Transfer Report

Katherine Niehaus
Trinity College
University of Oxford

Supervised by
Dr. David Clifton
Professor Timothy Peto

August 28, 2014
Abstract

The increasing adoption of electronic health records means that repositories of patient information are becoming available at an unprecedented scale. At the same time, the decreasing cost of DNA sequencing means that genomic information is also becoming increasingly accessible. However, currently, there is limited clinical understanding of how to relate genomic information with patient phenotypes; the analysis tools for modelling patient phenotypes, relating phenotype to genomic information, and using these models for medical prognosis have not yet been fully developed.

Machine learning techniques within the field of clinical informatics are focused upon precisely these questions. We present an overview of the current state-of-the-art in this intersection of machine learning and medicine, before introducing examples from our work that have used machine-learning approaches to perform inference using genomic and clinical data to answer questions of clinical significance. We illustrate our ability to predict *Mycobacterium tuberculosis* drug resistance based upon the bacterial genome using supervised learning methodologies, with improved accuracy in comparison to approaches using current clinical knowledge. We also identify novel subgroups of inflammatory bowel disease patients using unsupervised techniques. Our long-term goal is the development of new tools that can be used to perform inference with such heterogeneous data.

Publications and posters from 2014:


1 Introduction

The increasing abundance of medical data has the potential to greatly advance our understanding of human pathophysiology and to improve both diagnosis and treatment. Such data sources include electronic health records (EHRs), which typically contain clinical and physiological measurements, as well as genomic information, where the latter is becoming more accessible through the decreasing cost of DNA sequencing. There is an urgent need to develop the tools that perform inference with such heterogeneous data. Much work within the field of clinical informatics focuses upon the usage of human genetic information (i.e., genotype) to predict or augment our understanding of a patient’s phenotype. Here, we define phenotype to include patient symptoms, physiological measurements, and clinical observations. Throughout this report, we will describe the development of tools based on machine learning that can be used to answer clinically-relevant questions in two example domains: infectious disease (ID) and inflammatory bowel disease (IBD), integrating both EHR and genomic data.

We will begin in Chapter One with an introduction to EHR and genomic data sources, followed by a literature review of machine learning techniques that have been used in clinical informatics. We will start in Chapter Two with an investigation of genomic data, demonstrating our supervised machine learning methodology with an application in drug susceptibility prediction for tuberculosis patients. We then extend our analysis in Chapter Three to include EHR data, with an example of using unsupervised clustering to define clinical phenotypes for patients with IBD. Finally, in Chapter Four, we set out a plan for the remainder of the DPhil.

1.1 Electronic health record data

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. For the purposes of this report, we will define EHR systems in the broadest sense, referring to any large scale electronic compilation of patient health information. The EHR market is dominated by six large providers: McKesson (USA), Cerner Corporation (USA), Siemens (Germany), Epic Systems (USA), Allscripts (USA), and General Electric (USA) [1,2].
EHR systems are being widely adopted around the world. For instance, through the passage of the “Health Information Technology for Economic and Clinical Health” (HITECH) act in the USA in 2009, every healthcare provider is now incentivised to illustrate their “meaningful use” of EHR systems. From 2001 to 2013, the adoption of EHR systems by office-based physicians increased from 18% to 78% [3], with hospital adoption of EHR systems (with at least basic functionality) increasing from 9.4% in 2008 to 44.4% in 2012 [4]. In the UK, while the programme to create an NHS-coordinated EHR system (“Connecting for Health”) was unsuccessful, EHR adoption in the UK has historically been very high: 89% of primary care practices had an EHR with advanced functionality in 2009, compared to 26% in the USA [5]. It is clear that the trend is towards increasing collection and storage of healthcare data [6].

EHRs contain a number of different data types. These generally include diagnosis codes (e.g., International Classification of Diseases - ICD10); procedural codes (e.g., the USA uses Current Procedural Terminology - CPT); laboratory results; data such as vital-sign summary values; and free text. While some information, such as laboratory data, is generally stored in a structured form, other information, such as the prescription and dosage of medication, can be much more difficult to ascertain. Specific characteristics of the data are often unique to individual EHRs (and even between different specialties within the same hospital), making data extraction methods difficult to standardise across systems.

In addition, because EHR data are collected for clinical care, patient records are often partially missing, incorrect, systematically misleading, and contradictory. The process of extracting meaningful EHR data that reflects true patient physiology (i.e., “phenotyping” the patient) is not a simple task. Hripcsak and Albers outline several of the associated challenges [7]. As various authors have identified, within the scope of machine learning analysis, a pertinent drawback of EHR data is that the labels required for supervised learning may be confounded by the fact that care is being provided [8] [9].

Much work within the growing field of EHR analysis has focused upon the problem of retrieving useful information from the record - this largely falls within the domain of natural language processing (NLP). The large amount of free text, and the frequent misspellings and abbreviations that occur within medical records, means that NLP processing is an important step in automated EHR analysis. Such work has included efforts to identify drug prescriptions
from unstructured text and to combine billing codes into larger general concepts [10] [7] [11].

While much work remains to make EHR systems interoperable across hospitals and easily incorporated into clinical workflows, EHR data provide a rich source of clinical patient information from which to perform inference and to understand disease and its treatment.

1.2 Genomic data

The declining cost of human genome sequencing means that genomic information has the potential to inform routine clinical practice. Human genetic variation is encoded in our DNA, which is composed of a long series of nucleotide base pairs (adenine, thymine, cytosine, and guanine). In the process of transcription, portions of DNA are converted into single-stranded RNA. Triplets of RNA bases (codons) 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. Only about three percent of the 3 billion nucleotide bases in the human genome constitute genes. Understanding the role of the rest of the genome is still a work in progress, but it is well-established that parts are involved in regulating gene expression (i.e., protein production). Heritable (and non-heritable) gene expression regulation can also be accomplished through changes to the macro-structure of the DNA through chemical alterations such as methylation. Such changes are termed “epigenetic” because they result in changes in gene expression without altering the underlying DNA sequence.

Human genomic variation can be due to a variety of sources. A single nucleotide polymorphism (SNP) is a single-base change in the DNA. 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) and copy-number variants are also sources of inter-human variation. Copy-number variants are large portions of the genome that have been deleted or duplicated; they constitute approximately 12% of human DNA [12].

Variation can thus be assessed at many levels - from the genetic code to epigenetic modifications - through various different “omics” technologies. At the genetic level, microarrays can be used to examine the presence of SNPs at certain pre-defined genomic locations. 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. 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 (USA), the current market leader in sequencing platforms, recently announced the sale of the first machines capable of the long-heralded “one-thousand dollar” human genome in early 2014. Other sequencing companies, such as Pacific Biosciences, Oxford Nanopore, and Ion Torrent, all claim to soon be offering even cheaper sequencing alternatives. While gene sequences provide an understanding of some underlying human variation, we are coming to realise how strongly post-transcriptional modifications, post-translational modifications, and epigenetics can impact cellular phenotypes. Microarray expression data can be used to examine gene expression at single or multiple time points across experiments. Proteomics techniques such as mass spectrometry can be used to assess protein levels in samples.

The growing ability to obtain omics data is exciting from a research perspective, but it still remains prohibitive to include genome sequencing results in EHRs. Not only are the costs still too high for routine practice, but the clinical knowledge of how to act given a particular patient’s sequence remains to be developed. For instance, in 2014, Dewey et al. described their methodology for analysing the risk of disease in 10 healthy volunteers through WGS. The predominant finding of their study was that WGS is not yet ready for use in clinical practice, given the inconsistency of results, inadequate base coverage in key disease-related areas of the genome, and rudimentary understanding of how the majority of SNPs may affect the risk of certain diseases.

1.3 Combining Genomic and EHR data

In an effort to accelerate the methodological advances required to incorporate genomic data into clinical practice, a number of institutions have begun programmes to link genomic and EHR data. In 2007, the Electronic Medical Records and Genomics (eMERGE) network was formed in the USA, which is a consortium of nine medical centres. The group combines microarray-based genetic data with phenotypes derived from EHRs. To date, most of the

$1 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.
Algorithm work has focused on defining and extracting phenotypes from EHRs. One of the outputs of this consortium has been the Phenotype KnowledgeBase (PheKB), a repository of such algorithms for defining phenotypes from electronic health records [16].

The Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) is another such effort. The RPGEH has obtained biosamples from 100,000 members of the Kaiser health plan, which it has used to investigate associations between genes and phenotypes. The “i2b2” service is another similar project, which has been created by the National Center for Biomedical Computing in the USA. This service allows researchers to create their own database extractions from a central source, which includes both medical record and genomic data [17].

Jensen et al. provide an excellent high-level overview of the challenges and opportunities involved with combining EHR and genomic data sources to improve clinical care [18].

1.4 Patient consent

When working with EHR and genomic data, it is of utmost importance to ensure patient privacy, consent, and confidentiality. There has been work showing that even anonymised genetic data can be traced back to the person from which they originate [19]. In the UK, the 1997 Caldicott review placed restrictions on how data from patients can be used in medical research, with the aim of increasing the protection of patient privacy [20]. The UK Biobank, which aims to create a long-term repository of patient information accessible to a wide range of research projects, has dealt with patient consent through the principle of “consent to consent.” 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 [21].

1.5 Supervised machine learning in clinical informatics

Machine learning constitutes a family of data analysis techniques that has been widely adopted in many fields. Its application towards clinical informatics has been more recent, partially due to the previously-mentioned challenges of obtaining sufficient data to train and validate models. Many clinical questions can be framed as standard supervised machine learning problems. Such questions often centre upon prediction of patient risk (e.g. risk for developing various

2The “Caldicott2” review in 2013 highlighted the need to balance these concerns with adequate sharing of information between health professionals.
conditions; risk of deterioration after diagnosis), with the training examples consisting of a
set of patients who have gone on to deteriorate, and a set of patients who did not. Available
data to answer these questions typically include patient demographics, the presence of various
comorbidities, blood test values, specific SNP presence or absence, and sensor-based data (e.g.
[22,24]). Methodological research often centres on how to derive useful features from these
data. Particularly for those input variables that vary over time (e.g., laboratory values, vital
signs), it is not necessarily straightforward to identify relevant discriminatory information. As
Hripcsak et al. caution, the apparent time series in an EHR is a reflection not merely of patient
measurements, but also of clinical workflow [25]. Once features have been derived, classifiers
that are frequently used include Bayesian product of marginals (BPM)3, logistic regression
(LR), support vector machines (SVM), random forests (RF), and neural networks (NN).

1.5.1 Predicting patient risk with EHR data

A common application of supervised machine learning techniques is in predicting patient risk
for developing a given disease or condition. For instance, Balkau et al. used LR to predict the
risk of diabetes in a prospective study of over 3,000 French patients (of which approximately 200
went on to develop diabetes) [26]. They included clinical variables such as the circumference
of the waist, hypertension, smoking; biologic variables such as fasting glucose and triglyceride
levels; and genetic polymorphisms at four sites. In their analysis, genetic information was
not found to improve predictive ability, but clinical and biologic variables together provided
good prediction of diabetes risk, with AUROC statistics of 0.85 and 0.92 for men and women,
respectively; however, no separate test set was used.

In 2013, Leung et al. used machine learning to predict the risk of diabetic kidney disease
(DKD) within a Chinese population of type-2 diabetes patients, based upon both clinical and
genotypic factors [22]. The group compared a number of different machine learning algorithms,
including SVMs, RFs, BPMs, and NNs. The set of features that they used included 107 SNPs
from 65 genes associated with inflammation, metabolism, etc; as well as clinical variables such
as blood pressure, retinopathy, lipid levels, etc. They found that SVM-based prediction with a
radial basis kernel function yielded the best performance of all models considered, with accuracy

3BPM is often also commonly referred to as naïve Bayes (NB), with the “naïvieté” coming from its
strong assumptions of independence.
exceeding 95%. The group also used the weightings of the various features selected for use by the different classifiers to produce a final set of ten features that yielded similar levels of predictive performance to the full-feature model (94% accuracy).

Huang et al. used a set of 35,000 patient records to predict the diagnosis of depression in patients up to a year before the actual clinical diagnosis [27]. From the EHR of each patient, the authors extracted 10,600 binary features, such as demographics, ICD-9 codes, drug ingredients, etc. They then selected the subset of 50 features that gave the greatest information, and used these to train a BPM classifier. They were able to attain sensitivities and specificities on par with those achieved by primary care physicians, as assessed in previous studies.

Wu et al. similarly investigated their ability to predict heart failure before clinical diagnosis using various EHR variables as input [28]. They used binary values to denote the presence or absence of variables such as smoking and laboratory orders; continuous values for variables such as body mass index and patients’ most recent laboratory test measurements; and time since events such as the most recent visit to the doctor and the diagnosis of comorbidities. The authors compared the performance of several machine learning models applied to this task, including LR, SVMs, and Adaboost; they applied various feature selection methods such as the Bayesian information criterion (BIC). They concluded that LR combined with BIC-based feature selection was best able to predict heart failure, with an AUROC of 0.76.

Johnson, et al. created a severity score for ICU patients that requires fewer variables than alternative existing methods, but which retains similar predictive performance [29]. The authors used a genetic algorithm to pre-select features and particle swarm methods to assign weights in an LR model. Input features included the most abnormal laboratory test values for a panel of lab tests, the most abnormal vital sign measurements, and binary clinical characteristics (such as whether the patient was admitted to the ICU from a general unit, was ventilated on the first day of admittance, received thrombolytic drugs, etc).

1.5.2 Predicting patient risk with time-series vital-sign data

While the previous examples have primarily used EHR data, together with a few pre-identified genomic markers, as input, vital signs acquired via sensors have also been used to predict the risk of deterioration for patients in intensive care facilities, where such information is routinely collected.
Saria et al. used vital signs (heart rate, respiratory rate, and oxygen saturation) from 138 preterm infants to create a predictive risk score for severe comorbidities [30]. The authors first preprocessed the time-series data to obtain estimates of long-term trends and short-term variability, and then they took the mean and variance of these signals. The resulting summary features were then modeled using long-tailed candidate distributions: exponential, Weibull, lognormal, normal, and gamma. The distribution which had the highest likelihood after cross-validation was selected for the final model. The log-odds ratio for individual patients were then used to train an LR classifier to distinguish between low- and high-morbidity infants. The resulting scoring system attained an AUROC of 0.92 for predicting high morbidity, in comparison to alternative available risk scores, which had AUROCs in the range of 0.70 – 0.85. The authors found that adding laboratory test values as features did not improve prediction.

In a more exploratory analysis, these authors developed a generative model they term a Time Series Topic Model (TSTM) to examine common features within their time-series data [31]. Their TSTM draws upon NLP approaches for clustering documents by topic, with the topics defining distributions over words. By analogy, they assume that there are an underlying set of “topics” within the neonatal vital-sign signals, and that any given time series is formed through a transition matrix over these topics; any topic is associated with a probability of “words,” which are features of the signal. These words are also learned during construction of the model. A hierarchical Dirichlet process was used as a prior over topics to allow for words to be shared by the different time series.

Gaussian processes (GPs) are a nonparametric method for modelling data in a series. Time-series sensor data is often incomplete and subject to artefacts caused by movement and sensor dislocation, making the Bayesian approach of GPs particularly suitable. Clifton et al. illustrate how patient-specific GP regression can be used to identify patient deterioration much earlier than would be possible using traditional methods [23]. Using wearable ECG and pulse oximetry sensors to acquire data from ambulatory patients recovering from surgery, the authors use GPs to model the time-series of each vital sign. This work has been further extended by Pimentel et al. to identify common trajectories of heart rate and breathing rate following surgery [32]. 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 data were compared to the time-series clusters to determine if the test data were similar to “normal” or “abnormal” clusters.

More work remains to allow for similar analytical methods to be applied to alternative time-series healthcare data, such as laboratory biomarkers. In one novel approach, Wiens, et al. first created a day-by-day patient risk score for becoming infected by Clostridium difficile [33]. This risk score was derived from an SVM classifier with >10,000 features from the patient EHR as input. Features included the reason for admission, demographics, lab results, room location, vital sign measurements, etc (binary features were created from categorical variables, which explains the high dimensionality). They then modelled this risk score as a time-series, using three different approaches (extracted features, similarity metrics, and HMMs) to perform classification. Their methods were able to predict patient risk more successfully than traditional approaches of taking aggregate or daily features, with AUROCs of up to 0.79 in contrast to the traditional approach’s AUROC of 0.69.

1.5.3 Predicting patient risk using dynamic Bayesian networks

Another method for incorporating temporal information into disease prognosis is through Dynamic Bayesian Networks (DBNs). DBNs are an extension of probabilistic graphical models to allow modelling over a sequence. The nodes of a DBN correspond to the random variables of interest, edges indicate the relationship between these random variables, and additional edges model the time dependency.

Temporal Bayesian networks such as DBNs have the desirable property that they allow for interpretation of the interactions between different variables, which is not the case for “black box” methods such as SVMs and NNs. For instance, van der Heijden et al. used temporal Bayesian networks to model variables such as sputum volume, temperature, and blood oxygen saturation for ten patients with chronic obstructive pulmonary disease (COPD) in order to predict exacerbation events [34]. They used a score-based search procedure to learn the model structure (via expectation maximisation, EM) to deal with missing data. Due to their small sample size, they used bootstrapped samples to assess the uncertainty of their resulting model with respect to the data. For comparison, they also used a technique they term a “temporal
naïve Bayes classifier” (which was essentially the same, but without the final bootstrapping step), and heuristic models developed by experts. The proposed method was able to obtain up to 0.90 AUROC for prediction of exacerbation within 24 hours using a validation dataset of 13 COPD patients; they obtained very similar performance with both the full bootstrapped model and with the temporal naïve Bayes classifier.

Nachimuthu et al. also used temporal Bayesian networks for modelling insulin dosing in an ICU within the Salt Lake City Intermountain Healthcare System [35]. Many ICU patients experience hyperglycaemia in the ICU, even if not diabetic; current sentiment is that close control of these glucose levels leads to improved patient outcomes. To predict future insulin requirements, Nachimuthu, et al. used an expert-informed Bayesian network structure, with the values of its parameters determined using an EM algorithm (to accommodate missing data).

1.5.4 Disease discrimination

Another common clinical question where machine learning approaches have been applied is that of using biomarkers to discriminate between different diseases, or to predict the response of patients’ condition to various drugs. Biomarkers may include genetic markers or compounds that can be measured in the blood. For instance, Plevy et al. used a RF algorithm to classify patients as having Inflammatory Bowel Disease (IBD) or not, and, within IBD-predicted patients, as having ulcerative colitis (UC) or Crohn’s disease (CD) [24]. As features, the authors used the concentrations of eight serological antibodies, four pre-defined SNPs, and five inflammatory marker concentrations. They obtained 73.6% sensitivity and 89.6% specificity in IBD diagnosis, as well as 88.9% sensitivity, 81.0% specificity for diagnosing CD and 97.7% sensitivity, 83.5% specificity in diagnosing UC. The gastrointestinal microbiome (i.e., bacterial populations within the gut) has also been used as a set of input features to distinguish between IBD and non-IBD pediatric patients [36].

Walgee et al. report on their use of a random forest algorithm to predict drug response for IBD patients. The group attempted to predict responses to thiopurines based upon patient age and blood tests [37]. Blood marker features included levels of neutrophils, white blood cell count, total bilirubin, sodium, haemoglobin, the distribution of red blood cell width, and the mean corpuscular volume. The group was able to obtain 72% sensitivity and 77% specificity in predicting response, which is better than that possible with the only reproducible alterna-
tive method of using the levels of the 6-TGN and 6-MMP metabolites (62% sensitivity, 72% specificity was obtained in a meta-analysis).

Further work on genomic-based biomarker disease discrimination is discussed in Section 1.7.

1.6 Anomaly detection

Anomaly detection is a specific approach towards data analysis that is frequently used when clinicians are to be alerted of some notable event, such as patient deterioration. Anomaly detection has been widely used to detect abnormal behavior based upon continuous sensor data [38]. A full review of anomaly detection methods was covered by Pimentel et al. in 2014 [39]. Within medical applications, one-class SVMs [40], multivariate extensions of extreme value theory [41], the addition of “novel” factors into Factorial Switching Kalman Filters [42], and outlier detection methods [43] have all been employed for use with vital-sign data. These methods typically create some sort of model of normality (e.g., a density estimate of the distribution of representative training data via kernel methods) and attempt to identify patients deviating from this model. Models are typically based on vital signs such as heart rate, breathing rate, and oxygen saturation, and they are then used to define criteria as to which values are “anomalous” in a multivariate setting.

1.7 Genomic analysis approaches

While human genetic data is rarely included in EHR systems today, such information will be available in the future. The previous studies presented have used only a small number of established genetic markers as input, if genetic information is included at all. To enable healthcare systems to exploit the growing volumes of genomic data becoming available, it is necessary to establish methods for combining large-scale omic data sets with clinical data from the EHR. We first examine common approaches used within genomics research today.

1.7.1 Genome-wide association studies

A dominant paradigm for identifying genomic associations with human phenotypes is through genome-wide association studies (GWAS). Typically, several hundred thousand SNPs are assessed for a group of people 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 are present in about 5% of the population\[44\].

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 \[45\]. In addition, corrections are often made for population structure through adjustment by a genomic inflation constant\[ or using principal component analysis (PCA) \[46\].

To date, more than 1,600 papers have published GWAS results, identifying 2,000 genotype-phenotype associations \[47\]. However, these associations typically together account for less than 10% of the phenotype’s heritability\[ 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, copy-number variants, interactions across SNPs that are not currently captured, epigenetic factors, and environmental influences \[44, 47\]. 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. \[48\]).

Case-control studies have also been designed with gene expression data as the input data source. These approaches have been found to be very useful, particularly for cancer treatment. For instance, the association between gene expression levels for breast cancer cells and patient prognoses \[49\] was found to provide a large improvement over the use of existing clinical markers \[50\]. The authors used a supervised learning algorithm, in which they selected highly-associated genes by iteratively adding and removing them from the classifier to optimise prognosis accuracy.

The large phenotype databases accrued by the eMERGE and RPGEH consortiums (see Section \[1.3\]) have also allowed for a new type of genomic-based analysis, sometimes called “phenome-
wide association studies” (PheWAS). Such studies 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 [11].

1.7.2 Machine learning with genomic data

The focus of much current research with GWAS is to better identify relevant marker SNPs. Machine-learning approaches have been employed for this purpose. Roshan et al., for instance, used chi-squared rankings to identify a subset of SNPs, ranked these SNPs using SVMs and RFs, and then selected the top-ranking SNPs as input into an LR prediction model [51]. To assess the stability of these predictions, the group computed the correlation coefficient between the rankings in jacknifed studies. They found that SNP rankings obtained with an SVM in this manner were the most accurate in simulated studies.

Wei et al. in 2013 reported on their ability to predict IBD risk based upon Immunochip genotyping [52]. Using a sample size of over 17,000 CD cases, together with over 22,000 control subjects, they performed a three-staged machine learning pipeline. In the first stage, 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 $< 0.01$. The second stage used $L^1$ penalised LR, with 10-fold cross-validation, to further select a subset of predictive SNPs. In the third stage, the authors tested the performance of the best-performing regularised classifier using the test set. The final sensitivity and specificity were 71% and 83%, respectively. The predictive performance using LR was similar to that obtained with a non-optimised radial basis SVM.

1.7.3 Machine learning for bacterial and viral genomics

While human genetic information is not yet available in most EHR and clinical decision systems, bacterial and viral DNA analysis is much more manageable (due to the smaller size of such genomes when compared with the human genome) and has already started to be incorporated into hospital systems. Machine learning techniques have been applied to predict bacterial and viral phenotypes from the genotype.

Prediction of viral drug resistance is a pressing problem for many viruses, such as Human

7The minor allele is that allele found less commonly in the population.
Immunodeficiency Virus (HIV). Both rule-based methods (e.g., ANRS, Rega, and Stanford HIVdb [53]) and machine-learning techniques (e.g., geno2pheno [54]) have been developed to improve genotypic prediction of HIV drug-susceptibility. Machine-learning methods have been found to predict more accurately the response of patients to drugs in retrospective analysis than do rule-based methods used for the same task [55]. Geno2pheno, a publically-available tool, uses a combination of various machine learning algorithms. As features, every codon within the protease and reverse transcriptase HIV genes are mapped onto 20 indicator variables (one for each amino acid), yielding a total of approximately 6,380 features [54]. Partially due to the success of these tools, genotypic susceptibility testing is now more commonly used than phenotypic testing for HIV [56].

Machine learning techniques have also been used to predict virulence profiles of clinically-relevant micro-organisms. In 2014, Laabei et al. used whole-genome data to predict the virulence of methicillin resistant S. aureus [57]. The authors performed several pre-processing steps to obtain the features (SNPs and indels) used for classification. First, they filtered the original 3,060 SNPs to include only those with both a minor allele frequency of > 5% and significantly associated with toxicity ($p < 0.05$, adjusted for genomic inflation), which left 122 SNPs/indels. They then removed those SNPs/indels in linkage disequilibrium to obtain a total of 51 features. Phenotypes were assessed by measuring the T-cell lysis and vesicle lysis (i.e., ability to destroy cellular machinery) activity of the 90 strains and splitting the isolates into three categorical levels. Using an RF model, the authors were able to predict correctly low- and highly-toxic strains (in a test set of 30 isolates, based upon a training set of 60 isolates), but were unable to distinguish medium-toxicity strains from highly-toxic strains.

1.8 Clustering analysis

While we often have labeled outcome data to describe different possible patient outcomes, in other cases the goal is to examine whether there are new, undetected groups of similar patients within the dataset. This type of exploratory analysis requires unsupervised learning methods. The latter have been used widely in the bioinformatics literature to cluster the output of gene expression data, but are less commonly used within the broader field of clinical informatics, 8Protease and reverse transcriptase are enzymes involved in HIV replication that are targets for many antiretroviral drugs.
using EHR data. We discuss below those various approaches that have been used both in genomics studies and clinical informatics studies.

1.8.1 Gene expression clustering

In 2000, Friedman et al. laid the groundwork for gene expression clustering when they constructed Bayesian networks associated with \textit{Saccharomyces cerevisiae} (yeast) gene expression data \cite{friedman2000bayesian}. The basic model assumes that there are clusters of related genes and clusters of similar experimental conditions (e.g., corresponding to the different physical microarray chips themselves), and that these together give rise to the pattern of expression seen in microarray experiments. As explained in a 2004 review \cite{friedman2004gene}, this work was extended to infer gene regulatory networks, which are the ways in which certain proteins regulate the expression of other gene products. It is assumed that genes with similar regulatory patterns will share similar expression patterns in various experimental conditions. Within the Bayesian network, this was modelled by incorporating an additional regulatory layer \cite{friedman2004gene}. To improve the interpretability and tractability of this modelling approach, constraints have been applied such as considering only a limited number of regulating proteins or assuming that a panel of regulatory proteins regulate genes in the same way \cite{friedman2004gene, friedman2000bayesian}. It is difficult to disentangle co-expression and regulation, and novel networks must always be confirmed experimentally. However, the graphical modelling approach for gene regulation has had promising results: the networks that have been inferred from yeast microarray experiments correspond well with networks already presented in the literature, and there are several instances of new findings being confirmed with gene-knockout experiments \cite{friedman2004gene}.

Microarray experiments may also be run for a set length of time, leading to time-series data. Methods to allow for inconsistent sampling across experimental replicates was developed through the use of hierarchical GPs by Hensman, et al. \cite{hensman2015gaussian}. Here, the underlying expression profile, $g_n(t)$, of a given gene during experiment $A$ is assumed to be drawn from a GP with a certain covariance function. The expression profile of this gene during a replicate trial of experiment $A$ is drawn from a GP with mean $g_n(t)$. To infer clusters, the authors add an additional layer into their hierarchical model defining the cluster, and they use an iterative EM-type algorithm to assign the genes different clusters. As illustrated using gene expression patterns of \textit{Drosophila melanogaster}, the fruit fly, this approach allows for the imputation of
both random and systematically-missing data.

1.8.2 Clustering of genomic clinical data

Gene expression clustering has also been used to examine whether particular biomarkers may be able to distinguish between disease conditions. For instance, Noble et al. looked at gene expression in CD patients and controls, using both univariate methods and unsupervised hierarchical clustering to identify common subgroups within the dataset [64]. They found several upregulated and downregulated genes in CD patients, in comparison to controls. Hierarchical clustering of the expression results according to the sex of the patient revealed a distinction between CD patients and controls, though the authors did not attempt a predictive analysis.

As mentioned previously in Section 1.7, gene expression patterns cluster according to breast cancer subtypes; van ’t Veer et al. performed an unsupervised hierarchical clustering of their data in addition to their supervised analysis [49].

The proteome has also been used for unsupervised hierarchical clustering by Han, et al. in a small pilot study of 9 IBD patients and 3 controls. CD patients and UC patients with active and inactive disease were consistently separated based upon their colon mucosal protein profiles, though it was unclear whether these results may be confounded by multiple samples being taken from each patient [65].

1.8.3 Clustering of EHR data

Chen et al. frame clinical EHR data in terms of a “clinarray,” using the analogy to microarrays to guide their analysis. For instance, they used time-averaged laboratory values for patients with cystic fibrosis and Crohn’s disease to perform a hierarchical clustering that differentiated between severe and less-severe patients (with severity measured by a surrogate marker: the frequency at which lab measures were taken) [66]. They also used independent component analysis (ICA) decomposition of clinarrays for a number of diseases to identify physiological factors in common, as indicated by the corresponding laboratory test measurements [67].

Though they did not use EHR data, Lewis et al. used $k$-means clustering to identify subtypes of the early progression of Parkinson’s disease [68]. Roque et al. mapped diagnostic codes onto a cohort of psychiatric patients using text mining and performed hierarchical clustering to define patient subgroups [69]. The authors then examined common comorbidities and, using a
systems biology approach, constructed protein networks based on possible co-involved proteins.

1.8.4 Clustering of time-series data

Clustering methods for time-series differ from those of static data in that the distance metric between two time-series is often less well-defined. Numerous such distance metrics have been proposed, including options such as the Euclidean distance, Pearson’s correlation factor, and dynamic time warping [70]. As categorised by Liao, there are three different approaches for clustering time-series data: using the raw time-series as input; using features extracted from the raw data; or by presuming an underlying model of the data. Time-series clustering has been less commonly explored in medical applications. The work of Pimentel, et al. and Saria, et al., presented in Section 1.5.2, provide a few examples of using vital-sign data and unsupervised clustering methods to examine possible underlying latent groups of patients [31] [32].

1.9 Combining different data sources

Several of the machine learning approaches described previously have drawn upon disparate sources of data. However, further work is needed to understand how to best combine larger scale genomic data with clinical data. For instance, Bayesian networks were used by Gevaert et al. in 2006 to combine microarray and clinical data together in a cohesive, probabilistic manner to predict breast cancer prognosis [71]. The authors compared methods for combining the datasets, finding that building separate networks for each dataset (i.e., microarray and clinical data) and then combining their overall prediction together with a weighting system proved to be the most predictive (with a resulting AUROC of 0.793). They modelled both datasets as multinomial distributions with Dirichlet priors, finding parameter estimates using maximum a posteriori (MAP), and obtaining the network structure through the K2 algorithm (a commonly-used approach for learning network structure). However, their risk prediction performance with the addition of clinical variables did not improve in comparison to prediction using only gene expression data.

Methods must also be developed for integrating genomic, clinical, and physiological time-series data within unsupervised analyses. Some such methods have already been proposed. Kirk et al., for example, used a Dirichlet-multinomial allocation (DMA) mixture model (a finite approximation to a Dirichlet Process mixture model) to integrate multiple datasets [72]. The
underlying idea is that the clustering within one dataset informs the clustering in other datasets, which the authors refer to as “correlated clustering.” The study used GP models (for time-series gene expression data) combined with multinomial models (for discrete gene expression data), with comparable performance to other clustering methods, but with the advantage of being able to incorporate more than two distinct data-types. Kirk et al. applied this method to identify genes with similar behaviour across yeast datasets, but such a method could also conceivably be used to identify clusters of patients with similar disease trajectories. Phan et al. provide a review of large collaborative research projects attempting to integrate genomic, clinical, and imaging data sources, with the presented applications primarily focused upon cancer [73].

1.10 Conclusions
The goal of our research in the field of clinical informatics is to better understand and predict patient disease progression, so as to better provide appropriate care. The EHR and genome sequencing both provide a growing source of information at a scale that has not previously been accessible. We have presented an overview of the approaches that have been used to analyse and interpret clinical patient data from a machine learning perspective. Current research involves the identification of useful features from clinical, laboratory, and genomic data sources, as well as more sophisticated methods of finding similarities across patients. Much work in the development of analysis methodology remains. For example, time-series analysis techniques have long-been applied to vital-sign data such as the ECG, but their application to changes in laboratory values over time for patients with chronic disease is still unexplored. Informative clustering techniques for identifying latent subgroups of patients, based upon both time-changing and static features, also need to be further developed.

As always, the appropriate analysis approach for a given dataset depends upon the clinical question at hand, but with increasing numbers of patients and increasing granularity of data, the types of questions that are possible to answer are only growing. This review has evaluated different approaches and methodologies for analysing this data, in an attempt to make it useful and informative for future patients.
2 Machine learning to predict drug susceptibility for *Mycobacterium tuberculosis*

2.1 Clinical challenge: prediction of drug resistance

Tuberculosis (TB), caused by the bacterium *Mycobacterium tuberculosis* (MTB), infects over one-third of the human population and claims over one million lives each year [74]. While TB caused by a drug-susceptible bacterium is completely curable through antibiotics, drug resistance is increasing worldwide. As warned in a 2014 WHO report, “drug-resistant TB threatens global TB control and is a major public health concern in several countries” [75]. Multi-drug resistance (MDR) is defined as being resistant to the most effective first-line antibiotics, which are isoniazid (INH) and rifampicin (RIF). MDR now accounts for about 3.6% of all new TB cases and 20.2% of all previously-treated cases worldwide. About ten percent of these MDR cases are extensively drug-resistant (XDR), which is defined as also resistant to two different classes of second-line drugs [75].

Determining the drug susceptibility profile, or antibiogram, of a new bacterial isolate is of paramount importance in order to prescribe appropriate drugs. Otherwise, the prescribed drugs will not cure the patient, the patient may develop further resistance, and the patient will go on spreading the (possibly now MDR) infection to others. Current methods for testing susceptibility include phenotypic and genotypic methods. Phenotypic methods involve growing the MTB isolate in media impregnated with antibiotics. The gold-standard phenotypic method is the “proportion method” on sloped Löwenstein-Jensen (LJ) solid media [76]. This method, performed in specialised reference labs, compares bacterial growth with and without the presence of an antibacterial drug. However, MTB’s slow growth-rate means that the LJ proportion method can require up to two months to obtain results.

Bacterial drug resistance arises due to mutations in the bacterial genome that enable it to avoid damage caused by the antibiotic. Several such resistance-conferring mechanisms are known, and genotypic line-probe assays have been developed to identify the presence of these known SNPs in a bacterial sample. The “MTBDRplus” is a line-probe assay created by Hain
(Germany), which tests for the primary mutations associated with resistance to INH and RIF. The Cepheid (USA) “Xpert” system is able to detect resistance to RIF within two hours. However, these methods only are available for a subset of antibiotics, require further testing to confirm their results, and only probe for the most common resistance-conferring mutations. A much more flexible approach lies in the incorporation of whole genome sequencing (WGS) into the clinical diagnostic pathway [74], which offers the opportunity to identify the presence of any known mutation in a bacterial sequence with a single assay.

However, some resistant bacterial isolates lack an established resistance-conferring mutation, suggesting that unknown mechanisms of resistance remain. Between 10%-20% of INH-resistant isolates, for instance, lack a mutation in a known resistance gene [77]. Furthermore, other isolates are phenotypically susceptible despite having an established mutation. It is possible that some of this discrepancy may be explained by epistatic interactions (i.e., two or more SNPs may be required to gain drug resistance) or because some of the established mutations do not actually cause resistance. Further analysis is needed in order to provide both improved predictive power for drug resistance and to identify new mechanisms of resistance. Machine learning provides a principled method for such analysis. We therefore set out to use a machine learning framework to predict MTB drug susceptibility to the four first-line MTB antibiotics, using whole genome sequencing data.

2.2 Methods

2.2.1 Phenotypic assessment

We obtained 1,835 frozen sputum samples from patients diagnosed with active TB from University Hospitals, Birmingham, UK; the Oxford Hospitals NHS Trust, UK; and clusters of cases from the surrounding regions. In England, all MTB infections must be reported to local authorities and samples are held within Public Health England facilities. These samples were made available under the umbrella of service-delivery assessment for the purposes of public health follow-up and protection, as is allowed under English law (Public Health Act 1984, 2003 Health Protection Agency Act, and 2002 Statutory Instrument 1438). Following standard UK clinical practice, bacterial MTB colonies were grown for 1-3 weeks from each sample. Phenotypic drug susceptibility for each drug was determined by performing an initial screen for resistance in
liquid culture, which was then confirmed using LJ methods. In this analysis, we examined the four common first-line drugs: INH, RIF, ethambutal (EMB), and pyrazinamide (PZA).

2.2.2 DNA sequencing

DNA from the bacterial cultures was extracted using the QuickGene DNA Tissue Kit S (Fujifilm, Japan). Standard lab and Illumina (USA) protocols were used to randomly fragment the DNA. Sequencing was performed with a HiSeq 2000 (Illumina). MTB has a very stable genome, meaning that most of its variation is due to chromosomal point mutations, rather than through plasmids moving in or out of the bacteria. As such, we mapped sequencing reads to the reference MTB strain H37Rv (NCBI reference sequence NC_000962.2) using the commonly-employed software Stampy to obtain an aligned sequence for each isolate. Nucleotide bases were determined (or “called”) using standard filters, which take into account the sequencing and alignment quality, as well as the number of reads for each base. After these steps, the nucleotide bases at some locations cannot be called with confidence due to low sequencing or alignment quality (i.e. “null calls”). We considered these null calls to be SNPs if the base with the highest number of reads did not agree with the reference. For example, if the reference base at a given position was adenine (A), and there were 100 mapped reads with an A at this position and 20 mapped reads with thymine (T) at this position, we would not consider this to be a SNP.

2.2.3 Features

All SNPs found within those 23 genes suspected to be involved in resistance mechanisms (a representative selection of which are listed in Table 2.1) and their 100 base-pair upstream regions were identified. Upstream regions were included so as to capture SNPs that may potentially be involved in gene regulation. The resulting set of 300 SNPs constituted the feature set for subsequent analysis. These features were binary variables indicating the presence (= 1) or absence (= 0) of the corresponding SNP in the isolate. The average number of SNPs per isolate was 5.0, ranging between 0 and 23.

We also performed analyses in which we limited the included features to: (i) polymorphisms

1 Plasmids are small circular rings of DNA that are able to replicate independently from a cell’s chromosomal DNA. They often carry drug-resistant and other fitness-enhancing genes.
2 At any given position in the reference genome, there may be hundreds of small pieces of sequenced DNA that map to this location; each of these short 100-300 base pair sequences is called a “read”.

21
Table 2.1: A selection of genes suspected to be involved in resistance mechanisms. Starred genes contain specific loci previously documented in the literature as being associated with drug resistance. Chart compiled primarily from [78, 79].

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Relevant drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>ahpC*</td>
<td>Oxidative stress</td>
<td>INH</td>
</tr>
<tr>
<td>eis*</td>
<td>Cell surface involvement</td>
<td>Aminoglycosides</td>
</tr>
<tr>
<td>embB*</td>
<td>Cell wall bisynthesis</td>
<td>EMB</td>
</tr>
<tr>
<td>gyrA*</td>
<td>Enzyme for DNA coiling</td>
<td>Fluoroquinolones</td>
</tr>
<tr>
<td>inhA</td>
<td>Fatty acid biosynthesis</td>
<td>INH</td>
</tr>
<tr>
<td>iniA*</td>
<td>Likely transmembrane protein</td>
<td>EMB, INH</td>
</tr>
<tr>
<td>pncA*</td>
<td>Intermediary metabolism</td>
<td>PZA</td>
</tr>
<tr>
<td>rmlD</td>
<td>Sugar biosynthesis</td>
<td>EMB</td>
</tr>
<tr>
<td>rpoB*</td>
<td>Transcriptional enzyme</td>
<td>RIF</td>
</tr>
<tr>
<td>thyA</td>
<td>Virulence; methylation</td>
<td>Aminoglycosides</td>
</tr>
</tbody>
</table>

- found on genes already thought to be involved in resistance for a given drug; (ii) polymorphisms that were already suspected to confer drug resistance; and (iii) polymorphisms that were not previously suspected to confer drug resistance.

2.3 Classifiers

We assessed several different supervised classification algorithms for the prediction of isolates as being susceptible or resistant to each of the four first-line drugs. This comparison allowed us to understand how well the assumptions of each (e.g., linear combinations of features; independent features) were substantiated in the data. We examined LR, SVM, RF, and BPM classifiers.

We will consider a subset of $N$ isolates to represent a training set of examples $x_1...x_N$ with labels $\ell_1...\ell_N$, $\ell \in \{0, 1\}$, with 1 indicating drug resistance for a given drug and 0 indicating susceptibility. Each example $x_i$ is composed of a vector of $D$ binary features indicating the presence ($x_{ij} = 1$) or absence ($x_{ij} = 0$) of a given SNP.

2.3.1 Logistic regression

LR is a linear classification method that optimises a set of weights $w$ 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:

$$P(\ell_n = 1|x_n, w) = h(x_n) = \frac{1}{1 + e^{-w^T x_n}} \quad (2.1)$$

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 $h(x_n)$ is greater than or less than a set threshold.
We define a cost function that includes a penalty when the hypothesis is incorrect:

\[ f(w) = -\frac{1}{N} \sum_{n=1}^{N} \left[ \ell_n \log(h(x_n)) + (1 - \ell_n) \log(1 - h(x_n)) \right] \]  

(2.2)

Adding an \( L^2 \) regularisation term to discourage the weights from overfitting the data by penalising large values in \( w \), the final equation to be minimised is:

\[ f(w)_R = f(w) + \lambda \frac{\lambda}{2N} \sum_{i=1}^{D} w_j^2 \]  

(2.3)

where \( \lambda \) is an adjustable parameter 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 \( L^1 \) norm \( ||w|| = \sum_i |w_i| \) does not exceed some threshold value. From a Bayesian perspective, this is equivalent to putting a zero-mean Laplace prior on the feature weightings, meaning that the prior assumption is that the feature is not important until the training data shows otherwise.

### 2.3.2 Support vector machine

The SVM is a classification algorithm that attempts to separate two groups by the widest margin possible in some feature space. The 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 a support vector machine begins by considering the distance of each training example \( x \), from the hyperplane \( y(x_n) = w^T x_n + b \), where \( b \) is a bias parameter and \( w \) is again a vector of weights. This distance is written as \( \frac{|y_n(x_n)|}{||w||} \). This is subject to the constraint that \( y_n(w^T x_n + b) \geq 1 \) 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, \( \xi_n \), and a cost parameter, \( C \). The slack variable \( \xi_n = 0 \) if the example datapoint \( x_n \) lies on or within its correct boundary, and \( \xi_n = |\ell_n - y(x_n)| \) otherwise. The parameter \( C \) penalises misclassified examples. \( C \) is analogous to a regularisation parameter in that lower values of \( C \) correspond to more slowly-changing decision boundaries.
(because misclassifications are not penalised heavily), and vice-versa. The constraint is therefore \( \ell_n y(x_n) \geq 1 - \xi_n \). As the goal is to maximise the distance between the hyperplane and the closest training example, which requires maximising \( ||w||^{-1} \), this is equivalent to minimising \( ||w||^2 \), where the square is introduced to avoid taking the root in \( ||w|| \). This therefore requires the minimisation of \( f(w) = C \sum_{n=1}^{N} \xi_n + \frac{1}{2} ||w||^2 \), 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:

\[
f(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{j,k=1}^{N} \alpha_j \alpha_k \ell_j \ell_k x_j^T x_k, \quad 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i \ell_i = 0 \quad (2.4)
\]

where \( \alpha \) is another vector of weights, with \( w = \sum_{i=1}^{N} \alpha_i \ell_i x_i \). 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 \( x_n \), a kernel function that describes the features of each example in relation to each other is used instead. Through Mercer’s theorem, any positive semi-definite kernel function corresponds to a high-dimensional space, for which \( k(x, x') = \phi(x) \phi(x') \) and where \( \phi(x) \) 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. 2.4 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:

\[
k(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right) \quad (2.5)
\]

### 2.3.3 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, each of which is formed from a random subset of the available features and a random subset of the available training set. After all of the trees have been built, the classifier’s prediction is based upon majority-voting of the trees. For problems involving genomic loci as features, building

\[^3\text{In deriving the dual form from the Lagrangian, } \alpha \text{ is the vector of Lagrange multipliers.}\]
40 – 400 trees and using a random selection of half of the features has been found to be a suitable means of initialising the various parameters \[80\].

2.3.4 Product of marginals

Unlike the classifiers described previously, BPM is a generative model, and it assumes that all features are independent. This assumption rarely holds, but BPM nevertheless often produces good classification performance. With our training examples \( X = \{x_n\}_{n=1...N} \), \( \pi \) as the prior probability of being in a class, \( \theta_j \) as the probability that a given feature is “on” (i.e., it has value 1), and again taking \( \ell_1...\ell_N \), \( \ell \in \{0, 1\} \), as labels, we have \( p(x_i, \ell_i) = p(\ell_i|\pi) \prod_j p(x_{ij}|\theta_j) \). Including both classes and taking the logarithm, we obtain

\[
\log p(X|\theta) = \sum_c N_c \log \pi_c + \sum_{j=1}^D \sum_c \sum_{i:1(y_i=c)} \log p(x_{ij}|\theta_{jc})
\]

with \( 1 \) as the indicator function and \( N_c = \sum_i 1(y_i = c) \); i.e., \( N_c \) is the number of examples in class \( c \). We could then use a maximum likelihood estimate (which would involve differentiation of the log likelihood, introduction of Lagrange multipliers, and solving for the parameters), but here we will instead obtain full distributions over the model parameters by introducing a set of prior distributions\[4\]. We will place a \( Beta(\beta_0, \beta_1) \) prior over each \( \theta_{jc} \) and a \( Dirichlet(\alpha) \) prior for each \( \pi \). This then leaves us with

\[
p(\theta|X) = p(\pi|X) \prod_j \prod_c p(\theta_{jc}|X)
\]

where \( p(\pi|X) = Dirichlet(N_1 + \alpha_1, ... N_c + \alpha_c) \), \( p(\theta_{jc}|X) = Beta(a, b) \), \( a = N_{jc} + \beta_0 \), and \( b = N_c - N_{jc} + \beta_1 \).

The BPM approach provides a probability distribution over the probability that an isolate in class \( c \) has the given SNP. After training, predictions are made on new data by calculating

\[4\]The cumulative distribution function (cdf) is the probability that a given random variable will have a value less than or equal to \( x \), while the probability density function (pdf) is the relative likelihood that a given random variable will take on the value \( x \). The cdf may also be called the “distribution function,” while the pdf may also be called the “density.” However, the machine learning literature tends to use the word “distribution” to refer to the pdf, as we will in this report.

\[5\]The Beta distribution provides a distribution over the interval \([0, 1]\). It is parameterised by \( a > 0 \) and \( b > 0 \), which determine the distribution’s shape. The Dirichlet distribution is distributed over \( K \) random variables and has a single \( K \)-dimensional parameter, \( \alpha > 0 \). The beta distribution is a special case of the Dirichlet distribution, in which \( K = 2 \).
the probability of the class label, given the new example and the training data. The class label with the highest probability is the final prediction. This probability is formulated as

\[ p(\ell = c|x, X) \propto p(\ell = c|X) \prod_{j=1}^{D} p(x_j|y = c, X). \]

Expanding out, this becomes

\[
p(y = c|x, X) \propto \frac{N_c + \alpha_c}{N + \alpha_0} \prod_{j=1}^{D} \bar{\theta}_{jc}^{x_j=1}(1 - \bar{\theta}_{jc}^{x_j=0})
\]

with \( \bar{\theta}_{jc} \) being the mean value of the fitted parameter distribution, equal to \( \frac{N_{jc} + \beta_0}{N_c + \beta_0 + \beta_1} \).

We used a U-shaped Beta prior, \( \text{Beta}(0.5, 0.5) \), for every \( \theta_{jc} \) except for the established SNPs. For these, we used a \( \text{Beta}(1, 0.25) \) prior for the resistant class, which shifts the prior distribution towards \( \theta_{jc} = 1 \), and a \( \text{Beta}(0.25, 1) \) prior for the susceptible class, which shifts the prior distribution towards \( \theta_{jc} = 0 \), as is illustrated in Fig. 2.1 (A). We used a uniform Dirichlet distribution as a prior over each class.

![Figure 2.1](A) Illustration of the prior distributions over the parameters governing features (SNPs). Established SNPs were given a \( \text{Beta}(1, 0.25) \) prior for isolates in the resistant class and a \( \text{Beta}(0.25, 1) \) prior for isolates in the susceptible class. All other SNPs were given a \( \text{Beta}(0.5, 0.5) \) prior. (B) Phenotypic labels of resistance and susceptibility to each drug. The dataset is composed primarily of susceptible isolates. As is illustrated, because our dataset comes from the clinical record, second-line drugs are typically only tested for resistance if the isolate is already found to be resistant to a first-line drug.

### 2.3.5 Direct association method

For comparison, we implemented a baseline algorithm, which we term the Direct Association (DA) method. This method uses prior clinical knowledge and essentially represents the best predictive performance that could be obtained based upon those clinical associations already
identified in the literature. Drawing upon the list of established Hain mutations and a database of MTB mutations\cite{78,79}, we assembled a list of mutations that have been previously associated with resistance in clinical and experimental studies. The loci contained in this list of “established” mutations correspond to locations within the starred genes in Table 2.1. We applied an “OR” rule: if any of the established mutations was present for a given isolate, the isolate was classified as being resistant to that drug.

2.4 Machine learning pipeline

Fig. 2.1 (B) illustrates the pattern of phenotypic drug resistance found within the dataset. There were many more susceptible isolates than resistant isolates; of the 1,835 isolates, only 266, 97, 47, and 59 isolates were resistant to INH, RIF, EMB, and PZA, respectively. Therefore, to assemble a balanced dataset for training a classifier, such that both “resistant” and “susceptible” classes were equally represented (to avoid bias in the classifier), we randomly selected a subset of susceptible isolates equal to the number of resistant isolates (e.g., for INH analysis, we selected 266 susceptible isolates). We then trained each algorithm on 80% of this balanced dataset and tested on the held-out 20%. During training, we used internal five-fold cross-validation to determine $\sigma$, the SVM RBF width (see Eq. 2.5), $C$, the SVM soft margin parameter (see Eq. 2.4), and $\lambda$, the LR regularisation parameter (see Eq. 2.3). 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. This process was repeated $M = 100$ times: random samplings from the pool of susceptibility examples were performed $M$ times. To provide a fair comparison, we also assessed the performance of the Direct Association method on the test set with each of the $M$ iterations. The mean and standard deviation of the accuracy, sensitivity, and specificity of each method was subsequently calculated across the $M$ iterations, allowing an assessment of the variation in the process due to the stochastic selection of the training and test data.

2.4.1 Investigation of biomarkers

Isolates were then identified that were classified as being susceptible by the Direct Association method, despite being phenotypically resistant; i.e., they were “false negatives” because they
had none of the known resistance mutations. Since the resistance of these false-negative isolates could not be explained by the known resistance mutations, the predictive capability of any new SNPs found within this group was investigated. “Promising” new mutations were defined as those (a) found in at least two isolates resistant to the given drug, (b) found in at least one isolate lacking any alternative established resistance-conferring mutation, and (c) having a positive predictive value (PPV) of 1.0 (i.e., every isolate with the SNP was drug-resistant).

2.5 Results

2.5.1 Direct Association results

Confusion matrices for the DA prediction method, using the entire dataset, are shown in Table 2.2. In general, the DA predictions had high specificity, with 100%, 100%, 97%, and 99% for INH, RIF, EMB, and PZA, respectively. With the large number of susceptible isolates in the dataset, this resulted in high overall accuracy. Among the 1,808 isolates tested for INH resistance, for instance, overall accuracy of the DA method was 98%. For most drugs this came with a trade-off in sensitivity, however, with 89%, 86%, 96%, and 37% sensitivity for INH, RIF, EMB, and PZA, respectively.

**Table 2.2: Direct Association confusion table. The counts of isolates that were misclassified are shown in bold. Phenotype is the gold-standard comparison method; DA refers to the Direct Association method; R and S correspond to resistant and susceptible, respectively.**

(a) INH  (b) RIF  (c) EMB  (d) INH

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>DA R</th>
<th>DA S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>Phenotype</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>Phenotype</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>Phenotype</td>
<td>R</td>
<td>S</td>
</tr>
</tbody>
</table>

2.5.2 Prediction comparison

Prediction results using the baseline feature set (i.e., all SNPs found across the 23 genes of interest) are shown in Table 2.3. Results are shown as averages across the N = 100 balanced experiments. As is evident, machine learning prediction achieves improved predictive performance for all drugs except for EMB. There was similar performance across all examined classifiers; no single classifier consistently outperformed the others.
Table 2.3: Drug resistance predictive performance with the baseline feature set. Means across the $M = 100$ experiments are presented, with the standard deviation in parentheses. Predictive accuracies using machine learning methods that are significantly different from the performance using the DA method are starred (*). (Mann-Whitney $U$ test at the $\alpha = 0.01$ level, Bonferroni-corrected by six comparisons). Note, the difference may be either in a positive or negative direction.

<table>
<thead>
<tr>
<th></th>
<th>INH</th>
<th></th>
<th>RIF</th>
<th></th>
<th>EMB</th>
<th></th>
<th>PZA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>DA</td>
<td>94</td>
<td>89</td>
<td>99</td>
<td>92</td>
<td>85</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(1)</td>
<td>(3)</td>
<td>(7)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>LR</td>
<td>95</td>
<td>93</td>
<td>98</td>
<td>94</td>
<td>92</td>
<td>96</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>(2)*</td>
<td>(3)</td>
<td>(2)</td>
<td>(4)</td>
<td>(6)</td>
<td>(5)</td>
<td>(7)</td>
</tr>
<tr>
<td>LR L*</td>
<td>96</td>
<td>93</td>
<td>98</td>
<td>95</td>
<td>94</td>
<td>97</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>(2)*</td>
<td>(4)</td>
<td>(2)</td>
<td>(3)</td>
<td>(5)</td>
<td>(3)</td>
<td>(5)</td>
</tr>
<tr>
<td>SVM</td>
<td>96</td>
<td>97</td>
<td>95</td>
<td>95</td>
<td>96</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>(2)*</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(7)</td>
</tr>
<tr>
<td>RF</td>
<td>96</td>
<td>97</td>
<td>96</td>
<td>94</td>
<td>94</td>
<td>94</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>(2)*</td>
<td>(3)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>BPM</td>
<td>94</td>
<td>92</td>
<td>96</td>
<td>91</td>
<td>91</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(3)</td>
<td>(4)</td>
<td>(6)</td>
<td>(6)</td>
<td>(6)</td>
</tr>
</tbody>
</table>

2.5.3 Feature comparison

We now present results according to the use of different subsets of features. The accuracy of the single best-performing classifier for each feature subset is shown in Table 2.4. When features were further limited to just the established mutations (row 3 in the table) or the limited subset of associated genes (row 4), predictive performance generally improved in comparison with the performance of the baseline feature set. This illustrates that the predictive SNPs are primarily captured within these already-established regions (PZA is a notable exception).

However, when only those SNPs that were not previously established to cause resistance were used as features (row 5 in the table), predictive accuracy is still high: we observed 84% predictive accuracy for PZA (compared to 70% obtained with the DA approach). This illustrates the possible presence of previously unidentified resistance-conferring mutations, as well as the fact that phylogenetic relationships between isolates can provide predictive utility through linkage disequilibrium. The results obtained when the input examples are reduced to susceptible isolates and only those resistant isolates that lack any established mutation (row 6) provide consistent findings. Here, though the accuracy of resistance prediction decreases, there is still clearly some predictive information that remains in the genome.

2.5.4 Probability distribution over features

The BPM models produce a distribution over the probability of a given feature (i.e., SNP) being present within the resistant and susceptible classes. The probability distribution of an isolate having a mutation at katG_*315*, an established resistance-conferring mutation for INH, is shown in Fig. 2.2. As is apparent, the probability of this mutation being present in resistant
Table 2.4: Comparison of feature performance, with accuracy and standard deviation in parentheses. 1) DA: Direct Association method; 2) Baseline feature set: all SNPs found across 23 genes; 3) Established mutations only: including only the established mutations as features; 4) Associated genes only: including as features all SNPs found across the genes previously associated with resistance for the given drug; 5) All "new" SNPs: all SNPs found across 23 genes, except for the established SNPs; 6) Limited sample: same feature set as (5), but only including phenotypically resistant isolates lacking any established resistance mechanism in the resistant class.

<table>
<thead>
<tr>
<th></th>
<th>INH</th>
<th>Rif</th>
<th>EMB</th>
<th>PZA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DA</td>
<td>94</td>
<td>92</td>
<td>98</td>
<td>70</td>
</tr>
<tr>
<td>2. Baseline</td>
<td>96</td>
<td>95</td>
<td>95</td>
<td>89</td>
</tr>
<tr>
<td>3. Est. muts only</td>
<td>96</td>
<td>97</td>
<td>98</td>
<td>89</td>
</tr>
<tr>
<td>4. Assoc. genes only</td>
<td>95</td>
<td>97</td>
<td>97</td>
<td>88</td>
</tr>
<tr>
<td>5. All &quot;new&quot; SNPs</td>
<td>92</td>
<td>88</td>
<td>88</td>
<td>84</td>
</tr>
<tr>
<td>6. Limited sample</td>
<td>90</td>
<td>90</td>
<td>81</td>
<td>81</td>
</tr>
</tbody>
</table>

Figure 2.2: (A): Probability distributions illustrating the probability of isolates having the katG_*315* mutation, which confers INH resistance. (B): Probability distributions for pncA_*35*, which is an established resistance mutation for PZA. (C): Probability distributions for pncA_*125*, a promising new mutation (as defined in Sec. 2.4.1) that could confer PZA resistance.

isolates is very high, whereas it is nearly zero in susceptible isolates. This is in contrast to the probability distribution of having a mutation at pncA_*35*, which is an “established” mutation conferring resistance to PZA. In this dataset, contrary to expectation, two susceptible isolates had this SNP and it was not found in any resistant isolates. The probability distributions that arise through the BPM model provide easy visualisation of these comparisons and the degree of confidence in the estimates, even for small sample sizes.

2.5.5 Promising mutations

Across all drugs, 30 SNPs fit our definition of “promising” new mutations (see Section 2.4.1). Of these 30 SNPs identified, a subset are presented in Table 2.5. These may constitute new markers for resistance. The probability distribution of the SNP pncA_*125*, for instance, is shown in Fig. 2.2. There is more support in the data for this SNP as a resistance mutation than for other “established” clinical mutations.
Table 2.5: Promising new resistance-conferring mutations, with the number of false negative isolates harboring the mutation (n). † indicates that, in addition to the noted number of FN isolates with the mutation, an additional resistant isolate containing an established mutation also shared the SNP.

<table>
<thead>
<tr>
<th>INH SNP</th>
<th>fabG1_<em>-17</em></th>
<th>katG_<em>1899</em></th>
<th>inhA_<em>94</em></th>
<th>pncA_<em>194</em></th>
<th>pncA_<em>94</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>2†</td>
<td>2†</td>
<td>1†</td>
<td></td>
</tr>
<tr>
<td>RIF SNP</td>
<td>rpoB_<em>1290</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PZA SNP</td>
<td>pncA_<em>-12</em></td>
<td>pncA_<em>125</em></td>
<td>pncA_<em>309</em></td>
<td>pncA_<em>4</em></td>
<td>pncA_<em>7</em></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

2.6 Discussion

Our first aim was to evaluate whether machine learning approaches can be used to identify drug-resistant MTB isolates, based upon their genome. We found that for isolates resistant to the four first line drugs, we were able to predict resistance with high accuracy using machine learning methods - ranging from 89% to 96% accuracy on balanced datasets. Also promising was the fact that we obtained higher predictive accuracy using our methods than the DA method, the currently-available clinical alternative method of using established mutations for isolates resistant to INH, RIF, and PZA.

High predictive accuracy of the methods described in this chapter can be due to mechanistic mutations (i.e. SNPs that cause a functional change to a protein, which results in drug resistance), or to SNPs that confer no resistance, but which for some reason are enriched in resistant isolates. Improved prediction upon the DA method could be therefore due to the discovery of new resistance-conferring mutations, epistatic interactions between mutations that together cause resistance, phylogenetic associations, or the fact that isolates are commonly resistant to multiple drugs.

2.6.1 Interpretation of model performance

Some of the possibilities described above can be assessed based upon the comparison of classifier performance. For example, the consistently lower performance of the BPM classifier suggests that its assumption of feature independence is not valid (as anticipated), which also suggests that the other classifiers may be taking advantage of the inherited phylogenetic similarities in feature patterns. The similar levels of performance between LR and the nonlinear SVM and RF classifiers suggests that these interactions are probably linear. Yet, the improvements in
classifier predictive performance for INH, RIF, and EMB, after limiting the input features to the genes associated with resistance, suggests that the established mutations constitute the majority of the resistance in the dataset.

The improvement in PZA predictive accuracy from 70% with the DA method to 89% using our methods illustrates the current limited clinical understanding of PZA resistance mechanisms. As is shown by the distribution in Fig. 2.2(B), there is little evidence in this dataset for many of the established mutations. We examined the highest-weighted features of our machine-learning classifiers and often found that these features were established mutations causing resistance to other drugs. Models for estimating resistance to a particular drug could be built based upon the strong signals associated with resistance to other drugs. However, when analysis was limited to the pncA gene (associated only with PZA resistance), the accuracy of prediction remained high, indicating that improved prediction performance for PZA is not due merely to associations with other drugs’ resistance. It appears that there could be some interactions between the various mutations of this single gene.

The limited number of resistant isolates lacking any established mutation in this dataset (e.g., only 37 PZA-resistant isolates are available) makes it difficult to investigate new mechanisms of resistance with a machine-learning approach. Nevertheless, the strong predictive performance overall illustrates that the classifiers are able to identify the major resistance-conferring mutations, which holds promise for capturing the impact of new mechanisms as more data become available.

2.6.2 Biomarker identification

While a machine learning approach allows for the possibility of epistatic interactions across the input set of SNPs, we have also described an analysis to search for possible mutations that could be sufficient to cause resistance on their own. Our search for these “promising” mutations (see Table 2.5) yielded several new SNPs that could be resistance-conferring. The fact that several of them are found in common across numerous resistant isolates that do not have any other known mechanism for resistance is highly encouraging. These mutations will have to be validated through additional experimental analysis, but if found to cause drug-resistance, they could be used in the future in fast genomic screening tools, similar as the Hain line probe assays.
3 Unsupervised clustering to identify sub-types of Inflammatory Bowel Disease

3.1 The clinical need

Inflammatory bowel disease (IBD) is a chronic disease marked by inflammation of the gastrointestinal (GI) tract - primarily the small and large intestine. IBD consists of two sub-disorders: Crohn’s disease (CD) and ulcerative colitis (UC). 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 GI tract for CD. An illustration of the different GI areas that may be involved is shown in Fig. 3.1. UC inflammation is typically contiguous, whereas inflammation may occur in discontinuous regions in CD. Classically, CD patients tend to exhibit transmural inflammation (through the entire intestinal wall), while UC is more marked by mucosal ulceration [81]. Cases in which there is inflammation of the colon, but not the small intestine, and in which there is no further definitive histological evidence, are termed to be “IBD - undetermined” (IBDu) [82]. Approximately 620,000 patients in the UK have IBD, with estimates of healthcare costs per patient ranging from £631 to £3,000 per year [83]. The majority of these costs are due to in-hospital management of episodes of acute disease flare-up, illustrating the importance of better understanding of how to predict and mitigate these events [83].

Figure 3.1: Illustration of the physical areas where IBD patients may exhibit inflammation.

The etiology of CD and UC remains to be further elucidated, but it is apparent that both genetic and environmental factors play a role. 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. One of the most widely-used classifications is the Montreal classification (MC), which is based upon the disease phenotype. The categorisations for CD and UC are outlined in Tables 3.1 and 3.2 respectively (drawn from [82]).

Table 3.1: Crohn’s Disease Montreal classifications. Starred (*) criteria are used as modifiers for the previous descriptions (e.g., A patient with ileal and upper alimentary canal inflammation would be classified as “L1L4.”)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Code</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis</td>
<td>A1</td>
<td>less than 17 y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>17 to 40 y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>greater than 40 y</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>L1</td>
<td>ileal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>colonic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>ileocolonic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>upper alimentary canal</td>
<td>*</td>
</tr>
<tr>
<td>Behavior</td>
<td>B1</td>
<td>non-stricturing, non-penetrating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>stricturing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>penetrating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>perianal</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Ulcerative Colitis Montreal Classifications

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Code</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of disease</td>
<td>E1</td>
<td>Ulcerative proctitis</td>
<td>Only rectum</td>
</tr>
<tr>
<td></td>
<td>E2</td>
<td>Distal/left-sided UC</td>
<td>Part of colorectum distal to splenic flexure</td>
</tr>
<tr>
<td></td>
<td>E3</td>
<td>Pancolitis/extensive UC</td>
<td>Proximal to splenic flexure</td>
</tr>
<tr>
<td>Severity</td>
<td>S0</td>
<td>Clinical remission</td>
<td>No symptoms</td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>Mild UC</td>
<td>Passage of less than five stools/day; no systemic illness; normal inflammatory markers</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>Moderate UC</td>
<td>Passage of at least five stools/day; limited systemic toxicity</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>Severe UC</td>
<td>Passage of at least six bloody stools/day; pulse of ≥ 90 bpm; temperature ≥ 37.5°C; hemoglobin ≤ 10.5g/100mL; ESR ≥ 30 mm/h</td>
</tr>
</tbody>
</table>

The MC was established by a working party composed of a group of experts. Though the collective experience of the group was considerable, the resulting classification was not based upon a data-driven examination of patient characteristics. It is possible that unrealised patterns in inflammation may exist, which only become apparent when looking across larger patient cohorts. Unsupervised clustering analysis provides a set of tools for conducting such investigation. Previous clustering analysis for IBD has focused primarily upon diagnosis and distinguishing between CD and UC. Data sources used in such analyses have ranged from gene expression data and proteomics to exhaled gases [64] [65].

However, given the highly heterogeneous nature of patient response to therapy and the wide
variation in severity, it seems likely that there are additional, latent subtypes of IBD even within the two major clinical classifications of CD and UC. The possible evidence of such subtypes has not previously been examined. Whether the extent of inflammation in IBD patients exhibits wider ranges of variation than currently captured by the MC scheme is an open question. As such, we performed an unsupervised machine-learning investigation based upon the locations of inflammation in CD and UC patients.

3.2 Machine learning methodology

3.2.1 IBD Clinical Dataset

1,268 patients diagnosed with IBD who have been treated at the John Radcliffe (JR) Hospital, Oxford, were included in this study. All patients consented to the use of their records and tissue samples for research purposes (REC 09/H2104/30). 1,065 patients were diagnosed with CD; 1,037 with UC; and 126 with IBDu. We extracted information from the clinical notes (electronic or paper) to create a database of information for each patient. Such characteristics included gender, diagnosis, age of diagnosis, inflamed areas across the GI tract, previous and current medications prescribed, etc.

Our primary variable of interest was the location of inflammation. Each patient was annotated with the presence (= 1) or absence (= 0) of inflammation in each of the following areas: mouth, esophagus, duodenum, ileum, caecum, ascending colon, transverse colon, descending colon, sigmoid colon, rectum, anus, and pouch (as annotated in Fig 3.1). Pouch refers to a surgically-created pouch, located in the GI, in an effort to relieve symptoms. Patients are examined endoscopically as part of their diagnosis, which means that these inflammation variables are well-defined. We excluded patients without any indicated inflammation from our analysis: as intestinal inflammation is the defining feature of IBD, an absence of inflammation in the record most likely indicates that the full patient record was unable to be extracted, rather than that the patient truly did not have any inflammation.

3.2.2 Bernoulli Mixture Model

With the goal of uncovering latent groupings of patients with similar patterns of inflammation, we employed a mixture model approach. Our variable of interest was the binary presence or
absence of inflammation, and so we modelled our dataset using a Bernoulli mixture model (BMM). Given a set number of $K$ latent groups, a BMM assumes that all data are derived from a mixture of these $K$ underlying multivariate Bernoulli distributions.

A given patient’s vector of data $x$ with dimension $D$ can be modelled with a single multivariate Bernoulli as $p(x|\mu) = \prod_{i=1}^{D} \mu_i^{x_i}(1 - \mu_i)^{(1-x_i)}$. The values $\mu_i$ are the probability of the feature being “on” ($= 1$). A mixture of Bernoullis for the entire set of $N$ patients, with $X = x_1...x_N$ and $\pi = \pi_k, k = 1...K$, is then formulated as

$$p(X|u, \pi) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(x_n|\mu_k)$$ (3.1)

with $\pi_k$ being the mixture coefficient, defining the probability of belonging to group $k$. Taking the logarithm of the above, the optimal choice of values for $\mu$ and $\pi$ cannot be derived in closed form due to the summation inside the logarithm. We therefore introduce a latent variable $z$ to define the assignment of each patient $x_n$ into group $k$ and solve using the Expectation-Maximisation (EM) algorithm. Introducing the latent variable $Z = z_{nk}, n = 1...N, k = 1...K$, taking the logarithm, and simplifying, produces the complete-data log-likelihood:

$$\log p(X, Z|\mu, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[ \log \pi_k + \sum_{i=1}^{D} x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \right]$$ (3.2)

Our goal is then to take the expectation with respect to the latent variable $Z$’s posterior distribution; this is the expectation step. The expectation of $z_{nk}$, which we define as $\gamma(z_{nk})$, is also termed the “responsibility,” it is the posterior probability of component $k$, given $x_n$. Calculating the expectation therefore involves calculating $E[z_{nk}]$, which is

$$\gamma(z_{nk}) = E[z_{nk}] = \frac{\pi_k p(x_n|\mu_k)}{\sum_{j} \pi_j p(x_n|\mu_j)}$$ (3.3)

We then perform the maximisation step: we use the newly-calculated expected values of $z_{nk}$ to maximise the complete-data log likelihood, in order to obtain the optimal values for $\mu$ and $\pi$. We therefore take the derivative of the complete-data log-likelihood in terms of these parameters. For $\pi$, we also must enforce the constraint that all $\pi_k$ values must sum to one, and so we use Lagrange multipliers. After setting the derivatives to zero and simplifying, we obtain:
\[
\mu_k = \frac{\sum_N \gamma(z_{nk})x_n}{\sum_N \gamma(z_{nk})}, \quad \pi_k = \frac{\sum_N \gamma(z_{nk})}{N}
\] (3.4)

We then iteratively alternate between the expectation and maximisation steps with the updated parameters until we reach convergence. We defined convergence as a less than a 0.01% proportional change to the log likelihood. This produces a final set of parameters for each group: \(\mu\) (size \(K \times D\)) describing the probability of each feature being “on” (= 1) in the group, and \(\pi\) (size \(N \times K\)) describing the probability of each patient belonging to each group.

### 3.2.3 Model selection

We do not know the number of groups \textit{a priori}, and since this is an unsupervised problem we had no “gold standard” labels. We therefore used the Akaike Information criterion (AIC) and the Bayesian Information criterion (BIC) in our selection of the number of groups. 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). The BIC and AIC allow us to estimate this model posterior probability by implementing a penalty term for increasing the number of clusters.

The BIC approximates model selection based upon Bayes factors, which are ratios of the posterior odds of the models. It asymptotically selects the model with the “best” structure, assuming the best model is in the pool of candidates. 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 Kullback discrepancy between the true model and the approximating model; it asymptotically selects the model within the 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 criterion.

### 3.2.4 Analysis outline

We performed unsupervised BMM clustering within each IBD sub-disorder. Our features were the binary presence or absence of inflammation within the \(D = 12\) GI regions. We varied the number of groups from \(K = 2\) to 10. The EM algorithm is guaranteed to always reach a local minimum, but it may not reach the global minimum, depending upon the initialisation. We
therefore randomly initialised the the values of the parameters ten times for each group size and calculated the lowest AIC and BIC across these ten iterations. The group size with the lowest AIC was selected as being the optimal number of groups.

We assigned patients to a latent group if the patient had a $\geq m\%$ probability of belonging to the given group, with $m$ being a cutoff variable. Any patients with a $< m\%$ probability of belonging to a group were noted as “poorly defined” patients. We varied $m$ from 0.5 to 1.0 to assess the number of poorly defined patients at different cutoffs.

### 3.3 Clustering results

#### 3.3.1 CD results

After removing patients lacking data describing the extent of their intestinal inflammation, we had 708 CD patients. Fig. 3.2 (A) shows the AIC, BIC, and model evidence (i.e., marginal likelihood) upon varying the number of groups within the BMM from two to ten for these patients. The decreasing numerical values of both the AIC and BIC as the number of groups increased up to eight groups indicates the models’ improving fit to the data. The fitted feature parameters for these eight groups, which represent the probability of inflammation in the region, are shown for each group in Fig. 3.3 (A).

![Figure 3.2](image)

**Figure 3.2:** (A) Model evidence, BIC, and AIC values as the number of groups in the CD patient BMM is increased from 2 to 10. (B) Proportion of patients who do not belong to a group with a greater than $m$ probability, as $m$ is varied from 0.5 to 1.0.

Fig. 3.2 (B) shows the proportion of “poorly defined” patients as the cutoff for belonging to a group was varied. The proportion of misclassified CD patients decreased steadily as the value
Figure 3.3: (A) Probability of each feature being “on” (i.e., each area being inflamed) in each group for CD. Number of patients calculated by summing over the posterior probabilities for each group. (B) Patterns of inflammation displayed by CD patients with >90% probability of belonging to the indicated group. Features that are “on” for a given patient are displayed in colour. Patients with less than 90% probability of belonging to a group are shown in the final graph.

of the cutoff decreased. Maps displaying the binary feature values for each patient assigned to the noted group are shown in Fig. 3.3 (B). As shown, the patients who are assigned to each group exhibit very consistent and distinct patterns of inflammation.

3.3.2 UC and IBDu results

Upon repeating the same analysis for UC and IBDu patients, we analysed 677 UC and 93 IBDu patients. The optimal number of groups according to the minimum AIC was eight and ten for UC and IBDu, respectively. As is shown in Fig. 3.2 (B), almost all UC and IBDu patients belonged to a group with high probability (>90%); patients with UC and IBDu appear to be more well-defined by the groups identified by the mixture model than patients with CD. Maps illustrating the features that are exhibited by each patient are shown in Fig. 3.4.

3.4 Discussion

Our understanding of distinct clinical subtypes of IBD is still in its early stages. Here, we have shown that patients may be subdivided into well-defined clusters based upon the extent of their GI inflammation. Clinical CD phenotypes are based upon the MC, which ascribes only
Figure 3.4: (A) Patterns of inflammation displayed by UC patients with > 80% probability of belonging to the indicated group. Patients with less than 80% probability of belonging to a group are displayed in the final graph (UC patients are illustrated at an 80% rather than 90% cutoff because of the second “elbow” bend in the proportion of unassigned patients at m = 80%, as shown in Fig. 3.2). (B) IBDu patients with a > 90% probability of belonging to the indicated group; poorly defined patients are again shown in the final graph.

three primary phenotypic categories: ileal, colonic, and ileocolonic inflammation, any of which may be combined with an upper GI modifier. As shown in Fig. 3.3, we have illustrated that further subtypes probably exist. While ileal (CD group 1), colonic (CD group 4), ileocolonic (CD group 2), and ileocolonic with upper GI involvement (CD group 7) patient subgroups are easily identifiable, there are also patients with ileal and lower colonic inflammation (CD group 3), lower colonic inflammation only (CD group 5), ileal and upper colonic inflammation (CD group 6), and primarily upper GI involvement (CD group 8). These additional subgroups are not insignificant: over one-quarter of CD patients were assigned to these “new” subtypes.

UC patients are divided into three categories through the MC: rectum only, the descending colon through to the rectum, and the entire colon. Consistent with the MC, patients with rectal inflammation (UC group 1), inflammation in the descending colon through the rectum (UC group 3), and pancolitis (UC group 2) are the primary clinical subgroups. However, we have again displayed that more subgroups of inflammation patterns probably exist within our patient cohort. As shown in Fig. 3.4, there are also patients with ileal involvement (UC group 6)
and upper colonic involvement only (UC group 8). Inflammation of surgically-created pouches is also a meaningful clinical feature: patients may have inflammation only in this area (UC group 5), or pouch inflammation combined with colonic inflammation (UC group 7).

Our results also illustrate a number of patient subgroups within the IBDu patient population. The IBDu cohort constitutes a smaller group of patients in our dataset, meaning that only small numbers of patients belong to any single BMM-assigned group with high probability. However, the groups that we do find allow easy identification of patients who may share more similarities with either CD or UC patients.

Studies attempting to correlate phenotypic characteristics of IBD with medication efficacy or genetic factors have historically struggled to find strong associations. However, these more finely-detailed phenotypic sub-groupings of IBD patients suggest that there may indeed be important differences in IBD patients that have not previously been identified. With improved phenotypic understanding of the nature of the disease, we may be able to gain further understanding of which subtypes of the disease respond well to particular therapies and whether these are partially due to underlying genetic predispositions.
4 Conclusions and Future Work

4.1 Summary of work to date

This work comprised a series of investigations applying machine learning techniques towards clinical problems involving both genomic and EHR data sources. For the application of predicting bacterial drug resistance, it was desirable to improve on our ability to accurately predict bacterial resistance and to identify new possible mechanisms of resistance. Using various supervised classification algorithms, we were able to provide improved prediction accuracy in comparison to the best alternative clinical rule-based approach for three of the four first-line drugs. We also identified several possible new resistance-conferring mechanisms and used generative probabilistic models to better visualise and understand putative resistance-conferring mutations.

IBD, the second clinical area of focus, has been studied for decades, but there is still only rudimentary understanding of possible disease sub-phenotypes. We performed a probabilistic clustering of IBD patients based upon their extent of GI inflammation in order to identify possible new subtypes of disease. We found that consistent patterns in inflammation are evident, beyond those that have already been identified and used in clinical practice. Our next steps will involve correlating these phenotypic subgroups with genotypic and other clinical data.

4.2 Immediate future work

Work over the coming months will be focused upon modelling of clinical time-series data, a very commonly-encountered type of data in medical applications, but which may be seldom used for analysis. The initial application will be the use of laboratory measurements from the patients within our IBD cohort (the same patients as identified in Chapter Three). In addition to the clinical data already presented, we have access to over 20 years of laboratory measurements taken across the Oxfordshire region for these patients. Once diagnosed, IBD is typically a lifelong 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 [83]. The clinical
question is therefore whether we can use the pattern of laboratory measurements surrounding the time period of diagnosis to predict future severity of the condition.

We have begun preliminary feature-based analyses of this data, finding that we are able to predict severity (as defined through surrogate variables from the EHR) with 69% accuracy. In this initial work, we have used metrics including the median, maximum, minimum, standard deviation, and proportion of abnormal laboratory values for each of the 17 laboratory tests of interest as features. Our next steps will involve two complementary approaches for modelling the underlying time-series data, which are relevant given that the laboratory data are characterised by frequent large spikes in values. The first approach will be to model the joint probability distribution of the entire dataset and to assess whether extreme values in relation to this multivariate distribution are indicative of patient severity. Secondly, we will model the outlier (i.e., above “normal” range) values using a generalised Pareto distribution, which is an appropriate way to model large exceedances of some high threshold, being based on extreme value statistics. While the first approach models the dataset as a whole, the second assumes that it is the distribution of the “extreme but normal” values that contains useful discriminatory information. We will also investigate a Gaussian process (GP) modelling approach to define patient-specific abnormal values, where such methods have been previously demonstrated to model the dynamics of time-series vital sign data [23].

While we can use features derived from these approaches to predict severity and non-severity (or alternative patient labels for other applications), we will also investigate possible latent clusters within the data. For instance, previous methods have clustered time series based upon their GP model hyperparameters or through their similarity using a likelihood-based metric [32]. We will further develop these methods and examine their applicability towards our datasets, in conjunction with the approaches investigated in the work described by this report.

While appropriately modelling and clustering the time-varying data is one aspect of the anticipated future work, we also plan to examine how disparate data types can be combined together to inform our predictions. Within our IBD dataset, we have clinical variables and time-varying laboratory information, as well as patient Immunochip (Illumina, USA) data. While small amounts of such data-types have been frequently combined in supervised predictive

1The Immunochip is a microarray that probes for SNPs associated with a wide range of inflammatory conditions.
analysis, appropriate methods for clustering still remain to be developed. We plan to draw upon the unsupervised clustering work of Kirk et al., which involved the correlated clustering across datasets including time-varying gene expression and other static cell markers \[72\]. However, while the Kirk et al. dataset was “well-behaved,” being derived from controlled experiments, our laboratory measurements are subject to the many challenges associated with working with realistic EHR data (as introduced in Chapter One). A key contribution will therefore be the extension of such methods to cope with our complex, realistic (and noisy) datasets - and to perform such inference at the very large scale required for clinical use.

4.2.1 Datasets in the pipeline

Future work will be focused upon developing analysis methods to model our time-series data appropriately, cluster these time-series, and to combine these clustering methods across different types of data. We will apply these techniques to our two primary application areas of IBD and infectious disease (ID). Within IBD, as mentioned, we already are in possession of a rich set of laboratory measurements, with lab results for 22 different tests for over 20 years. In addition to the Immunochip genetic data, we will gradually obtain full exome\(^2\) sequences for these patients over the next two years, which will be combined with existing data for these patients.

Within ID, we plan to examine the antibiograms of all *Escherichia coli* cultures collected in the Oxford University Hospitals NHS Trust from 1997 through 2014. We plan to use clustering approaches to analyse whether resistance to certain antibacterials occurs in clusters, which will allow us to investigate the presence of possible “resistance cassettes.” Resistance cassettes refer to groups of resistance genes that are all typically inherited together, either within a recombinant section of the genome or through a plasmid.

Also for *E. coli*, we plan to use EHR data from the past decade to identify whether incidence of antibiotic resistance decreased prior to implementation of strict infection control measures. This tackles a similar question as that previously investigated for *Staphylococcus aureus* \[84\]. We plan to compare both conventional analysis techniques (which have been used in the previous analysis) and GP-based time-series techniques.

Finally, for ID, should the data be available within the hospital records, we will develop methods for more sophisticated phenotyping of patients with bacterial infection. This will

\(^2\)The exome is the protein-coding region of the genome.
involve analysis of the EHR data available for patients who either enter the hospital with infection or become infected during their stay. In a similar manner to the IBD analysis, we again plan to analyse time-series laboratory data to obtain finer understanding of infection severity. We will correlate this with the whole genome sequences of the bacteria. Similar preliminary work has been performed using conventional medical statistics for *Clostridium difficile* [85], but we now also have whole genome sequences to bolster our genetic analysis, motivating our use of principled machine learning methods.

### 4.3 Plan until completion

Fig. 4.1 presents a Gantt chart illustrating the planned activities through to D.Phil completion. The remainder of 2014 will be focused upon developing methods for the clustering of our clinical IBD time-series data and Bayesian clustering of our ID antibiogram data. This will then be supplemented with further development of clustering algorithms that are able to combine both time series and single time-point data, which is planned through until the end of 2015. A study-abroad experience via the Centre for Doctoral Training (CDT) programme with a complimentary research focus is anticipated for the summer of 2015.

The DPhil is scheduled to be completed by September 2016. As much as possible, portions of the thesis will be written in parallel to the completion of new work. However, the focus will shift towards the writing of the thesis at the end of the winter months of 2016. Throughout the next two years, these activities will be interspersed with appropriate conference preparation and attendance to disseminate findings and to discuss methodology.

### 4.4 Risk assessment

There are no major anticipated risks associated with the above plan. We do not yet have access to the ID data, but they should be straightforward to obtain, as they are contained within the Infections in Oxfordshire Research Database (IORD). IORD has been approved for research use by the Oxford Research Ethics Committee and the National Information Governance Board, therefore only requiring an internal ethics application. We are currently preparing the ethics applications to access the data we require from this database.

Since we have not yet seen the data contained in IORD, this means that we cannot be certain what sort of information we may be able to extract. As such, we may require different
### Figure 4.1: Gantt chart illustrating activities from submission of transfer report through until DPhil completion in September 2016. The next nine months are shown in detail, together with plans for the following terms until completion.

models than those that we had originally anticipated. However, this is not so much a risk as an exciting challenge representative of the goals of this DPhil: we must develop methods with which to manage and perform inference on real-world hospital data.

Another possible risk is that, upon reaching the end of the DPhil, no new medically-relevant discoveries will have been made. This is unlikely, given that for both IBD and ID we have access to numerous data sources that have not before been available and analysed collectively. Nevertheless, our analysis approaches and methods will still be valid and can be applied to other clinical questions, even if they are not fruitful for the applications we have been investigating specifically.


