatively small feature sets (i.e., fewer than 20 features). Were we to consider larger feature sets, we may wish to introduce sparsity-

promoting regularisation terms into our model. In addition, the inverse-Wishart or Wishart as a prior introduces dependencies into the multivariate normal portion of each mixture component: the variance and correlation are modelled jointly[5]. This means that a high variance within a particular cluster will indicate high correlation amongst features.

An extension that could be applied to other applications is to weight certain features more heavily than others. In our current implementation, we put equal weighting over each data type. However, we could also introduce additional weighting hyperparameters over each data type within our likelihood equation to allow certain features to carry more weight in the clustering. This would allow us to create a clustering that is driven primarily by continuous Gaussian data, for instance, with the other features playing a smaller role (or vice versa).

Finally, while we have taken a straightforward relabelling approach to mange the label-switching problem, there have recently been developments of probabilistic approaches that incorporate the re-labelling directly into the model formulation itself (e.g., [175]). These latter methods could be considered in future work.

5.7.2 Clinical conclusions

When we applied our model to IBD patient data, we gained valuable insights into the patient data sub-structure. We first examined whether there were consistent patterns of gene expression within a small subset of patients with this form of data available. We found two subtypes of patients: one group with high expression of cytokines related to monocyte pathways combined with low expression of cytokines related to T-cell pathways, and one group with the reverse pattern. This is an important finding that suggests that perhaps these two groups of patients may have different underlying genetic backgrounds or that they may have distinct phenotypic progressions of disease. We believe it is worthy of future study in larger patient cohorts.

We then focused upon identifying subtypes of IBD patient "severity." We included a variety of clinical features typically associated with severity in this model, such as progression to GI surgery, medication prescriptions, and our previously-developed lab metrics. We examined a large clinical cohort of several hundred patients and found that across IBD as a whole, severity was split predominantly by disease diagnosis (e.g., UC versus CD). We also found several interesting subgroups, such as a small

[5]This is because the covariance is modelled with a single prior, though it can be decomposed as $\Sigma = Q\rho Q$, where Q is a diagonal matrix with entries of $\frac{1}{\sqrt{\lambda^d}}$, λ^d is the precision for feature d, and ρ is a matrix of correlations. If the precision of a particular feature increases, this must mean that the correlation between features also increases.

group of patients with very late diagnosis age (> 60 years old) and consistent severity features. Upon clustering within each disease categorisation, we found that within UC, there were clear sets of patients with differing severity profiles: a group of very mild patients ($C_{uc,116}$); a group with mild disease, but higher AUAS values ($C_{uc,101}$); a group with severe disease ($C_{uc,25}$); and a group with severe disease and higher rates of drug prescription ($C_{uc,25}$). Within CD, we also identified subgroups with progressing levels of severity, though the majority of patients did fall into a single subgroup ($C_{cd,198}$). There was a group of mild patients ($C_{cd,25}$), a group that did not exhibit a large lab-based response ($C_{cd,59}$), a group that likely progressed quickly to surgery after diagnosis ($C_{cd,22}$), and a group of very severe patients ($C_{cd,42}$).

These findings provide clarity to the IBD patient population's phenotypic patterns. We found that, in terms of severity phenotype, there is indeed a difference in CD and UC patients. However, there are some patients with CD who have very mild disease and some with UC with very severe disease; our severity clusters provide a methodology for grouping these patients together, which is relevant for investigating relationships between severity and genetics across IBD patients as a whole. In addition, we were able to identify several different subtypes of UC and CD patients, with distinctions between those who have a very mild disease course and those who, while not requiring surgery or biologic therapy, still have a large number of lab-related relapse events. It seems as though these subgroups may represent differing treatment patterns (which, again, reflects underlying severity). For instance, patients with low AUAS lab metrics combined with intensive medical therapy and/or surgery likely had very severe disease initially, but then were able to maintain remission.

We believe that as larger cohorts of patients are collected, our model's results will become increasingly useful in the clinical setting, as adequate patient numbers within disease diagnosis categories will allow the rich phenotypic structure that exists to be fully explored. In addition, we can relate these patient subgroups to other variables, such as genetic background, to determine whether there is a genetic basis for these differences (as is the subject of the next chapter).

Chapter 6

Associations between genotype and phenotype

6.1 Introduction

The purpose of the analysis in this chapter is to investigate the presence of univariate associations between the patient phenotype and the underlying genotype. In our previous chapters, we have been developing approaches for capturing and summarising relevant phenotypic characteristics. While the literature has reported some evidence of associations between patient genetic variance and IBD severity, as described in detail in Chapter 1, this has been primarily investigated in terms of the classical Montreal classifications or the requirement of last-line treatment, such as surgery or biologic therapy. As presented in Chaper 3, we have information on our IBD patient cohort from multiple different dimensions of enquiry: genetic, gene expression, and longitudinal clinical data. We therefore have the opportunity to investigate the relationship between genetic variation and intermediate measures of inflammation, which likely give rise to the final, more-easily observed outcomes of biologic therapy and surgery. In Chapter 4, we presented our point-process modelling approach to capture lab-based severity using a single metric. In Chapter 5, we explained how to combine various types of input phenotypic data (including our lab-based metrics and other clinical metrics such as drug prescriptions) to identify clusters of patients with similar severity profiles. The next step, in this chapter, is to relate both our lab-based metrics and severity subtypes to patient genetics. Though our current sample size is under-powered to detect significant associations via multiple hypothesis testing, we can use our methods as a hypothesis-generating step, which can then be further validated in future work.

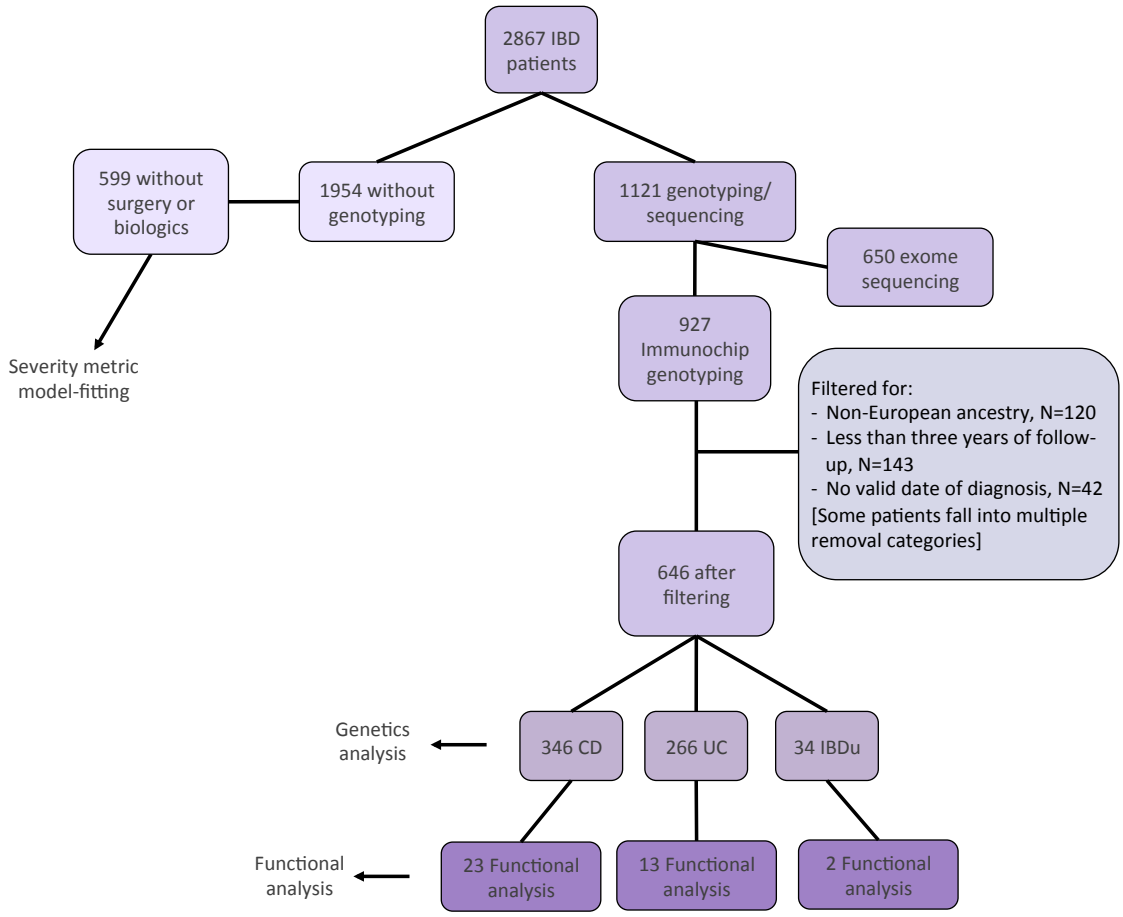


Figure 6.1: *Illustration of the patient numbers available for different portions of the analysis. Abbreviations: CD: Crohn's disease; UC: ulcerative colitis; IBDu: IBD, undetermined.*

6.2 Data

We began analysis with the patient cohort described earlier in Chapter 3. We limited our investigation to the patients of European ancestry so as to reduce any effects of confounding due to population structure. We also limited ourselves to non-relatives to reduce the effect of a shared environment on apparent genetic effects [1]. In addition, since we are investigating severity metrics, we disregarded records for patients who had fewer than three years of follow-up data, so as to not confound our analysis by those patients with insufficient data to characterise the phenotype. For the use of lab-based metrics to be meaningful, we also required a date of diagnosis for each patient, and so we also discarded records from patients without a valid data of diagnosis

[1] Identification of patients of non-European ancestry and relatives was performed by Luke Jostins at the Wellcome Trust Centre for Human Genetics, University of Oxford.

available at the time of analysis. The composition of our patient population is shown in Figure 6.1.

6.2.1 Genetic data

The source for our genetic data was from Immunochip genotyping and exome sequencing (as described in Chapter 3). We limited our analysis to those Immunochip SNPs that have previously been found to be associated with IBD risk [11] and to those rare variants[2] found in the exome sequence of the *NOD2* gene. Within the *NOD2* exome sequence, we included only those variants that were suspected of being deleterious through either SIFT or POLYPHEN [176, 177]. The latter are two algorithmic approaches for predicting whether an allelic substitution affects the protein structure and function; they are generally more sensitive than specific, but they provide reasonable guidance as to which variants are worthy of further investigation[3] [178]. Table 6.1 shows the list of the loci we considered as deleterious. The pre-processing steps for the Immunochip and exome data are outlined in Appendix A.3.

Table 6.1: *Rare* NOD2 *variants predicted as deleterious by SIFT or POLYPHEN*

rs104895424	rs104895452	rs149002807	rs50750582
rs104895425	rs104895467	rs199883290	rs5743277
rs104895427	rs117836686	rs41450053	rs5743293
rs104895431	rs140876663	rs50744850	rs61747625
rs104895443	rs143110172	rs50745041	rs61755182
rs104895444	rs146435555	rs50745821	rs75934273

We converted all SNPs under investigation in this analysis into "risk-format," by which we mean that we converted the readings so that a "0" indicates that a patient is homozygous for the allele that either is not associated with IBD susceptibility or confers protection; "2" means that a patient is homozygous for the disease susceptibility allele[4]. This conversion was necessary because the Immunochip does not always read the same alleles that have been used in association studies, and sometimes the associations found are protective. We performed this conversion so as

[2]Rare variants are those defined as having a mean allele frrequency of less than 0.01.

[3]Annotations of variants as being deleterious or not were performed by Eva Goncalves Serra at the Wellcome Trust Sanger Institute, Cambridge, UK.

[4]We remind the reader that an allele is one possible genetic variant at a particular loci. Being homozygous for a particular allele means having two copies of that allele, while being heterozygous means having a different allele on each chromosome; refer to Chapter 1 for more details.

to have a consistent meaning throughout our analysis. The logic of the procedure is shown in Algorithm 3, where for each SNP loci, s, there is an associated allele ($a_{m,s} \in [A, T, C, G]$) from previous meta-analysis, an odds ratio for being associated with disease (OR_s) for this allele from previous meta-analysis [11], the allele read by the Immunochip $a_{I,s} \in [A, T, C, G]$, and an Immunochip reading $R_{I,s} \in [0, 1, 2]$. The result is a converted reading, $R_{c,s}$. We have assumed that every loci is biallelic[5]. For instance, suppose we have a loci where the Immunochip reads $a_{I,1} = C$, and previous meta-analysis has yielded an $\text{OR}_1 = 1.3$ for $a_{m,1} = G$. For a patient with Immunochip reading of $R_{I,1} = 0$, this means that they have 0 copies of the Immunochip allele C and therefore two copies of the alternative possibility, G. From previous meta-analysis, G is associated with disease susceptibility, so $R_{c,1} = 2$.

Algorithm 3 Procedure for converting Immunochip SNP readings

```
for s=1:N do                                  % For every SNP,
    if OR_s > 1 then
        if a_{m,s} == a_{I,s} then            % If associated allele = Immunochip allele
            C = 2
        else
            C = 0
    else                                      % If OR is less than 1
        if a_{m,s} == a_{I,s} then            % If associated allele = Immunochip allele
            C = 0
        else
            C = 2
    if C == 0 then                            % Convert readings if necessary
        R_{c,s} = 2 − R_{I,s}
    else
        R_{c,s} = R_{I,s}
```

6.2.2 Longitudinal and clinical data

As before, in this analysis we focused upon the inflammatory marker CRP, haemoglobin (which reflects levels of blood loss in the gut), and various WBC components: neutrophils, monocytes, lymphocytes, eosinophils, and basophils. For each of our labs of interest, we extracted several features that capture relevant aspects of patient trajectories, as presented in Table 6.2.

In addition, we also screened for genomic associations with relevant clinical variables: gender, ileal location of disease[6], requiring surgery, requiring biologics, requiring either surgery or biologics (CS, as defined in Chapter 5), severity as defined by our

[5] Biallelic means that there are only two potential alleles at a given genetic loci.

[6] We remind the reader that disease localisation, as outlined in Chapter 1, is used for

Table 6.2: *Lab features*

- Median
- Minimum
- Maximum
- Proportion of values above or below "normal" threshold
- Number of times labs exceeded risk-threshold (only counted once per 20 days)
- AUAS composite severity metric (see Chapter 4)
- Median of values within risk-threshold
- Value at diagnosis (closest within one year)
- Median of first three months following diagnosis
- Maximum of first three months following diagnosis
- Minimum of first three months following diagnosis

CCCS (collapsed composite clustered severity, as described in Chapter 5), and age at diagnosis. We screened for associations with the gene expression clusters that we also described in Chapter 5. However, because only 38 patients had this gene expression data available, we note that our power to detect any associations was small.

6.3 Association analysis

We were interested in uncovering statistical associations between our different patient data dimensions. We first performed univariate statistical tests between each longitudinal lab marker or severity outcome and each genomic loci of interest. Our dataset was of multiple types (e.g., continuous, binomial, binary), and so multiple different statistical tests were required. For any particular SNP loci, we will define gA = patients with 0 risk variants, gB = patients with 1 risk variant, and gA = patients with 2 risk variants. We used a Kruskal-Wallis H-test (KW) to investigate associations between genetic SNPs and continuous clinical data [179]. This allowed us to determine whether there were differences between the distributions of continuous variables across the three possible categories: homozygous for the risk allele, heterozygous for the risk allele, and homozygous for the non-risk allele. We then performed a screening step, whereby for any significant association, we required the medians of the groups to have either a positive or negative directionality (i.e., a "dose effect"). The KW test is an extension of the Mann-Whitney U test (MW) for the case in which the number of groups under comparison is greater than two; it examines whether the samples from all examined groups are from the same underlying distribution. It is a nonparametric test because it employs the ranks of the samples, rather than their numerical values.

phenotypically defining disease in clinical practice, and ileal localisation has been associated with severity previously [60].

The KW test identifies whether there is a difference between any of the included groups. For example, when examining median CRP levels across our three groups, a significant Kruskal-Wallis result indicates that median CRP levels differ between at least two of the groups. However, it does not determine between which groups this difference exists. We therefore performed follow-up testing with a MW test between groups for those situations in which the Kruskal-Wallis test identified a difference.

We used a Cochran-Armitage test for trend to examine associations between genetic SNPs and binary clinical data. In this case, we calculated $\mathbf{y}$ as the proportion of samples in each group $g = [0, 1, 2]$ (representing $[gA, gB, gC]$, respectively) whose value for the binary variable b is equal to one. In other words,

$$y(0) = \frac{\sum_i^{N_{gA}} \mathbb{1}(gA_{i,b} = 1)}{N_{gA}} \tag{6.1}$$

with N_{gA} being the number of examples in gA. The Cochran-Armitage procedure is a test of the null hypothesis: $m = 0$, where m is the slope of $\mathbf{y} = m\mathbf{g}$ [180]. Essentially, this is determining whether there is an additive association between having additional SNPs that is different between examples with $b = 0$ versus those with $b = 1$.

We illustrate these steps in Algorithm 4, where we define D as the number of dependent variables under examination, C as the set of continuous dependent variables, and α as our threshold for significance.

Algorithm 4 Univariate testing procedure

for $s = 1 : N$ **do** % For every SNP,
  Define groups: $gA = 0$ risk variants, $gB = 1$ risk variant, $gA = 2$ risk variants
  for $d = 1 : D$ **do** % For every dependent variable,
    if $d \in C$ **then** % If d is a continuous variable
      Calculate p_k using KW
      if $p_k < \alpha$ **then** % If significant
        if $\text{med}(gA) < \text{med}(gB) < \text{med}(gC)$ or $\text{med}(gA) > \text{med}(gB) > \text{med}(gC)$ **then**
          Calculate p_f using MW test
    else % If d is a binary variable
      Calculate p_c using Cochran-Armitage test for trend

Multivariate

It has been well-established that the localisation of IBD inflammation (i.e., ileal/ileocolonic/colonic) has a large impact upon disease severity, with ileal disease being more severe than non-ileal [60, 181]. Recent work has also suggested that the phenotypes of

ileal CD, ileocolonic CD, colonic CD, and UC constitute a continuum of disease [10]. In addition, variables such as age and gender can have an effect on a patient's lab measurements. We therefore additionally sought to control for these covariates in our analysis by performing a multivariate statistical analysis through logistic regression. We modelled our genomic SNP count as a binary outcome variable (grouping having one or more risk-associated alleles together: 0 risk-SNPs versus 1 or 2 risk-SNPs), with each lab metric being used as the input variable in turn, and using {inflammation localisation, current age, gender, age at diagnosis} as additional covariates. We report significant covariates and lab metrics using the Wald statistic. The Wald statistic assesses a ratio between the given parameter's coefficient in the model and the coefficient's standard error.

Multiple hypotheses

For each of these sets of tests, we determined the false discovery rate (FDR) using the Benjamini-Hochberg-Yekutieli procedure, which is appropriate when the significance tests are dependent [182]. In our case, fluctuations in the time series lab data are often correlated, meaning that this was necessary. The FDR is a metric that provides more power to identify associations than corrections such as the Bonferroni correction, at the expense of larger numbers of false positives. However, while we calculated the corrected p-value cutoff corresponding to FDR-adjusted significance, we also tabulated results that met nominal significance[7], as this analysis was intended as a hypothesis-generating step. In other words, even if we found no significant associations upon multiple-hypothesis correction, the most-strongly associated findings were still examined for biological plausibility and potential validation.

Step-up procedure

Testing across the entire set of variables obtained from the Immunochip or exome would leave us with either very low power (if correcting for multiple hypotheses) or high numbers of false positives (if using nominal associations in a hypothesis-generating manner). We therefore limited our analysis to those SNPs that had already been previously associated with IBD from GWAS meta-analysis [11]. Such an approach is useful and reasonable because any identified SNPs will already have confirmation of being involved in IBD pathogenesis. Within these SNPs, we constructed three tiers of analysis: (a) targeted candidate loci, (b) protein-coding and eQTL loci, (c) all IBD-risk SNPs.

[7]Nominal significance refers to the significance level that would be apparent if only a single statistical test were performed [183].

Our targeted candidate loci (a) consisted of Immunochip SNPs found in the *IL23R* and *ATG16L1* genes, as well as the three common *NOD2* SNPs. All of these SNPs have either been long-established as being IBD-related or are primary cytokines in inflammatory pathways [13, 184]. This means that our prior belief over their potential involvement in disease severity was larger than that for other genetic loci. As mentioned previously, we supplemented the three common *NOD2* SNPs, which were available through the Immunochip, with additional rare variants identified in *NOD2* through exome sequencing. Only a subset of patients had exome sequencing available, and so we only analysed exome variants via an aggregate "any *NOD2* variant" variable. Our second tier of genetic loci (b) consisted of those IBD-risk SNPs that are either exonic or have evidence of being eQTLs[8]. We obtained annotations of exonic SNPs from [12]. Our final tier (c) consisted of all remaining IBD-risk SNPs, as found in large meta-analysis of European populations [11]. We separated tier (b) from tier (c) because exonic or eQTL SNPs are more easily interpretable and may have larger effect sizes in terms of influencing disease severity than other SNPs. A full overview of our approach is shown in Figure 6.2.

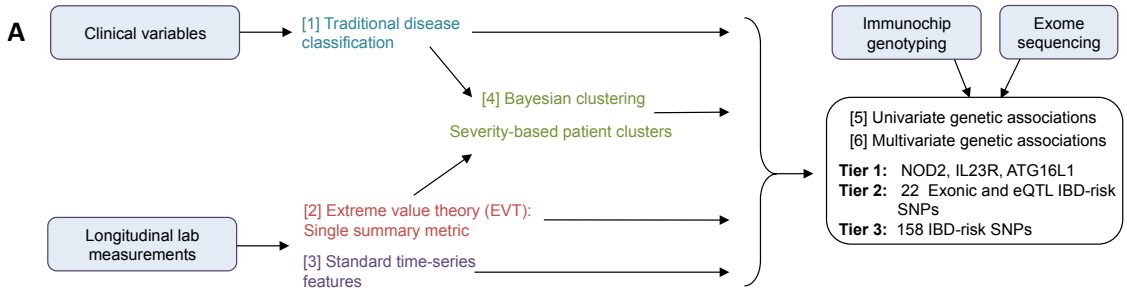


Figure 6.2: *Overview of approach. (1) Clinical variables, such as components of the classical Montreal classification, IBD medications, the requirement for surgery, etc. (2) AUAS severity metrics derived from longitudinal inflammatory data such as CRP and WBC count, etc. (3) Standard time-series features, such as maximum, median, proportion above normal range, etc. for the same CRP, WBC, etc. (4) Clusters obtained from Bayesian clustering of severity-related features, including several of feature set (1) and feature set (2). (5) Univariate and (6) multivariate association study between the above variables (1), (2), (3), (4), and genetic variables. The genetic variables were examined in a tiered manner, with increasing rates of false positives expected with increased numbers of SNPs analyzed.*

[8] As previously described in Chapter 3, exonic SNPs are those that are located within protein-coding regions of the genome. eQTLs are loci that have evidence of regulating the expression of genes elsewhere in the genome.

6.4 Results

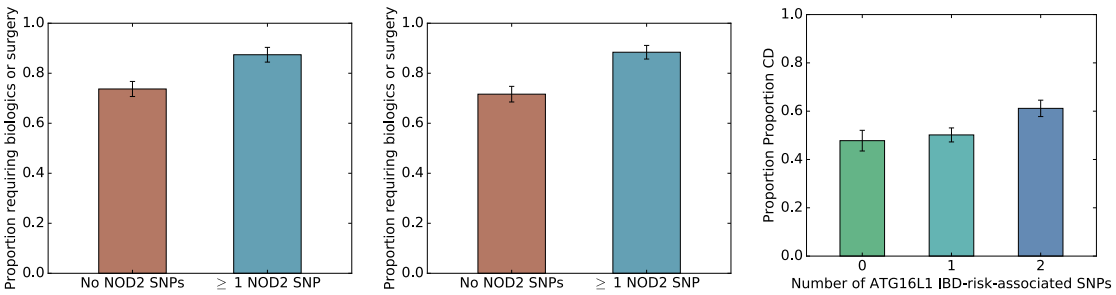
6.4.1 Tier One: NOD2, IL23R, ATG16L1

We identified several expected associations in our dataset, validating previous findings. For instance, as shown in Figure 6.3(a), we found that within CD, *NOD2* variants were associated with the CS metric of requiring surgery or biologics ($p = 0.0043$). We also found (Figure 6.3(c)), again supporting previous findings, that the rs12994997 SNP in *ATG16L1* was associated with CD, rather than UC ($p = 0.0092$). Notably, even these well-established associations with a large effect size did not reach the FDR threshold for multiple-comparison significance. However, we used this stage of the analysis to generate potential new genetic variants for additional follow-up; all p-values presented are uncorrected for multiple hypothesis testing.

We found, as also shown in Figure 6.3(b), that incorporating the rare *NOD2* variants available through exome sequencing further increased the association between *NOD2* and severe outcomes. In other words, patients with rare variants in their *NOD2* gene that were suspected of having a deleterious effect also tended to have more severe outcomes ($p = 0.0004$). This suggests that it is not just the three most common *NOD2* SNPs that are involved in disease pathogenesis; SNPs elsewhere in the gene indeed also have an pathogenic effect on the protein.

CRP at diagnosis

Within our candidate SNPs from genetic analysis (a), we were additionally able to establish new relationships between genetic associations and longitudinal lab metrics. The first of these that we will discuss is the finding that CRP measurements



(a) *NOD2*, 3 common (b) *NOD2*, with rare (c) *ATG16L1*

Figure 6.3: *Association between (a), (b)* NOD2 *and severity and (c)* ATG16L1 *and diagnosis. As shown in (b), inclusion of rare* NOD2 *variants increases the association between* NOD2 *and severe outcomes. Cochran-Armitage tests.*

surrounding the period of diagnosis were closely related to *NOD2* status. In results from our univariate analysis, patients with any of the three common *NOD2* variants had higher CRP levels around diagnosis, as shown in Table 6.3, where we present p-values for associations between NOD2 variants and various CRP-derived features: value at the time point closest to diagnosis, the median of all measurements over the first three months after diagnosis, and the minimum of all measurements over the first three months. In a similar manner to the association with CS, the significance of this CRP association became stronger when the rare *NOD2* variants were included (e.g., for the association between CRP diagnosis value and *NOD2*, $p = 0.032$ with the common SNPs only, while $p = 0.013$ with rare variants included). When we controlled for disease localisation, the significance of these assocations became stronger (as shown in Table 6.3). Notably, the association between *NOD2* and CRP levels was not significant when we examined the overall patient CRP median and CRP AUAS. This could be because patients with *NOD2* variants tend to have more severe initial presentations, but that their resulting disease progression is not as strongly related to *NOD2*. Perhaps more likely is that these longer-term severity effects are difficult to disentangle from treatment effects.

Interestingly, the clearest distinction between patients carrying *NOD2* variants and other patients was in their minimum CRP value within the first three months of diagnosis: in localisation-adjusted LR analysis, the coefficient for minimum CRP was 0.59, versus 0.42 for the diagnosis CRP value and 0.44 for the median CRP value. This suggests that patients with *NOD2* variants were not necessarily severe in terms of extreme levels of inflammation, but rather that their inflammation was persistent and prolonged (i.e., they had no low CRP measurements over the first few months of disease), as compared to other patients. This association between minimum CRP and having any *NOD2* variant disappeared at six months (though it was maintained

Table 6.3: *p-values obtained from associations between CRP features and* NOD2*, Kruskal-Wallis test (univariate) and Wald test (multivariate)*

	Univariate		Disease localisation adjusted	
CRP features:	Common	With rare	Common	With rare
Diagnosis value	0.032	0.013	0.011	0.012
Median over 3 months	ns	0.089	0.023	0.026
Minimum over 3 months	0.024	0.011	0.012	0.0099
Med/min; AUAS; 3-month max.	ns	ns	ns	ns

ns: nonsignificant, var: variants

for the rs2066844 *NOD2* variant alone, $p = 0.0098$).

Though having one or more *NOD2* variants was related to CRP levels (an effect that remained significant even after controlling for disease localisation), the effect of the defective *NOD2* protein appeared to be much stronger in patients with ileal disease. Figure 6.4 shows in more detail the relationship between disease localisation, *NOD2* status, and median CRP level from the first three months following diagnosis. As is apparent in (a), median CRP levels after diagnosis generally were highest for CD patients with ileal disease, with a median of 19.0 mg/L. Median CRP levels after diagnosis decreased steadily by disease localisation, with a median of 14.0 mg/L for CD patients with colonic disease only and 8.0 mg/L for UC patients. As shown in Figure 6.4 (b), CD patients with a *NOD2* variant likewise had higher CRP levels than those without, with a median of 17.5 mg/L and 13.00 mg/L, respectively ($p = 0.053$). The opposite (nonsignificant) trend was apparent for patients with UC, where patients with *NOD2* variants had a median CRP level of 3.5 mg/L and those without had a median of 8.0 mg/L.

Figure 6.4 (c) shows the combined effect of *NOD2* and ileal disease. Though *NOD2* variants were associated with ileal disease, CD patients with ileal disease and a *NOD2* variant had higher CRP levels than those without a *NOD2* variant. The median CRP of the former group in the first three months after diagnosis was 27.5 mg/L, while the latter group had a median of 12.8 mg/L ($p = 0.04$). The effect of *NOD2* variants was much smaller in the set of CD patients with only colonic disease, where the median CRP of patients with and without *NOD2* variants was 15.0 mg/L and 13.8 mg/L, respectively (no significant difference). Again, the opposite (nonsignificant) trend was observed for patients with colonic UC, where colonic UC patients with a *NOD2* variant had a median of 4.0 mg/L and those without had a median of 8.0 mg/L.

Neutrophils throughout disease

We also found an interesting inverse association between CD patients with the *NOD2* p.L1007fsX1008 variant (rs2066847; frameshift; L1007fsinsC) and extreme neutrophil levels. Figure 6.5 shows the distribution of patient neutrophil AUAS scores, maximum neutrophils over the course of their disease, and the number of above-threshold neutrophil measurements recorded (normalised by follow-up time). As is evident, patients with the *NOD2* p.L1007fsX1008 variant had lower levels of extreme neutrophils in univariate analysis. For instance, the median neutrophil AUAS for patients without any *NOD2* p.L1007fsX1008 variant was 2.6, while it was 2.2 for those with one copy of the variant and 1.7 for those with two copies. Likewise, the maximum recorded

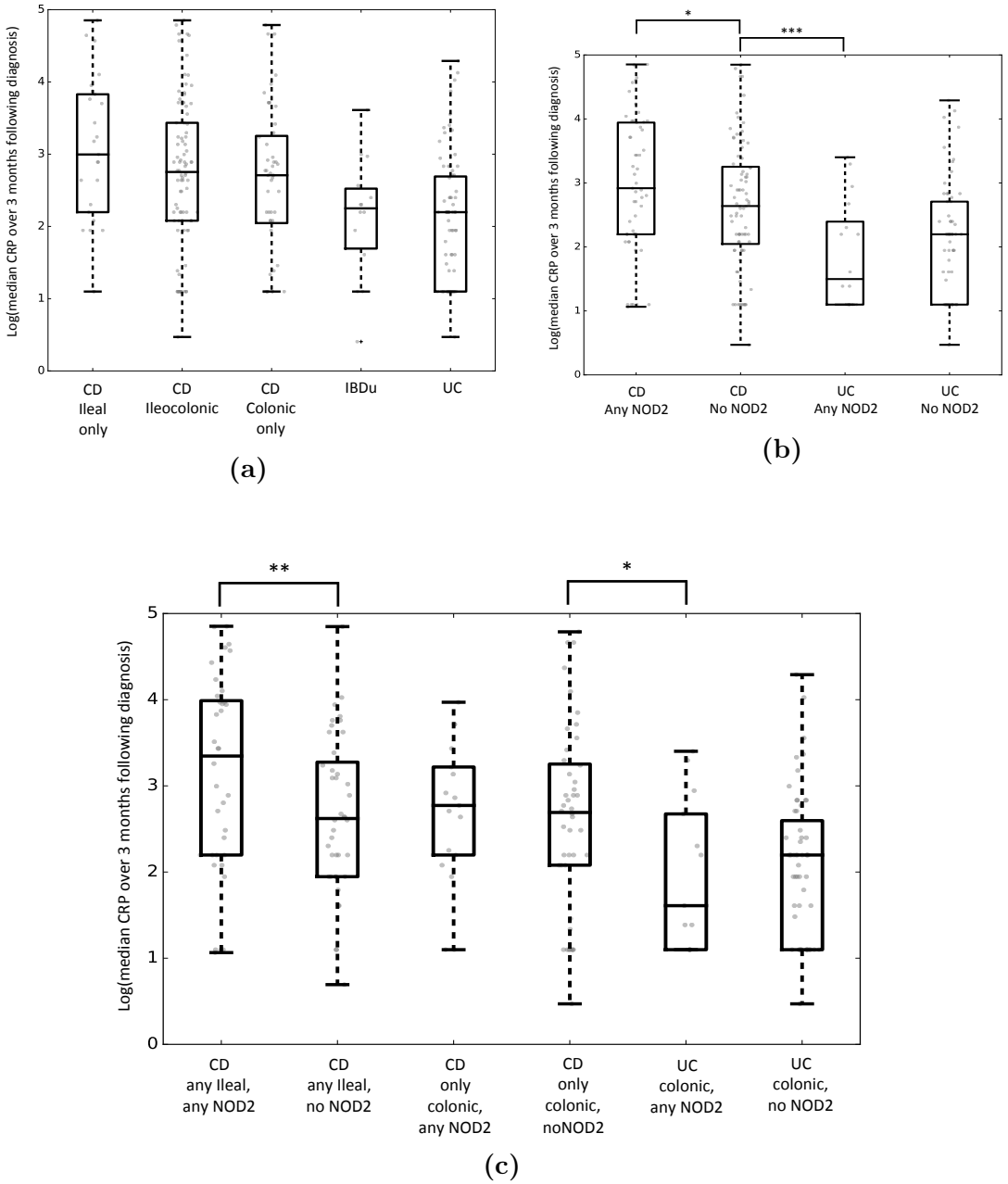
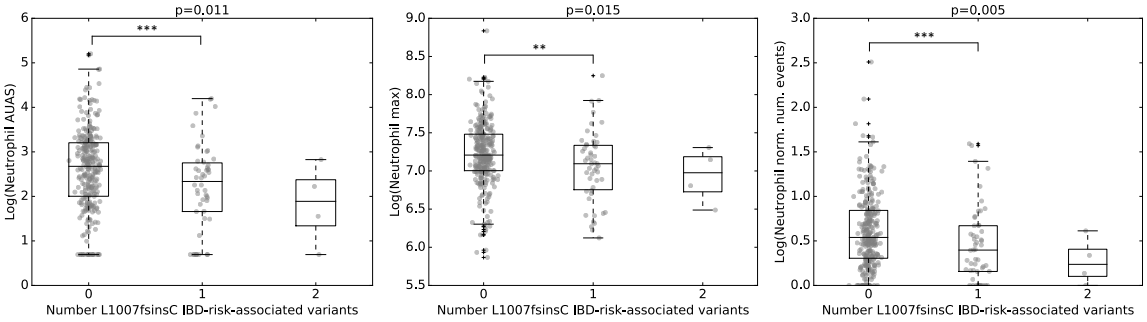


Figure 6.4: *Relationship between* NOD2*, disease localisation, and CRP levels. (a) Disease localisation only; (b)* NOD2 *status only; (c) Both disease localisation and* NOD2 *status. Significance indicators only shown between successive groups; one, two, and three stars indicate* $p < 0.1$*,* $p < 0.05$*,* $p < 0.01$*, respectively (MW test).*

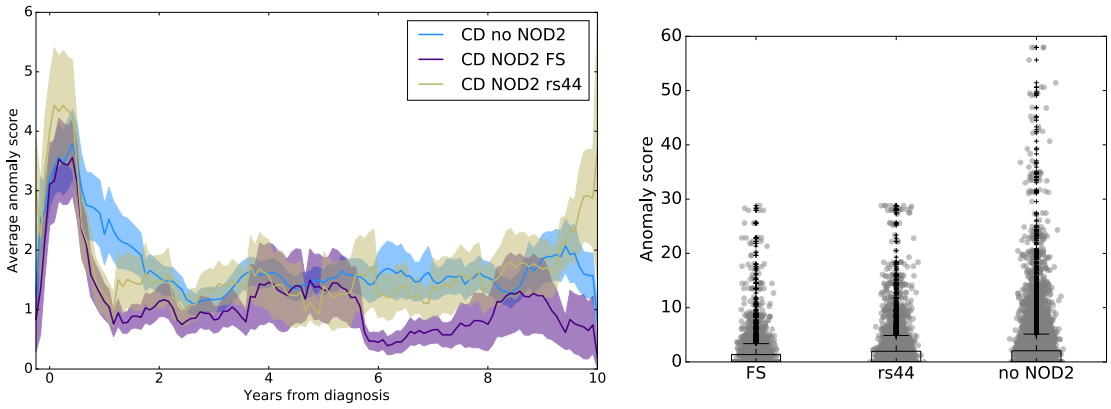


(a) AUAS (b) Maximum (c) Number of events

Figure 6.5: *Inverse association between the* NOD2 *p.L1007fsX1008 variant and elevated neutrophil levels. Significance indicators between successive groups: one, two, and three stars indicate* $p < 0.1$*,* $p < 0.05$*,* $p < 0.01$*, respectively (MW test).*

measurement was a count of 1349×10^7, 1204×10^7, and 1085×10^7 cells, and the number of extreme events per year was 0.71, 0.49, and 0.27, for each of these groups, respectively. When we examined the associations after adjusting for disease localisation, our findings remained consistent, with a negative association between having any *NOD2* p.L1007fsX1008 variants and neutrophil AUAS (p=0.0054), as well as the number of above-threshold neutrophil events (p=0.038).

Because this finding was counter-intuitive - we would expect patients with the *NOD2* p.L1007fsX1008 variant, which is severity-inducing, to have more severe lab measurements - we investigated this relationship further by examining the anomaly scores trajectories for these patients. Figure 6.6 shows the average anomaly scores for CD patients with the *NOD2* p.L1007fsX1008 variant ($N = 56$), the p.R702W (rs2066844) *NOD2* variant ($N = 68$), and no *NOD2* variants (including rare variants; $N = 208$). We did not include patients with the p.G908R (rs2066845) *NOD2* variant because the number of such patients was too small to be informative ($N = 22$). We performed a permutation test to assess the difference in the area under the curve for these average anomaly scores; we call this metric the "area under the average anomaly score" (AUAAS). We found that the AUAAS was smaller for patients with the *NOD2* p.L1007fsX1008 variant than for patients without any *NOD2* variant ($p = 0.03$). It therefore seems as though, on average, patients with the *NOD2* p.L1007fsX1008 variant tended to have lower neutrophil levels during relapse events through the course of their disease. We caution that Figure 6.6 (a) should not be interpreted as the actual patient trajectories, as each patient generally only contributes to a subset of the examined time period (patients have different lengths of follow-up and may not have lab data available until several years after diagnosis). Rather, it provides a



(a) (b)

Figure 6.6: *Illustration of patient anomaly scores, by* NOD2 *status. (a) Plot of average neutrophil anomaly scores, by* NOD2 *group, over time for CD patients. (b) Boxplots of the distribution of all anomaly scores, colored by* NOD2 *group. Abbreviations: FS: frameshift (p.L1007fsX1008* NOD2 *variant); rs44: rs2066844 (p.R702W* NOD2 *variant); no NOD2: no NOD2 variants.*

smoothed representation of the relative degree of how extreme patients' neutrophils are at various time periods after diagnosis.

We also examined the distribution of all of the anomaly scores for the time periods when patients had lab data available, as shown in Figure 6.6 (b). For clarity, we emphasise that Figure 6.6 (b) depicts the distribution of all anomaly scores for the patients within each group; each patient contributes multiple values to the distribution. Again, the median of anomaly scores for patients with the *NOD2* p.L1007fsX1008 variant was lower than the median anomaly scores for either patients with the *NOD2* p.R702W variant ($p < 0.0001$, MW test) or patients without any *NOD2* variants ($p < 0.0001$, MW test). We also found that there were no significant differences in neutrophil levels at the time of diagnosis between *NOD2* p.L1007fsX1008 variant patients and other patients. These results together suggest that the differences in AUAS that we have observed are due to differences in disease trajectories over time, rather than the initial presentation.

6.4.2 Tiers two and three

Within our genetic analyses (b) and (c) (i.e., extending the association study to the remainder of the IBD risk-SNPs), we again identified several associations with promising potential for future follow-up. Two of these are shown in Figure 6.7. IBD patients carrying one or more copies of the rs4656958 risk allele seemed to have less

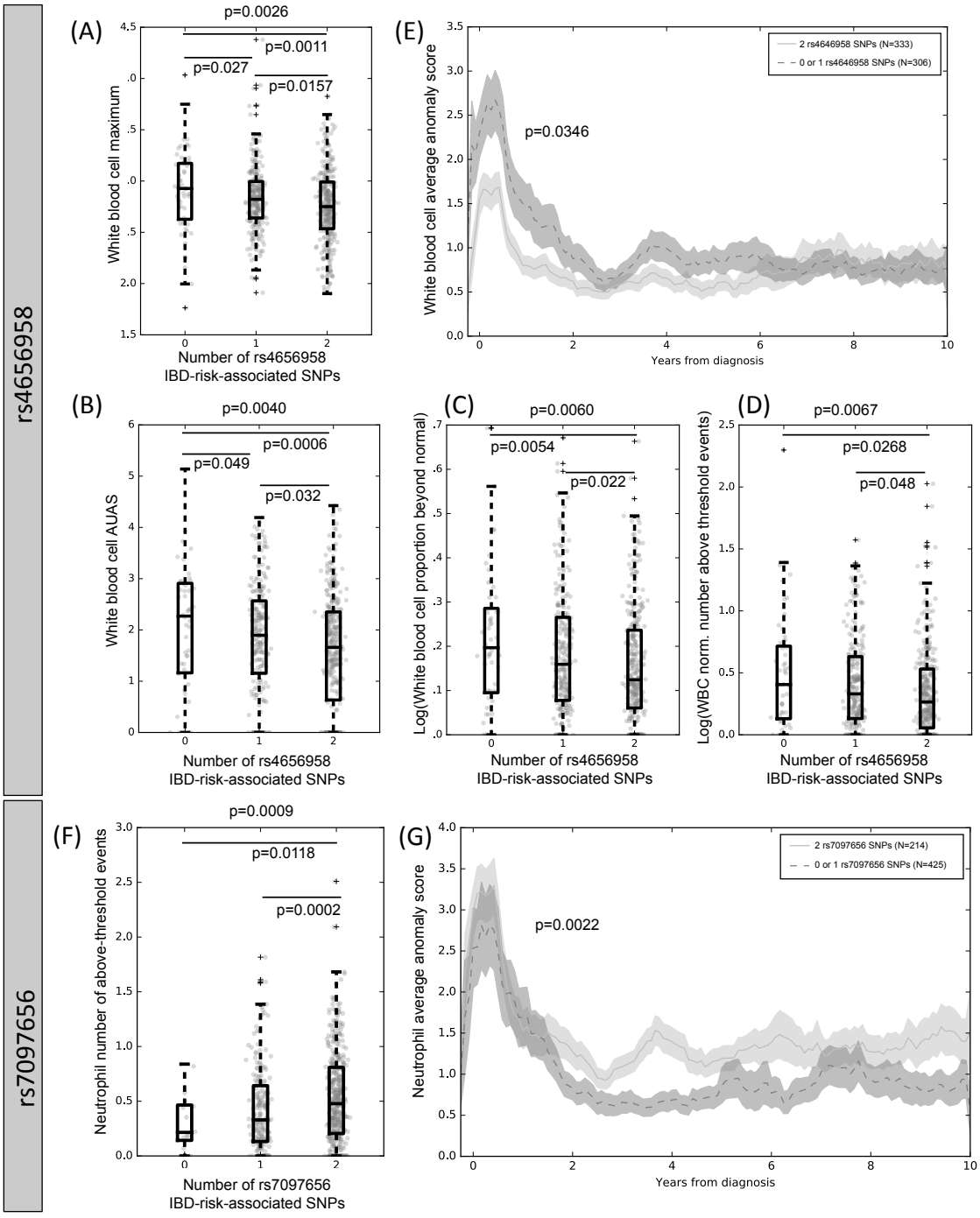


Figure 6.7: *Selection of nominally significant Tier 2 and Tier 3 associations. (A)-(D) Supporting figures for associations identified for rs4656958. (F), (G) Supporting figures for associations identified for rs70977656. P-values on top of box plots: Kruskal Wallis test. Remaining p-values: permutation tests.*

extreme white blood cell levels, as shown by the univariate association between having more copies of the SNP and, as shown in Figure 6.7(A), a lower WBC maximum, Figure 6.7(B) AUAS, Figure 6.7(C) proportion of abnormal measurements, and Figure 6.7(D) extreme events. For instance, the median WBC AUAS was 2.27, 1.90, and 1.66 for patients with 0, 1, and 2 copies of the rs4656958 risk allele, respectively. For the same groups, the median of the maximum observed WBC was 17.7×10^9/L, 15.8×10^9/L, and 14.7×10^9/L, respectively, and the median percentage of abnormal measurements was 21.8%, 17.3%, and 13.2%, respectively.

When adjusted for disease localisation, the association with WBC levels remained for CD patients, with an association between having any rs4656958 risk alleles and the number of extreme events ($p = 0.0038$) and the proportion of abnormal measurements ($p = 0.009$). When we examined the average anomaly scores for patients with this SNP, as shown in Figure 6.7(E), it appeared as though this difference was largely present around the time of diagnosis. After this period, WBCs tended to be around the same level as patients without any copies of the risk allele.

The SNP rs4656958 is not located within a gene-coding region, but it is physically near several genes, including the gene *CD244*, which encodes a receptor that is expressed on the cell surface of natural killer cells (a type of lymphocyte) as well as other WBCs [185]. This receptor is involved in modulating cell killing. Given the potential for linkage disequilibrium, the effect we observe could be actually due to a variant in the *CD244* gene. The rs4656958 loci is also an eQTL for *USF1*, a gene that encodes upstream stimulating factor 1 (USF1). USF1 is a transcription factor in reticular epithelial cells; these cells screen T cells to make sure that they won't attack the body's own cells [186].

For SNP rs7097656, as shown in Figure 6.7(F), we found that patients with increasing copies of the risk allele had higher neutrophil levels, with a median 0.24, 0.39, and 0.61 above-threshold events per year for patients with 0, 1, and 2 copies of the risk allele, respectively. When we examined the median anomaly scores over time, as shown in Figure 6.7 (G), we found that patients with two copies of this SNP initially had similar levels of neutrophils as other patients, but later in their disease course these patients tended to have increased levels of extreme neutrophils. We hypothesise that perhaps these patients do not maintain their response to medication, resulting in frequent and severe relapse events. The rs7097656 SNP is intronic and located within the *TSPAN14* gene, which encodes the Tetraspanin-14 (Tspan-14) protein [187, 188]. Tspan-14 is involved in enzyme binding [186], but it is unclear how this variant (which causes a missense mutation) may be related to immune regulation.

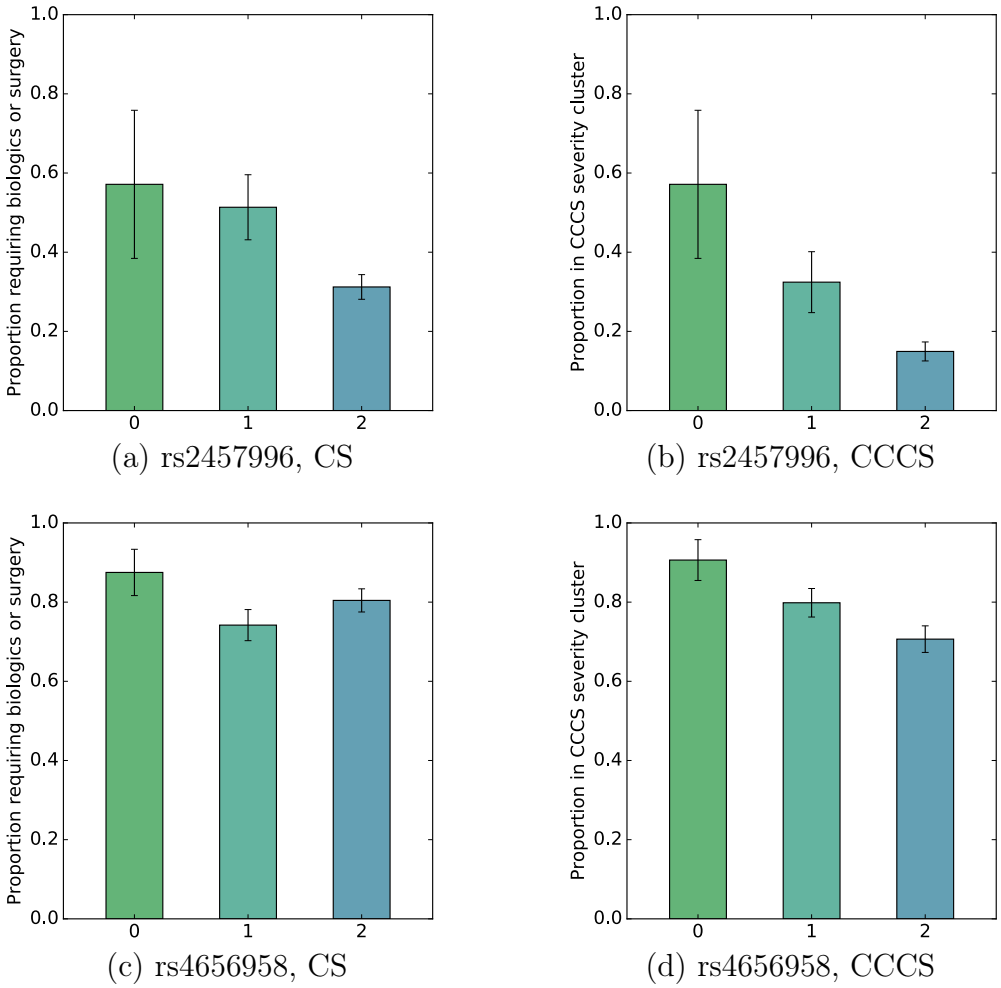
6.4.3 Associations with severity clusters

We were interested in how the results obtained from our clustering algorithm showed an association with underlying genetics, as we developed these metrics with the goal of capturing complex patient phenotypes in a manner that could be related to genetics. We found that in some cases, our CCCS results offered clarity in trends that were beginning to be apparent in the CS definitions, as shown in Figure 6.8 (a) and (b) for SNP rs2457996. Here, there was a trend towards decreasing CS for UC patients with additional copies of the risk-associated allele ($p = 0086$). As shown in Figure 6.8 (a), 57.1%, 51.4%, and 31.2% of patients with 0, 1, and 2 rs2457996 risk alleles, respectively, required surgery or biologics. This trend became even stronger when examined in terms of CCCS, with 57.1%, 32.4%, and 14.9% of patients classified as "severe" according to our clustering algorithm, respectively ($p = 0002$).

In other cases, the CCCS groups showed trends that were not apparent in the traditional metrics, as shown in Figure 6.8 (c) and (d) for SNP rs4656958. Here, there was an inconsistent relationship between additional copies of the risk allele and CS, with 87.5%, 74.2%, and 80.4% of CD patients with 0, 1, and 2 rs4656958 risk-SNPs requiring surgery or biologics, respectively ($p = 0.97$). However, a negative association was evident in terms of the CCCS groups, with 90.6%, 79.8 %, and 70.7% of these same patients falling into the "severe" cluster ($p = 0.0059$). In such situations, the differing results suggest that the SNP may be more strongly associated with the intermediate inflammatory response, rather than outcomes such as progression to surgery or biologic therapy.

Notably, for *NOD2* variants, the prevalence of having one or more variants was not different between our "severe" and "nonsevere" CCCS groups, as shown in Figure 6.9. This was likely because of the interesting patterns we have observed between *NOD2* and systemic inflammation, where some patients with the p.L1007fsX1008 *NOD2* variant actually had less severe long-term inflammatory markers. The effect is also further diluted by the fact that we collapsed our multiple clusters down into two groups (i.e., severe and nonsevere), thereby losing the resolution present in the original clustering results (which had five clusters, as described in Section 5.6.2).

We did find one SNP that was associated with the previously-identified gene expression clusters. Of the patients in the low monocyte/high T-cell cluster, 16 (89%) had at least one risk-associated allele at the loci rs1260326, while 8 patients (40%) in the high monocyte/low T-cell cluster had at least one of these risk alleles ($p = 0.0018$, Cochran-Armitage test). This SNP causes a missense mutation in glucokinase (hexok-



(a) rs2457996, CS (b) rs2457996, CCCS

(c) rs4656958, CS (d) rs4656958, CCCS

Figure 6.8: *For two example SNPs: (a), (c) proportion of patients who are severe according to CS; (b), (d) proportion of patients who are severe according to CCCS. Cochran-Armitage test.*

inase 4) regulator (GKCR), and most previous studies have associated its effect with triglyceride metabolism [187]. However, it could potentially be involved in immunse responses in some other way, or it could be in linkage disequilibrium with another loci that is.

6.5 Survival analysis

In addition to assessing the relationship between genomic loci and outcomes such as surgery or biologics, we used a survival analysis to examine differences in time-to-event for patients with and without specific variants and disease localisations. We defined survival as time free of biologic prescription or GI surgery. However, our analysis was limited here because the time stamps of surgeries and biologic prescription were

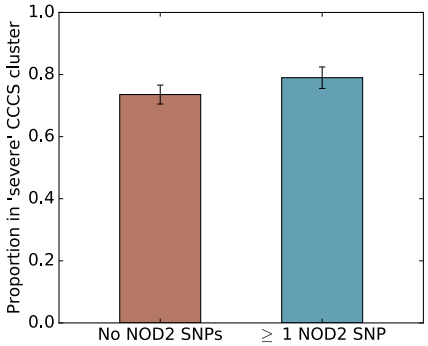


Figure 6.9: *Proportion of CD patients in the severe CCCS group, by* NOD2 *status. Compare to Figure 6.3 for the relationship with CS. Cochran-Armitage test.*

only available from 2007. We therefore only included patients diagnosed after 2007 in this analysis so as to have a true representation of time-to-event (otherwise, a patient diagnosed in 1987 may have had multiple surgeries early in their disease course, but a surgery recorded in 2010 would mistakenly suggest that the patient survived 23 years before surgery). Some patients had a record of surgery or biologics (as extracted manually from the record), but no recorded date associated with this. Therefore, to account for surgeries and biologics that may have occurred outside of the Oxford University Hospitals (the source of our data), we did not include these patients in our survival analysis.

Limiting our sample to only those patients diagnosed after 2007 substantially decreased our sample size (from 346 CD patients to 81, and from 266 UC patients to 80), and so we also performed the same analysis limited by patients diagnosed after 2000 (resulting in 210 CD and 151 UC patients). The composition of this patient sample will be biased by the lack of timing information for patients prior to 2007, meaning that survival periods will appear longer than they actually were. Nevertheless, loosening our date of diagnosis restriction in this way provided some insight into the patterns expected with a larger patient sample.

We used Kaplan-Meier curves and log-rank tests to assess differences in survival [189]. This method estimates a nonparametric survival function through maximum likelihood, the result of which is $\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$, where n_i is the number of patients surviving at time t_i (i.e., no severe outcomes and also not yet censored) and d_i is the number who reach a severe outcome at time t_i. Significant differences between two survival curves can be assessed through the log-rank test. A log-rank test (also termed the Mantel-Cox test) employs the hypergeometric distribution to determine the expected number of events that occur at each time point, each of which

is compared with the observed number. The resulting statistic is approximately Gaussian-distributed, then allowing the computation of a p-value.

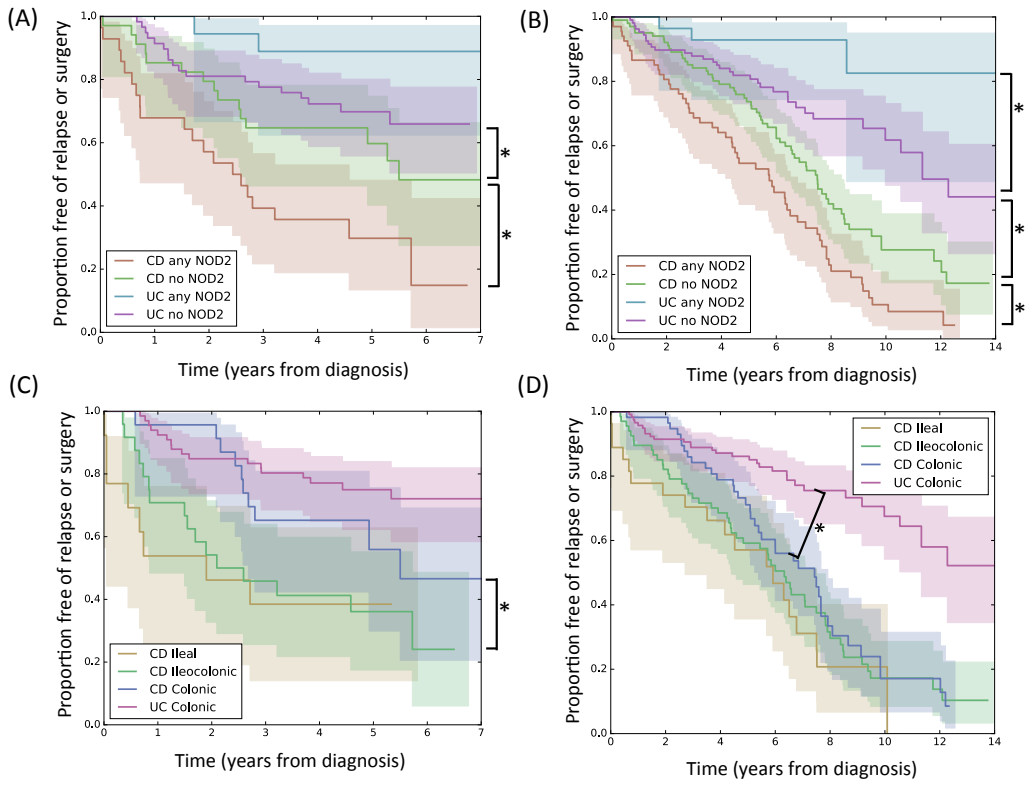


Figure 6.10: *Survival curves, with the included patients limited by date of diagnosis (DOD). (A)* NOD2 *status for patients with DOD after 2007; (B)* NOD2 *status for all patients with DOD after 2000; (C) disease localisation status for patients with DOD after 2007; (D) disease localisation status for patients with DOD after 2000. Stars indicate significant differences between successive patient groups* ($\alpha = 0.05$, *log-rank test).*

In our survival analysis, we found clear differences in time to severe outcomes based upon *NOD2* status, as shown in Figure 6.10. We see from (A) that survival for CD patients with *NOD2* variants was shorter than other patient groups, with 30% surviving free of surgery or biologics after five years, compared to 60% for CD patients without deleterious *NOD2* variants ($p = 0.012$). On the other hand, survival for UC patients with *NOD2* variants was longer than for the remaining UC patients, with 89% survival after five years compared to 70%, respectively ($p = 0.09$). When we increased our sample size by including patients diagnosed between 2000-2007 in Figure 6.10 (B) (rather than limiting those diagnosed after 2007; which, as mentioned, biases us towards longer survival times), this difference between UC patients with and without *NOD2* variants became more significant (p=0.029, log-rank test).

When we looked at the effects of disease localisation in Figure 6.10 (C), we found

that CD patients with only colonic disease localisation survived much longer without surgery or biologics (median survival of 5.5 years) compared to CD patients with ileal disease localisation (median survival of 1.9 years and 2.1 years for ileal and ileocolonic localisation, respectively, $p = 0.05$ for both). We did not see any significant difference in survival time between CD patients with colonic disease and UC patients. This difference became larger (and significant) when we expanded the sample size to include patients diagnosed between 2000-2007, as shown in Figure 6.10 (D). We also found that there was no longer a significant difference between CD patients based upon disease localisation with this larger cohort, with median survivals for all CD groups ranging from 5.9 years to 7.4 years. Nevertheless, the lack of a difference between colonic CD and UC patients in our post-2007 patient subset (as shown in (C)) suggests that in the early years of disease, there is not a large difference in severe outcomes for these two patient groups.

6.6 Discussion

We have undertaken a large association study between genomic loci, intermediate inflammatory markers, and disease outcomes. The purpose of such analysis was to investigate whether there are variations in genetics that are linked to disease phenotypes. Previous studies have only been able to examine outcome measures such as requiring surgery/biologic therapy or the conventional Montreal classification scheme [10, 145]. Here, we have used our methods to summarise laboratory measurement patterns over time, allowing us insight into how genetics may effect intermediate markers of inflammation.

We have replicated previous findings relating *NOD2* variants to severity, with the intention of demonstrating the validity of our approach [46, 58, 60]. In addition, we have demonstrated that patients with *NOD2* variants have higher CRP levels at diagnosis. While this relationship has already been described in children with IBD [46], we have confirmed this effect for adults and further described that the relationship between *NOD2* and CRP is modulated by the location of inflammation, with the largest effects occurring in patients with ileal disease. Indeed, in UC patients, where inflammation is typically confined to the colon, we see that *NOD2* variants have a protective effect on CRP levels of inflammation at diagnosis. This trend holds even when we look at outcomes such as biologic therapy and surgery: UC patients with *NOD2* variants tend to have longer survival times until such severe events.

Recent unpublished work suggests that the relationship between *NOD2* variants and CRP in healthy volunteers is the inverse of that observed in CD disease states

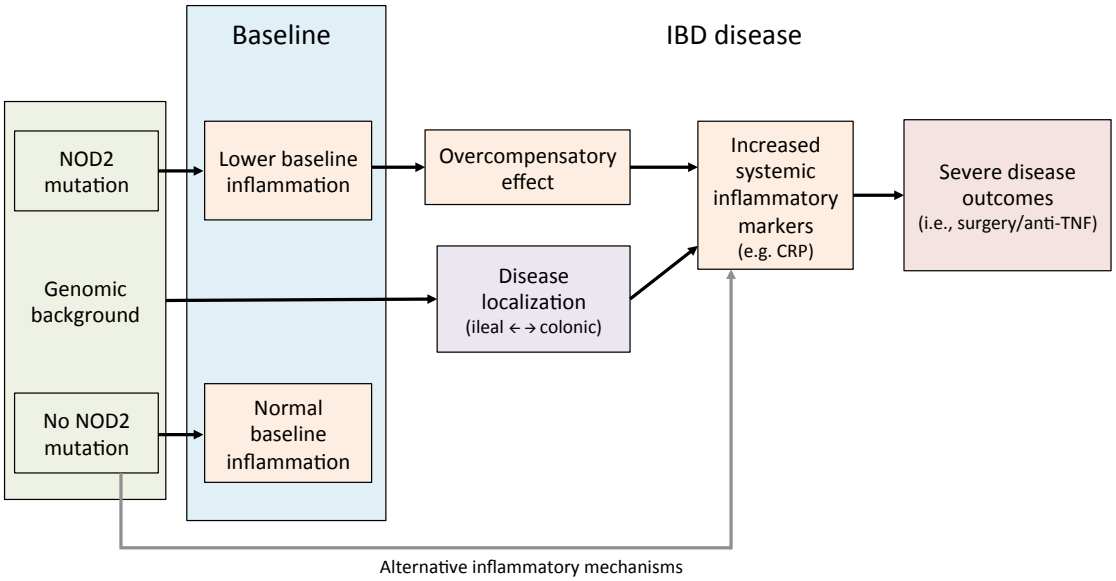


Figure 6.11: *Potential schema illustrating mechanism of CRP and* NOD2 *relationship*

(personal communication, Dr. Luke Jostins, Wellcome Trust Centre for Human Genetics). In other words, healthy subjects with a *NOD2* variant have very slightly lower baseline CRP levels than subjects without a variant in this gene. This is similar to the effect that we have observed for our UC patient data.

This information, together with our analysis, suggests that the slightly lower baseline level of inflammation in patients with *NOD2* may contribute towards disease development, as outlined in Figure 6.11. Lower initial inflammatory responses to bacteria in the gut may result in a lack of bacterial clearing. This then requires a more substantial compensatory inflammatory response, which in some patients becomes disregulated and results in IBD symptoms. For patients in whom this disregulation is localised in the ileum (which may be due to both other genetic and environmental factors), the *NOD2* variant is particularly damaging and results in high systemic inflammation (i.e., high CRP). Physicians are trained to treat patients based upon observing high levels of CRP, due to the nature of CRP as an inflammatory marker. As a result, this may be the reason that we see the relationships we have observed whereby CD patients with *NOD2* variants progress more quickly to severe outcomes. Meanwhile, the CRP response is still lower for patients with colonic-isolated disease, resulting in a slower progression to severe outcomes. This indicates that the role of *NOD2* may be quite distinct in the ileum as compared to the colon; its behaviour apparently depends upon the local tissue environment.

Because the time period of diagnosis is the "cleanest" portion of the data - i.e., patients have not yet had any medications - it provides the clearest mechanistic insight and a representation (albeit from a single point in time) of disease severity if left untreated. Lab measurements from later periods are affected by patient medication as well as the disease course, and must therefore be interpreted with caution: other, non-modelled factors may be explain the effects we have observed. For instance, we found that CD patients with the *NOD2* p.L1007fsX1008 variant have less extreme neutrophil levels over the course of their disease. This is a counter-intuitive finding, given that (as discussed) CD patients with *NOD2* variants typically have more severe disease.

Given that there are no differences in neutrophil levels at the time of diagnosis between *NOD2* p.L1007fsX1008 variant-carrying patients and other patients, this relationship appears to be something that arises after treatment. It is unlikely that the lower levels of extreme measurements are simply because of higher rates of immunosuppressant prescription, as CD patients with the frameshift variant have similar rates of immunosuppressant prescription to other CD patients (21.7%, compared to 26.2% across other CD patients). It could be, however, that patients who present with more severe disease, as is the case for patients with *NOD2* variants, may be treated and followed-up more aggressively, resulting, in practice, in a more mild disease course in terms of lab-based measurements. Alternatively, it could be that these patients have lower levels of neutrophils in the peripheral blood (where our measurements were obtained) because they have all been recruited to the gut. Another potential explanation is that the neutrophil response is simply impaired in these patients. Finally, of course, given the number of associations that we examined, this finding could be a false positive; it therefore needs to be replicated in an independent cohort of patients.

Another notable finding from our study is that the inclusion of rare variants, as is possible through exome sequencing, can improve our ability to resolve genomic relationships with disease outcomes. For instance, inclusion of *NOD2* rare variants previously identified as being potentially deleterious increased the association of *NOD2* with severe outcomes. While we only included the *NOD2* exome sequencing data for our analysis, it would be beneficial in future work to include the exome information for other key genes (e.g. *IL23R*, *ATG16L1*) known to be involved in inflammatory pathways.

Throughout this analysis, we have used standard metrics to describe patient lab trajectories (e.g., overall median; maximum measurement within the first three months of diagnosis), along with our anomaly score (AUAS) and severity-based clus-

tering metrics (CCCS). We have found that the AUAS proved to be a useful variable for our association here, as it combined information across multiple standard metrics, which often increased the power of the association. For example, once adjusted for localisation, the neutrophil AUAS value was strongly associated with the *NOD2* frameshift variant, while other neutrophil metrics (e.g., maximum neutrophils) were no longer associated or associated only weakly (e.g., number of above-threshold events).

Our anomaly scores were also useful for visualising the trajectories of patients over time, as in Figure 6.6. These visualisations represent average scores over time, and so they provide a very high-level vantage point of "typical" inflammation for the patient group, as patients relapse at different times. Though the use of AUAAS values should be viewed in this light, they can be informative in understanding where the difference in patient severity has originated, as in the difference between an early (e.g., rs4656958) and late (e.g., rs7097656) effect shown in Figure 6.7.

Our patient clusters were less informative than our anomaly scores, with very few associations appearing across the dataset. This is likely due to the fact that we collapsed our clusters into two groups (i.e., severe and nonsevere) to increase power. As a result, we lost some of the homogeneity in the patient clusters, making it less likely to observe genetic associations. In addition, as we have discovered through our analysis here, long-term lab trends may be more reflective of medication and treatment patterns than patient genetics; future studies where medication time stamps are available would help to resolve the contributing effects. Nevertheless, our more-inclusive definitions of severity can prove useful when examined in combination with classical severity definitions: when both are significantly associated with a genetic variant, it suggests that the driving differentiating factor is the outcome of surgery or biologics. When they differ, it suggests that the intermediate inflammatory markers are potentially playing a larger role. In either case, the severity clustering still provides a useful way to group together patients with severe disease in terms of either progression to surgery or biologics, or severe lab measurements.

6.6.1 Assumptions and challenges

As always in such retrospective analysis of patient records, our results are subject to bias and to being confounded by unobserved effects. For instance, while in our survival analysis we attempted to limit our patient population to only those patients with observed surgeries and biologics, there is always the possibility that patients received either of these treatments outside of the Oxofrd University Hospitals. Similarly, our

lab measurements were only reflective of the data collected within the Oxfordshire area. Patients who may be students (a significant portion of the Oxford population) may have a large majority of their measurements recorded elsewhere, thereby giving a false representation of either severe or nonsevere disease. Though this missingness of data for patients who received healthcare elsewhere is certainly present, we feel that for patients with long period of follow-up, any missingness is generally reflective of periods of patient health. It is when people are sick that they come to the clinic, and it is then that we begin observing measurements.

As mentioned already, our analysis of disease trajectories, as captured through the AUAS, is certainly limited by the fact that treatment is being provided. Unfortunately, as we do not know when different treatments are prescribed (except for a subset of anti-TNF therapies, as is the subject of Chapter 7), we cannot include this information in our analysis. However, in the future, if available, such information could be incorporated, for instance through an approach using hidden Markov models, where patients progress through disease states partially defined through medications.

The most important next step for this work is an increase in patient sample sizes. An increased patient sample would increase our power to detect real associations and would also allow us to correct for multiple hypothesis testing. Larger patient samples would also allow us to examine associations with each of our identified severity clusters independently, thus increasing the likelihood of finding true associations with severity phenotypes. Indeed, given the large number of associations examined, some of our identified associations are almost certainly false positives.

Another interesting avenue of future research is the multivariate combination of many SNPs to predict either disease outcomes or inflammatory markers. While we pursued such analysis (i.e., using many SNPs to predict severity status), we did not meet with success. This is not surprising, given that previous work has also found that most of the predictive effect of genetics on patient severity is driven by *NOD2* in European populations and that the overall predictive power is very small [10]. However, a potentially more promising route is to identify SNPs that likely work together in inflammatory pathways and examine their ability to predict inflammatory markers. Such analysis is challenging because the current understanding of IBD inflammatory pathways is still incomplete, and the actual function of many of the IBD risk-associated SNPs is not understood, even after fine-mapping efforts [12]. However, recent work has directly identified many potential new eQTLs amongst the IBD risk-SNPs [190]; this information could be used to construct more accurate pathways than has been possible in the past.

While our analysis here was subject to a limited patient population, this framework can be applied to larger cohorts when available. We acknowledge that our current findings are hypothesis-generating only and need to be further investigated. Indeed, validation of our finding that CRP levels at diagnosis are associated with *NOD2* status and disease localisation is currently underway.

Chapter 7

Biologic drug comparison

7.1 Introduction

One of the key molecules involved in IBD inflammatory pathways is TNF-α. High levels of this cytokine can result in the over-inflammation that is characteristic of IBD (see Figure 3.4). Identification of this molecule has allowed for the development of biologic anti-TNF-α therapies, which have been found to be very effective for IBD patients with severe disease. Anti-TNF-α antibodies are believed to act through two major mechanisms. The first is by blocking the interaction between TNF-α and its receptors, thereby blocking the resulting inflammation cascade. Anti-TNF-α can also bind to transmembrane TNF, which can induce cell death (e.g., in activated T-cells and monocytes); this is desirable given the over-abundance of inflammatory cells [191, 192].

IFX, the first such anti-TNF-α biologic therapy, was first approved for use in 1998 in the USA (1999 in the European Union, EU), and since that time it has become a well-established treatment option for severe CD and UC [193]. In 2007, both the USA and EU approved another anti-TNF therapy for use in severe CD patients: ADA. IFX is a chimeric mouse-human monoclonal antibody and must be administered intravenously (Remicade, MSD, Whitehouse Station, NJ), whereas ADA is a fully human monoclonal antibody and is administered subcutaneously (Humira, Abbvie bv, North Chicago, IL) [194]. This means that ADA can be administered at home, avoiding the need for infusion in the hospital (as is required for IFX). Since this time, several additional anti-TNF medications have been developed, such as certolizumab pegol (not approved in the EU for IBD in 2008, though approved in the USA) and golimumab (approved for UC in the EU in 2009). There have also been several IFX biosimilars (i.e., with the same underlying molecule, after the original

patent expires) approved in the EU, including Flixabi (2016), Inflectra (2013), and Remsima (2013) [195]. At the time of writing, ADA and IFX are the only two anti-TNF-α antibodies used with high frequency in clinical practice in the UK.

Both ADA and IFX have been found to be effective individually in multiple clinical trials [196–199]. Typically, clinical trials evaluate efficacy in terms of the CDAI or HBI (see Section 1.8). A clinical response is usually defined as a reduction in CDAI or HBI, while clinical remission is defined by achieving a score below a specific threshold[1] [200]. In initial trials, outcomes are usually measured at four weeks after the drug is first received [196, 198], with later maintenance trials assessing efficacy at longer time points. Hanauer et al., for example, found that the median time to loss of response to IFX was more than 54 weeks (the trial was 54 weeks long) [197], and Colombel et al. found that between 36-41% (depending upon dosage) of ADA-treated patients maintained remission after 56 weeks.

Systematic reviews and meta-analyses have evaluated the effect of each of the drugs individually and as an aggregate biologics class [201], again finding both to be effective. However, there have not been many direct comparisons between the drugs. Van Assche et al. conducted a small randomised controlled trial (RCT) with a total of 73 patients, which compared patients continuing on IFX and with those switching from IFX to ADA [202]. They found that the patients who switched to ADA required dosage intensification and were more likely to become unresponsive to the drug. However, patients reported more favourable experiences with ADA (perhaps due to its subcutaneous administration, rather than intravenous). While RCTs are extremely valuable for evaluating treatment options, it is also important to evaluate the efficacy in real-world situations.

Kestens, et al. conducted a retrospective study comparing the two drugs in CD patients, finding no significant differences between the response rates of their 100 ADA and 100 matched IFX-treated patients after two years [194]. The primary outcome indicator was rate of steroid-free disease remission, which was defined as not requiring steroids, not being hospitalised for IBD-related surgery, and not discontinuing treatment due to side effects or loss of efficacy. The authors also examined rates of infection and side effects. Similarly, two studies in Italy found no differences between the two drugs for CD patients. The first of these, with 93 patients (44 IFX; 49 ADA) from secondary/tertiary care centers, also used steroid-free remission as their outcome measure [203]. The second study, with 44 patients (20 IFX; 24 ADA) from

[1]For CDAI, the reduction required for a clinical response is typically 70-100 points, or 25% reduction in score; remission is typically defined as a score of less than 150 [200].

primary care centers, used the HBI to assess effectiveness [204]. Both assessed patient response for up to 1.5 years following therapy initiation.

While anti-TNF-α therapies are effective for many patients, approximately 40% of patients do not respond to the drug [200]. Patient anti-TNF response can be subdivided into two categories: primary nonresponse and secondary nonresponse. In clinical trials, primary nonresponse is usually assessed after 3 to 3.5 months. In clinical practice, however, the drugs are often used for six months before declaring a patient nonresponsive. Clinical trials report a 40% primary nonresponse rate, though this has been found to be lower in clinical practice. This may be due to less stringent criteria for defining response in clinical practice. Response is usually defined, as discussed above, as a reduction in CDAI or HBI. Secondary loss of response occurs for patients who are responding initially, but who subsequently lose this response. It may be defined in terms of an increase in CDAI or the requirement for additional therapy. In an ADA meta-analysis, the risk of secondary nonresponse was found to be 20.3% per patient, per year; for IFX, this risk has been estimated at 13% per patient, per year [200]. In general, however, definitions of response and loss of response vary widely, making it difficult to compare studies and consolidate findings [205].

Several previous studies have investigated genomic factors associated with biologic therapy response, finding some evidence for SNPs within the *TNF-α* (rs1799724), TNF receptor 1B (rs3397, rs1061624), and Lymphotoxin Alpha (rs909253) genes to be associated with CD patient response [206]. However, identification of these loci has not been reproduced in follow-up studies [200], meaning that there is no current set of genomic predictors for anti-TNF-α response. Numerous other loci have been associated with anti-TNF-α response for other immune-related diseases such as rheumatoid arthritis and psoriasis [206]. In addition to genomic factors, more traditionally-measured metrics have also been evaluated for their ability to predict anti-TNF-α response. Generally, better responses are found for patients who are younger (despite young age of onset being associated with more severe disease), those who have had the disease for two years or less, those with higher CRP levels prior to drug initiation, and those who are non-smokers. In addition, there is some evidence that the disease behaviour (luminal in contrast to fistulising[2]) and not having prior surgery are also associated with a better response [200, 207].

For physicians seeking to determine which drug to prescribe to patients, it is

[2]As described in Chapter 1, luminal disease is confined to inflammation of the surface of the intestine, while fistulising disease involves the creation of channels between different parts of the intestine or between the intestine and other tissues.

important to understand the effects of biologic therapy and whether there are differences between IFX and ADA. Previous retrospective studies have very appropriately focused upon steroid-free remission as the outcome of interest and have typically examined follow-up times of one year. Here, we add to this literature by examining the effects of biologic therapy over longer time periods and describe the progression of patient disease using a variety of metrics. Largely, we use common biochemistry lab measurements to capture patient physiology. (For more details on these patient lab measurements, see Chapter 3.) We focused most closely upon CRP, which is the systemic inflammatory marker that has been shown to be a strong indicator of IBD disease activity, and which we have discussed at length in previous chapters. We examined a variety of endpoints to compare treatment efficacy between the two drugs under study, including time to relapse or surgery, relapse rates, and CRP decline after initial drug prescription. In addition, to assess patient-specific responses to anti-TNF-α therapies, we examined the severity of relapse events before and after medication initiation, as well as severity trajectories over the treatment period.

Because patients exhibit heterogeneous responses to biologic therapy, we also used these various outcome metrics to identify those with a strong response and those without. We performed an exploratory univariate genetic association study with the IBD risk loci, as well as a supervised analysis to determine whether we could predict drug response using both previously-identified features and pre-specified genomic loci.

Our study therefore serves the purpose of evaluating, in a real-world setting: (1) the effects of biologic therapy on patient physiology in a more holistic way than has been previously described, and (2) the difference in efficacy between ADA and IFX. In addition, we (3) illustrate how our methods may be used to understand genetic factors that underlie patient anti-TNF-α response. Our cohort is unusual in that ADA is prescribed more commonly in Oxfordshire than in most other hospitals, providing us with the opportunity to study patient response to biologics with a larger ADA-treated cohort than has been available for previous studies.

7.2 Patient cohort

From our IBD cohort described in Chapter 3, we identified all CD patients who had dates of ADA or IFX prescription available in the EHR. We did not limit the sample by age of diagnosis or current patient age. We attempted to include only those patients without a history of biologic treatments (i.e., if a patient was treated with ADA and then switched to IFX, we only included the treatment with ADA in our study). However, many patients had records of both ADA and IFX prescriptions,

while pharmacy prescription dates were only available for one of the drugs. We included the available dates for these patients, despite not being able to ascertain whether this was their first biologic prescription.

We first assessed whether there were differences between the two groups of patients who received IFX and ADA, as this could influence their later response to the drugs. We therefore examined whether there were significant differences between the two groups in terms of disease localisation, medication prescription, age at diagnosis, gender, current age, etc. For the subset of patients with Immunochip data available, we also assessed whether there was a difference in the prevalence of the three common NOD2 mutations. For the binary features of interest (such as gender), we examined whether there were differences between IFX and ADA patient groups using a chi-squared test. For the continuous features of interest (such as age at diagnosis), we used a MW test (as noted previously, this is a nonparametric test of whether two samples come from a different population).

Of the 233 ADA-treated and 143 IFX-treated patients in our IBD cohort with dates of biologic therapy, 215 ADA and 78 IFX patients were diagnosed with CD. Of these, 193 ADA and 70 IFX patients had a valid date of diagnosis. Patient characteristics from our two cohorts are shown in Table 7.1. We did not find any significant differences between the two treatment groups in terms of any of the examined covariates. The largest difference between the two groups was in pre-biologic therapy CRP values. ADA-treated patients had an average pre-biologic CRP level of 13.2 mg/L, while IFX-treated patients had a median pre-biologic CRP level of 4.8 mg/L ($p = 0.015$, uncorrected).

Many of our analyses involved limiting patient numbers based upon length of follow-up, and so it is informative to examine the distribution of biologic starting dates. These are shown in Figure 7.1, together with the time from diagnosis to biologic therapy initiation and patient age at therapy initiation. As is evident, most patients with time stamps available in our IBD cohort began biologic therapy between 2009 and 2013. For the majority of patients, this was within ten years of diagnosis, and very few patients were older than 60 years old at this time point.

7.3 Survival analysis

Our first investigation sought to reveal whether there was a difference in time to severe outcomes between the two drugs. Our severe outcomes included GI surgery and relapse. We defined relapse as a CRP measurement greater than 80 mg/L. As described in Chapter 3, "normal" CRP measurements are less than 8 mg/L; levels

Table 7.1: *Summary statistics for anti-TNF-α-treated CD patient cohort.*

	ADA		IFX		
	Median	IQR	Median	IQR	p-value
Age at start of biologic therapy	30.9	15.5	32.5	20.8	ns
Age at diagnosis	21.0	9.7	21.0	14.5	ns
Years from diagnosis until biologic therapy	8.0	9.2	7.3	11.2	ns
CRP measurements					
Pre-biologic CRP value	13.2	30.5	4.8	21.8	0.015
Median CRP in year following diagnosis	16.0	22.5	19.0	21.5	ns
Maximum CRP in year following diagnosis	71.5	87.5	55.0	116.0	ns
	Prop.		Prop.		
Male	0.42		0.50		ns
Required surgery prior to biologic therapy	0.27		0.29		ns
Description of disease					
Inflammatory	0.39		0.43		ns
Stricturing	0.31		0.34		ns
Penetrating	0.19		0.21		ns
Location of inflammation					
Ileum	0.60		0.50		ns
Caecum	0.46		0.44		ns
Ascending	0.37		0.39		ns
Transverse	0.45		0.40		ns
Descending	0.44		0.37		ns
Sigmoid	0.47		0.54		ns
Rectum	0.40		0.41		ns
Anus	0.24		0.29		ns
Locations summary					
Ileal only	0.12		0.10		ns
Ileocolonic only	0.47		0.40		ns
Colonic	0.76		0.73		ns
Colonic only	0.28		0.33		ns
Medication ever prescribed					
5-aminosalicylates	0.42		0.41		ns
Prednosolone	0.63		0.59		ns
Azathioprine	0.78		0.79		ns
6-mercaptopurine	0.16		0.19		ns
Cyclosporine	0.02		0.03		ns
Methotrexate	0.11		0.09		ns
Mycophenolate mofetil	0.16		0.13		ns
Intravenous hydrocortisone	0.23		0.30		ns
Genetic mutations					
rs2066847 (NOD2 p.L1007fsX1008)	0.08		0.06		ns
rs2066845 (NOD2 p.G908R)	0.04		0.01		ns
rs2066844 (NOD2 p.R702W)	0.10		0.20		0.064

IQR=interquartile range; ns=not significant, with $\alpha = 0.1$; p-value is uncorrected and is for difference between ADA and IFX, as described in text.

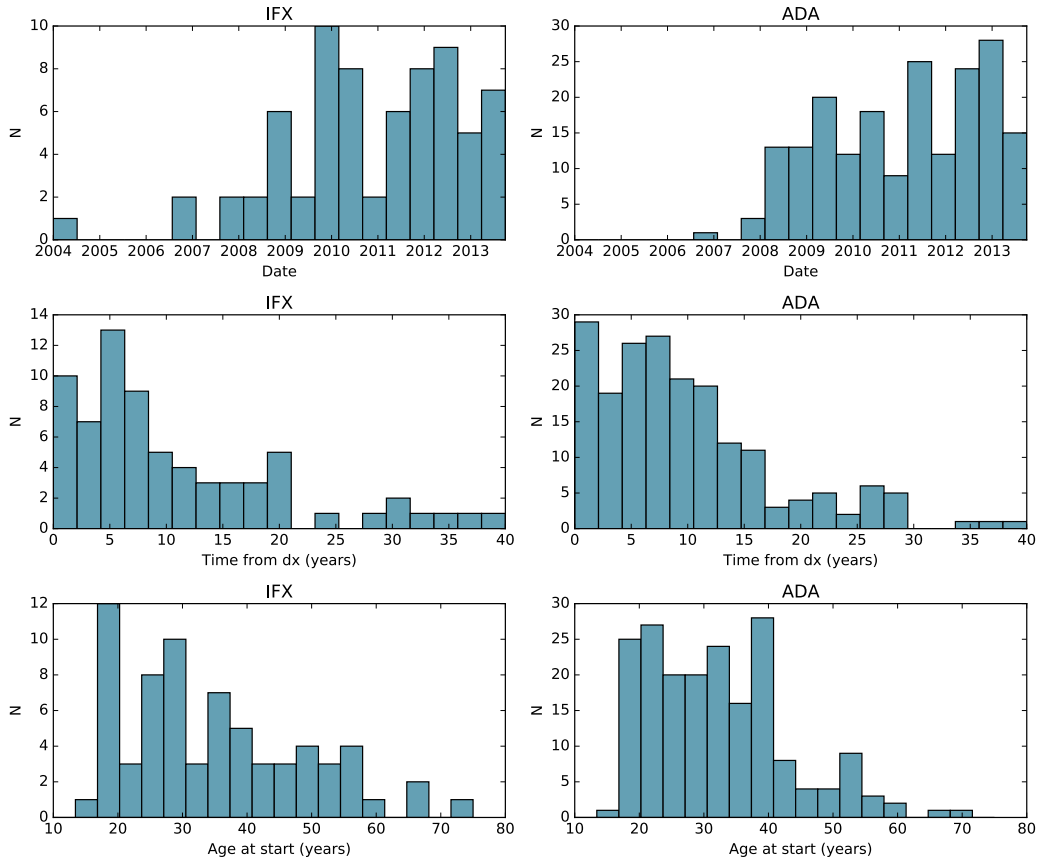


Figure 7.1: *Summary histograms of relevant timing information regarding biologic therapy initiation. Top: calendar date. Middle: time in years from diagnosis. Bottom: patient age*

of up to 40 mg/L are indicative of the inflammation encountered in a typical viral infection. Higher levels indicate more substantial levels of inflammation [208]. We also assessed survival with a CRP relapse-defining threshold of 40 mg/L and 100 mg/L as a sensitivity analysis. Patients could be censored from the investigation due to conclusion of the study period, which was taken to be November 30, 2015, to cover the time period during which we had laboratory measurements available for the patients. For the surgery sub-analysis, the conclusion was taken to be June 30, 2014, the time period over which we had surgery time stamp data.

We did not include the conclusion of biologic therapy as a severe outcome because, in an ideal treatment course, a patient would reach disease remission and no longer require the biologic. If relapse occurred, this would already be captured through our analysis. In addition, we did not include switching to a different biologic (i.e., an ADA-treated patient changing to IFX treatment) because in our pharmacy data, we only had time stamp data for four patients switching from ADA to IFX and eight

Table 7.2: *Years of survival relapse and surgery free. Cases in which patient follow-up was not long enough to assess median survival are demarkated with ().*

		Both	ADA	IFX
Relapse and surgery	Median survival	5.9	5.9	6.7
	75% survival	1.0	1.0	1.2
Relapse	Median survival	6.7	()	8.8
	75% survival	1.6	1.6	1.8
Surgery	Median survival	()	()	()
	75% survival	2.9	2.7	3.1

patients switching from IFX to ADA. With such a small sample size, we chose not to investigate this outcome separately.

We used Kaplan-Meier survival curves to assess time-to-severe-outcomes for all patients (as introduced in Section 6.5). We also examined how many patients responded to biologic therapy. Whereas previous studies have defined primary response based on response within the first three months and secondary nonresponse as a later loss of response, our longer available patient follow-up allowed us to examine longer time frames. Because all patients had at least two years of follow-up (between biologic therapy initiation in 2013 and lab measurements taken through 2015), we defined patients who had a relapse or surgery within two years of biologic therapy as "nonresponders" and those who did not as "responders."

Our survival analysis allowed us to assess both the effectiveness of the biologics together and to compare the relative efficacy of ADA and IFX. Figure 7.2 shows the Kaplan-Meier survival curves, where it may be seen that the overlap is such that there was no significant difference between ADA and IFX survival when examining relapse and surgery together or separately. Table 7.2 shows the median and 75% survival for each of these outcomes. Patients tended to succumb to relapse more quickly than surgery: 75% percent of patients remained free from relapse for 1.6 years, free from surgery for 2.9 years, and free from either for 1.0 years. This would be expected, as the decision for surgery is a major treatment decision. As we can also see, the median and 75% survival between ADA and IFX is similar. For instance, by one year after treatment, 25% of ADA-treated patients have had either a relapse or surgery; the same proportion of IFX-treated patients have had relapse or surgery by one year and three months.

When we varied the CRP threshold for defining relapse between 40 mg/L and 100 mg/L, we found that the resulting overall median survivals were 3.5 years and 6.3 years, respectively. The corresponding survival curves are shown in Appendix B.3.

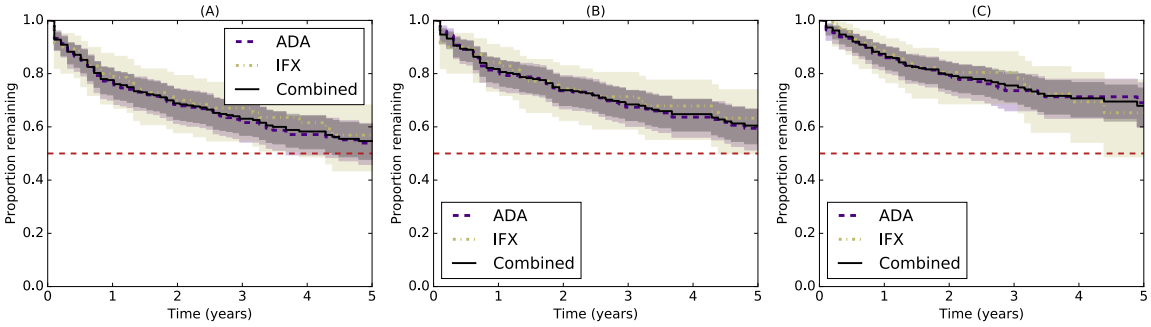


Figure 7.2: *Patient survival curves until (A) relapse or surgery; (B) relapse alone; (C) surgery alone. Time zero is defined by the initiation of biologic therapy.*

Based upon our definitions of responders and nonresponders, 181 patients were responders and 82 were nonresponders. The proportion of responders was similar between the two treatment groups: 132 ADA-treated patients (68%) and 49 IFX-treated patients (70%) were responders.

7.4 Rates of relapse

While survival analysis is a traditional approach for situations such as drug efficacy, we also have the opportunity to investigate additional aspects of the therapy's effects on the body. For instance, we would expect that if the drugs are successful, overall rates of relapse and surgery would decrease in the time period following therapy, as compared to prior time periods. To assess this, we compared rates of relapse and surgery for the periods before (t_{pre}, years) and the periods after (t_{post}, years) drug treatment. Initiation of biologic therapy is typically triggered by a relapse or treatment failure, so we excluded the 3-month window (t_{init}) both before and after drug initiation so as to not bias the rates by the initiating event, as illustrated in Figure 7.3. We only included patients who initiated drug therapy at least t_{post} years before the study end-period so as to not bias our sample with patients who did not have a relapse after therapy simply because they were not followed-up for as long.

While we included both surgery and relapse events in our baseline analysis, we also examined relapse events separately because surgical time stamps only became available in 2007. As a result, many patients will be lacking surgical events that occurred prior to biologic therapy simply because the data are not available. This decreases the number of expected surgical events in the pre-biologic period t_{pre}, which biases our results towards seeing a larger ratio of post/pre events (whereas our hypothesis is that this ratio of post/pre events would be less than one because biologic therapy is intended to decrease levels of inflammation). Therefore, this missingness

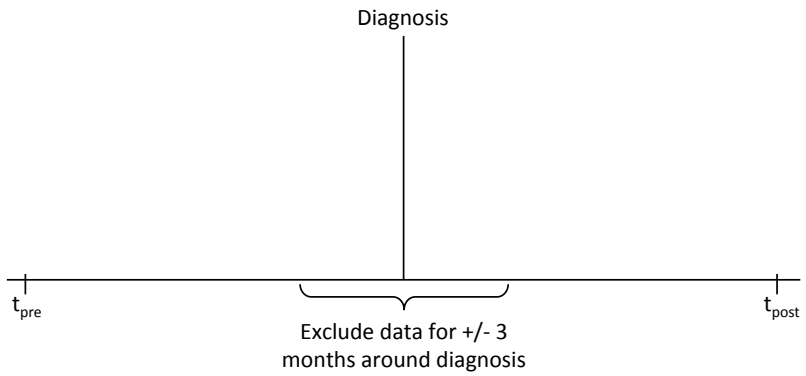


Figure 7.3: *Illustration of the range of data included for our relapse rate study.*

of the surgical data means that our estimates of any significantly larger numbers of pre-events will be conservative.

Once again, we defined a relapse event for a patient as a CRP measurement greater than 80 mg/L. A relapse event will typically be accompanied by many measurements in quick succession over the course of immediate treatment, and so we disambiguated a relapse event by disregarding all measurements within +/- 10 days of the maximum measurement over the cutoff of 80 mg/L.

We then used a proportion test to ascertain the probability of observing the given number of relapse or surgery events in the time period following biologic therapy. We assumed a null hypothesis of a uniform distribution of events before and after therapy, therefore our null hypothesis proportion was $p_0 = \frac{t_{post} - t_{init}}{t_{pre} + t_{post} - 2t_{init}}$. To compare the two drugs, we took the difference between the ratio $\frac{N_{post}}{N_{pre}}$ for each drug, where N_{pre} and N_{post} are the number of relapse or surgery events prior to and post biologic therapy, respectively. We then performed a permutation test where we randomly shuffled the drug label of each relapse and computing the resulting difference in ratios.

From the initial set of biologic-treated patients, 80 ADA-treated and 33 IFX-treated patients (113 total) had follow-up for at least five years following therapy initiation. When we examined whether there were differences in the rates of relapse and surgery before and after biologic therapy for these patients, we found that there were more severe (i.e., surgery and relapse) events before biologic therapy than afterwards ($p < 0.001$), with 179 events in the time period prior to therapy and 142 events afterwards (excluding t_{init}). This difference was also maintained when relapse events were examined alone ($p < 0.001$), with 147 relapse events in the time period prior to therapy and 91 events afterwards (again excluding t_{init}). When we compared rates of surgery and relapse between ADA and IFX-treated patients, we saw no significant

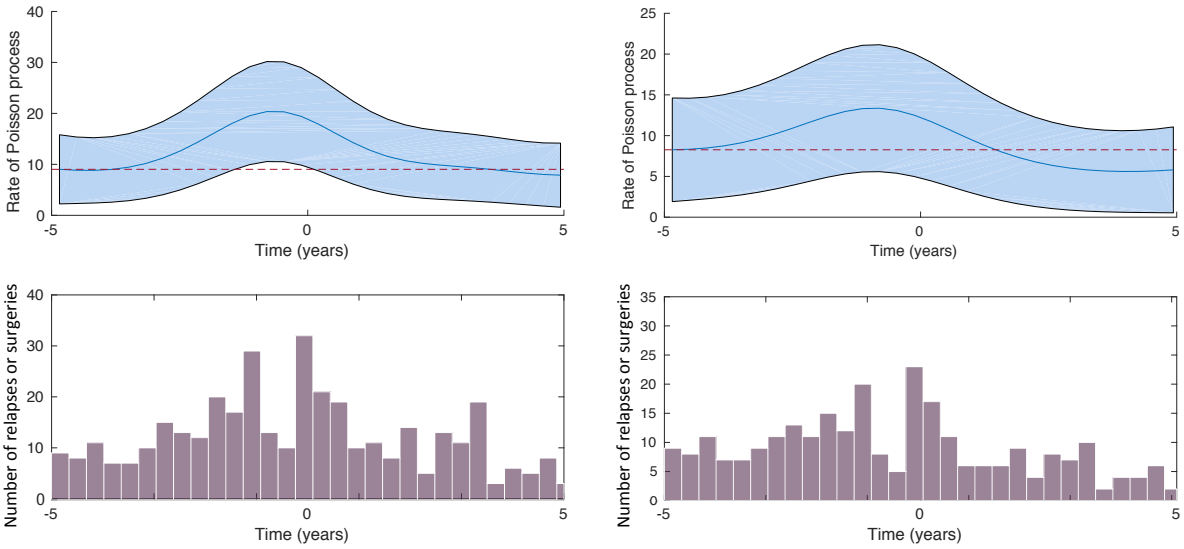
differences.

For visualisation purposes, we also used a log Gaussian Cox process (LGCP) to estimate the changing rate of relapse and surgery over the time horizon of $[-t_{pre}, t_{post}]$. The LGCP is a nonparametric approach for estimating the rate, $\lambda(t)$, of a point process. With our relapses and surgeries as points occuring over time, we were able to evaluate how the rate changed over the time period before and after therapy. Noteably, we did not exclude the period surrounding therapy initiation here, as we wished to visualise the relapse/surgery rate throughout the entire time period. Conceptually, a LGCP is similar to kernel density, except that it enforces a smoothness constraint via a Gaussian process. More details on this methodology are available in Appendix A.4.

Figure 7.4 shows the LGCP visualisations of the changing rate of relapse and surgery events over time. For reference, a rate of 10 would mean that on average there were 0.26 events per year per patient, or equivalently, each patient would have about 1 event every four years (we split the time range of +/- five years surrounding biologic initiation into thirty windows). As is apparent, even five years after biologic prescription, the rate of relapses had decreased from the original baseline of 8.3 (i.e., 0.22 relapses per year per patient) five years prior to biologic prescription, to 5.8 (i.e., 0.15 relapses per year per patient). We also see that the rate of relapses and surgeries increased for a long period preceding therapy, with a maximum of 20.4 (i.e., a rate of one relapse/surgery in a period of two years, on average per patient), which suggests that for many patients, progression towards the prescription of biologics may have been building for some time.

7.4.1 Analysis of relapse event severity

While the occurrance of relapse is one metric by which to assess the efficacy of therapy, we also sought to investigate a more holistic description of a patient's physiological state at relapse. It could be that relapses, while still occurring for some patients, become less severe after biologic therapy. We examined this by also considering patient levels of haemoglobin, WBCs, neutrophils, and monocytes at times of relapse. To compare the values of these additional labs during times of relapse, we used the GPD (as introduced in Chapter 4) to model the probability of extreme measurements of these additional labs. We used the GPD rather than our previously-developed time-window-based anomaly score, as we were interested in the likelihoods of individual points (i.e., single lab measurements), rather than the likelihood of a series of points in a window. We employed the same model-fitting procedure as we described previously



(a) Surgery and relapse (b) Only relapse

Figure 7.4: *LGCP results. Top: estimated rate of severity-event point process over time course of biologic therapy, with time zero being the start of therapy. Bottom: histogram of raw data.*

in Chapter 4 to ascertain the GPD parameters $\{\mu, \sigma_p, \xi\}$. We then evaluated the GPD negative log-likelihood, $S_{GPD,i,l}$, of the lab l measurement taken at the time i closest to relapse for each patient, for each relapse.

Having obtained a set of $S_{GPD,i,l}$ scores at each patient's relapse events before and after therapy, for each lab type, we created a composite GPD anomaly score, $S_{GPD,i}$, for each relapse event by summing each of these scores together. We then compared the median of these scores, per patient, before therapy and after therapy using a Wilcoxon signed-rank test (nonparametric paired T-test), where each patient's pre measurements are paired with his or her post measurements.

We found that there were no significant differences in the severity of relapse events before and after biologic therapy, in terms of our GPD lab-based metric.

7.5 Population-level CRP progression

While rates of surgery and relapse events provide us with important insight into how biologic therapy effects patients in terms of discrete events at a population level, we were also interested in the population-level progression of response in terms of systemic inflammation. We assessed this again via CRP levels, but in this case we examined all CRP measurements, not solely those that constituted relapse events. We calculated median CRP levels for each patient during set time intervals before

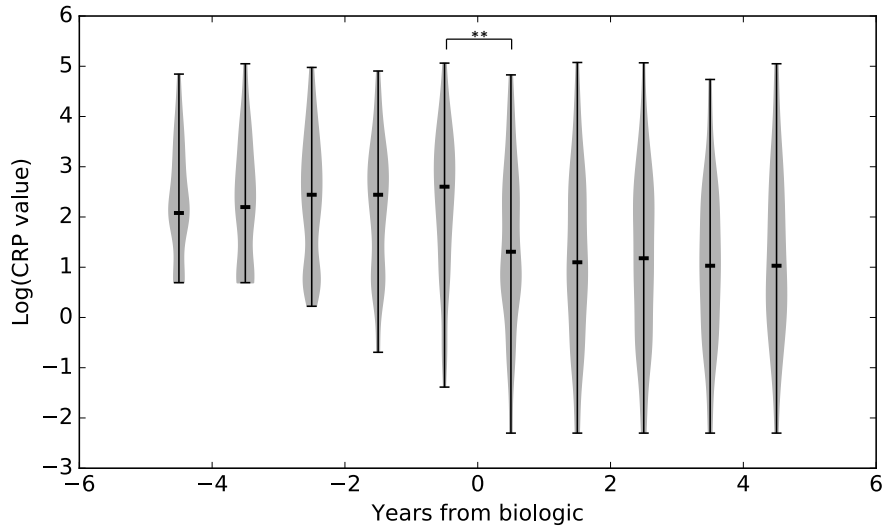
and after therapy.

While in our previous analysis of changes in relapse rates before and after therapy (Section 7.4), we were careful to include only those patients with adequate follow-up, here we allowed measurements from all patients. This was because here we were interested in studying the distribution of CRP measurements at various time points before and after therapy. A study of the difference in rates would be biased by patients lacking adequate follow-up because then the rates of relapse would be artificially reduced. Here, however, because all patients have a synchronisation time (i.e., the time of therapy initiation), any additional data will be informative in describing the population-level distribution of lab measurements at a particular time window. We used a MW to test for differences between the population medians at each of the successive time-windows. In the same manner, we examined differences between ADA and IFX in terms of median CRP levels within a given time window.

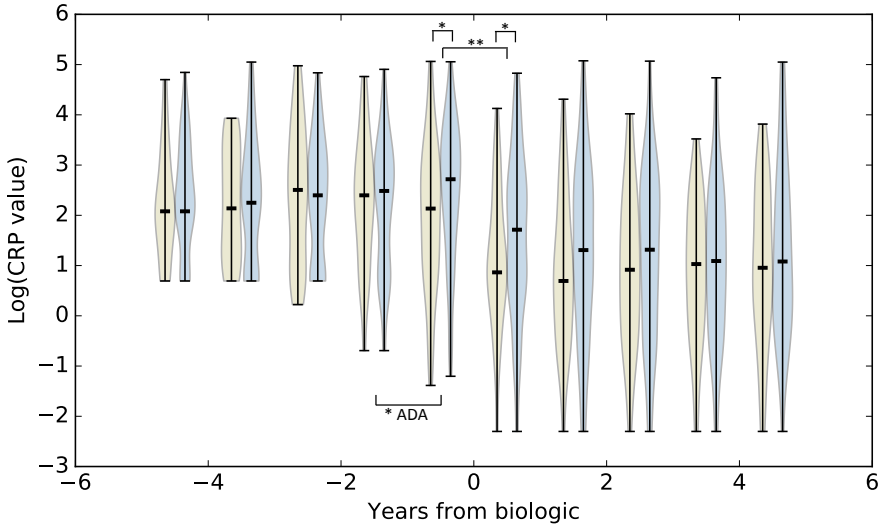
Figure 7.5 shows the median CRP measurements for all biologic-treated patients over the time period of five years prior to and five years after therapy. There was a significant difference in population-level CRP levels before and after initiation of biologics, with a median of 13.5 mg/L before therapy and 3.7 mg/L after ($p < 0.001$). Notably, this depression was sustained for the five years following therapy, with the population medians ranging from 2.8 to 3.3 mg/L (all of these successive time points were significantly different from the pre-therapy time points, $p < 0.001$).

When we compared ADA and IFX (Figure 7.5(b)), we found a difference of 6.8 mg/L in median CRP levels between the two populations receiving the drugs for the time window before therapy ($p = 0.013$), with ADA-treated patients having a median CRP of 15.2 mg/L and IFX-treated patients having a median of 8.4 mg/L. This difference decreased in subsequent time windows to 3.2 mg/L for the year immediately after therapy and to 1.7 mg/L in the following year ($p = 0.028$ and $p = 0.064$ for a difference in median CRP between the two drugs for these two time windows, respectively). We note that these presented p-values are uncorrected for multiple hypothesis testing.

However, while the median CRP levels decreased following therapy, we also found that the range increased, with some patients still having very high CRP levels. For instance, though the median CRP level immediately after therapy was 5.5 mg/L for patients treated with ADA, the maximum was 125 mg/L (the 75^{th} percentile was 14.2 mg/L). During the next two time windows, the observed maximum increased to 160 mg/L. This was not surprising, given our previous findings that many patients do not sustain a response. However, this does suggest that whether or not a biologic



(a) ADA and IFX combined



(b) ADA and IFX separately

Figure 7.5: *CRP decline for time period from five years prior to biologic therapy to five years after. Left: ADA and IFX combined. Right: Yellow=IFX (violins with left offset from center); Blue=ADA (violins with right offset from center). Significance markers: * = $p < 0.05$, ** = $p < 0.01$; shown for successive time windows and between two drugs.*

is effective for a particular patient, at the population level, patient inflammatory levels reach a sustained decline after initiating therapy (this of course may be due to additional or other therapies or surgery, following a failed biologic).

7.5.1 Responders and nonresponders

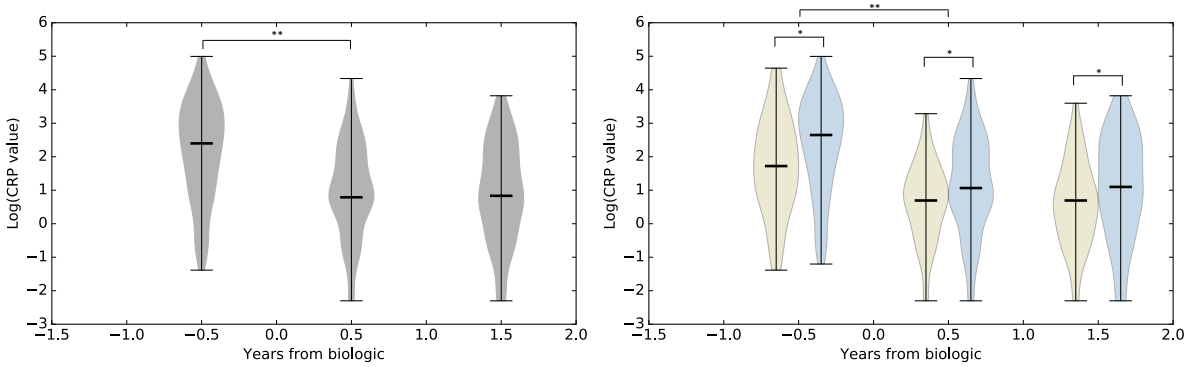
We then conducted this same analysis within two subgroups: those patients who adequately responded to biologic therapy (i.e., "responders" as defined in Section 7.3) and those who did not. We saw similar trends, as shown in Figure 7.6. Biologic responders had a very clear decrease in CRP levels following therapy: the median CRP before therapy was 11.0 mg/L, while it decreased to 2.2 mg/L in the following year afterwards. Prior to the start of treatment ADA-treated responders had slightly higher CRP levels, with a median of 14.0 mg/L, than IFX-treated responders, with a median of 6.75 mg/L ($p = 0.033$). However, this difference decreased after treatment to a median of 3.25 mg/L for ADA and 2.0 mg/L for IFX patients ($p = 0.063$ for a difference between IFX and ADA in the post-therapy time period).

It is interesting to note that even non-responders exhibited a (nonsignificant) decrease in CRP after therapy started, as shown in Figure 7.6 (b), and that levels continued to decrease after the first year. However, this decrease was not as pronounced as for the responders; median CRP decreased from 15.4 mg/L prior to therapy to a median of 6.6 mg/L in the year following. In addition, the continuing decrease after the first year to a median of 3.1 mg/L in the second year was likely to be because the treating physicians were trying alternative methods of therapy. We also found that non-responders seemed to have higher pre-therapy median CRP levels, of 15.5 mg/L, as compared to 11.0 mg/L for responders (we will return to this point in Section 7.7). Within non-responders, ADA and IFX-treated patients had very similar median CRP levels both before and after therapy (no significant differences).

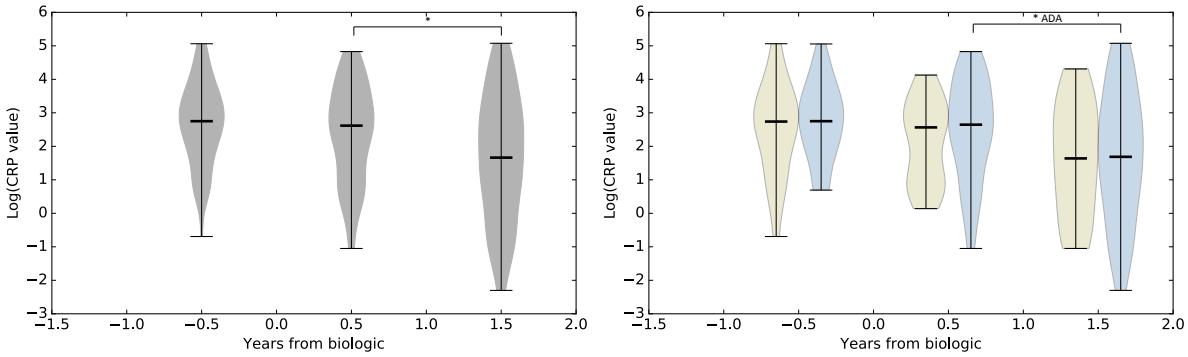
7.5.2 CRP responders

However, it is known from the clinical literature that some patients do not mount a CRP response [200]. Therefore, to represent more clearly the patient CRP response to biologics, we also split our population into two groups: those with any history of elevated CRP (before or after biologic drug prescription) and those without. We defined elevated CRP as a measurement of 40 mg/L or higher. We termed the former group "CRP responders" ($N = 221$) and the latter group "CRP non-responders" ($N = 42$).

For both groups, a decrease in median CRP levels was evident following biologic



(a) Biologic responders

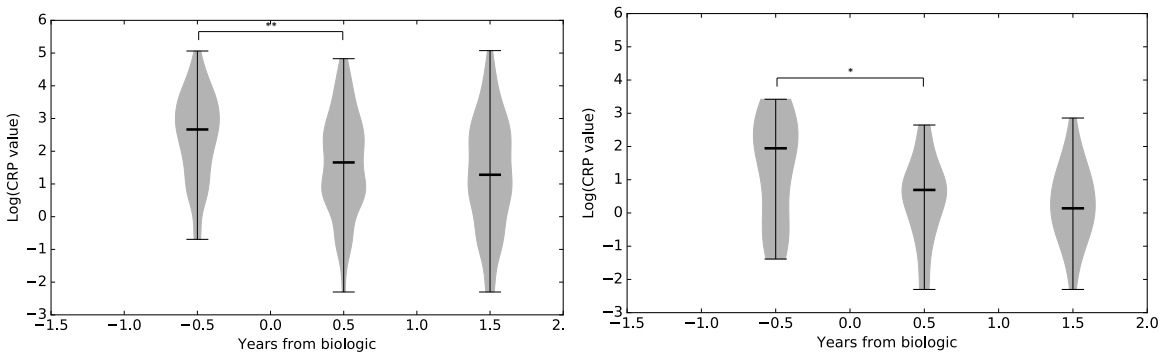


(b) Biologic non-responders

Figure 7.6: *Biologic responding (a) and non-responding (b) patients: CRP decline for time period from one year prior to biologic therapy to two years after. Left: ADA and IFX combined. Right: Yellow=IFX (violins with left offset from center); Blue=ADA (violins with right offset from center). Significance markers: * = $p < 0.05$, ** = $p < 0.01$; shown for successive time windows and between two drugs.*

therapy, even for those patients with depressed CRP levels. This is shown in Figure 7.7, where the median CRP for responders was 14.4 mg/L before therapy and 5.2 mg/L afterwards, for a difference of 9.2 mg/L ($p < 0.001$). For nonresponders the difference was 5.0 mg/L, where the median CRP before therapy was 7.0 mg/L and 2.0 mg/L afterwards ($p = 0.03$).

Figure 7.8 shows the subset of patients who were both CRP responders and biologic responders. Again, for this group, the depression of CRP in the years following biologic therapy was well-maintained, with a median CRP of less than 3.0 mg/L for all time windows up to five years after therapy (versus a median of 13.5 mg/L in the period before therapy). This indicates that the trend in Figure 7.5 was not simply because of patients who may not be able to mount a CRP response. The maximum per-patient median CRP was depressed for this subset of patients initially following



(a) CRP responders (b) CRP non-responders

Figure 7.7: *CRP responding (a) and non-responding (b) patients: CRP decline for time period from one year prior to biologic therapy to two years after. Significance markers: * = $p < 0.05$, ** = $p < 0.01$; shown for successive time windows.*

therapy (maximum of 76.3 mg/L and 45.6 mg/L in the first and subsequent year following therapy, respectively, as compared to a maximum of 147.7 mg/L before therapy). However, the maximum CRP observed steadily increased over time, with a maximum of between 90-148.8 mg/L in the following years. In other words, though at a group-level, systemic CRP decreases were maintained, as time went on some patients unfortunately progressed to relapse. ADA and IFX-specific results for CRP responders are shown in Appendix B.3.

7.6 Individual patient trajectories

In the previous section we described our approach for looking at CRP progression for the entire population. However, we were also interested in whether there are sub-types of response to biologics at an individual patient level. We returned here to our anomaly scores, as presented in Chapter 4, as a means for capturing trajectories over time. These scores serve to identify and accentuate variations in extreme lab measurements, while ignoring sub-threshold measurements. While the previous analysis gave us a representation of the entire range of CRP measurements over time, the anomaly scores allow us to focus upon the severe measurements, which are more useful for distilling and understanding individual patient trajectories.

We used the same methodology as presented in Chapter 4. Briefly recapping, we created anomaly scores for each patient by evaluating the log-likelihood of extreme lab point patterns in successive time windows. We considered data lying in the range spanned from six months before initiation of therapy to three years afterwards; we moved our one-year time windows forward in one-month intervals, resulting in $L = 31$

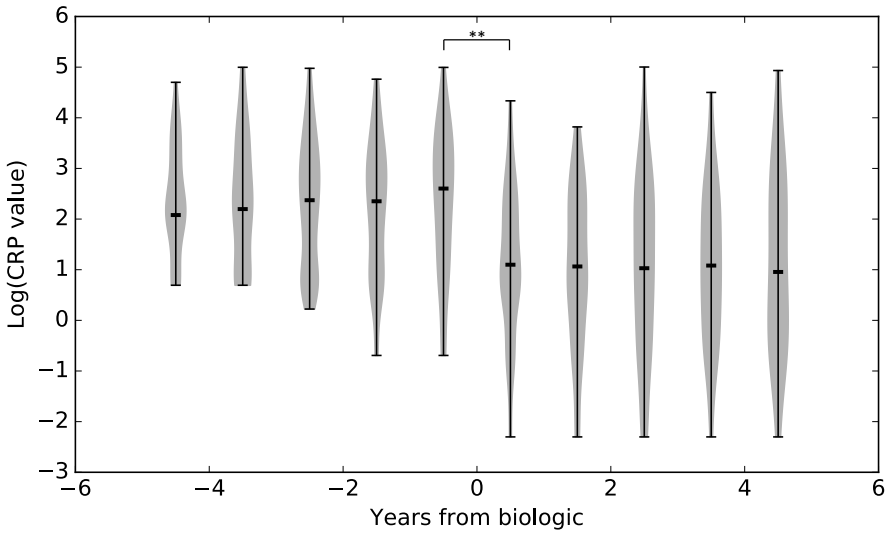


Figure 7.8: *Patients who were both CRP responders and biologic responders: CRP decline for time period from five years prior to biologic therapy to five years after. Significance markers: * = $p < 0.05$, ** = $p < 0.01$; shown for successive time windows.*

time windows for each patient. To capture patient trajectories not simply in terms of CRP, but also in terms of other labs, we created anomaly scores for hemoglobin, WBC, and neutrophil levels in addition. We thus obtained a vector of AUAS values of length L (one AUAS value for each time window). We concatenated these scores together into a single feature vector per patient, of length $L \times 4$. We did this so as to preserve differing trajectories that may exist within each lab (rather than, for instance, adding them together to create a composite severity score at each time point). Additionally, we created a "surgery anomaly score" to capture surgical events. We constructed this surgical anomaly score by inserting an offset, $F = 100$, at any point when a patient had surgery:

$$S_{surg}(t) = \begin{cases} F, & \text{if surgery in month } t. \\ 0, & \text{otherwise.} \end{cases} \tag{7.1}$$

We set $F = 100$ so as to make the surgical offset of a comparable scale to the range of observed AUAS scores. We then concatenated this score to the lab-derived scores, resulting in a final vector of length $L \times 5$ for each patient.

Having obtained such a vector for each patient, we then used hierarchical clustering with Ward's criterion and a Euclidean distance metric to identify clusters of patients with consistent trajectories. We employed hierarchical clustering because we were

interested in differences in the lab dynamics over time. Having derived our evenly-sampled anomaly scores, a Euclidean distance measure provided an effective way to compare trajectories. We chose the number of clusters for display by manually examining the resulting dendrogram.

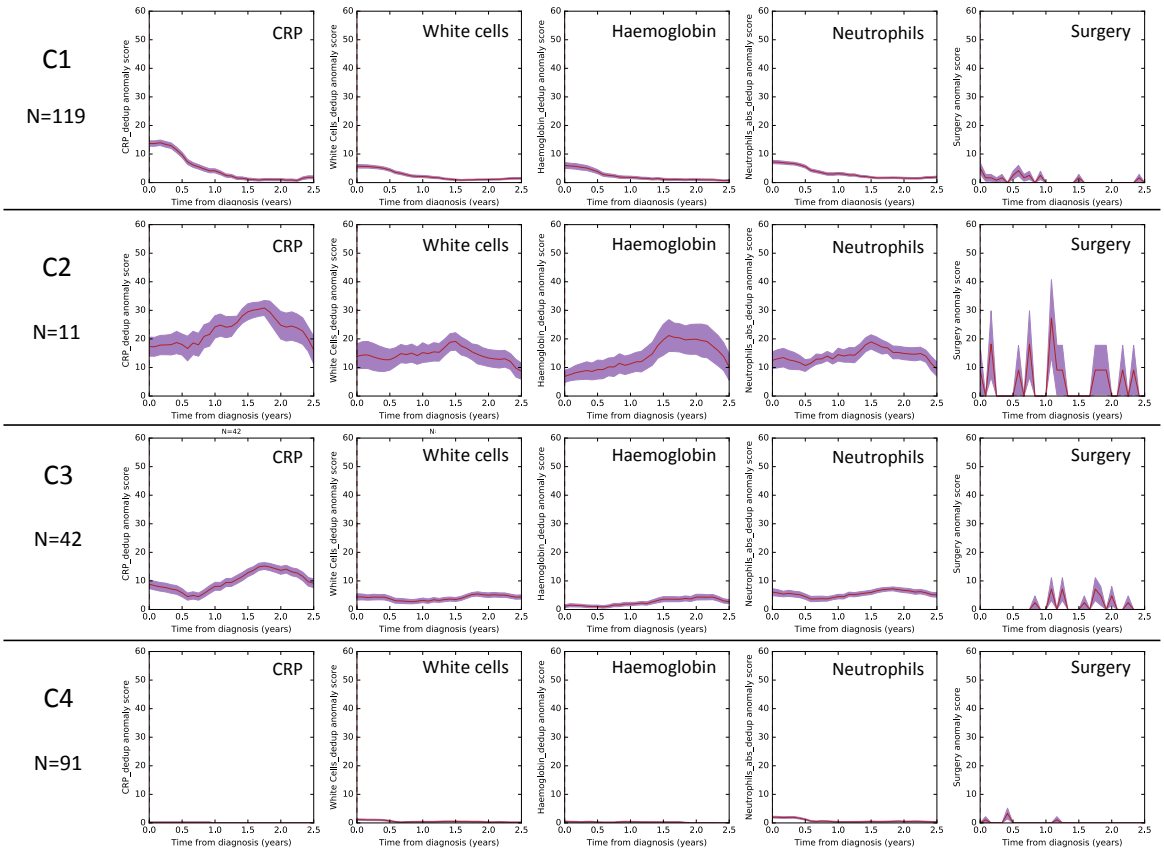


Figure 7.9: *Means and standard error of biologic-treated patient anomaly score trajectories for each input feature set, as labeled on y-axis. Each row represents a distinct cluster, C1, C2, C3, C4, respectively.*

Upon clustering our anomaly score feature vectors for the entire patient population, we found four distinct patient subgroups. The first and largest of these, which we will term C1 ($N = 119$), exhibited steadily decreasing anomaly scores over time. Some of these patients may have quickly not responded to their biologic (e.g. progressed quickly to surgery), and others may have taken some time to gain clinical remission. The second group, C2 ($N = 11$), consisted of a small group of patients who had a very severe disease course, even following biologic therapy. The third group, C3 ($N = 42$), seemed to respond initially to the drug, but they quickly progressed towards more severe phenotypic behaviour. The final group, C4, had a very strong response to anti-TNF therapy, with almost no abnormal laboratory measurements

across the $N = 91$ patients and very few surgery events.

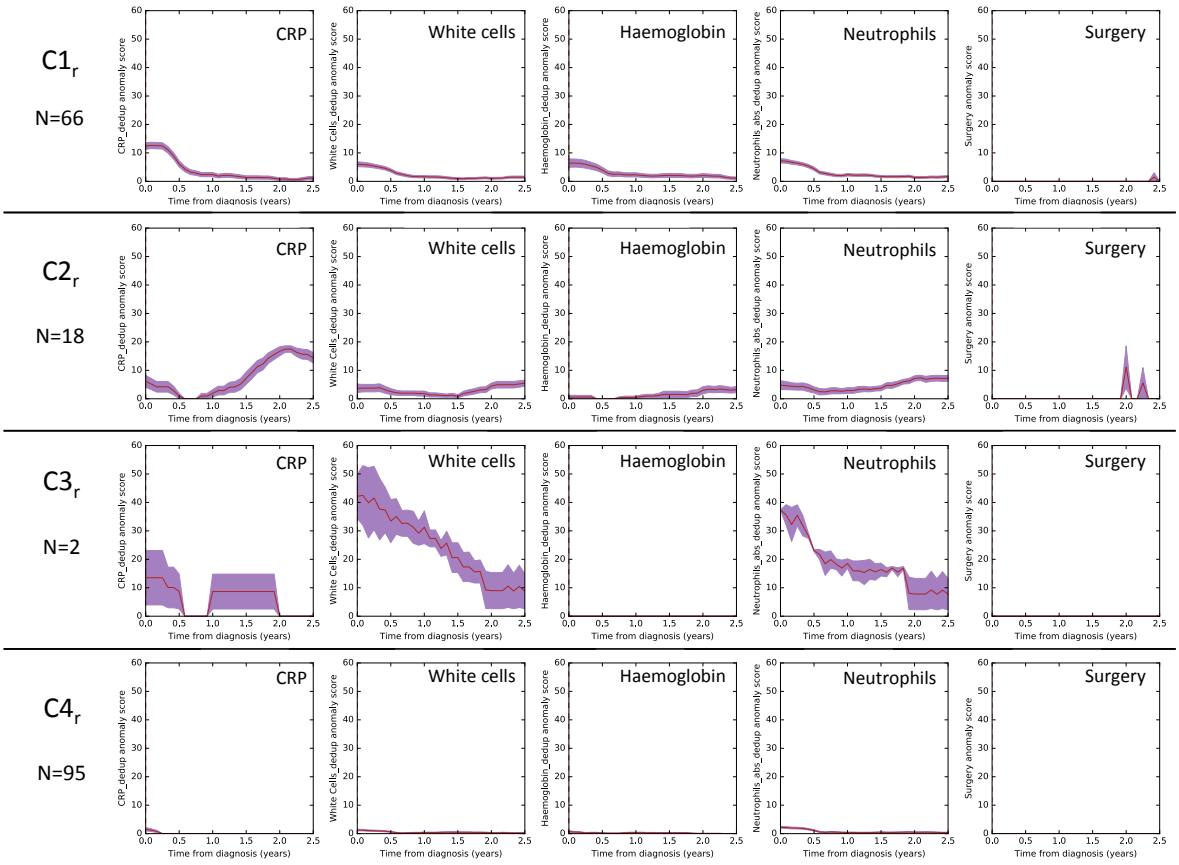
When we compared patients in each of these subgroups with the responders we defined in Section 7.3, we obtained the confusion matrix shown in Table 7.3. This fits with what we would expect given the lab and surgery anomaly scores: group C4 (very mild trajectories) consisted almost entirely of responders and group C2 (aberrent trajectories) was almost entirely nonresponders. C1 was comprised predominantly of responders, as well as some nonresponders who likely quickly progressed to either relapse or surgery. C3 was a mixture of responders and nonresponders, where it is likely that the responders will soon progress to a severe outcome.

Table 7.3: *Confusion matrix between patient trajectory clusters and responder status.*

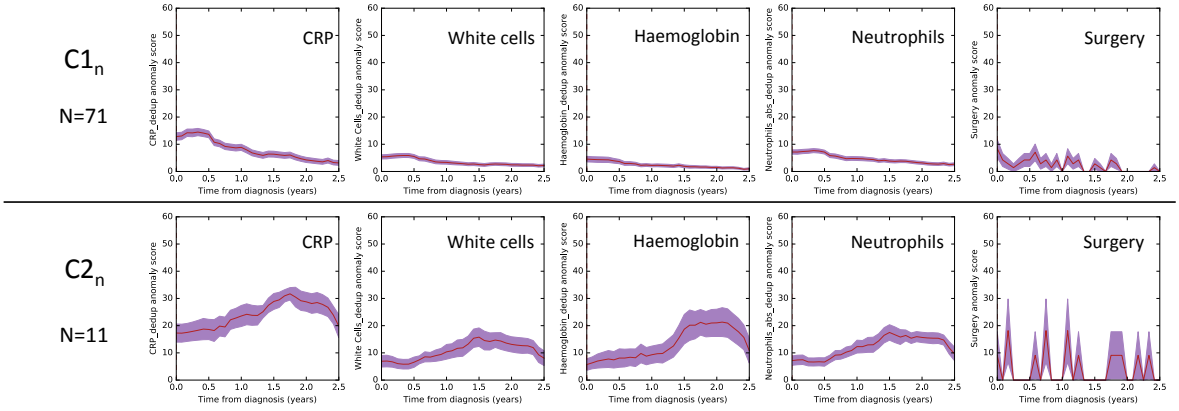
	C1	C2	C3	C4
Nonresponder	46	10	22	4
Responder	73	1	20	87

To better understand the patterns within the responder and nonresponder classes, we re-clustered within each of these subgroups separately. Four primary clusters were apparent within the responder patients. The first of these, $C1_r$ ($N = 66$), mirrored the pattern found for C1, where patients began with initially extreme lab measurements, but quickly progressed to the normal range. $C2_r$ ($N = 18$) mirrored the pattern for C3, with patients who initially seemed to respond but then progressed to severe labs or surgery. Indeed, we found that some of these patients had surgery just after the two year cutoff we have used to define responders. Again, with a similar pattern to C4, patients in $C4_r$ ($N = 95$) responded quickly and maintained this response. Finally, there was also a set of two patients ($C3_r$) who, while not having a CRP-defined relapse or surgery, nevertheless have very severe WBC and neutrophil measurements for a long period following biologic initiation.

Within the nonresponder patients, there were two primary subgroups. The first of these, $C1_n$ ($N = 71$), followed the pattern we have previously described, which is that of nonresponders having elevated lab measurements but nevertheless progressing in a positive direction towards more normal values. This is, again, likely due to intensification of alternative therapy or surgical removal of the affected portion of the GI tract. A second cluster of patients, $C2_n$ ($N = 11$), did not ever attain a lab-based response, and instead progressed towards even more severe measurements over time. These patients likely have very severe disease.



(a) Biologic responders



(b) Biologic non-responders

Figure 7.10: *Clustering results for biologic responding (a) and non-responding (b) patients: Mean anomaly score trajectories. Each row represents a distinct cluster. (a): $C1_r$, $C2_r$, $C3_r$, $C4_r$; (b): $C1_n$,$C2_n$, respectively.*

7.7 Predicting patient drug response

Our final analysis in this chapter involved the prediction of patient drug response. We took two approaches here: a statistical analysis of whether any of our examined covariates were significantly associated with therapy responsiveness, and a predictive analysis to determine whether we could create a useful clinical tool for predicting therapy response. For input features, we included those clinical covariates presented in Section 7.2, as well as genomic loci that are known to be associated with IBD risk [11]. For our analysis of genomic loci, we were obliged to limit our sample to those patients with Immunochip genotyping data available.

For statistical analysis of the enrichment of genetic loci in the responder and non-responder class, we used a chi-squared test. To aid in our power to detect these associations, we collapsed the alleles into two categories by merging the least-frequent homozygous allele with the heterozygous allele[3]. For the clinical covariates, we employed the same univariate approach as described in Section 7.2 (i.e., chi-squared tests for binary variables and MW tests for continuous variables). We then used a LR to identify covariates that were statistically associated with the outcome as a responder. Our predictive analysis followed the same methodology as that employed in Chapter 4. In brief, we used standard supervised machine learning classifiers (i.e., RFs and LR with both L1 and L2 norms) to predict the outcome of interest with internal cross-validation. However, here we additionally performed an internal feature selection via recursive feature removal. We trained the classification algorithm on the training set, removed the t lowest-ranked features, and repeated this process m times. Having then removed a total of $t \times m$ features, we tested the final algorithm on the test set. For the RF classifier, we used the feature importance (I_{RF}, as introduced in Section 4.4.3) as our ranking metric, and for LR, we used the absolute value of the feature coefficient ($|\boldsymbol{w}|$).

To label patients as responders or not in our predictive analysis, we employed three different definitions. The first of these, label (a), was according to our definition in Section 7.3, which we took as our baseline. This involved defining patients as responders if they did not have a severe event within the first two years of therapy initiation. We will refer to these patient sets as $P_{a,r}$ and $P_{a,n}$ for responders and nonresponders, respectively. We then refined this definition using our trajectory clustering results from Section 7.6. For label (b) in the responder class, we took the

[3]We remind the reader that alleles are potential genetic variants at a given loci; homozygous refers to having two copies of the same allele, while heterozygous refers to having a different copy on each chromosome.

overlap of those patients in Cluster C4 and the baseline (a) responder class (see Table 7.3, which follows in the Results): $P_{b,r} = P_{a,r} \cap \text{C4}$. For our label (b) nonresponder class, we took the overlap of patients in Clusters C1, C2, and C3 and the baseline (a) nonresponder class: $P_{b,n} = P_{a,n} \cap (\text{C1} \cup \text{C2} \cup \text{C3})$. Finally, for label (c), we defined our responder set in the same manner as with label (b): $P_{c,r} = P_{b,r}$. We limited our nonresponder set further, as $P_{c,n} = P_{a,n} \cap (\text{C2} \cup \text{C3})$. Our purpose in these various responder labels was to assess whether our trajectory phenotype clustering could be informative in isolating more homogeneous responder classes.

We first examined whether there were any univariate associations between genomic loci or clinical variables and our outcome of interest: responder status. For the genomic analysis, 114 ADA-treated patients and 42 IFX-treated patients (59% and 60%, respectively) had Immunochip data available, for a total of 156 patients. A single SNP loci, imm_17_35165903_G (rs12946510), was found to have an association with responder status (uncorrected $p = 0.0072$). The nonresponder class was enriched in this SNP, with 44 patients having at least one SNP (compared to the expected number of 37); 73 responders had at least one SNP, in comparison to the expected number of 80 (according to chi-squared expectation). A survival curve illustrating the patient survival times free of surgery or relapse is shown in Figure 7.11. There was a significant difference in survival between the two groups defined as those patiens with at least one copy of the rs12946510 risk allele and those without any copies ($p = 0.026$). We were unable to use the genomic features successfully to predict responder status, either alone or with the clinical features. This was not surprising, given our lack of univariate association findings and the large number of features (195 SNPs) for our patient cohort size (156 patients).

When we looked at our clinical variables (here we included the entire patient cohort), which included the prescribed biologic (i.e., ADA or IFX), there were no univariate associations between these features and responder status ($\alpha = 0.01$). This further supports the finding of equivalent performance between the two drugs.

Our AUROC results for prediction of patient drug responder status using clinical input features are shown in Table 7.4. As is apparent, our clinical covariates were not predictive of our baseline definition of response (i.e., labels (a)), with no AUROCs exceeding 0.57. LR found no features to be significant, though several had trends towards significance: the requirement of surgery prior to biologic therapy (coefficient of -0.29, $p = 0.06$), the pre-therapy CRP value (coefficient of -0.27, $p = 0.08$), and the maximum CRP at diagnosis (coefficient of 0.46, $p = 0.06$).

However, we did have some predictive power when we refined our patient responder

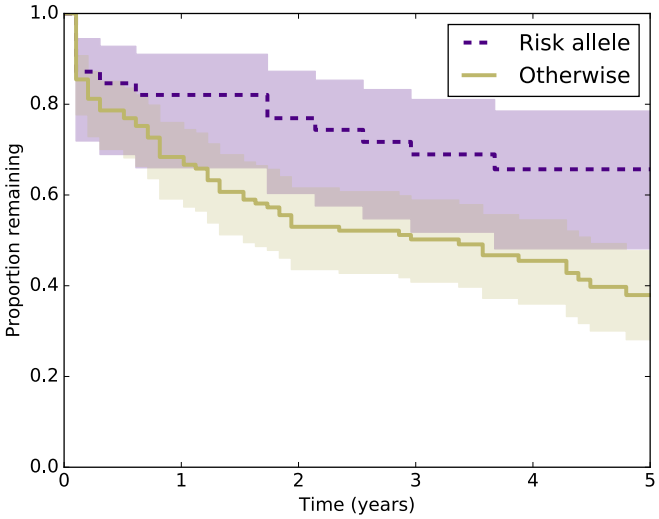


Figure 7.11: *Survival curves when split by patients with and without the IBD-risk allele SNP imm_17_35165903_G (rs12946510).*

Table 7.4: *Baseline prediction of patient responder status, presented as mean AUROC (standard deviation)*

	N resp.	N nonresp.	LR, L1	LR, L2	RF
Labels (a)	181	82	52.7 (9.7)	50.6 (10.2)	**56.2 (11.5)**
Labels (b)	87	78	66.5 (7.4)	**68.9 (7.1)**	66.9 (7.7)
Labels (c)	87	32	62.6 (14.3)	**65.2 (14.3)**	62.8 (13.4)

sets by labels (b) and (c), as shown in Table 7.4, where we were able to attain a mean AUROC of 68.9 (7.1 standard deviation) and 65.2 (14.3 standard deviation), respectively. This suggests that using our trajectory methods to more carefully clean patient cohorts can be beneficial in understanding relevant predictors of outcomes.

Upon performing recursive feature removal in predicting labels (b), our predictive performance was maintained with a much smaller feature set. The progression of AUROCs for models with increasing numbers of features internally removed is shown in Figure 7.12. We see that predictive AUROC was generally well-maintained with reduced sets of features, with performance actually increasing with only three features for the LR classifiers.

The features that consistently remained after the recursive removal procedure were: pre-drug CRP value, age at initiation of biologics, and whether surgery was required prior to biologic therapy. These features overlap with those features found to be significant in statistical logistic regression, which were: pre-therapy CRP (coefficient of -1.7, $p < 0.001$) and whether or not the patient required surgery (coefficient

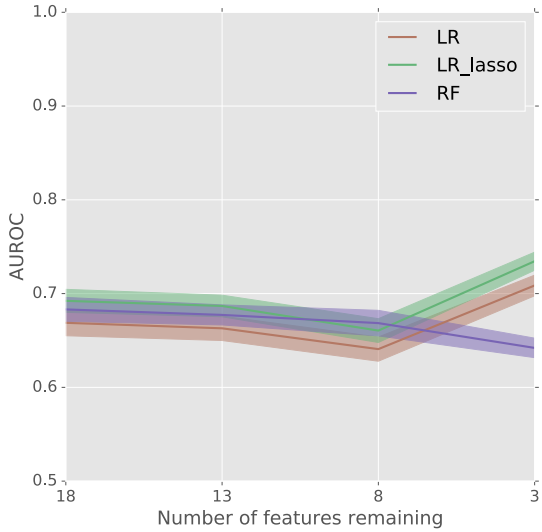


Figure 7.12: *AUROC for various classifiers as features were progressively removed within recursive feature elimination scheme.*

of -0.45, $p = 0.03$). These results suggest that high levels of CRP prior to biologic therapy and having had GI surgery prior to biologic therapy are associated with poorer outcomes over two years. While our surgery finding is consistent with previous research, our CRP finding was the opposite: previous studies have found a high level of pre-therapy CRP to be associated with positive outcomes [200]. It is possible that this difference could be due to the longer time range of follow-up in our study in comparison to previous work, or this could be an anomaly in our patient cohort.

7.8 Discussion

Anti-TNF-α therapy is very expensive and also can have side effects ranging from infection and rashes to anaphylactic shock and lymphoma [200, 209]. It is therefore important to understand how patients physiologically respond to therapy, whether there are differences between the available medication options, and to identify which patients will not respond. In the analysis presented here, we have examined these questions in our cohort of 263 CD patients treated with biologics.

We have found that in our patient cohort, median survival until relapse (defined in terms of extreme CRP measurements) or GI surgery was nearly six years and that there was no difference in survival between patients treated with ADA versus IFX. Approximately 70% of our patients reached two years without relapse or surgery. This number is roughly in-line with previous studies that have looked at primary and

secondary response to biologics, with reported rates of nonresponse ranging between 13% and 40% over time periods from eight weeks to two years [200]. Our survival findings are quite informative in understanding patient pathways after biologic prescription, as most studies look at follow-up periods of one or two years. The fact that median survival was approximately six years indicates that biologic therapy can have long-term positive outcomes for patients with severe CD.

When we examined actual rates of relapse/surgery, which is a high-level metric of long-term therapy effectiveness, we again found sustained improvement after biologic therapy even over periods of five years: rates of relapse and surgery were significantly lower after biologic therapy than before. We again found no significant differences between the effect of the two drugs. This could be because if biologic therapy fails, physicians may recommend more drastic surgeries than would have been recommended before anti-TNF was attempted. In other words, the decreased rate of severe events may be due to treatment intensification through other means, rather than because of the biologic drug actually being effective. However, the fact that we found long periods free of either relapse or surgery suggests that, for the patient population as a whole, biologics are actually very effective in reducing severe outcomes for several years.

Our population-level analysis of systematic inflammation levels (i.e., CRP measurements) over time also supports this finding. As a population, patients experienced a notable and significant decline in CRP levels after biologic therapy initiation, even within patients who did not mount a strong CRP response. This population-level suppression was maintained over a five-year period. Patients treated with ADA tended to have higher pre-therapy CRP levels, and though this discrepancy was maintained for short periods following therapy (i.e., median CRP levels were higher for IFX-treated patients than ADA-treated patients), any significant difference disappeared in the data after two years of follow-up.

Upon refining this population-level analysis with a clustering analysis of subtypes of individual trajectories, we found several distinct types of patient trajectories following biologic prescription. Some patients have an immediate and sustained depression in physiologic markers of inflammation after therapy. Another subgroup of patients required approximately one year for their inflammation to reach normal levels. Another set of patients, while initially responding well, eventually began to have increasing blood levels of CRP and WBCs. Many of these patients then progressed to surgery or relapse. Finally, there was a small minority of patients who had such severe disease that their inflammatory state never reached normality. In examining

these progressions over time, we found that our various examined labs (i.e., CRP, haemoglobin, WBCs, and neutrophils) generally mirrored each other in terms of their trends over time. This sub-population analysis showed us that even non-responding patients tended to eventually reach (over longer time periods than responders) normal levels of systemic inflammation. This supports our notion that physicians continue to treat patients aggressively who fail biologic therapy, and it suggests that their efforts are generally successful.

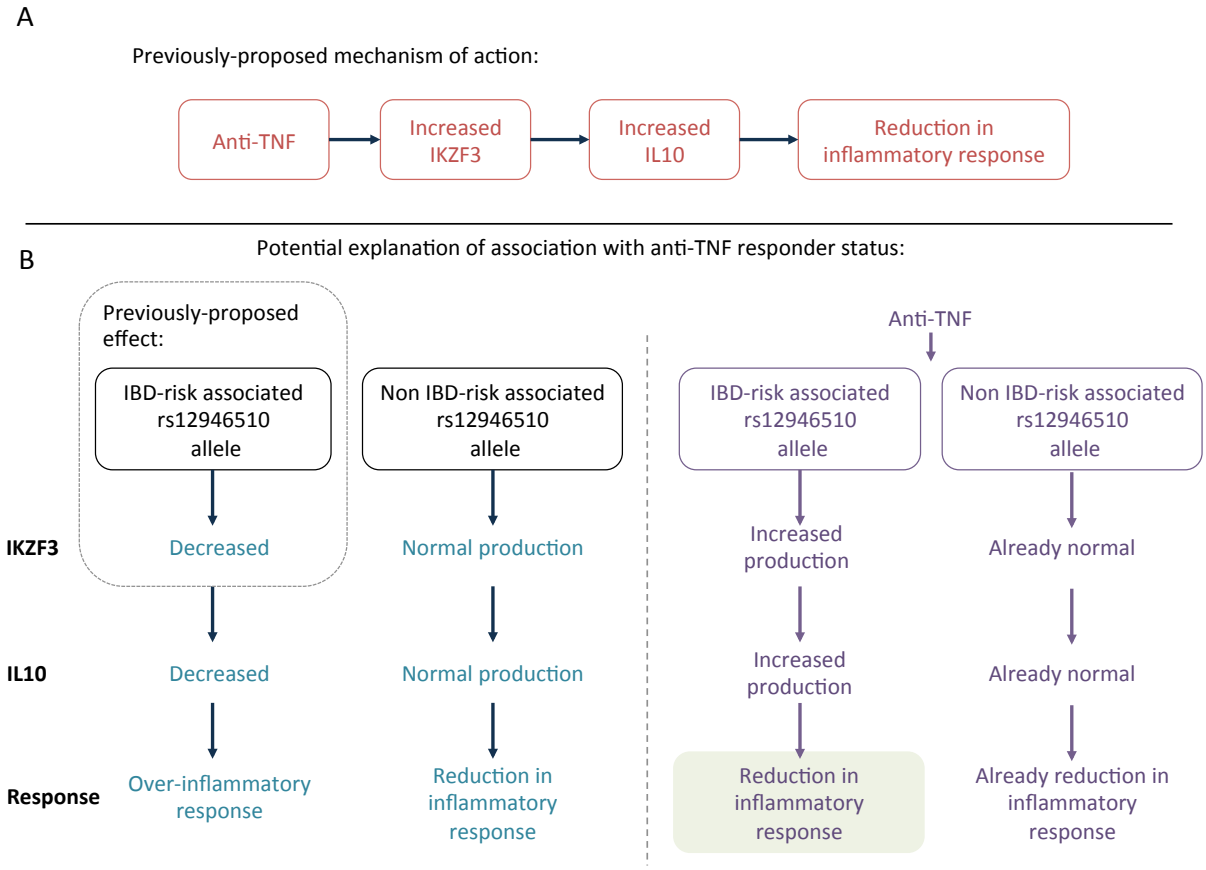


Figure 7.13: *Overview of potential mechanism of action of rs12946510, found to be nominally associated with responder status. A: previously-described effect of anti-TNF [210]. B: fitting pathway from (A) into previously-described eQTL effect of rs12946510 [211] and proposing potential mechanism for effect seen in anti-TNF patient response.*

Finally, in our prediction analysis, we found that we were unable to find strong predictive performance, precluding our models from being used in clinical practice to predict therapy response. However, the rs12946510 SNP we identified as being potentially associated with a positive drug response fits well into the existing literature, suggesting that this may in fact be involved mechanistically with anti-TNF

drug response. In a recent large autoimmune fine-mapping study, Farh et al. found that the risk-associated allele at the rs12946510 loci also acts as an eQTL to decrease expression of the Ikaros family zinc finger 3 (IKZF3) protein [211]. Separately, in their analysis of cellular responses to anti-TNF-α agents, Evans et al. found that anti-TNF-α acts via the IKZF3 gene to upregulate production of IL-10, an anti-inflammatory mediator [210]. This suggests that perhaps one mechanism of risk for patients with the CD risk-associated rs12946510 allele is because of lowered IKZF3 levels, which would result in lower-than-normal levels of anti-inflammatory IL-10. At the same time, since anti-TNF-α acts to increase the levels of this gene's expression, the drug may work more effectively for these patients because this pathway is already broken for them. Patients with the non-risk-associated allele already have adequate production of IL-10 (via this mechanism) and therefore anti-TNF is less effective through this means. This mechanism is illustrated in Figure 7.13.

7.8.1 Assumptions and future considerations

We have conducted a retrospective study on biologic drug effectiveness, with routinely-collected clinical information as our data source. Such a study is of course subject to bias by what is recorded in the patient record and confounding by clinical treatment. We have already touched upon the fact that the long-term favourable responses following biologic therapy may be due to physician intervention not recorded in the record, rather than actual biologic therapy effectiveness. In addition, we only have information recorded from the Oxfordshire region. If a patient sought treatment elsewhere, this information would not have been recorded, and we therefore may have missed relapse or surgery events. However, we have no reason to think that the rate of missingness of these events would differ over time, suggesting that rates of missingness before and after biologic therapy should be approximately equal. If anything, we would expect patients to be more closely monitored after biologic prescription, meaning that CRP would be more carefully recorded in the time period following this therapy than before. This would only increase the decreases in relapse rates and median CRP that we have observed.

Another challenge that we have not addressed here within our genomic association analysis was that some patients may progress quickly to relapse or surgery not because the drug was ineffective at reducing their IBD symptoms, but because they were intolerant or had severe side effects. While in both cases, nonresponse is the "correct" label in terms of patient treatment, the grouping together of these disparate reasons for treatment failure may reduce power to detect any true genetic associations. Future

work with larger patient numbers could subdivide by the reason for therapy failure.

Another aspect of our analysis that bears discussion is the use of our trajectory clustering results to inform our labels of responders and nonresponders. We found that further refining our labels from labels (a) to labels (b) and (c) produced improvement in our prediction performance. However, we also found that the key predictive variable was pre-therapy CRP. By refining labels via our cluster results, which took into account the time period from six months prior to therapy (a period which is very informative for purposes of understanding patient trajectories), we have effectively "cleaned" our patient pool of responders to include only those with very low pre-therapy CRP levels. Therefore, our refined prediction of responders and nonresponders may be more accurately described as prediction of patients with very low systemic inflammatory levels throughout and after their biologic therapy prescription period. In a final assessment, what we can confidently say is that in our dataset, a low CRP value prior to biologic therapy initiation is predictive of a mild CRP-related disease course following therapy, for up to two years.

7.8.2 Conclusion

In summary, we find no evidence for a difference in effectiveness between ADA and IFX, in terms of any of our examined metrics. Our analysis of outcomes for a real-world set of patients should lend confidence to physicians deciding between the two drugs that their effects are roughly equivalent. While at a population level patients tend to improve in terms of systemic inflammation over time, our approach has given us insight into differing progressions following therapy. Our findings show that even patients who do not initially have an immediate reduction in lab-based inflammatory levels may still attain long-term benefits from biologic therapy. In addition, it appears that some patients who eventually relapse or require surgery may have a progression in their lab-based severity leading up to these events. Future studies that systematically measure labs at specified intervals following therapy may be useful to further understand these patient trajectories. While we have outlined our methodology for predicting which patients may respond to therapy and for identifying underlying genomic factors, a larger patient cohort is needed in order to develop a robust clinical predictive tool. Additionally, it may be that our currently-evaluated metrics are insufficient and that alternative metrics (e.g., gene expression levels) will be required in order to obtain sufficient ability to predict which patients will respond to anti-TNF-α therapy before beginning.

Chapter 8

Conclusions and future work

Many diseases have heterogeneous trajectories. IBD is one such disease; patients vary greatly in their age of onset, the severity of their initial presentation, frequency of relapse, and response to medication and surgery. There are many other diseases and disorders (e.g., autism, asthma, heart disease) for which this heterogeneity is similar. Disease subtypes may be related to underlying genetics or to environmental factors that influence changes in gene expression patterns. Better understanding these subtypes will allow for both improved scientific understanding of disease physiology as well as progress towards "precision medicine" by enabling the delivery of personalised treatments. The work in this thesis contributes towards these efforts.

8.1 Summary of the thesis

In this thesis, we have used techniques from the field of machine learning to model and better understand IBD. We have started in Chapter 1 with a summary of relevant background material on clinical IBD and genetics. In Chapter 2, we provided an overview of previous work in IBD. There have been only a small number of studies that have sought to apply machine learning approaches to answer predictive questions related to IBD, such as distinguishing CD from UC. On the other hand, there has been a large body of literature focused on the association between genetics and IBD susceptibility. IBD statistical genetics has shifted more recently towards relating genetics to established IBD phenotypes. However, these studies have, to date, only looked at broad outcomes and phenotypes such as patient progression to surgery and the Montreal classifications. We have also reviewed the previous work on phenotypic modelling to uncover latent disease subtypes, as applied to other diseases. This field is new but growing; our review is therefore a useful contribution for summarising the current state-of-the-art. We found that while many studies have applied standard

clustering techniques to questions of disease sub-typing, fewer have directly modelled disease progression in a time-varying manner. In addition, there have been very few examples of subtyping studies that have combined hetergeneous gene expression, genetic, and clinical data.

In Chapter 3, we have described our patient cohort and the different types of data under investigation. Both clinical (e.g., demographic, lab measurement, medications, surgeries, IBD-specific variables, etc) and genetic data, as well as gene expression data for a small subset of patients, were available in our study. In Chapter 4 we introduced a cornerstone of our analysis in this thesis: an EVT model to summarise patient lab trajectories. To our knowledge, EVT methods have never before been used to model patient extreme lab measurements. This novel application of the methodology is well-suited for our clinical situation, where patients have periods of relapse and remission and more extreme labs are indicative of clinically-relevant increases in inflammation. In addition, because both single large deviations from normality and many repeated less-extreme above-threshold measurements are both indicative of severity, our model is well-suited to capture both aspects of "anomalous" behaviour. Because IBD patients tend to exhibit large, and less-informative, jumps in measurements within the clinically normal range, other methodologies that we examined to model the trajectories were less successful. Focusing on the extremes is a more informative way to compare patients in terms of disease progression.

Having explained and applied our EVT model to simulated and real patient data, we then introduced in Chapter 5 our clustering model for combining binary, binomial, and continuous data. We only used the binary and continuous portions of the model in the results presented in the thesis, as we wished to perform the subsequent genetic association study presented in Chapter 6. However, genomic variants could also be directly included in the model in future studies or other disease applications. After applying the clustering model to our gene expression dataset, we found two clear subsets of patients with differing gene expression patterns. These distinct phenotypes are promising for future study. We also found multiple subtypes of patients in terms of clinical "severity," where we included both classical severity metrics (such as medications required) and intermediate inflammatory markers (i.e., our previously-derived lab metrics). When we performed clustering across the entire IBD cohort, we found that the severe clusters were overwhelmingly composed of CD patients (rather than UC). This finding emphasises that, in general, CD patients have more severe disease phenotypes than UC patients. It also shows that it is useful to view the two diseases as a continuum, as some CD patients do have very mild trajectories while

some UC patients have very severe disease. When clustering within CD patients and UC patients separately, we found subgroups of patients with progressively increasing disease severity.

Turning then to the relationships between genetics and our EVT-derived lab severity metrics and these identified patient clusters, we conducted a large association study in Chapter 6. The results of this study provided clarity to the relationship between CRP levels, disease localisation, and the *NOD2* gene, where we found that *NOD2* variants appear to have a much larger effect in terms of acute inflammation in patients with ileal disease localisation. On the other hand, *NOD2* variants appear to confer a protective effect in terms of acute inflammation for UC patients. We are currently conducting a validation study of this finding. In addition, we put forth numerous other potential associations that may be worthy of further study. For instance, we found a counter-intuitive association between decreased neutrophil response during relapse events for patients with the *NOD2* p.L1007fsX1008 variant.

In Chapter 7, we used both standard clinical methods and our newly-developed EVT methods to conduct a concise study into anti-TNF-α therapy. This study provides a case example of how our methods can be used to answer open clinical questions. We examined several questions regarding this class of drugs, such as whether there are differences in patient response between different drug options and whether it is possible to predict patient response prior to therapy initiation. We found several prototypical ways that patients respond to the drugs, which is informative from both a descriptive understanding of patient progressions and as a means to isolate patient groups for genomic associations.

8.1.1 Key Contributions

The key contributions of this thesis therefore consist of:

- A review of the machine learning approaches for patient subtyping (no such review currently exists in the literature)
- Application of classical and new EVT methods for the modelling of chronic disease (these methods have not before been used in this application, to our knowledge)
- The implementation of a Bayesian clustering algorithm capable of handling various data types
- IBD patient subtypes description, using both our lab-based EVT severity metrics and classical metrics as input (such clustering has not previously been

performed for IBD patients)

- Results from a hypothesis-generating genomic association study relating genetics to disease severity (no study has before examined such a comprehensive representation of IBD severity)

- The finding that anti-TNF-α agents IFX and ADA are equally effective, in retrospective analysis of real patient data; the establishment of prototypical patient trajectories in response to these therapies

8.1.2 Discussion

Our work contributes to the development of methods to convert the overwhelming influx of medical data into usable and interpretable information. The tools we have developed can be utilised in many ways. Our EVT-based anomaly scores provide us with a means to represent patient lab-based severity in a principled, probabilistic manner. Our anomaly score trajectories provide not only useful visual summaries of patient lab-based abnormality over time, but also a means through which patient longitudinal patterns can be compared. This has application to many clinical questions; we illustrated such an approach within our analysis of response to biologic drugs. As we have also shown, the AUAS metrics derived from our scores represent a useful summary of the information relating to classical severity that is contained within patient lab measurements. This motivates their use as features for further phenotypic modelling (as discussed below regarding our clustering algorithm).

In addition, though we did not pursue such an approach in this thesis, our EVT model could be used to evaluate the lab-based severity of disease in an online manner. For example, as new patient lab measurements are taken during routine follow-up visits, these could be analysed to produce summary severity scores (based upon measurements take over the previous several months). Such scores could help physicians recognise progression towards relapse sooner than otherwise possible. Our scores could also be used as outcome measures within clinical trials to compare responses to new drugs, surgery alternatives, and differences in response when stratified by age, disease localisation, etc.

The identification of subgroups through our clustering methodology can allow for more targeted genetic association studies. One of the challenges with traditional GWAS is that case phenotypes are often defined broadly, resulting in heterogeneous patient groups. This introduces additional noise and also combines together many patient endotypes: distinct mechanisms may be obscured because they may only

be present in small subsets of the examined population. The discovery of new or particular genetic associations, as possible with more cohesive definitions of disease phenotype, can contribute towards improved understanding of IBD biology. Better understanding of the underlying biological mechanisms can then allow for the development of new drugs. In addition, the identification of disease endotypes (i.e., patient subgroups with distinct underlying mechanisms), together with targeted therapeutics, will allow for medications to be prescribed specifically to those patients in whom a positive response is much more likely (i.e., a precision medicine approach).

An exciting aspect of our models is that their use is not limited to modelling IBD lab-based severity. We could also use our EVT techniques to represent gene expression changes over time, microbiome population changes over time, or other relevant biomarkers. Similarly, our clustering model can be used for other relevant data types, depending upon the clinical question. For instance, while we have focused on severity-based subtypes within this thesis, subtypes based upon disease localisation (as will be discussed further in Section 8.2.2), environmental exposures, or co-morbidities may also be interesting to investigate.

In addition to developing models with widespread applicability, we have presented findings that improve current understanding of IBD symptomology. To our knowledge, principled clustering techniques have never before been applied to the question of identifying IBD patient severity subtypes. Our findings show that while some very severe patients exhibit both extreme lab measurements and require intensive medication, there are also patient subgroups with differing degrees of severity in terms of the two primary axes of lab-based and treatment-based severity. This could be due to differing treatment patterns or genetic factors, such as a predispositions towards decreased systemic inflammation or disease mechanisms that result in differing inflammatory responses in the gut.

8.2 Future work

From the initial work presented in this thesis, there are many avenues of future development. As noted, our presented methods and approaches are applicable to larger datasets when they become available. Larger patient numbers will likely lead to the identification of additional subtypes and will increase our power to detect the genomic associations that we began investigating in this thesis in Chapter 6.

8.2.1 Laboratory modelling extensions

To date, we have taken a population-modelling approach for our labs, where we have modelled the nonsevere patient population to determine the parameters defining "normality," and we have then used these population-defined parameters to assess the severity of new individuals. While this is useful for defining relative severity, it does not take into account individual-level differences. Some patients, for instance, may have lower baseline CRP levels and a small increase may be just as clinically notable as a large increase for another patient. Future models could take into account individual-specific effects, perhaps in a combined population/individual approach as by Liu et al. [90] (in which the patient-specific residuals from the population mean levels were modelled directly), or through a set of covariates in a generative model, as by Schulam et al. [103] (in which potential individual covariates were modelled within the generative framework).

An extension of our base univariate PP EVT model to the multivariate case would also be relevant for our application. This could allow us to model jointly our various labs, rather than fitting model parameters to each lab individually. One way in which we could model multiple labs at once is through a multi-task GP approach, in a similar manner to [86]. The individual EVT scores from our multiple labs would represent the multiple output tasks (which would be beneficial, as some labs are not measured as frequently as others). We could then also incorporate individual patient covariates as additional input variables (i.e., as a multivariate model).

The EVT approach we have presented in this thesis represents one way in which lab measurements over time can be modelled. We feel that our approach is motivated by our application and the desired question of summarising patient trajectories. However, additional modelling approaches are possible and may be more suitable for slightly different clinical questions. For example, the Hawkes process [212] is a self-exciting point process, meaning that when one measurement is observed, the likelihood of observing another measurement is increased. While in our work we have "deduplicated" the IBD patient lab measurements to induce i.i.d. behaviour, another approach is to model the repeated lab measurements explicitly through such a process. Hawkes processes have been used in finance applications (Laub et al. provide introduction to the methodology and examples [213]) and have recently been adopted by the machine learning community (e.g., for analysing social network activity [214]). These processes would be particularly useful in a predictive setting, where we wish to predict future patient behaviour, as they explicitly model the fact that in IBD,

the observation of one measurement increases the likelihood of observing another measurement.

NNs represent another potential methodological approach. Recent uses of NNs for clinical time series data [80,82,87,89] suggest that such approaches could be applicable for learning underlying patterns of IBD lab trajectories. The learned features within the hidden layers of a NN would be informative for understanding patient disease dynamics and patterns present across patients. While promising, such approaches would require an order of magnitude more patients; Kale et al., for instance, found difficulty fitting their models with densely-sampled ICU data for $8,000$ patients.

8.2.2 Clustering extensions

In Chapter 5 we discussed several extensions that could be made to our mixture model. One improvement that could be particularly beneficial would be a VB inference procedure. VB involves a deterministic algorithm, and so the label-switching problem is less relevant (though still relevant if multiple re-starts are performed). Another extension, especially as patient populations grow larger, is to implement a nonparametric mixture model through a Dirichlet process prior over the group probabilities, rather than using the Dirichlet distribution. In our current model implementation, we set our maximum number of groups, K, to a much larger number than we expect to observe in the dataset, which allows the data to choose the appropriate number of groups in an empirical Bayes manner. However, such an approach is a special case of the use of a Dirichlet process prior, which accurately models the creation and collapse of new potential groups, again based upon the complexity of the data [109,215].

An area of future work is the design of a clustering model that directly acts upon the input time series data rather than on extracted features. Again, a GP framework may be the most useful means for developing such a model, where we may consider a set of several latent functions, with our patient measurements being realisations from these functions. This approach was used, for example, by [103], [96], [99], and [109]. Hensmen et al. have developed fast inference methods to allow the discovery of latent clusters of genes across experimental replicates for each gene, for the purpose of identifying genes with similar expression patterns over time [216,217]. We could apply these methods by forming GPs over our EVT-derived PP likelihoods, where we could model our various labs as "replicates" for each patient to discover latent trends of disease progression over time.

As mentioned, we have focused on the question of patient severity in this thesis. However, various other feature sets could also be considered. For instance, a

clustering over the disease localisations could yield intesting insight into how the current localisation portion of the Montreal classifications should perhaps be adjusted to reflect the actual localisation patterns present in the data. We have provided preliminary results for this question in Appendix B.4, where we have identified 11 patient subgroups. These groups reflect the Montreal classifications to some degree, but they also illustrate that a much wider range of common phenotypes exist than can be represented with the current terminology.

8.2.3 Additional data sources and questions

While, due to space and time constraints, we have presented a limited number of analyses in this thesis, there are many other potential clinical questions that could be answered with our existing dataset. For instance, recent work has shown that it is possible to predict whether acute severe UC patients will re-present with acute severe symptoms within three years, based upon their extent of disease, CRP, and haemoglobin levels at diagnosis [218]. With the power of nonlinear machine learning classifiers and additional disease metrics, it is possible that we could further improve upon performance in this predictive task. The ability to predict disease course at diagnosis would be very useful, particularly if subtyping analysis reveals informative genetic loci. Such tools could guide clinical decision-making by helping physicians to decide which patients to treat more aggressively. Other prediction questions, such as response to surgery, are also possible to investigate with our dataset. We could address such questions by applying various supervised learning techniques, as presented in Chapter 4, combined with feature selection techniques (for example, recursive feature removal, as presented in Chapter 7). Having identified such features, we could also evaluate their performance once converted into easily-used clinical scoring metrics (such as the point system employed in the acute severe UC score described above).

Beyond extensions to our current underlying methods and current dataset, there are different approaches that could be taken if additional types of data become available. Our ability to identify patient subtypes would be much improved if we could incorporate a patient's current medication prescriptions at every time point. For instance, while we could perform patient trajectory clustering for all patients (i.e., a similar analysis as in Chapter 7, but from diagnosis, rather than just following biologic therapy), such an analysis would be very exploratory with our current data. This is because medication timing information is absent in performing this clustering, meaning that it is difficult to determine whether periods of disease remission are due to concurrent usage of strong immunosuppressants or due to a mild disease

course. While prescription does not indicate compliance, it is more informative than not having this information at all. We could then cluster patient trajectories just within patients prescribed similar medications (for instance comparing responses as in Chapter 7 for biologics), or we could directly incorporate medication information as a covariate in a time series model (as discussed in Section 8.2.1).

With additional follow-up time for the current patient cohort, we could also then begin to use patient endoscopies as another dimension to represent severity. Currently, our endoscopy data is very limited due to the short time that the hospital endoscopy department has had electronic reporting software (i.e., since 2013). With longer follow-up, we can begin to look at patient inflammatory patterns over time in terms of endoscopy results. We can also examine how and whether lab patterns are related to endoscopy behaviour. We could then combine both of these temporal metrics of severity to create a composite severity score. One way to combine these metrics may be to develop a score within the context of a clinical trial, during which we could obtain physician-defined (or patient-defined) "gold standard" scores of patient severity at designated time intervals. We could then determine appropriate weightings for our various lab-based and endoscopy-based features in a supervised framework. However, part of the challenge of such an approach is defining severity; as discussed in Chapter 1, IBD patient severity is very difficult to define, and numerous clinical scores exist in the literature.

In addition, functional gene expression data could be incorporated into the severity analysis as more patient samples undergo these experiments. Functional data could more clearly point towards disregulated inflammatory pathways within more and less severe patients. For example, to look at patterns present at both the level of gene expression and the level of the severity phenotype, we could use methods similar to the work of Kirk et al. (Bayesian clustering) or Wang et al. (similarity networks), both of which induce dependencies between different datasets when clustering [109, 112]. This is in contrast to the clustering we have performed in this thesis, in which we modelled each feature (regardless of data type) as contributing equally towards the determination of patient subgroups.

8.3 Conclusion

We see our work in this thesis as part of a larger trend towards better understanding chronic disease progression and patient-specific differences. We believe that with the growing collection of IBD patient phenotypic data and the rapidly advancing methods for analysing this data, we will begin to be able to identify subtypes of disease

that will be informative for patient care. Such findings may eventually motivate new phenotypic classification systems, potentially supplanting the current Montreal classifications and CD/UC distinction.

The importance of our work is underscored by recent efforts in both the USA and the UK to develop large national IBD-specific biobanks. In June 2015, an initiative called the IBD Plexus was announced in the US to aggregate clinical information for 40,000 IBD patients, 7,000 of whom will have biosamples collected over time [219]. In the UK, a new national initiative termed the IBD BioResource will be starting to recruit patients this year (2016), with an eventual goal of reaching 25,000 patients [220, 221]. Again, genetic information and blood samples will be obtained from consented patients. These programs are exciting, but the cohorts will not be fully recruited for several years. The work of this thesis lays the groundwork for future analyses - we present methods that can be readily applied to these larger patient cohorts to identify trends that may only be apparent with larger sample sizes.

References

[1] I. S. Kohane, "Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases," *Genome Biology*, vol. 15, no. 5, p. 115, 2014.

[2] M. W. State and N. Šestan, "The emerging biology of autism spectrum disorders," *Science (New York, NY)*, vol. 337, no. 6100, p. 1301, 2012.

[3] J. Lötvall, C. A. Akdis, L. B. Bacharier, L. Bjermer, T. B. Casale, A. Custovic, R. F. Lemanske, A. J. Wardlaw, S. E. Wenzel, and P. A. Greenberger, "Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome," *Journal of Allergy and Clinical Immunology*, vol. 127, no. 2, pp. 355–360, 2011.

[4] G. W. De Keulenaer and D. L. Brutsaert, "The heart failure spectrum time for a phenotype-oriented approach," *Circulation*, vol. 119, no. 24, pp. 3044–3046, 2009.

[5] J. Loscalzo, I. Kohane, and A.-L. Barabasi, "Human disease classification in the postgenomic era: A complex systems approach to human pathobiology," *Molecular Systems Biology*, vol. 3, no. 1, 2007.

[6] J. Satsangi, M. Silverberg, S. Vermeire, and J. Colombel, "The Montreal classification of inflammatory bowel disease: Controversies, consensus, and implications," *Gut*, vol. 55, no. 6, pp. 749–753, 2006.

[7] A. N. Ananthakrishnan, "Epidemiology and risk factors for IBD," *Nature Reviews Gastroenterology & Hepatology*, 2015.

[8] C. Abraham and J. H. Cho, "Inflammatory bowel disease.," *The New England Journal of Medicine*, vol. 361, no. 21, pp. 2066–2078, 2009.

[9] I. S. Group, "IBD standards: Standards for the healthcare of people who have inflammatory bowel disease (IBD), 2013 update," 2013.

[10] I. Cleynen, G. Boucher, L. Jostins, L. P. Schumm, S. Zeissig, T. Ahmad, V. Andersen, J. M. Andrews, V. Annese, S. Brand, *et al.*, "Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study," *The Lancet*, vol. 387, no. 10014, pp. 156–167, 2016.

[11] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, *et al.*, "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease," *Nature*, vol. 491, no. 7422, pp. 119–124, 2012.

[12] H. Huang, M. Fang, L. Jostins, M. U. Mirkov, G. Boucher, C. A. Anderson, V. Andersen, I. Cleynen, A. Cortes, F. Crins, *et al.*, "Association mapping of inflammatory bowel disease loci to single variant resolution," *bioRxiv*, p. 028688, 2015.

[13] B. Khor, A. Gardet, and R. J. Xavier, "Genetics and pathogenesis of inflammatory bowel disease," *Nature*, vol. 474, no. 7351, pp. 307–317, 2011.

[14] W. Strober, I. Fuss, P. Mannon, *et al.*, "The fundamental basis of inflammatory bowel disease," *The Journal of Clinical Investigation*, vol. 117, no. 3, pp. 514–521, 2007.

[15] E. Stange, S. Travis, S. Vermeire, C. Beglinger, L. Kupcinkas, K. Geboes, A. Barakauskiene, V. Villanacci, A. Von Herbay, B. Warren, *et al.*, "European evidence based consensus on the diagnosis and management of Crohn's disease: Definitions and diagnosis," *Gut*, vol. 55, no. suppl 1, pp. i1–i15, 2006.

[16] R. Xavier and D. Podolsky, "Unravelling the pathogenesis of inflammatory bowel disease," *Nature*, vol. 448, no. 7152, pp. 427–434, 2007.

[17] J. D. Lewis, "The utility of biomarkers in the diagnosis and therapy of inflammatory bowel disease," *Gastroenterology*, vol. 140, no. 6, pp. 1817–1826, 2011.

[18] D. Burger and S. Travis, "Conventional medical management of inflammatory bowel disease," *Gastroenterology*, vol. 140, no. 6, pp. 1827–1837, 2011.

[19] A. N. Ananthakrishnan, "Personalizing therapy for inflammatory bowel diseases," *Expert Review of Gastroenterology and Hepatology*, 2013.

[20] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for Mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[21] M. V. Rockman and L. Kruglyak, "Genetics of global gene expression," *Nature Reviews Genetics*, vol. 7, no. 11, pp. 862–872, 2006.

[22] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype - phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[23] A. D. Goldberg, C. D. Allis, and E. Bernstein, "Epigenetics: A landscape takes shape," *Cell*, vol. 128, no. 4, pp. 635–638, 2007.

[24] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annual Review of Medicine*, vol. 61, pp. 437–455, 2010.

[25] P. C. Phillips, "Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems," *Nature Reviews Genetics*, vol. 9, no. 11, pp. 855–867, 2008.

[26] A. Cortes and M. A. Brown, "Promise and pitfalls of the Immunochip," *Arthritis Res Ther*, vol. 13, no. 1, p. 101, 2011.

[27] H. D. VanGuilder, K. E. Vrana, and W. M. Freeman, "Twenty-five years of quantitative PCR for gene expression analysis," *Biotechniques*, vol. 44, no. 5, p. 619, 2008.

[28] S. F. Kingsmore, "Multiplexed protein measurement: Technologies and applications of protein and antibody arrays," *Nature Reviews Drug discovery*, vol. 5, no. 4, pp. 310–321, 2006.

[29] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[30] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLOS Computational Biology*, vol. 8, no. 12, p. e1002822, 2012.

[31] P. I. de Bakker, M. A. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight, "Practical aspects of imputation-driven meta-analysis of genome-wide association studies," *Human Molecular Genetics*, vol. 17, no. R2, pp. R122–R128, 2008.

[32] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 549–558, 2013.

[33] F. Sabb, A. Burggren, R. Higier, J. Fox, J. He, D. Parker, R. Poldrack, W. Chu, T. Cannon, N. Freimer, *et al.*, "Challenges in phenotype definition in the whole-genome era: Multivariate models of memory and intelligence," *Neuroscience*, vol. 164, no. 1, pp. 88–107, 2009.

[34] A. Geremia, P. Biancheri, P. Allan, G. R. Corazza, and A. Di Sabatino, "Innate and adaptive immunity in inflammatory bowel disease," *Autoimmunity Reviews*, vol. 13, no. 1, pp. 3–10, 2014.

[35] J. Z. Liu, S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, T. Shah, *et al.*, "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations," *Nature Genetics*, vol. 47, no. 9, pp. 979–986, 2015.

[36] M. D. Teare and J. H. Barrett, "Genetic linkage studies," *The Lancet*, vol. 366, no. 9490, pp. 1036–1044, 2005.

[37] W. E Ek, M. D'Amato, and J. Halfvarson, "The history of genetics in inflammatory bowel disease," *Annals of Gastroenterology*, vol. 27, no. 4, p. 294, 2014.

[38] W. Strober, N. Asano, I. Fuss, A. Kitani, and T. Watanabe, "Cellular and molecular mechanisms underlying NOD2 risk-associated polymorphisms in Crohn's disease," *Immunological Reviews*, vol. 260, no. 1, pp. 249–260, 2014.

[39] J. Wehkamp, J. Harder, M. Weichenthal, M. Schwab, E. Schäffeler, M. Schlee, K. Herrlinger, A. Stallmach, F. Noack, P. Fritz, *et al.*, "NOD2 (CARD15) mutations in crohn's disease are associated with diminished mucosal α-defensin expression," *Gut*, vol. 53, no. 11, pp. 1658–1664, 2004.

[40] E. Li, C. M. Hamm, A. S. Gulati, R. B. Sartor, H. Chen, X. Wu, T. Zhang, F. J. Rohlf, W. Zhu, C. Gu, *et al.*, "Inflammatory bowel diseases phenotype, C. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition," *PLOS ONE*, vol. 7, no. 6, p. e26284, 2012.

[41] M. H. Wanderås, B. A. Moum, M. L. Høivik, and Ø. Hovde, "Predictive factors for a severe clinical course in ulcerative colitis: Results from population-based studies," *World Journal of Gastrointestinal Pharmacology and Therapeutics*, vol. 7, no. 2, p. 235, 2016.

[42] A. J. Walsh, R. V. Bryant, and S. P. Travis, "Current best practice for disease activity assessment in IBD," *Nature Reviews Gastroenterology & Hepatology*, vol. 13, no. 10, pp. 567–579, 2016.

[43] W. R. Best, J. M. Becktel, J. W. Singleton, and F. Kern, "Development of a Crohn's disease activity index," *Gastroenterology*, vol. 70, no. 3, pp. 439–444, 1976.

[44] E. M. Yoshida, "The Crohn's disease activity index, its derivatives and the inflammatory bowel disease questionnaire: A review of instruments to assess Crohn's disease.," *Canadian Journal of Gastroenterology*, vol. 13, no. 1, pp. 65–73, 1998.

[45] M. Henriksen, J. Jahnsen, I. Lygren, N. Stray, J. Sauar, M. H. Vatn, B. Moum, I. S. Group, *et al.*, "C-reactive protein: A predictive factor and marker of inflammation in inflammatory bowel disease. Results from a prospective population-based study," *Gut*, vol. 57, no. 11, pp. 1518–1523, 2008.

[46] R. K. Russell, H. E. Drummond, E. E. Nimmo, N. Anderson, L. Smith, D. C. Wilson, P. M. Gillett, P. McGrogan, K. Hassan, L. T. Weaver, *et al.*, "Genotype-phenotype analysis in childhood-onset Crohn's disease: NOD2/CARD15 variants consistently predict phenotypic characteristics of severe disease," *Inflammatory Bowel Diseases*, vol. 11, no. 11, pp. 955–964, 2005.

[47] R. K. Weersma, P. C. Stokkers, A. A. van Bodegraven, R. A. van Hogezand, H. W. Verspaget, D. J. de Jong, C. van der Woude, B. Oldenburg, R. Linskens, E. Festen, *et al.*, "Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort," *Gut*, vol. 58, no. 3, pp. 388–395, 2009.

[48] W. S. Mow, E. A. Vasiliauskas, Y.-C. Lin, P. R. Fleshner, K. A. Papadakis, K. D. Taylor, C. J. Landers, M. T. Abreu-Martin, J. I. Rotter, H. Yang, *et al.*, "Association of antibody responses to microbial antigens and complications of small bowel Crohn's disease," *Gastroenterology*, vol. 126, no. 2, pp. 414–424, 2004.

[49] C. A. Solem, E. V. Loftus, W. J. Tremaine, W. S. Harmsen, A. R. Zinsmeister, and W. J. Sandborn, "Correlation of C-reactive protein with clinical, endoscopic, histologic, and radiographic activity in inflammatory bowel disease," *Inflammatory Bowel Diseases*, vol. 11, no. 8, pp. 707–712, 2005.

[50] L. Beaugerie, P. Seksik, I. Nion-Larmurier, J.-P. Gendre, and J. Cosnes, "Predictors of Crohn's disease," *Gastroenterology*, vol. 130, no. 3, pp. 650–656, 2006.

[51] F. L. Wolters, M. G. Russel, J. Sijbrandij, T. Ambergen, S. Odes, L. Riis, E. Langholz, P. Politi, A. Qasim, I. Koutroubakis, *et al.*, "Phenotype at diagnosis predicts recurrence rates in Crohn's disease," *Gut*, vol. 55, no. 8, pp. 1124–1130, 2006.

[52] I. C. Solberg, M. H. Vatn, O. Høie, N. Stray, J. Sauar, J. Jahnsen, B. Moum, I. Lygren, I. S. Group, *et al.*, "Clinical course in Crohn's disease: results of a norwegian population-based ten-year follow-up study," *Clinical Gastroenterology and Hepatology*, vol. 5, no. 12, pp. 1430–1438, 2007.

[53] M. Henriksen, J. Jahnsen, and I. Lygren, "Are there any differences in phenotype or disease course between familial and sporadic cases of inflammatory bowel disease? results of a population-based follow-up study," *Am J Gastroenterol*, vol. 102, pp. 1955–1963, 2007.

[54] O. Höie, F. Wolters, L. Riis, G. Aamodt, C. Solberg, T. Bernklev, S. Odes, I. A. Mouzas, M. Beltrami, E. Langholz, *et al.*, "Ulcerative colitis: patient characteristics may predict 10-yr disease recurrence in a european-wide population-based cohort," *The American Journal of Gastroenterology*, vol. 102, no. 8, pp. 1692–1701, 2007.

[55] C. Loly, J. Belaiche, and E. Louis, "Predictors of severe Crohn's disease," *Scandinavian Journal of Gastroenterology*, vol. 43, no. 8, pp. 948–954, 2008.

[56] I. C. Solberg, I. Lygren, J. Jahnsen, E. Aadland, O. Høie, M. Cvancarova, T. Bernklev, M. Henriksen, J. Sauar, M. H. Vatn, *et al.*, "Clinical course during the first 10 years of ulcerative colitis: results from a population-based inception

cohort (ibsen study)," *Scandinavian Journal of Gastroenterology*, vol. 44, no. 4, pp. 431–440, 2009.

[57] A. M. Schoepfer, C. Beglinger, A. Straumann, M. Trummler, S. R. Vavricka, L. E. Bruegger, and F. Seibold, "Fecal calprotectin correlates more closely with the simple endoscopic score for Crohn's disease (SES-CD) than CRP, blood leukocytes, and the CDAI," *The American Journal of Gastroenterology*, vol. 105, no. 1, pp. 162–169, 2010.

[58] I. Cleynen, J. R. González, C. Figueroa, A. Franke, D. McGovern, M. Bortlík, B. J. Crusius, M. Vecchi, M. Artieda, M. Szczypiorska, *et al.*, "Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: Results from the IBDchip European Project," *Gut*, pp. gutjnl–2011, 2012.

[59] C. Siegel, H. Horton, L. Siegel, K. Thompson, T. Mackenzie, S. Stewart, P. Rice, J. Stempak, S. Dezfoli, T. Haritunians, *et al.*, "A validated web-based tool to display individualised crohn's disease predicted outcomes based on clinical, serologic and genetic variables," *Alimentary Pharmacology & Therapeutics*, vol. 43, no. 2, pp. 262–271, 2016.

[60] L. Beaugerie and H. Sokol, "Clinical, serological and genetic predictors of inflammatory bowel disease course," *World Journal of Gastroenterology*, vol. 18, no. 29, pp. 3806–13, 2012.

[61] J.-M. Benitez and E. Louis, "Can we predict the high-risk patient?," *Digestive Diseases*, vol. 32, no. 4, pp. 328–336, 2014.

[62] S. Black, I. Kushner, and D. Samols, "C-reactive protein," *Journal of Biological Chemistry*, vol. 279, no. 47, pp. 48487–48490, 2004.

[63] S. Vermeire, G. Van Assche, and P. Rutgeerts, "Laboratory markers in IBD: useful, magic, or unnecessary toys?," *Gut*, vol. 55, no. 3, pp. 426–431, 2006.

[64] M. Cesarini, G. Collins, A. Ronnblom, A. Santos, L. Wang, D. Sjoberg, M. Parkes, S. Keshav, and S. Travis, "Predicting the individual risk of acute severe colitis at diagnosis," *Journal of Crohn's and Colitis*, 2016.

[65] A. M. Schoepfer, C. Beglinger, A. Straumann, E. Safroneeva, Y. Romero, D. Armstrong, C. Schmidt, M. Trummler, V. Pittet, and S. R. Vavricka, "Fecal calprotectin more accurately reflects endoscopic activity of ulcerative colitis than the Lichtiger index, C-reactive protein, platelets, hemoglobin, and blood leukocytes," *Inflammatory Bowel Diseases*, vol. 19, no. 2, pp. 332–341, 2013.

[66] S. Travis, J. Farrant, C. Ricketts, D. Nolan, N. Mortensen, M. Kettlewell, and D. Jewell, "Predicting outcome in severe ulcerative colitis.," *Gut*, vol. 38, no. 6, pp. 905–910, 1996.

[67] G. Ho, H. Lee, G. Brydon, T. Ting, N. Hare, H. Drummond, A. Shand, D. Bartolo, R. Wilson, M. Dunlop, *et al.*, "Fecal calprotectin predicts the clinical course of acute severe ulcerative colitis," *The American Journal of Gastroenterology*, vol. 104, no. 3, pp. 673–678, 2009.

[68] J. D. Ryan, M. Silverberg, W. Xu, L. A. Graff, L. Targownik, J. R. Walker, R. Carr, I. Clara, N. Miller, L. Rogala, *et al.*, "Predicting complicated Crohn's disease and surgery: Phenotypes, genetics, serology and psychological characteristics of a population-based cohort," *Alimentary Pharmacology & Therapeutics*, vol. 38, no. 3, pp. 274–283, 2013.

[69] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, *et al.*, "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease," *Nature Genetics*, vol. 40, no. 8, pp. 955–962, 2008.

[70] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[71] S. Saria and A. Goldenberg, "Subtyping: What it is and its role in precision medicine," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 70–75, 2015.

[72] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, "Integrative methods for analyzing big data in precision medicine," *Proteomics*, vol. 16, no. 5, pp. 741–758, 2016.

[73] D. Clifton, K. Niehaus, P. Charlton, and G. Colopy, "Health informatics via machine learning for the clinical management of patients," *Yearbook of Medical Informatics*, vol. 10, no. 1, p. 38, 2015.

[74] O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, S. C. Sanderson, J. Kannry, R. Zinberg, M. A. Basford, *et al.*, "The electronic medical records and genomics (eMERGE) network: Past, present, and future," *Genetics in Medicine*, vol. 15, no. 10, pp. 761–771, 2013.

[75] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of Biomedical Informatics*, vol. 48, pp. 148–159, 2014.

[76] eMERGE, "PheKB: Phenotype KnowledgeBase," 2012.

[77] R. Howard, M. Rattray, M. Prosperi, and A. Custovic, "Distinguishing asthma phenotypes using machine learning approaches," *Current Allergy and Asthma Reports*, vol. 15, no. 7, pp. 1–10, 2015.

[78] M. R. Boland, G. Hripcsak, Y. Shen, W. K. Chung, and C. Weng, "Defining a comprehensive verotype using electronic health records for personalized

medicine," *Journal of the American Medical Informatics Association*, vol. 20, no. e2, pp. e232–e238, 2013.

[79] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, *et al.*, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLOS Computational Biology*, vol. 7, no. 8, p. e1002141, 2011.

[80] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PLOS ONE*, vol. 8, no. 6, p. e66341, 2013.

[81] J. C. Ho, J. Ghosh, and J. Sun, "Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 115–124, ACM, 2014.

[82] D. C. Kale, Z. Che, M. T. Bahadori, W. Li, Y. Liu, and R. Wetzel, "Causal phenotype discovery via deep networks," in *AMIA Annual Symposium Proceedings*, vol. 2015, p. 677, American Medical Informatics Association, 2015.

[83] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun, "Rubik: Knowledge guided tensor factorization and completion for health data analytics," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1265–1274, ACM, 2015.

[84] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, "Learning probabilistic phenotypes from heterogeneous ehr data," *Journal of Biomedical Informatics*, vol. 58, pp. 156–165, 2015.

[85] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in the ICU with sparse, heterogeneous clinical data," in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2015, p. 446, NIH Public Access, 2015.

[86] R. Dürichen, M. A. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, "Multitask gaussian processes for multivariate physiological time-series analysis," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 314–322, 2015.

[87] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[88] W. Lee, Y. Lee, H. Kim, and I.-C. Moon, "Bayesian nonparametric collaborative topic poisson factorization for electronic health records-based phenotyping," in *Twenty-fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[89] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Discovering disease phenotypes from a large database of inpatient records: A sepsis study," *Methods*, 2016.

[90] Z. Liu and M. Hauskrecht, "Learning adaptive forecasting models from irregularly sampled multivariate clinical data," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[91] S. Lewis, T. Foltynie, A. Blackwell, T. Robbins, A. Owen, and R. Barker, "Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 3, pp. 343–348, 2005.

[92] D. P. Chen, S. C. Weber, P. S. Constantinou, T. A. Ferris, H. J. Lowe, and A. J. Butte, "Clinical arrays of laboratory measures, or "clinarrays", built from an electronic health record enable disease subtyping by severity," in *AMIA Annual Symposium Proceedings*, vol. 2007, p. 115, American Medical Informatics Association, 2007.

[93] S. Saria, D. Koller, and A. Penn, "Discovering shared and individual latent structure in multiple time series," *arXiv preprint arXiv:1008.2028*, 2010.

[94] M. H. Cho, G. R. Washko, T. J. Hoffmann, G. J. Criner, E. A. Hoffman, F. J. Martinez, N. Laird, J. J. Reilly, and E. K. Silverman, "Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation," *Respiratory Research*, vol. 11, no. 1, p. 1, 2010.

[95] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 389–398, ACM, 2012.

[96] J. Ross and J. Dy, "Nonparametric mixture of Gaussian processes with constraints," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1346–1354, 2013.

[97] M. A. Pimentel, D. A. Clifton, and L. Tarassenko, "Gaussian process clustering for the functional characterisation of vital-sign trajectories," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pp. 1–6, IEEE, 2013.

[98] D. C. Belgrave, R. Granell, A. Simpson, J. Guiver, C. Bishop, I. Buchan, A. J. Henderson, and A. Custovic, "Developmental profiles of eczema, wheeze, and rhinitis: Two population-based birth cohort studies," *PLOS Medicine*, vol. 11, no. 10, p. e1001748, 2014.

[99] J. C. Ross, P. J. Castaldi, M. H. Cho, and J. G. Dy, "Dual beta process priors for latent cluster discovery in chronic obstructive pulmonary disease," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–162, ACM, 2014.

[100] F. Doshi-Velez, Y. Ge, and I. Kohane, "Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis," *Pediatrics*, vol. 133, no. 1, pp. e54–e63, 2014.

[101] S. J. Shah, D. H. Katz, S. Selvaraj, M. A. Burke, C. W. Yancy, M. Gheorghiade, R. O. Bonow, C.-C. Huang, and R. C. Deo, "Phenomapping for novel classification of heart failure with preserved ejection fraction," *Circulation*, pp. CIRCULATIONAHA–114, 2014.

[102] P. J. Castaldi, J. Dy, J. Ross, Y. Chang, G. R. Washko, D. Curran-Everett, A. Williams, D. A. Lynch, B. J. Make, J. D. Crapo, *et al.*, "Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema," *Thorax*, pp. thoraxjnl–2013, 2014.

[103] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[104] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley, "Identification of type 2 diabetes subgroups through topological analysis of patient similarity," *Science Translational Medicine*, vol. 7, no. 311, pp. 311ra174–311ra174, 2015.

[105] P. Schulam and R. Arora, "Disease trajectory maps," *arXiv preprint arXiv:1606.09184*, 2016.

[106] R. Y. Coley, A. J. Fisher, M. Mamawala, H. B. Carter, K. J. Pienta, and S. L. Zeger, "Bayesian joint hierarchical model for prediction of latent health states with application to active surveillance of prostate cancer," *arXiv preprint arXiv:1508.07511*, 2015.

[107] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[108] Y. Yuan, R. S. Savage, and F. Markowetz, "Patient-specific data fusion defines prognostic cancer subtypes," *PLOS Computational Biology*, vol. 7, no. 10, p. e1002227, 2011.

[109] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, "Bayesian correlated clustering to integrate multiple datasets," *Bioinformatics*, vol. 28, no. 24, pp. 3290–3297, 2012.

[110] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types," *The Annals of Applied Statistics*, vol. 7, no. 1, p. 523, 2013.

[111] P. Ray, L. Zheng, J. Lucas, and L. Carin, "Bayesian joint analysis of heterogeneous genomics data," *Bioinformatics*, vol. 30, no. 10, pp. 1370–1376, 2014.

[112] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[113] G. Carlsson, "Topological pattern recognition for point cloud data," *Acta Numerica*, vol. 23, p. 289, 2014.

[114] Ayasdi, "TDA and machine learning: Better together," 2015.

[115] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[116] X. Rong, "word2vec parameter learning explained," *arXiv preprint arXiv:1411.2738*, 2014.

[117] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873–882, Association for Computational Linguistics, 2012.

[118] S. Saria, D. Koller, and A. Penn, "Learning individual and population level traits from clinical temporal data," in *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine workshop*, Citeseer, 2010.

[119] M. K. Titsias and N. D. Lawrence, "Bayesian gaussian process latent variable model.," in *AISTATS*, pp. 844–851, 2010.

[120] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 1, pp. 193–197, 2013.

[121] J. A. Quinn, C. K. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *Pattern Anal-*

ysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 9, pp. 1537–1551, 2009.

[122] J. Wiens, J. Guttag, and E. Horvitz, "Patient risk stratification for hospital-associated *C. Diff* as a time-series classification task," *Neural Information Processing Systems Foundation (NIPS)*, 2012.

[123] K. E. Niehaus, J. M. Banda, J. W. Knowles, and N. H. Shah, "FIND FH - a phenotype model to identify patients with familial hypercholesterolemia," in *AMIA 2015 Annual Symposium. 2nd Workshop on Data Mining for Medical Informatics: Predictive Analytics*, AMIA, 2015.

[124] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, "Limestone: High-throughput candidate phenotype generation via tensor factorization," *Journal of Biomedical Informatics*, vol. 52, pp. 199–211, 2014.

[125] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[126] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[127] M. J. Van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[128] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[129] S. Zhao, C. Gao, S. Mukherjee, and B. E. Engelhardt, "Bayesian group latent factor analysis with structured sparse priors," *arXiv preprint arXiv:1411.2698*, 2014.

[130] J. C. Denny, "Mining electronic health records in the genomics era," *PLOS Computational Biology*, vol. 8, no. 12, p. e1002823, 2012.

[131] A. E. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, 2016.

[132] M.-A. Meuwis, M. Fillet, P. Geurts, D. De Seny, L. Lutteri, J.-P. Chapelle, V. Bours, L. Wehenkel, J. Belaiche, M. Malaise, *et al.*, "Biomarker discovery

for inflammatory bowel disease, using proteomic serum profiling," *Biochemical Pharmacology*, vol. 73, no. 9, pp. 1422–1433, 2007.

[133] C.-S. Chen, S. Sullivan, T. Anderson, A. C. Tan, P. J. Alex, S. R. Brant, C. Cuffari, T. M. Bayless, M. V. Talor, C. L. Burek, *et al.*, "Identification of novel serological biomarkers for inflammatory bowel disease using *Escherichia coli* proteome chip," *Molecular & Cellular Proteomics*, vol. 8, no. 8, pp. 1765–1776, 2009.

[134] E. Papa, M. Docktor, C. Smillie, S. Weber, S. P. Preheim, D. Gevers, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram, *et al.*, "Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease," *PLOS ONE*, vol. 7, no. 6, p. e39242, 2012.

[135] R. Duttagupta, S. DiRienzo, R. Jiang, J. Bowers, J. Gollub, J. Kao, K. Kearney, D. Rudolph, N. B. Dawany, M. K. Showe, *et al.*, "Genome-wide maps of circulating miRNA biomarkers for ulcerative colitis," *PLOS ONE*, vol. 7, no. 2, p. e31241, 2012.

[136] S. Plevy, M. S. Silverberg, S. Lockton, T. Stockfisch, L. Croner, J. Stachelski, M. Brown, C. Triggs, E. Chuang, F. Princen, *et al.*, "Combined serological, genetic, and inflammatory markers differentiate non-IBD, Crohn's disease, and ulcerative colitis patients," *Inflammatory Bowel Diseases*, vol. 19, no. 6, p. 1139, 2013.

[137] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. M. Visscher, *et al.*, "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," *The American Journal of Human Genetics*, vol. 92, no. 6, pp. 1008–1012, 2013.

[138] D. Mahapatra, P. Schueffler, J. A. Tielbeek, J. M. Buhmann, and F. M. Vos, "A supervised learning based approach to detect Crohn's disease in abdominal MR volumes," in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, pp. 97–106, Springer, 2012.

[139] P. J. Schüffler, D. Mahapatra, J. A. Tielbeek, F. M. Vos, J. Makanyanga, D. A. Pendsé, C. Y. Nio, J. Stoker, S. A. Taylor, and J. M. Buhmann, "A model development pipeline for Crohn's disease severity assessment from magnetic resonance images," in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, pp. 1–10, Springer, 2013.

[140] F. M. Vos, J. A. Tielbeek, R. E. Naziroglu, Z. Li, P. Schueffler, D. Mahapatra, A. Wiebel, C. Lavini, J. M. Buhmann, H.-C. Hege, *et al.*, "Computational

modeling for assessment of IBD: To be or not to be?," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3974–3977, IEEE, 2012.

[141] A. K. Waljee, J. C. Joyce, S. Wang, A. Saxena, M. Hart, J. Zhu, and P. D. Higgins, "Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines," *Clinical Gastroenterology and Hepatology*, vol. 8, no. 2, pp. 143–150, 2010.

[142] C. Gasche, J. Scholmerich, J. Brynskov, G. D'Haens, S. B. Hanauer, E. J. Irvine, D. P. Jewell, D. Rachmilewitz, D. B. Sachar, W. J. Sandborn, *et al.*, "A simple classification of Crohn's disease: Report of the Working Party for the World Congresses of Gastroenterology, Vienna 1998," *Inflammatory Bowel Diseases*, vol. 6, no. 1, pp. 8–15, 2000.

[143] M. S. Silverberg, J. Satsangi, T. Ahmad, I. D. Arnott, C. N. Bernstein, S. R. Brant, R. Caprilli, J.-F. Colombel, C. Gasche, K. Geboes, *et al.*, "Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology," *Canadian Journal of Gastroenterology and Hepatology*, vol. 19, no. Suppl A, pp. 5A–36A, 2005.

[144] L. Jostins, A. P. Levine, and J. C. Barrett, "Using genetic prediction from known complex disease loci to guide the design of next-generation sequencing experiments," *PLOS ONE*, vol. 8, no. 10, p. e76328, 2013.

[145] A. N. Ananthakrishnan, H. Huang, D. D. Nguyen, J. Sauk, V. Yajnik, and R. J. Xavier, "Differential effect of genetic burden on disease phenotypes in Crohn's disease and ulcerative colitis: Analysis of a North American cohort," *The American Journal of Gastroenterology*, vol. 109, no. 3, pp. 395–400, 2014.

[146] "UK Biobank ethics and governance framework," 2007.

[147] A. Latiano, O. Palmieri, L. Pastorelli, M. Vecchi, T. T. Pizarro, F. Bossa, G. Merla, B. Augello, T. Latiano, G. Corritore, *et al.*, "Associations between genetic polymorphisms in IL-33, IL1R1 and risk for inflammatory bowel disease," *PLOS ONE*, vol. 8, no. 4, p. e62144, 2013.

[148] C. N. Bernstein, A. Wajda, and J. F. Blanchard, "The clustering of other chronic inflammatory diseases in inflammatory bowel disease: A population-based study," *Gastroenterology*, vol. 129, no. 3, pp. 827–836, 2005.

[149] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2- $\delta\delta$CT method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

[150] G. Rogler and S. Vavricka, "Exposome in IBD: Recent insights in environmental factors that influence the onset and course of IBD," *Inflammatory Bowel Diseases*, vol. 21, no. 2, pp. 400–408, 2015.

[151] K. E. Niehaus, H. H. Uhlig, and D. A. Clifton, "Phenotypic characterisation of Crohn's disease severity," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 7023–7026, IEEE, 2015.

[152] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An Introduction to Statistical Modeling of Extreme Values*, vol. 208. Springer, 2001.

[153] K. E. Niehaus, S. Luca, and D. A. Clifton, "Point processes for identifying anomalous behaviour in clinical time series data," in *Under review*, 2016.

[154] S. Luca, M. A. Pimentel, P. J. Watkinson, and D. Clifton, "Point process models for novelty detection on spatial point patterns," *(Under review)*, 2016.

[155] B. A. Turlach *et al.*, *Bandwidth selection in kernel density estimation: A review*. Université Catholique de Louvain, 1993.

[156] A. Putto, O. Meurman, and O. Ruuskanen, "C-reactive protein in viral and bacterial infections," *Pediatric Research*, vol. 19, no. 10, pp. 1103–1103, 1985.

[157] J. P. Haran, F. L. Beaudoin, S. Suner, and S. Lu, "C-reactive protein as predictor of bacterial infection among patients with an influenza-like illness," *The American Journal of Emergency Medicine*, vol. 31, no. 1, pp. 137–144, 2013.

[158] H. He and E. A. Garcia, "Learning from imbalanced data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009.

[159] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[160] T. Warren Liao, "Clustering of time series data: A survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[161] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[162] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[163] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[164] B. E. Engelhardt and M. Stephens, "Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis," *PLOS Genetics*, vol. 6, no. 9, p. e1001117, 2010.

[165] E. Wit, E. v. d. Heuvel, and J.-W. Romeijn, "'all models are wrong...': an introduction to model uncertainty," *Statistica Neerlandica*, vol. 66, no. 3, pp. 217–236, 2012.

[166] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[167] H. Kamper, "Gibbs sampling for fitting finite and infinite gaussian mixture models," 2013.

[168] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," *def*, vol. 1, no. $2\sigma 2$, p. 16, 2007.

[169] A. Jasra, C. Holmes, and D. Stephens, "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, pp. 50–67, 2005.

[170] M. Stephens, "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 4, pp. 795–809, 2000.

[171] A. J. Cron and M. West, "Efficient classification-based relabeling in mixture models," *The American Statistician*, vol. 65, no. 1, pp. 16–20, 2011.

[172] A. Nobile and A. T. Fearnside, "Bayesian finite mixtures with an unknown number of components: The allocation sampler," *Statistics and Computing*, vol. 17, no. 2, pp. 147–162, 2007.

[173] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[174] B. Dai, S. Ding, G. Wahba, *et al.*, "Multivariate Bernoulli distribution," *Bernoulli*, vol. 19, no. 4, pp. 1465–1483, 2013.

[175] M. Sperrin, T. Jaki, and E. Wit, "Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models," *Statistics and Computing*, vol. 20, no. 3, pp. 357–366, 2010.

[176] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[177] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[178] S. E. Flanagan, A.-M. Patch, and S. Ellard, "Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations," *Genetic Testing and Molecular Biomarkers*, vol. 14, no. 4, pp. 533–537, 2010.

[179] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[180] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, 2006.

[181] J. Cosnes, C. Gower-Rousseau, P. Seksik, and A. Cortot, "Epidemiology and natural history of inflammatory bowel diseases," *Gastroenterology*, vol. 140, no. 6, pp. 1785–1794, 2011.

[182] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, pp. 1165–1188, 2001.

[183] D. of Statistics, "Glossary of statistical terms," 2015.

[184] S. A. Naser, M. Arce, A. Khaja, M. Fernandez, N. Naser, S. Elwasila, and S. Thanigachalam, "Role of ATG16L, NOD2 and IL23R in Crohn's disease pathogenesis," *World Journal of Gastroenterology*, vol. 18, no. 5, pp. 412–424, 2012.

[185] ImmunoBase, "Immunobase," 2016.

[186] E. B. I. European Molecular Biology Laboratory, "dbsnp," 2016.

[187] N. C. for Biotechnology Information, "dbsnp," 2016.

[188] A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, *et al.*, "Ensembl 2016," *Nucleic Acids Research*, vol. 44, no. D1, pp. D710–D716, 2016.

[189] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[190] A. F. Di Narzo, L. A. Peters, C. Argmann, A. Stojmirovic, J. Perrigoue, K. Li, S. Telesco, B. Kidd, J. Walker, J. Dudley, *et al.*, "Blood and intestine eQTLs from an anti-TNF-resistant Crohn's disease cohort inform IBD genetic association loci," *Clinical and Translational Gastroenterology*, vol. 7, no. 6, p. e177, 2016.

[191] S. Hu, S. Liang, H. Guo, D. Zhang, H. Li, X. Wang, W. Yang, W. Qian, S. Hou, H. Wang, *et al.*, "Comparison of the inhibition mechanisms of adalimumab and infliximab in treating tumor necrosis factor α-associated diseases from a molecular view," *Journal of Biological Chemistry*, vol. 288, no. 38, pp. 27059–27067, 2013.

[192] L. C. Silva, L. C. Ortigosa, and G. Benard, "Anti-TNF-α agents in the treatment of immune-mediated inflammatory diseases: Mechanisms of action and pitfalls," *Immunotherapy*, vol. 2, no. 6, pp. 817–833, 2010.

[193] A. Dignass, G. Van Assche, J. Lindsay, M. Lémann, J. Söderholm, J. Colombel, S. Danese, A. D'Hoore, M. Gassull, F. Gomollón, *et al.*, "The second European evidence-based consensus on the diagnosis and management of Crohn's disease: Current management," *Journal of Crohn's and Colitis*, vol. 4, no. 1, pp. 28–62, 2010.

[194] C. Kestens, M. G. van Oijen, C. L. Mulder, A. A. van Bodegraven, G. Dijkstra, D. de Jong, C. Ponsioen, B. A. van Tuyl, P. D. Siersema, H. H. Fidder, *et al.*, "Adalimumab and infliximab are equally effective for Crohn's disease in patients not previously treated with anti–tumor necrosis factor-α agents," *Clinical Gastroenterology and Hepatology*, vol. 11, no. 7, pp. 826–831, 2013.

[195] E. Medicines Agency, "European public assessment reports," 2016.

[196] S. R. Targan, S. B. Hanauer, S. J. van Deventer, L. Mayer, D. H. Present, T. Braakman, K. L. DeWoody, T. F. Schaible, and P. J. Rutgeerts, "A short-term study of chimeric monoclonal antibody cA2 to tumor necrosis factor α for Crohn's disease," *New England Journal of Medicine*, vol. 337, no. 15, pp. 1029–1036, 1997.

[197] S. B. Hanauer, B. G. Feagan, G. R. Lichtenstein, L. F. Mayer, S. Schreiber, J. F. Colombel, D. Rachmilewitz, D. C. Wolf, A. Olson, W. Bao, *et al.*, "Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial," *The Lancet*, vol. 359, no. 9317, pp. 1541–1549, 2002.

[198] S. B. Hanauer, W. J. Sandborn, P. Rutgeerts, R. N. Fedorak, M. Lukas, D. MacIntosh, R. Panaccione, D. Wolf, and P. Pollack, "Human anti–tumor necrosis factor monoclonal antibody (adalimumab) in Crohn's disease: the CLASSIC-i trial," *Gastroenterology*, vol. 130, no. 2, pp. 323–333, 2006.

[199] J.-F. Colombel, W. J. Sandborn, P. Rutgeerts, R. Enns, S. B. Hanauer, R. Panaccione, S. Schreiber, D. Byczkowski, J. Li, J. D. Kent, *et al.*, "Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the CHARM trial," *Gastroenterology*, vol. 132, no. 1, pp. 52–65, 2007.

[200] N. Ding, A. Hart, and P. De Cruz, "Systematic review: Predicting and optimising response to anti-tnf therapy in Crohn's disease–algorithm for practical management," *Alimentary Pharmacology & Therapeutics*, vol. 43, no. 1, pp. 30–51, 2016.

[201] A. C. Ford, W. J. Sandborn, K. J. Khan, S. B. Hanauer, N. J. Talley, and P. Moayyedi, "Efficacy of biological therapies in inflammatory bowel disease: Systematic review and meta-analysis," *The American Journal of Gastroenterology*, vol. 106, no. 4, pp. 644–659, 2011.

[202] G. Van Assche, S. Vermeire, V. Ballet, F. Gabriels, M. Noman, G. D'Haens, C. Claessens, E. Humblet, N. V. Casteele, A. Gils, *et al.*, "Switch to adalimumab in patients with Crohn's disease controlled by maintenance infliximab: Prospective randomised SWITCH trial," *Gut*, pp. gutjnl–2011, 2011.

[203] F. Zorzi, S. Zuzzi, S. Onali, E. Calabrese, G. Condino, C. Petruzziello, M. Ascolani, F. Pallone, and L. Biancone, "Efficacy and safety of infliximab and adalimumab in Crohn's disease: A single centre study," *Alimentary Pharmacology & Therapeutics*, vol. 35, no. 12, pp. 1397–1407, 2012.

[204] A. Tursi, W. Elisei, M. Picchio, and A. Penna, "Letter: Are infliximab and adalimumab similar for Crohn's disease in clinical practice?," *Alimentary Pharmacology & Therapeutics*, vol. 37, no. 7, pp. 763–764, 2013.

[205] S. Ben-Horin and Y. Chowers, "Review article: Loss of response to anti-TNF treatments in Crohn's disease," *Alimentary Pharmacology & Therapeutics*, vol. 33, no. 9, pp. 987–995, 2011.

[206] R. Prieto-Perez, T. Cabaleiro, E. Dauden, and F. Abad-Santos, "Gene polymorphisms that can predict response to anti-TNF therapy in patients with psoriasis and related autoimmune diseases," *The Pharmacogenomics Journal*, vol. 13, no. 4, pp. 297–305, 2013.

[207] T. Billiet, M. de Bruyn, K. Claes, V. Ballet, X. Liu, R. Kirkland, K. Drake, S. Lockton, F. Princen, S. Singh, *et al.*, "A matrix-based prediction model for primary response to infliximab in Crohn's disease patients," *Journal of Crohn's and Colitis*, vol. 8, no. Supplement 1, pp. S303–S304, 2014.

[208] S. Vermeire, G. Van Assche, and P. Rutgeerts, "C-reactive protein as a marker for inflammatory bowel disease," *Inflammatory Bowel Diseases*, vol. 10, no. 5, pp. 661–665, 2004.

[209] K. Bodger, T. Kikuchi, and D. Hughes, "Cost-effectiveness of biological therapy for Crohn's disease: Markov cohort analyses incorporating United Kingdom patient-level cost data," *Alimentary Pharmacology & Therapeutics*, vol. 30, no. 3, pp. 265–274, 2009.

[210] H. G. Evans, U. Roostalu, G. J. Walter, N. J. Gullick, K. S. Frederiksen, C. A. Roberts, J. Sumner, D. L. Baeten, J. G. Gerwien, A. P. Cope, *et al.*, "TNF-α blockade induces IL-10 expression in human CD4+ T cells," *Nature Communications*, vol. 5, 2014.

[211] K. K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, *et al.*, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature*, vol. 518, no. 7539, pp. 337–343, 2015.

[212] A. G. Hawkes and D. Oakes, "A cluster process representation of a self-exciting process," *Journal of Applied Probability*, pp. 493–503, 1974.

[213] P. J. Laub, T. Taimre, and P. K. Pollett, "Hawkes processes," *arXiv preprint arXiv:1507.02822*, 2015.

[214] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes.," in *AISTATS*, vol. 13, pp. 641–649, 2013.

[215] P. J. Green and S. Richardson, "Modelling heterogeneity with and without the dirichlet process," *Scandinavian Journal of Statistics*, vol. 28, no. 2, pp. 355–375, 2001.

[216] J. Hensman, N. D. Lawrence, and M. Rattray, "Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–12, 2013.

[217] J. Hensman, M. Rattray, and N. D. Lawrence, "Fast nonparametric clustering of structured time-series," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 383–393, 2015.

[218] M. Cesarini, G. Collins, A. Ronnblom, A. Santos, D. Sjoberg, M. Parkes, S. Keshav, and S. Travis, "Predicting the risk of acute severe colitis (ASC) at diagnosis of ulcerative colitis (UC): External validation," *Journal of Crohn's and Colitis*, vol. 9, no. 1, pp. S77–S443, 2015.

[219] Crohn's and C. F. of America, "CCFA launches development of world's largest research database for Crohn's disease and ulcerative colitis," 2015.

[220] C. N. BioResource, "New hope for people who suffer with IBD with the launch of a major new research resource," 2016.

[221] N. I. for Health Research, "IBD Resource," 2016.

[222] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nature Protocols*, vol. 5, no. 9, pp. 1564–1573, 2010.

[223] S. Turner, L. L. Armstrong, Y. Bradford, C. S. Carlson, D. C. Crawford, A. T. Crenshaw, M. Andrade, K. F. Doheny, J. L. Haines, G. Hayes, *et al.*, "Quality control procedures for genome-wide association studies," *Current protocols in Human Genetics*, pp. 1–19, 2011.

[224] M. E. Rentería, A. Cortes, and S. E. Medland, "Using PLINK for genome-wide association studies (GWAS) and data analysis," *Genome-Wide Association Studies and Genomic Prediction*, pp. 193–213, 2013.

[225] T. Shah, J. Liu, J. Floyd, J. A. Morris, N. Wirth, J. C. Barrett, and C. Anderson, "opticall: a robust genotype-calling algorithm for rare, low-frequency and common variants," *Bioinformatics*, vol. 28, no. 12, pp. 1598–1603, 2012.

[226] Y. Y. Teo, M. Inouye, K. S. Small, R. Gwilliam, P. Deloukas, D. P. Kwiatkowski, and T. G. Clark, "A genotype calling algorithm for the Illumina BeadArray platform," *Bioinformatics*, vol. 23, no. 20, pp. 2741–2746, 2007.

[227] C. E. Rasmussen and C. K. Williams, "Gaussian processes for machine learning. 2006," *The MIT Press, Cambridge, MA, USA*, vol. 38, pp. 715–719, 2006.

[228] S. Flaxman, A. Wilson, D. Neill, H. Nickisch, and A. Smola, "Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 607–616, 2015.

[229] C. Lloyd, T. Gunter, M. A. Osborne, and S. J. Roberts, "Variational inference for Gaussian process modulated Poisson processes," *International Conference on Machine Learning*, 2015.

[230] R. P. Adams, I. Murray, and D. J. MacKay, "Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 9–16, ACM, 2009.

Appendix A

Additional background material

In this Appendix we expand upon background material and methodological points that were previously mentioned. This material is organised by Chapter.

A.1 Additional background from Chapter 1

The molecular biology techniques that were used in the stimulation experiments presented in this thesis were RT-qPCR and ELISA. Additional details regarding these methods are presented below.

A.1.1 Reverse-transcriptase quantitative polymerase chain reaction description

RT-qPCR involves the reverse transcription of the mRNA into DNA so that it can be amplified using polymerase chain reaction (PCR). Throughout PCR, fluorescent reporters are released as the DNA is amplified, allowing the progression of the experiment to be measured "in real time." This is essential for mRNA measurement because when qPCR reactions reach completion, the result is nearly equivalent quantitites of the original inputs (due to the nature of the reaction). This means that any difference in the original concentrations of distinct mRNA will be lost if the concentrations are only measured at the end of the experiment. Instead, the measurement of the DNA concentrations throughout the course of the experiment allows for the distinction between concentrations to be elucidated, before the plateau phase is reached. To illustrate this, Figure A.1 shows the course of hypothetical real-time fluorescence measurements. There is an initial exponential phase of amplification, during which the levels of fluorescence are too small to be measurable. After reaching a threshold

point at which fluorescence is measurable, the increase is generally linear until reaching a plateau phase [27]. The expression levels during various treatments must be compared to a "housekeeping" gene's expression to allow for normalisation.

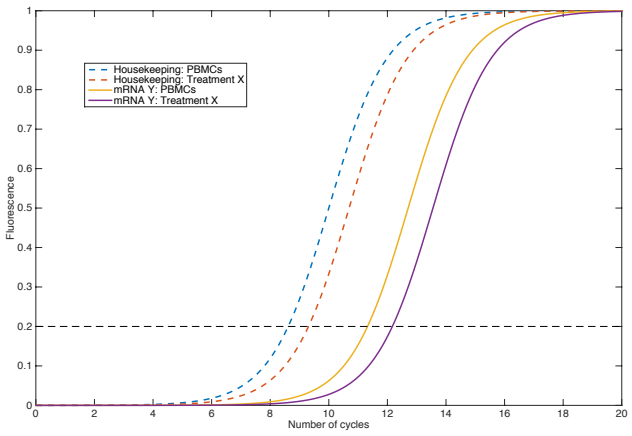


Figure A.1: *Example qPCR curves, with threshold shown as a horizontal dashed line.*

A.1.2 Enzyme-linked immunosorbent assay description

ELISA is a type of immunoassay technique in which antibodies to the proteins of interest have been bound to a solid substrate. In these "sandwich immunoassays" to detect cytokines, a liquid solution with the protein of interest, B, is then introduced, causing the desired proteins to bind to the antibodies, A, on the substrate surface. Another antibody, C, specific to the desired protein and also linked to an enzyme, D, is then added. The substrate, E, for the enzyme D is then added, which produces a color change or otherwise detectable signal, so that the presence of the protein can be measured [28]. This process is illustrated in Figure A.2. Of note, the accuracy of ELISA measurements depends on the availability of specific antibodies to the protein of interest that do not cross-react with similar species.

A.2 Additional background from Chapter 2

We presented a summary table of those studies covered in our literature review in Chapter 2. We provide additional information about these previous studies in Table A.1.

Table A.1: *Additional information regarding studies presented in the literature review*

Authors	Ref.	Year	∼N	Disease	Data	Sup./unsup.	Time-series pre-processing
Phenotyping and Time-series studies							
Roque et al.	79	2011	4,800	Pyschiatric	Diagnosis codes	Unsupervised	n/a
Lasko et al.	80	2013	4400	Leukemia/Gout	Serum uric acid	Unsupervised	Warped time series
Ho et al.	81	2014	10,000	Many	Diagnosis and procedure codes	Unsupervised	n/a
Kale et al.	82	2015	8500	Many	13 physiologic time series	Supervised	RTHR; impute missing
Wang et al.	83	2015	8000; 470,000	Many	Diagnoses, medications	Unsupervised	n/a
Pivovarov et al.	84	2015	10,000	Many	Free text, meds, diag. codes, labs	Unsupervised	n/a
Ghassemi et al.	85	2015	10,000	n/a	Clinician notes	Supervised	Topic modelling over notes
Durichen et al.	86	2015	Not specified	n/a	Vital signs	Regression	n/a
Lipton et al.	87	2015	10,000	Many	13 physiologic time series	Supervised	RTHR; forward/back-filling
Lee et al	88	2016	150,000	Many	Diagnoses, medications	Unsup. + sup.	n/a
Gligorijevic et al.	89	2016	1,100,00	Sepsis	Diagnosis codes	Unsup. + sup.	n/a
Liu et al.	90	2016	500	n/a	Complete blood counts	Regression	Linear interpolation
Subtyping studies							
Lewis et al.	91	2005	120	Parkinson's disease	PD-specific clinical features	Unsupervised	n/a
Chen et al.	92	2007	140-350	Chronic disease	Lab measurements	Unsupervised	Averaged across all times
Saria et al.	93	2010	140	Neonatal mortality	Heart rate signals	Unsupervised	n/a
Cho et al.	94	2010	300	COPD	COPD clinical variables	Unsupervised	n/a
Marlin et al.	95	2012	10,600	Mortality	13 physiologic time series	Unsup. + sup.	Discretise into hour-long intervals
Ross et al.	96	2013	40	COPD	Lung function clinical variable	Unsupervised	n/a
Pimentel et al.	97	2013	100	n/a	Vital signs	Unsupervised	n/a
Belgrave et al	98	2014	9,800	Asthma	Asthma-specific clinical variables	Unsupervised	n/a
Ross et al.	99	2014	850	COPD	7 Lung function clinical variables	Unsupervised	n/a
Doshi-Velez et al.	100	2014	5000	Autism	45 diagnosis codes	Unsupervised	Windowed code counts
Shah, et al.	101	2014	420	Heart failure	Cardiovascular features	Unsupervised	Features extracted
Castaldi et al.	102	2014	10,000	COPD	COPD clinical variables	Unsupervised	n/a
Schulam et al.	103	2015	Not specified	Sclerdoma	Sclerdoma-specific clinical markers	Unsupervised	n/a
Li et al.	104	2015	2,500	Type 2 diabetes	73 labs, diag. codes, meds., vital signs	Unsupervised	Static approach
Schulam et al.	105	2016	2,000	Sclerdoma	Sclerdoma-specific clinical markers	Unsupervised	n/a
Coley et al.	106	2016	870	Prostate cancer	PSA test results, biopsy results		
Omics studies							
Shen et al.	107	2009	40; 90	Cancers	CNV and gene expression	Unsupervised	n/a
Yuan et al.	108	2011	110; 150	Cancers	CNV and gene expression	Unsupervised	n/a
Kirk et al.	109	2012	550 genes	Yeast experiments	ChIP data, PPI data, gene expression	Unsupervised	n/a
Lock et al.	110	2013	230	Gliobastoma	miRNA and gene expression	Unsupervised	n/a
Ray et al.	111	2014	70	Ovarian cancer	Gene expression, CNV, methylation	Unsupervised	n/a
Wang et al.	112	2014	210	Gliobastoma	ChIP, mRNA & miRNA expression	Unsupervised	n/a

Abbreviations: CHiP: chromatin immunoprecipitation, CNV: copy number variation, COPD: chronic obstructive disease, diag: diagnosis, med: medications, PD: Parkinson's disease, PSA: prostate-specific antigen, RTHR: resample to hourly rate, Sup.: supervised, Unsup.: unsupervised.

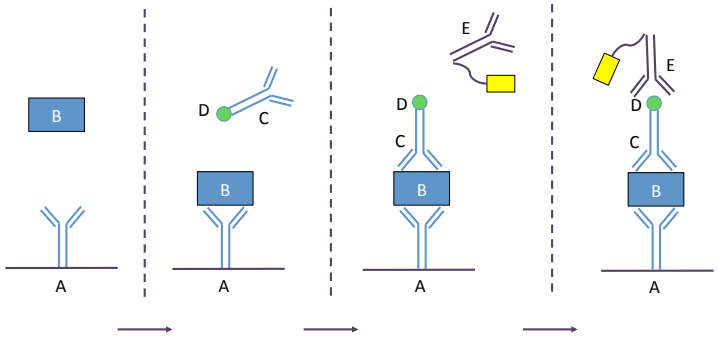


Figure A.2: *The steps in a sandwich immunoassay. A=antibody bound to substrate surface, B=protein of interest, C=second antibody, D=enzyme, E=substrate for enzyme, with fluorescence or other detectable property.*

A.3 Additional background from Chapter 3

A.3.1 ImmunoChip data processing

All of the processing steps described below were performed by Luke Jostins at the Wellcome Trust Center for Human Genetics, University of Oxford. As background, it is helpful to define several terms [222]. [223] and [224] provide overviews of the standard procedures required in large-scale GWAS.

- **Call rate**: proportion of SNPs that are non-missing. SNPs are "called" as either homozygous for the minor allele, heterozygous, or homozygous for the major allele. Whether a SNP is called depends upon the probe intensity on the physical microarray chip. A threshold for the required probe intensity must be set, which will result in a higher or lower call rate.

- **Haplotype map (HapMap)**: an international project to establish the baseline population structure of the genome. Based upon sets of alleles from several hundred individuals of African, Asian, and European ancestry.

- **Hardy-Weinberg equilibrium (HWE)**: occurs when two alleles from an individual are not statistically associated across a population. This is usually assessed through a chi-squared test.

- **Linkage disequilibrium (LD)**: regions of the genome that are statistically associated; i.e., rates of variants are not independent.

- **Minor allele frequency (MAF)**: frequency of the minor allele in the population

Gene-calling was performed using either Illuminus or optiCall [225, 226]. These methods use mixture models to estimate the probe intensities that define the various genotype possibilities at each loci. Based upon these algorithms, SNPs with a MAF of $< 1\%$, a call rate $< 90\%$, or HWE of $p < 1 \times 10^{-5}$ were removed. This resulted in the removal of 64,344 SNPs. A second round of quality control was then conducted, in which individuals with a call rate of less than 95% or low rates of heterozygosity (which is indicative of genotyping errors or inbreeding) were removed, resulting in the removal of 11 individuals. SNPs with a call rate of less than 95% were also removed, resulting in the additional removal of 7000 SNPs.

Duplicated and related samples were identified by comparing the genome similarity between samples, which resulted in the removal of 13 individuals. Non-European individuals were identified by first removing SNPs from high LD regions of the genome. The remaining SNPs were projected onto PCA components generated from HapMap individuals. Patients that did not overlap with the European clusters were removed from analysis; this resulted in the removal of 120 patients.

A.3.2 Exome sequencing data processing

All of the exome data pre-processing was performed by Eva Goncalves Serra at the Wellcome Trust Sanger Institute using the Genome Analysis Toolkit (GATK) best practices. Briefly, this protocol first maps the raw reads to a reference genome (using the Burrows-Wheeler Aligner toolbox); marks duplicate reads, as exact duplicates do not provide additional evidence for a particular variant (using the Picard toolbox); and realigns the sequence to take into account indels. Variants are then annotated with quality scores, where the scores are based on mixture models to assign each variant with a probability of being a "true" variant or not, in a two-step process. For our analysis, exome variants were additionally filtered for those that had a read depth of at least eight and genotype quality of at least 20. The functional consequences of the variants were predicted using Ensembl v75 Variant Effect Predictor, which uses SIFT and Polyphen, as described previously in Chapter 3.

A.4 Additional background from Chapter 7

We used a LGCP to model the changing rate of relapse and surgery over time. As mentioned, a LGCP is a model for estimating the rate, $\lambda(\mathbf{x})$, of a point process. The model assumes that $\lambda(\mathbf{x})$ varies smoothly over time $\mathbf{x}$, thus this rate is modelled as a GP. A GP is an infinite-dimensional distribution in which any finite collection of values forms a multivariate Gaussian [227]. We consider a univariate times series with

time points $\mathbf{x}$ and values $\mathbf{y}$, which are modelled by a latent GP $f(\mathbf{x}) \sim GP(m, k)$ with mean m and covariance k (also termed the kernel). In such a model we are interested in inferring this latent function f,

$$p(f|\mathbf{y}, \mathbf{x}) = \frac{p(\mathbf{y}|f, \mathbf{x})\ p(f|\mathbf{x})}{p(\mathbf{y}|\mathbf{x})} = \frac{p(\mathbf{y}|f, \mathbf{x})\ p(f|\mathbf{x})}{\int p(\mathbf{y}|f, \mathbf{x})\ p(f|\mathbf{x})\ \mathrm{d}f} \tag{A.1}$$

Because "any finite collection of values forms a multivarite Gaussian," $p(f|\mathbf{x}) \sim N(\mu, K)$, where $\mu_i = m(x_i)$ and $K_{ij} = k(x_i, x_j)$. In standard GP regression, the likelihood also takes on the functional form of a Gaussian, for instance as $p(\mathbf{y}|f) \sim N(f, \sigma_n^2 I)$ with some noise σ_n.

In the case of a LGCP, however, this likelihood is modelled according to the Poisson distribution: $p(\mathbf{y}|f) \sim \text{Poisson}(\exp(f(\mathbf{x})))$, where the exponential is taken to ensure a positive rate. Taking the exponential means that the logarithm of the rate function $\lambda(\mathbf{x})$ is modelled as a GP, thus motivating the name log Gaussian Cox process (a Cox process is an inhomogeneous Poisson process with a stochastic intensity). Other choices for enforcing a positive rate function (e.g. squaring) are also possible. Because instances of the GP form a multivariate normal, parameterised by covariance K, the relationship between nearby points is encoded by the choice of the kernel k. Choices of k such as the squared exponential encode a smoothness constraint on nearby points, which is exactly the behaviour we wish to model.

In our analysis, we used the method proposed and implemented by Flaxman et al. for inference in our LGCP model [228]. This method uses the Laplace approximation and Kronecker methods to estimate the latent rate function. Variational Bayes and MCMC approaches are also possible [229], [230].

Appendix B

Additional results

Supporting results from throughout the thesis are presented in this Appendix, organised by chapter.

B.1 Additional results from Chapter 4

There are several special cases that occur at particular values of ξ in the derivation of our EVT PP model, which we present below.

Case 1: $\xi < -1$

When $\xi < -1$, our GPD likelihood remains the same, but our inequality in the derivation of the region of integration is reversed because we are now dividing through by a positive (rather than a negative) value:

$$\sum_{i=1}^{k} \frac{1}{\xi} \log\Big(1 + \xi\big(\frac{x-\theta}{\sigma}\big)\Big) \;\leq\; \frac{-1}{1+\xi} \log\Big(\frac{\sigma_p^k v}{k! p_k}\Big) \tag{B.1}$$

which also means that our requirement for v shifts to $\frac{k! p_k}{\sigma_p^k} \leq v < \infty$. Again writing the left side as $\frac{u_i^2}{2} = \frac{1}{\xi}\log\Big(1 + \xi\big(\frac{x-\theta}{\sigma}\big)\Big)$ and the right-side as $c_k(v,\xi)^2 = \frac{-2}{1+\xi}\log\Big(\frac{\sigma_p^k v}{k! p_k}\Big)$, we can say:

$$X_v = \bigcup_{k \geq 1} \big\{ \mathbf{x}^e \mid \sum_{i=1}^{k} u_i^2 \leq c_k(v,\xi)^2 \big\} \tag{B.2}$$

As a result, this means that our region of integration is reversed, through until when we divide again by a positive number instead of a negative one. This returns the range of integration back to match exactly with the situation we have in the case of $\xi > 0$, $-1 < \xi < 0$, giving us a final density as in Equation 4.15.

Case 2: $\xi = 0$

For the case of $\xi = 0$, our GPD likelihood is now $p_k = \frac{1}{\sigma_p} e^{-}(\frac{x-\theta}{\sigma_p})$, meaning that we can write:

$$\sum_{i=1}^{k} \left(\frac{x-\theta}{\sigma}\right) \geq -\log\left(\frac{\sigma_p^k v}{k! p_k}\right) \tag{B.3}$$

with $0 \leq v \leq \frac{k! p_k}{\sigma_p^k}$. In this case, we define $\frac{u_i^2}{2} = \left(\frac{x-\theta}{\sigma}\right)$ and $c_k(v,\xi)^2 = -2\log\left(\frac{\sigma_p^k v}{k! p_k}\right)$. And again, $X_v = \bigcup_{k\geq 1} \left\{\mathbf{x}^e \mid \sum_{i=1}^{k} u_i^2 \geq c_k(v,\xi)^2\right\}$.

Our value for c_k, which defines the region of integration, is thus identical to the case of $\xi > 0$, except that the actual value of $\xi = 0$. Our u-substitution is now $\frac{u_i^2}{2} = \left(\frac{x-\theta}{\sigma}\right)$, meaning that $u_i du_i = \frac{1}{\sigma_p} dx_i$. Plugging this in, we again obtain Eq. 4.12. We can then proceed with the remainder of the proof, as before, meaning that no special alterations are needed in the analytical formula for the case of $\xi = 0$.

Case 3: $\xi = -1$

When $\xi = -1$, our GPD likelihoods each reduce to $\frac{1}{\sigma_p}$, meaning that our region of integration is:

$$\frac{k! p_k}{\sigma_p^k} \leq v \tag{B.4}$$

Going back to our equation for $\overline{G}(v)$ with the GPD likelihhoods for $\xi = -1$ entered, we have

$$\overline{G}(v) = \sum_{k\geq 1}^{\infty} p_k \int_{A_k(v,\xi)} \frac{1}{\sigma_p^k} d\mathbf{x}^e \tag{B.5}$$

Using our region $A_k(v,\xi)$ as defined above, we can write:

$$\overline{G}(v) = \sum_{k\geq 1}^{\infty} p_k \int_{0}^{v\sigma_p^k - k! p_k} \frac{1}{\sigma_p^k} d\mathbf{x}^e \tag{B.6}$$

For those k for which B.5 is valid, the integrals in B.6 are equal to one. For those for which B.5 is not valid, the integral is zero. Therefore, the final result is a sum over p_k for all of the k's for which B.5 is valid. The full distribution over Janossy densities can therefore be written as:

$$G(v) = \sum_{k\geq 1}^{\infty} \frac{p_k \overline{v}}{1 - e^{-\lambda}} \tag{B.7}$$

where

$$\overline{v}_i = \begin{cases} 0, & \frac{k! p_k}{\sigma_p^k} > v_i \\ 1, & \frac{k! p_k}{\sigma_p^k} \leq v_i \end{cases} \tag{B.8}$$

B.1.1 Supporting figures

Figure B.1 shows a PCA projection of data from CD patients, with each input feature highlighted individually. Only a selected set of features were shown in the main text. Similarly, Figure B.2 shows the PCA projection of data from UC patients.

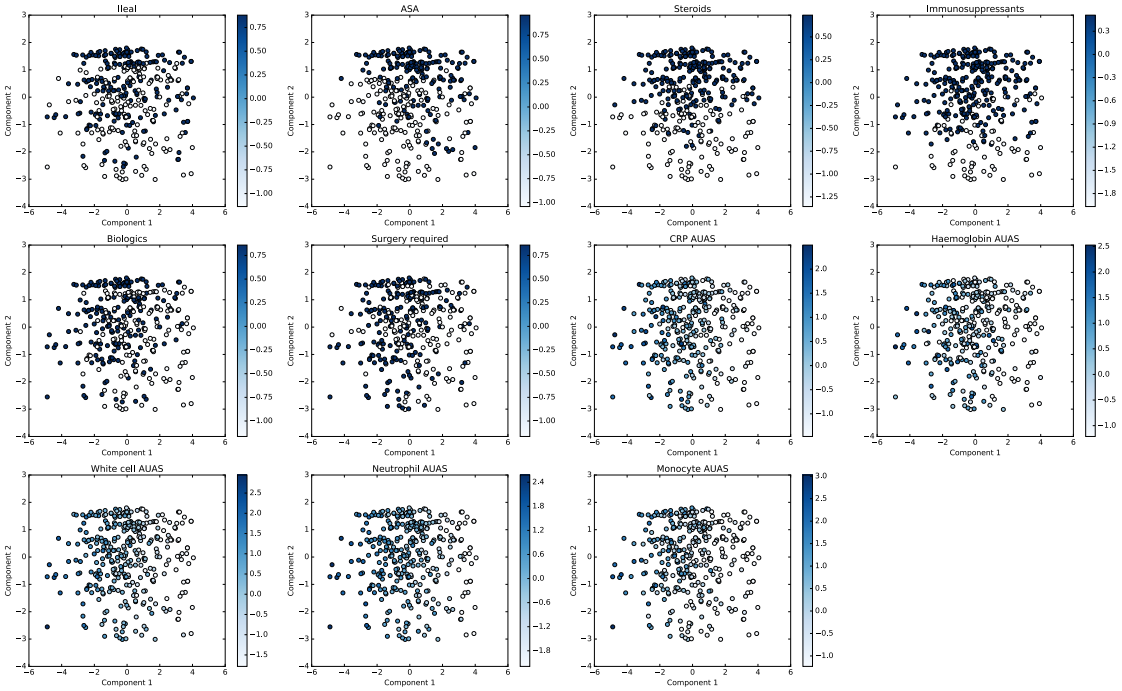


Figure B.1: *PCA projection of data from CD patients. First two components are shown. Each feature used as input in the PCA decomposition is shown in a subplot.*

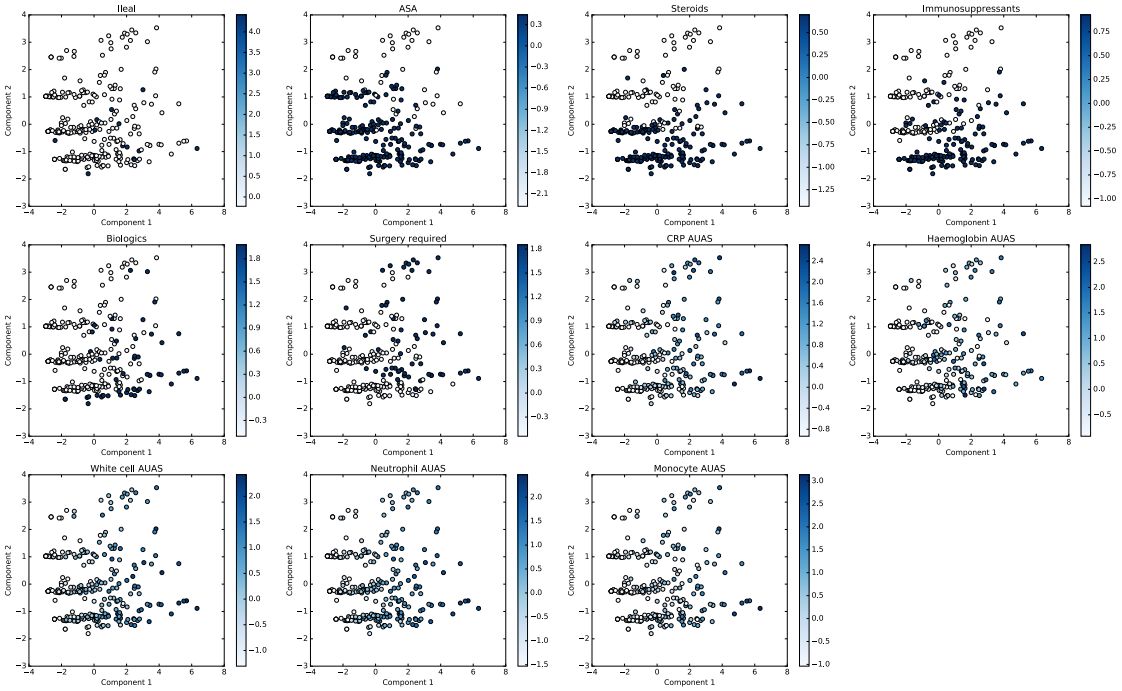


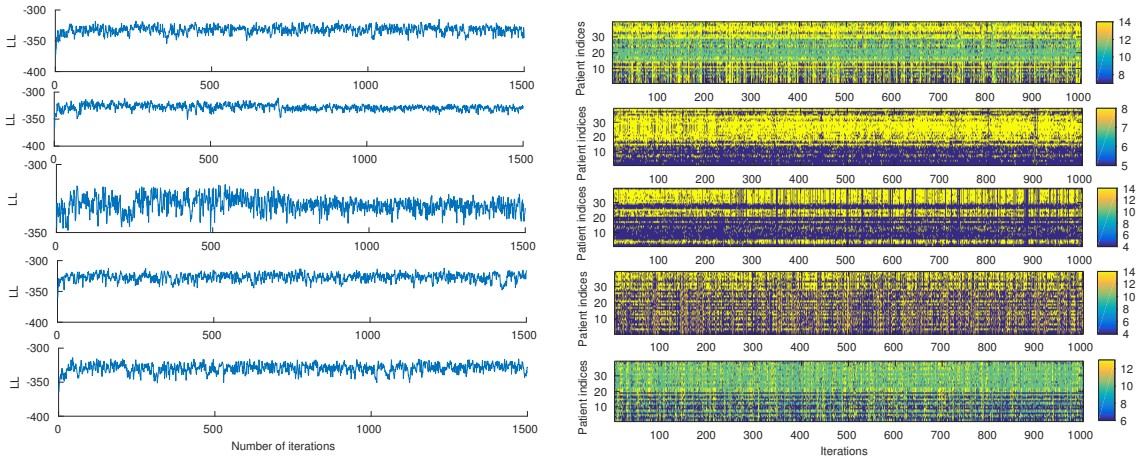
Figure B.2: *PCA projection of data from UC patients. First two components are shown. Each feature used as input in the PCA decomposition is shown in a subplot.*

B.2 Additional results from Chapter 5

The log-likelihood and patient assignments across iterations of our Gibbs sampler are shown in Figure B.3 for the case of clustering within our functional dataset.

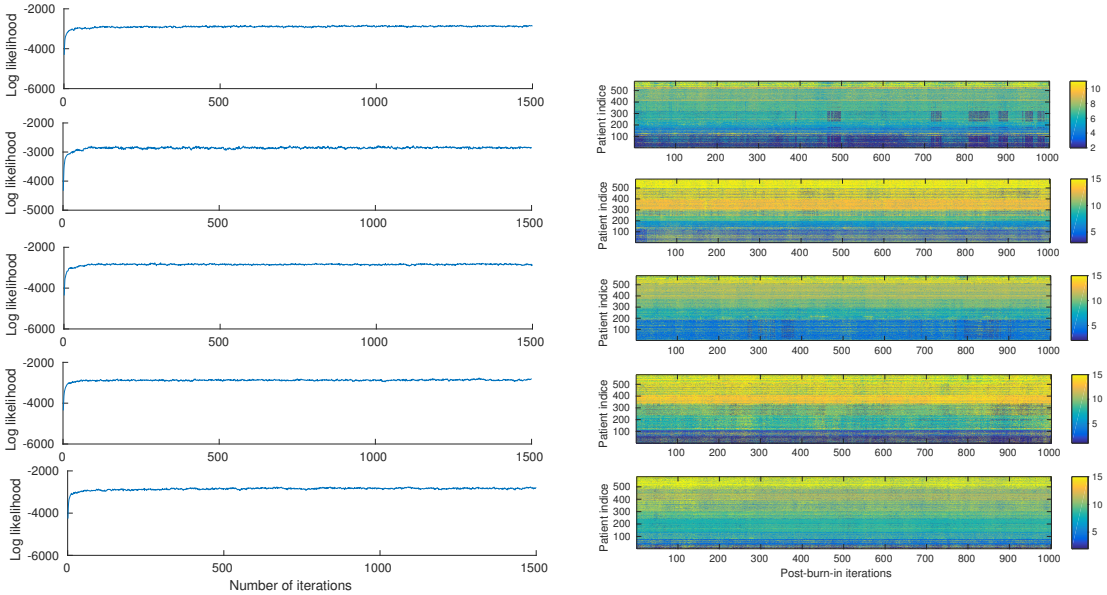
Similarly, these same diagnostic plots for the questions of clustering within all IBD patients, within just CD patients, and within just UC patients are shown in Figures B.4, B.5, and B.6, respectively.

The distributions for the collapsed severity clusters (i.e., combining cluster assignments into two composite "more" and "less" severe patient groups) for IBD, CD, and UC are shown in Figures B.7, B.8, and B.9, respectively.



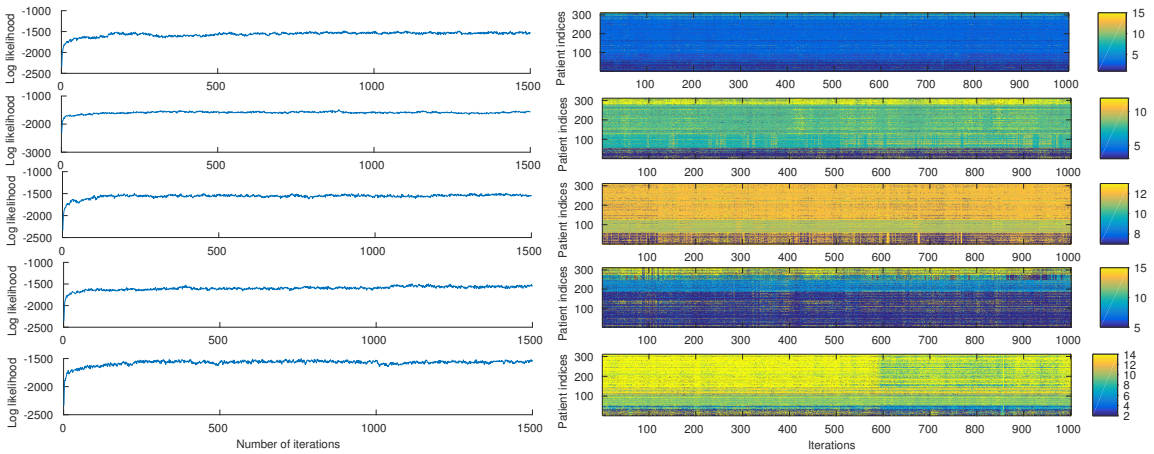
(A) Log-likelihood through sampling (B) Cluster assignments through sampling

Figure B.3: *Diagnostics of Gibbs sampling from five runs of clustering algorithm, for functional clustering problem. (A) Log-likelihoods across iterations. (B) Re-labelled patient assignments over iterations, post-burn-in. We note that the cluster assignments are shown only for those patients in the "training" set for the given cluster run, so the patient indices do not correspond across colormaps.*



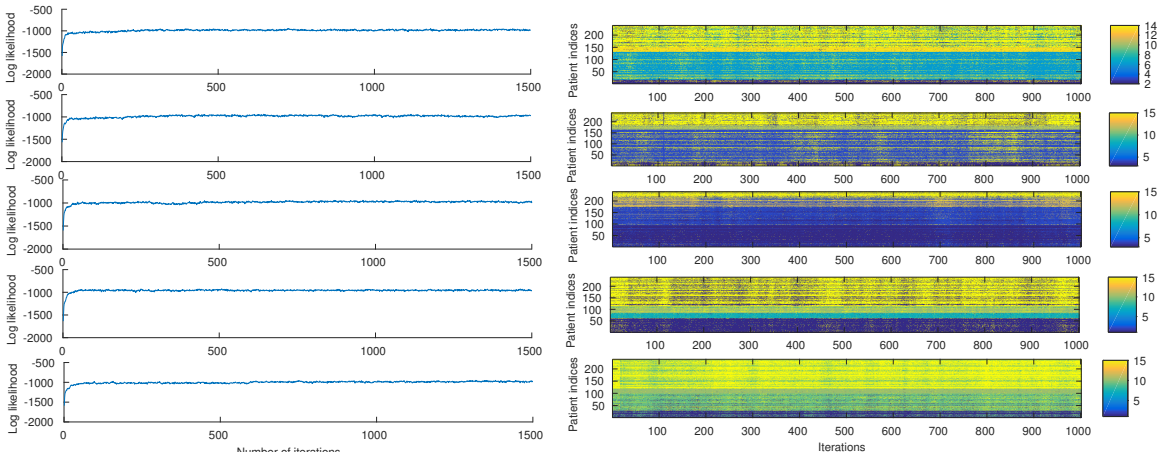
(A) Log-likelihood through sampling (B) Cluster assignments through sampling

Figure B.4: *Diagnostics of Gibbs sampling from five runs of clustering algorithm, for severity clustering problem for all IBD patients. (A) Log-likelihoods across iterations. (B) Re-labelled patient assignments over sweeps, post-burn-in. We note that the cluster assignments are shown only for those patients in the "training" set for the given cluster run, so the patient indices do not correspond across colormaps.*



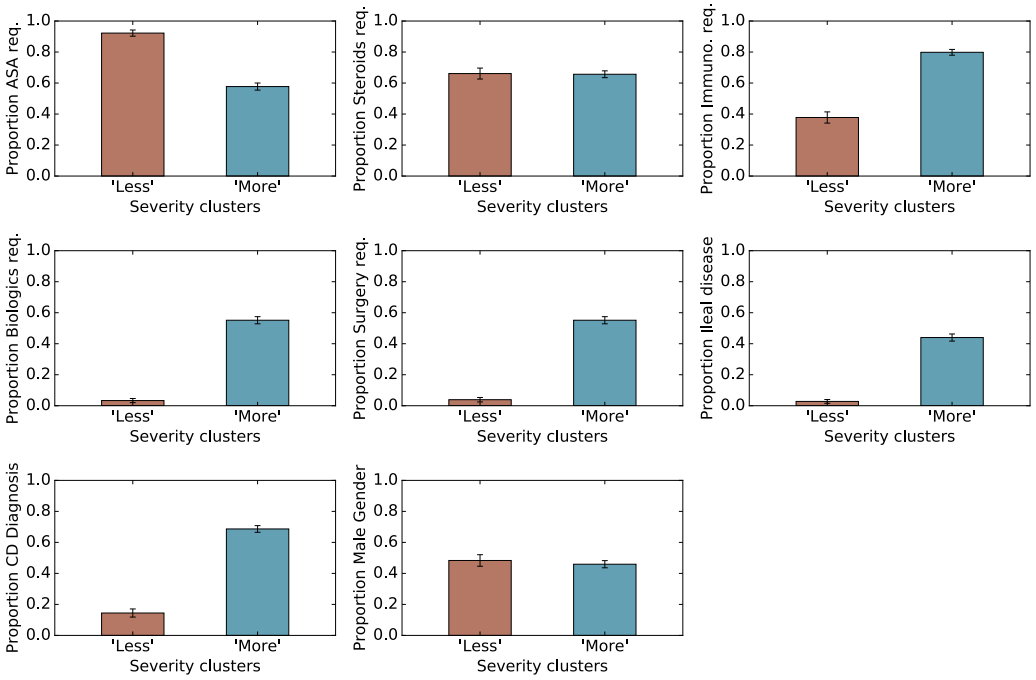
(A) Log-likelihood through sampling (B) Cluster assignments through sampling

Figure B.5: *Diagnostics of Gibbs sampling from five runs of clustering algorithm, for severity clustering problem for CD patients. (A) Log-likelihoods across iterations. (B) Relabelled patient assignments over iterations, post-burn-in.*

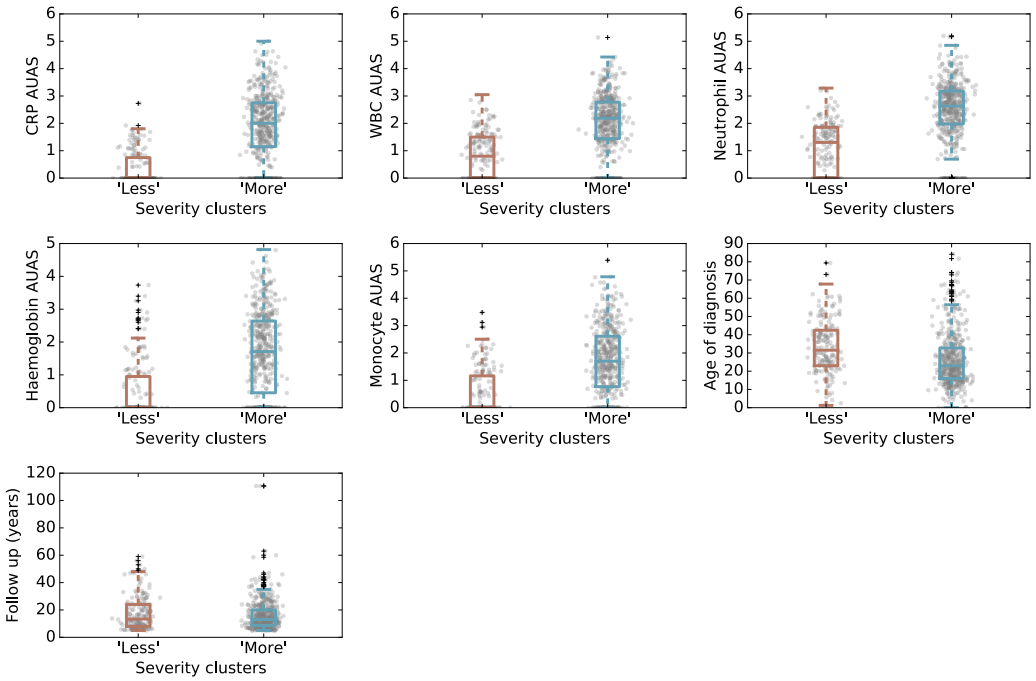


(A) Log-likelihood through sampling (B) Cluster assignments through sampling

Figure B.6: *Diagnostics of Gibbs sampling from five runs of clustering algorithm, for severity clustering problem for UC patients. (A) Log-likelihoods across iterations. (B) Relabelled patient assignments over iterations, post-burn-in.*

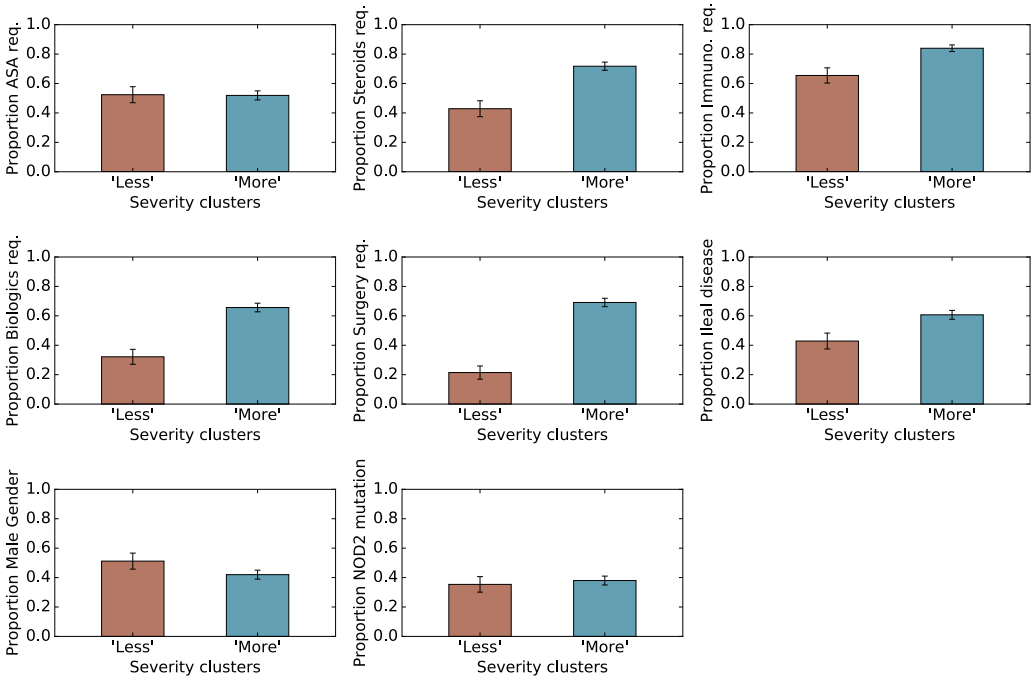


(A) Binary

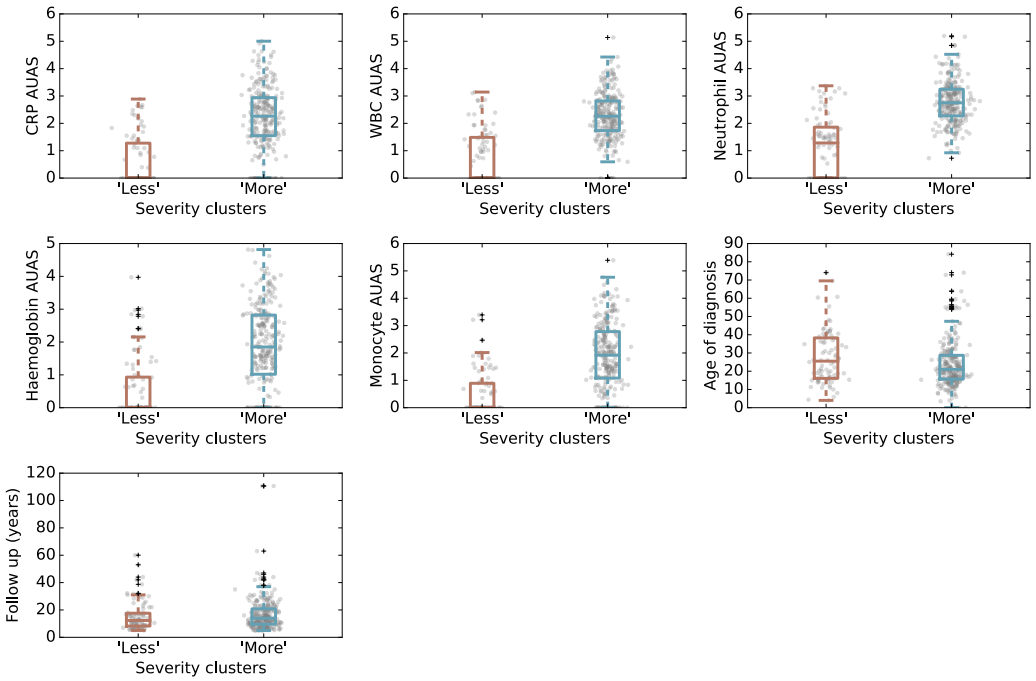


(B) Continuous

Figure B.7: *Feature distributions for (A) Binary features and (B) Continuous features for collapsed IBD severity clusters (CCCS).*

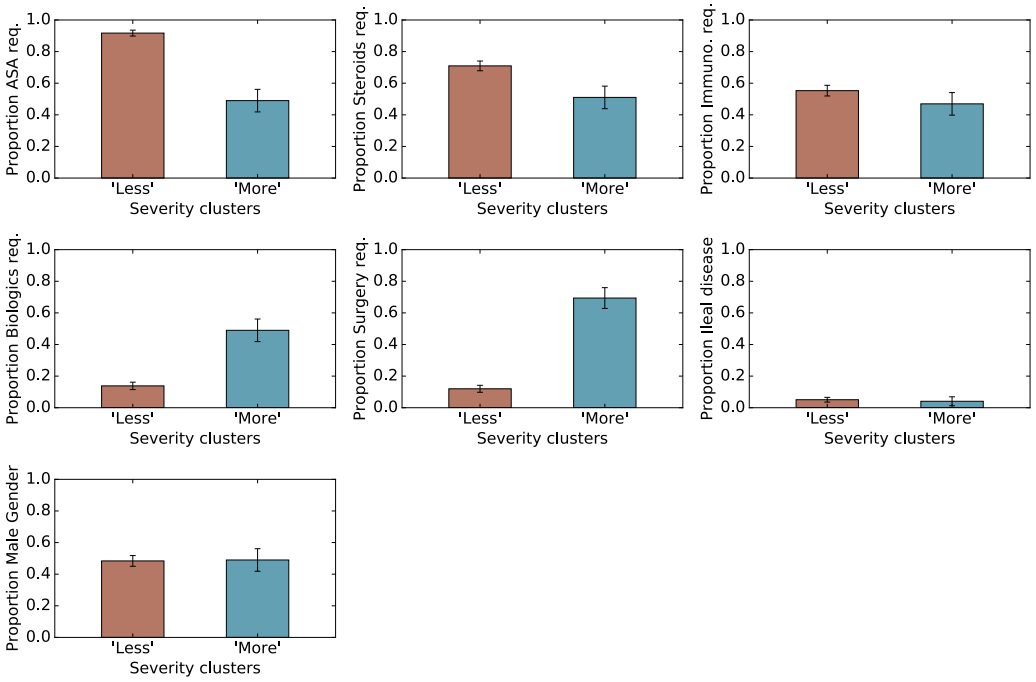


(A) Binary

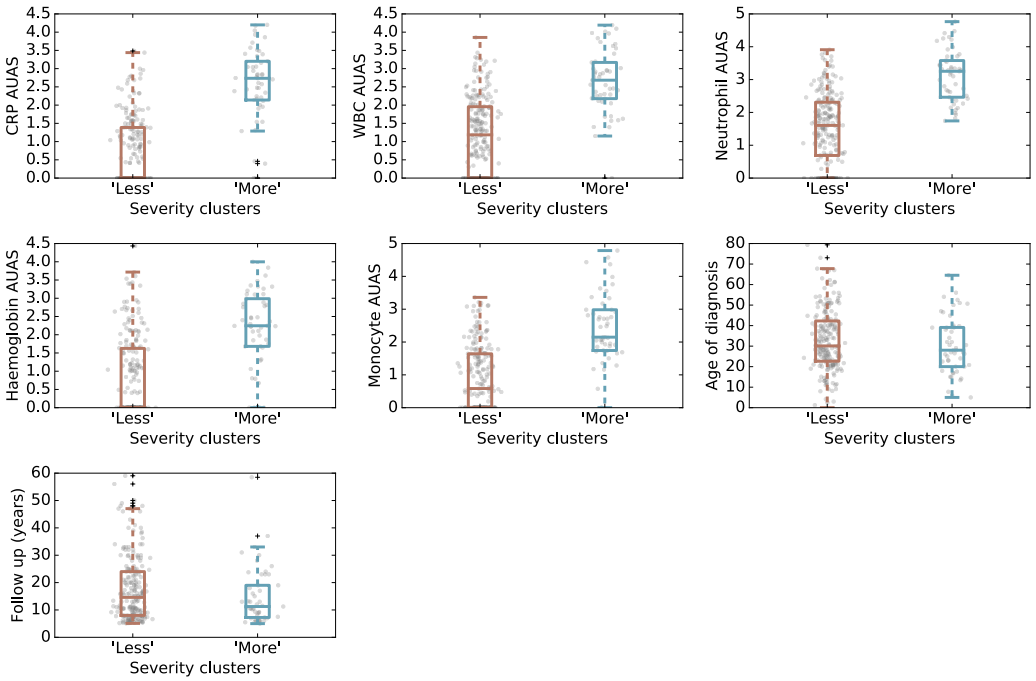


(B) Continuous

Figure B.8: *Feature distributions for (A) Binary features and (B) Continuous features for collapsed CD severity clusters (CCCS).*



(A) Binary

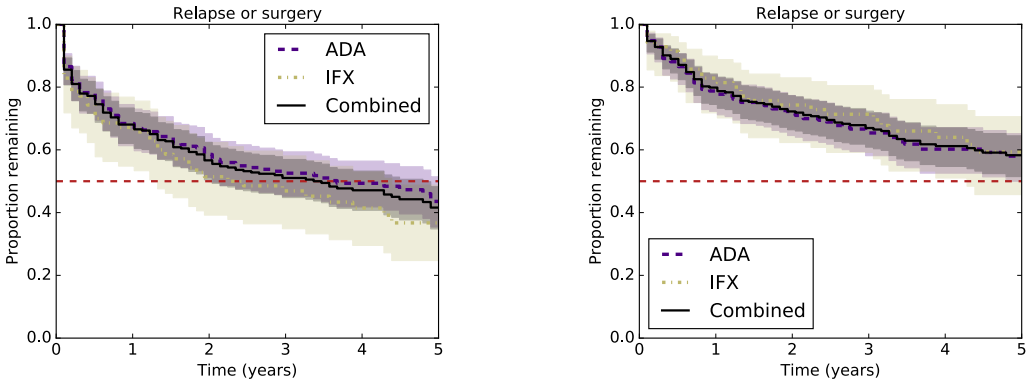


(B) Continuous

Figure B.9: *Feature distributions for (A) Binary features and (B) Continuous features for collapsed UC severity clusters (CCCS).*

B.3 Additional results from Chapter 7

In Chapter 7, we assumed a CRP threshold for defining a relapse event of 80 mg/L. Figure B.10 shows the survival curves for the case of a threshold of 40 mg/L (A) and 100 mg/L (B). As would be expected, a lower threshold for defining relapse results in a shorter time-to-severe events, and vice versa. While median survival with a 80 mg/L cutoff was 5.9 years, it was 3.5 years and 6.3 years, for thresholds of 40 mg/L and 100 mg/L, respectively.



(A) 40 mg/L CRP cutoff (B) 100 mg/L CRP cutoff

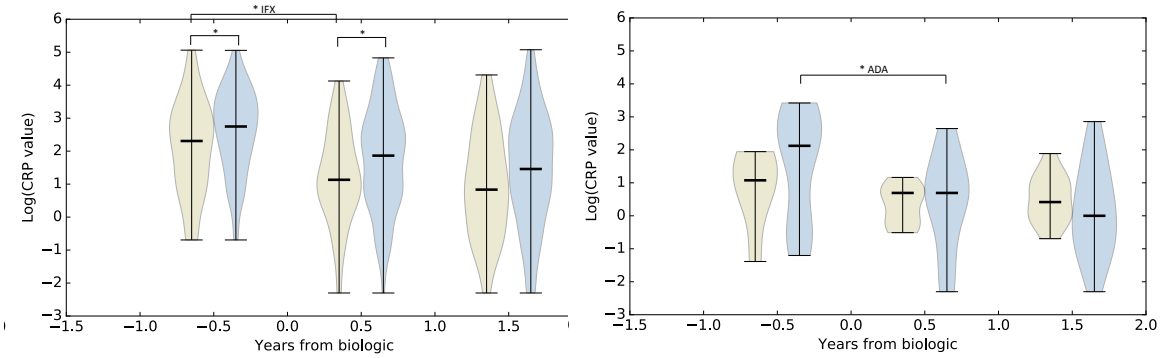
Figure B.10: *Patient survival curves until relapse or surgery, with the CRP threshold to define relapse set at 40 mg/L (A) and 100 mg/L (B).*

We presented the overall results for CRP responders in Chapter 7; we have split the results by ADA- and IFX-treated patients in Figure B.11. Similarly, we presented the results for ADA- and IFX-treated patients together for those patients who were both drug responders and CRP responders. In Figure B.12 we illustrate the results by these drugs individually.

B.4 Investigation into localisation subgroups

As mentioned in Chapter 8, we have also investigated the patterns of disease localisation present within our patient cohort. We used our clustering model, as presented in Chapter 5, to model the probability of inflammation in various locations of the GI tract: mouth, esophagus, duodenum, ileum, caecum, ascending colon, transverse colon, descending colon, sigmoid colon, rectum, anus, and the area from a pouch operation (see Figure 1.1). As these are all binary variables, we only required the binary portion of our combined-data-type mixture model.

Our clustering model identified 11 clusters. A visualisation of the inflammation



(A) CRP responders (B) CRP non-responders

Figure B.11: *CRP responding (A) and non-responding (B) patients: CRP decline for time period from one year prior to biologic therapy to two years after. Yellow=IFX; Blue=ADA. Significance markers: * = $p < 0.05$, ** = $p < 0.01$; shown for successive time windows and between two drugs.*

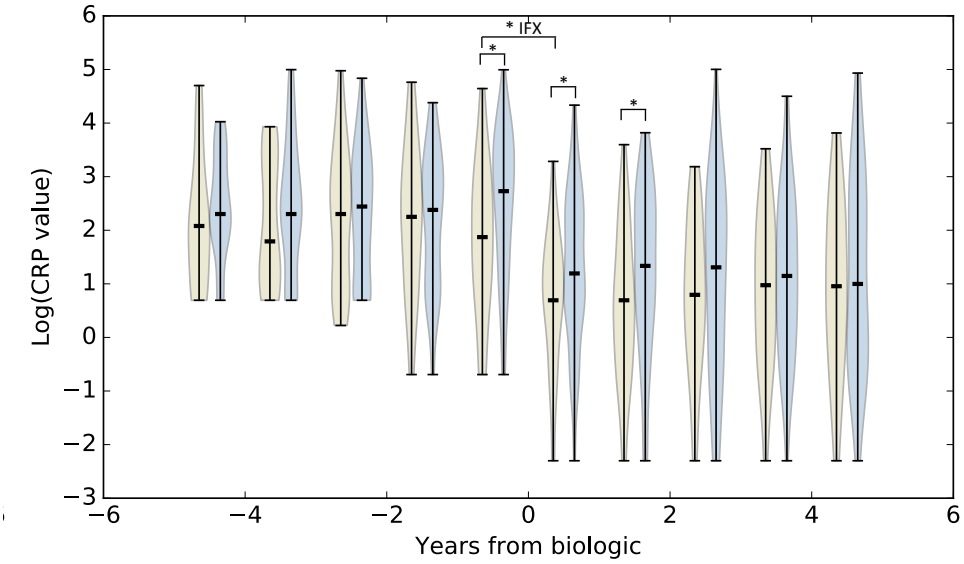


Figure B.12: *Patients who were both CRP responders and biologic responders: CRP decline for time period from five years prior to biologic therapy to five years after. Yellow=IFX (violins with left offset from center); Blue=ADA (violins with right offset from center). Significance markers: * = $p < 0.05$, ** = $p < 0.01$; shown for successive time windows and between two drugs.*

patterns evident for patients in each of these clusters is shown in Figure B.13. As in Chapter 5, we will refer to these clusters by the number of patients assigned to each. We see that a large number of patients (cluster $C_{loc,96}$ have no apparent inflammation. While possible, it is more likely that these patients lack any disease localisation information in the EHR. We then see groups of patients with only ileal disease ($C_{loc,76}$), ileal disease with caecum involvement ($C_{loc,55}$), predominantly colonic disease ($C_{loc,31}$), pancolitis ($C_{loc,54}$), upper colon inflammation ($C_{loc,13}$), entire colon ($C_{loc,102}$), lower colon ($C_{loc,3}$), lower colon with rectal involvement ($C_{loc,78}$), lower colon with upper GI and ileal involvement ($C_{loc,17}$), and finally just sigmoid and rectal disease ($C_{loc,121}$).

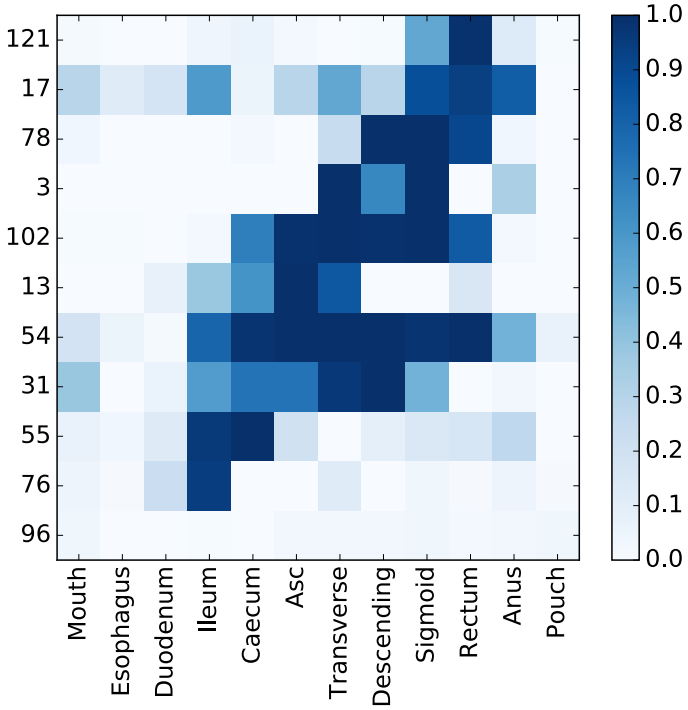


Figure B.13: *Probability of a particular portion of the GI tract being inflamed for each identified patient subgroup. The number of patients in each group is shown on the y-axis.*

There were not immediately apparent lab-based patterns that correlated with these localisation subgroups. However, medication usage, surgeries, and disease diagnosis (CD versus UC) did strongly correlate, as shown in Figure B.14. This is not surprising, as physicians commonly consider disease localisation when deciding upon treatment courses.

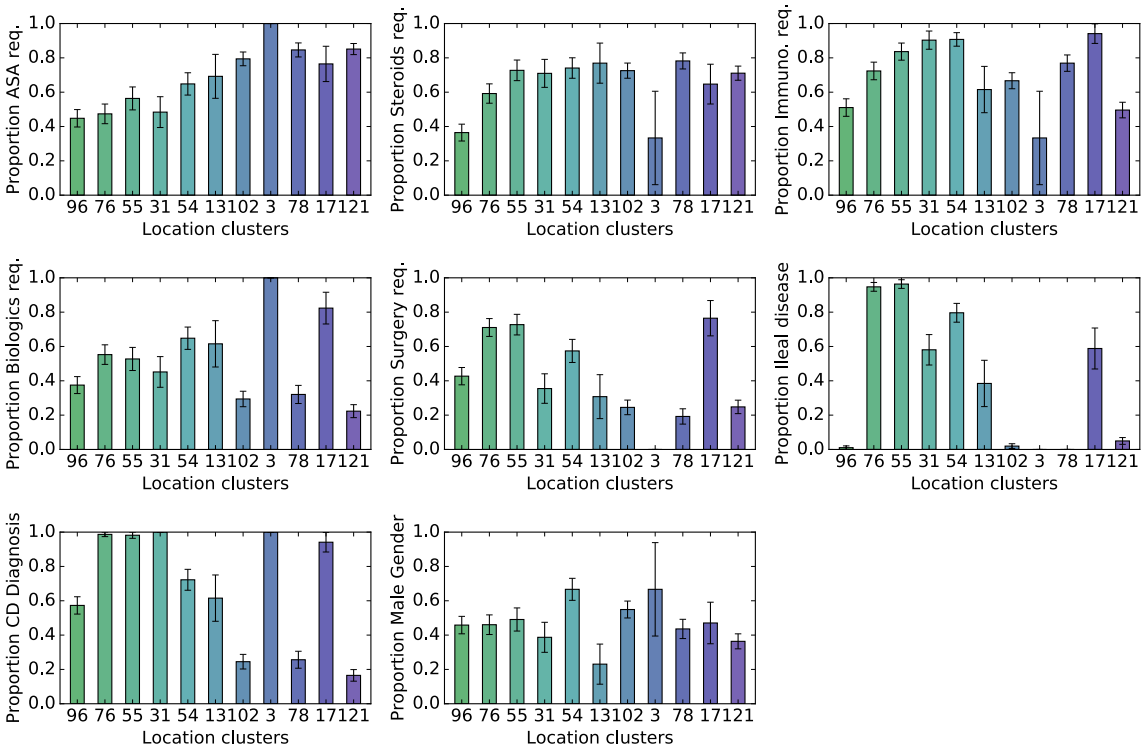


Figure B.14: *Distribution of prevalence of various features of interest for each localisation subgroup. The number of patients in each group is shown in the x-axis.*